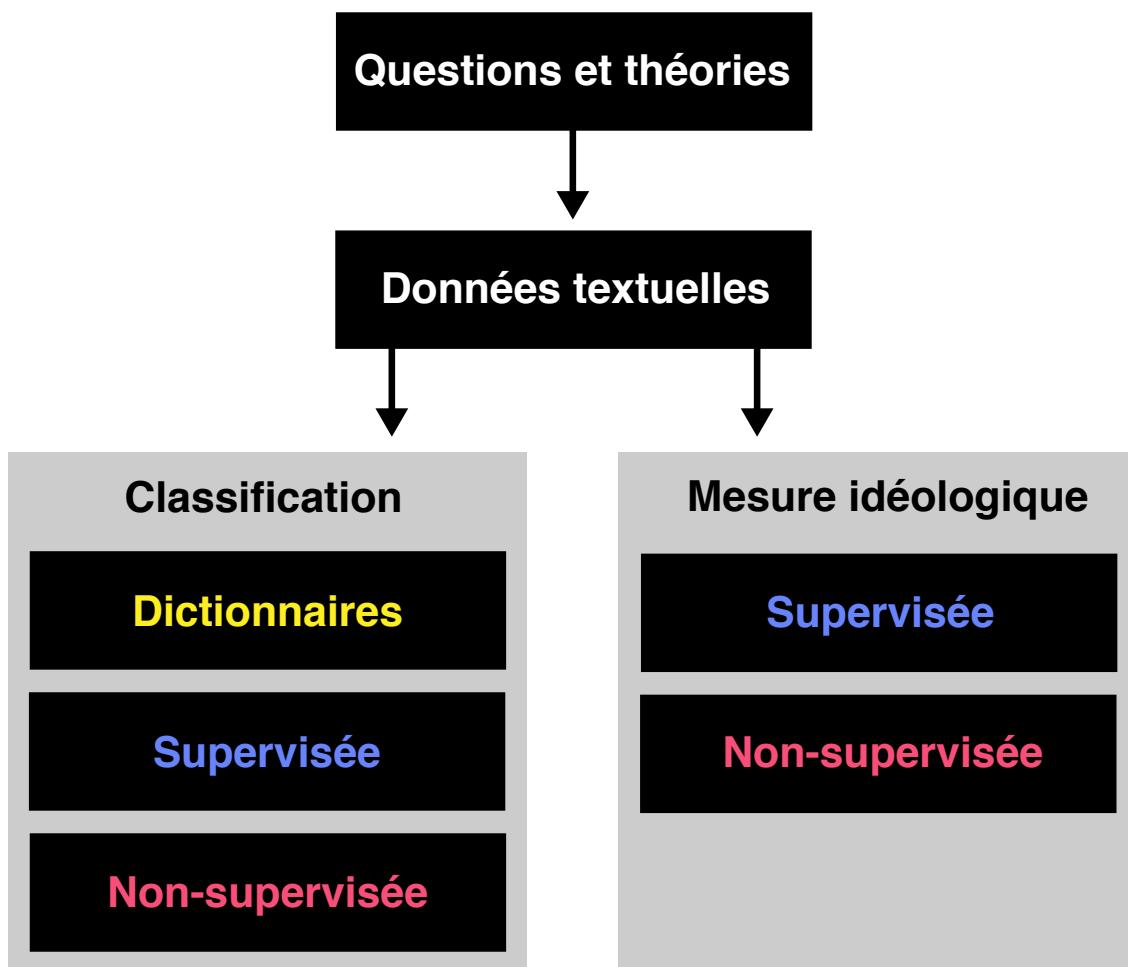


# Analyse textuelle

Sélection de textes



# Prélude

5

## Questions et théories

J. Grimmer and B. M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. <i>Political Analysis</i> , 21(3):267–297, 2013. . . . .	6
K. Krippendorff. Content analysis: An introduction to its methodology. Sage, 2012. Chapter 1: History. . . . .	37
C. C. Aggarwal and C. Zhai. Mining text data. Springer Science & Business Media, 2012. Chapter 1: Introduction to Text Mining. . . . .	51
B. O'Connor, D. Bamman, and N. A. Smith. Computational text analysis for social science: Model assumptions and complexity. 2011. . . . .	61
B. L. Monroe and P. A. Schrodt. Introduction to the special issue: The statistical analysis of political text. <i>Political Analysis</i> , 16(4):351–355, 2008. . . . .	69
J. Wilkerson and A. Casas. Large-scale computerized text analysis in political science: Opportunities and challenges. <i>Annual Review of Political Science</i> , 20:529–544, 2017. . . . .	74

## Données textuelles

K. A. Neuendorf. The content analysis guidebook. Sage, 2016. Chapter 1: Defining Content Analysis. . . . .	90
K. Benoit, D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov. Crowd-sourced text analysis: reproducible and agile production of political data. 2015. . . . .	125
H. Schütze, C. D. Manning, and P. Raghavan. Introduction to information retrieval, volume 39. Cambridge University Press, 2008. . . . .	143
M. F. Porter. Snowball: A language for stemming algorithms, 2001. . . . .	173
K. Welbers, W. Van Atteveldt, and K. Benoit. Text analysis in R. <i>Communication Methods and Measures</i> , 11(4):245–265, 2017. . . . .	186

## Dictionnaires

207

### Classification par dictionnaires

- Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010. . . . . 208
- P. S. Dodds and C. M. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, 2010. 239
- L. Young and S. Soroka. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231, 2012. . . . . 255
- M. Laver and J. Garry. Estimating policy positions from political texts. *American Journal of Political Science*, pages 619–634, 2000. . . . . 282

## Supervisées

298

### Mesure idéologique supervisée

- M. Laver, K. Benoit, and J. Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331, 2003. . . . . 299
- W. Lowe. Understanding Wordscores. *Political Analysis*, 16(4):356–371, 2008. . . . . 320
- W. Lowe, K. Benoit, S. Mikhaylov, and M. Laver. Scaling policy preferences from coded political texts. *Legislative Studies Quarterly*, 36(1):123–155, 2011. . . . . 336
- W. Lowe and K. Benoit. Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, 21(3):298–313, 2013. . . . . 369

### Classification supervisée

- D. Hillard, S. Purpura, and J. Wilkerson. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4):31–46, 2008. . . . . 385

- D. J. Hopkins and G. King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 2010. . . . . 401
- G. King, J. Pan, and M. E. Roberts. How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, 107(2):326–343, 2013. . . . . 420
- A. A. Jamal, R. O. Keohane, D. Romney, and D. Tingley. Anti-americanism and anti-interventionism in arabic Twitter discourses. *Perspectives on Politics*, 13(1):55–73, 2015. . . . 438

## Non-supervisées

457

### Mesure idéologique non-supervisée

- M. Denny and A. Spirling. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. 2017. . . . . 458
- J. B. Slapin and S.-O. Proksch. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722, 2008. . . . . 507
- S.-O. Proksch and J. B. Slapin. Position taking in European parliament speeches. *British Journal of Political Science*, 40(3):587–611, 2010. . . . . 525

### Classification non-supervisée I

- J. Grimmer and G. King. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, 108(7):2643–2650, 2011. . . . 550
- G. King, P. Lam, and M. E. Roberts. Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, 61(4):971–988, 2017. 558
- J. W. Mohr and P. Bogdanov. Introduction—topic models: What they are and why they matter, 2013. . . . . 576
- J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009. . . . . 601

- D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003. . . . . 610

## Classification non-supervisée II

- D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012. . . . . 640
- J. Grimmer. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, 18(1):1–35, 2009. . . . . 648
- K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228, 2010. . . . . 683
- M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082, 2014. . . . . 703

## Conclusion

722

- J. Grimmer. We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, 48(1):80–83, 2015. . . . . 723
- D. Lazer and J. Radford. Data ex machina: Introduction to big data. *Annual Review of Sociology*, 43:19–39, 2017. . . . . 727

# Prélude



## **Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts**

**Justin Grimmer**

*Department of Political Science, Stanford University, Encina Hall West 616 Serra Street,  
Stanford, CA 94305*  
*e-mail: jgrimmer@stanford.edu (corresponding author)*

**Brandon M. Stewart**

*Department of Government and Institute for Quantitative Social Science, Harvard University,  
1737 Cambridge Street, Cambridge, MA 02138*  
*e-mail: bstewart@fas.harvard.edu*

Edited by R. Michael Alvarez

Politics and political conflict often occur in the written and spoken word. Scholars have long recognized this, but the massive costs of analyzing even moderately sized collections of texts have hindered their use in political science research. Here lies the promise of automated text analysis: it substantially reduces the costs of analyzing large collections of text. We provide a guide to this exciting new area of research and show how, in many instances, the methods have already obtained part of their promise. But there are pitfalls to using automated methods—they are no substitute for careful thought and close reading and require extensive and problem-specific validation. We survey a wide range of new methods, provide guidance on how to validate the output of the models, and clarify misconceptions and errors in the literature. To conclude, we argue that for automated text methods to become a standard tool for political scientists, methodologists must contribute new methods and new methods of validation.

### **1 Introduction**

Language is the medium for politics and political conflict. Candidates debate and state policy positions during a campaign. Once elected, representatives write and debate legislation. After laws are passed, bureaucrats solicit comments before they issue regulations. Nations regularly negotiate and then sign peace treaties, with language that signals the motivations and relative power of the countries involved. News reports document the day-to-day affairs of international relations that provide a detailed picture of conflict and cooperation. Individual candidates and political parties articulate their views through party platforms and manifestos. Terrorist groups even reveal their preferences and goals through recruiting materials, magazines, and public statements.

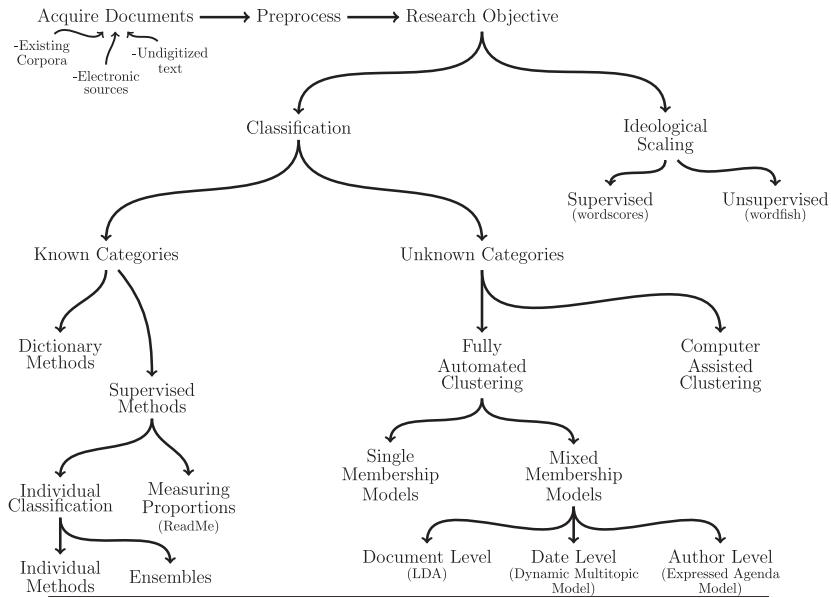
These examples, and many others throughout political science, show that to understand what politics is about we need to know what political actors are saying and writing. Recognizing that language is central to the study of politics is not new. To the contrary, scholars of politics have long recognized that much of politics is expressed in words. But scholars have struggled when using texts to make inferences about politics. The primary problem is volume: there are simply *too many* political texts. Rarely are scholars able to manually read all the texts in even moderately sized corpora. And hiring coders to manually read all documents is still very expensive. The result is that

Downloaded from <http://pan.oxfordjournals.org/> at London School of Economics on July 7, 2013

---

*Authors' note:* For helpful comments and discussions, we thank participants in Stanford University's Text as Data class, Mike Alvarez, Dan Hopkins, Gary King, Kevin Quinn, Molly Roberts, Mike Tomz, Hanna Wallach, Yuri Zhurkov, and Frances Zlotnick. Replication data are available on the *Political Analysis* Dataverse at <http://hdl.handle.net/1902.1/18517>. Supplementary materials for this article are available on the *Political Analysis* Web site.

© The Author 2013. Published by Oxford University Press on behalf of the Society for Political Methodology.  
All rights reserved. For Permissions, please email: journals.permissions@oup.com



**Fig. 1** An overview of text as data methods.

analyzing massive collections of text has been essentially impossible for all but the most well-funded projects.

We show how automated content methods can make possible the previously impossible in political science: the systematic analysis of large-scale text collections without massive funding support. Across all subfields of political science, scholars have developed or imported methods that facilitate substantively important inferences about politics from large text collections. We provide a guide to this exciting area of research, identify common misconceptions and errors, and offer guidelines on how to use text methods for social scientific research.

We emphasize that the complexity of language implies that automated content analysis methods will never replace careful and close reading of texts. Rather, the methods that we profile here are best thought of as *amplifying* and *augmenting* careful reading and thoughtful analysis. Further, automated content methods are *incorrect* models of language. This means that the performance of any one method on a new data set cannot be guaranteed, and therefore validation is essential when applying automated content methods. We describe best practice validations across diverse research objectives and models.

Before proceeding we provide a road map for our tour. Figure 1 provides a visual overview of automated content analysis methods and outlines the process of moving from collecting texts to applying statistical methods. This process begins at the top left of Fig. 1, where the texts are initially collected. The burst of interest in automated content methods is partly due to the proliferation of easy-to-obtain electronic texts. In Section 3, we describe document collections which political scientists have successfully used for automated content analysis and identify methods for efficiently collecting new texts.

With these texts, we overview methods that accomplish two broad tasks: classification and scaling. *Classification* organizes texts into a set of categories. Sometimes researchers know the categories beforehand. In this case, automated methods can minimize the amount of labor needed to classify documents. *Dictionary* methods, for example, use the frequency of key words to determine a document's class (Section 5.1). But applying dictionaries outside the domain for which they were developed can lead to serious errors. One way to improve upon dictionaries are

*supervised* methods (Section 5.2). These methods begin with human hand coding of documents into a predetermined set of categories. The human hand coding is then used to train, or *supervise* statistical models to classify the remaining press releases. But the performance of any one classifier can vary substantially across contexts—so validation of a classifier’s accuracy is essential to establish the reliability of supervised learning methods.

Classification methods can also be used to discover new ways of organizing texts. *Fully Automated Clustering* (FAC) algorithms simultaneously estimate the categories and then classify documents into those categories (Section 6.1). Applying FAC algorithms to texts may have mixed results—it is difficult to know *a priori* if any one method will provide useful clusterings in a particular context. We describe two approaches to improve the output of FAC algorithms. The first set of models, *mixed membership models*, include problem-specific structure to assist in the estimation of categories (Section 6.1.2). The second model, *Computer Assisted Clustering* (CAC), provides a method to explore thousands of potential clusterings (Section 6.2). Because there is no guarantee that unsupervised methods will return classes that are theoretically interesting, validation is essential. We describe several validation methods, including validations that combine supervised and unsupervised learning methods (Section 6.4).

Automated content methods can also estimate the location of actors in policy space, or produce a *scaling*. One method, *word scores*, relies on guidance from reference texts to situate other political actors in a space (Section 7.1). A second method, *word fish*, exploits an assumption about how ideology affects word usage (Section 7.2). We show how both methods rely heavily on the assumption that ideology dominates the language used in the text. When this assumption fits, when political actors are expressing their policy positions in texts, the models can perform well. When this assumption is violated, the model will place actors in a different and non-ideological space. We show that this space can be useful, but validations are necessary to establish its meaning.

No one article can hope to survey such a massive field. We have focused on methods that satisfy a diverse set of research objectives, focusing specifically on methods that analyze texts at the *document* level. This is most apparent with the absence of natural language processing methods to parse sentences and understand events (Jurafsky and Martin 2009). These methods are certainly useful in other fields, but have yet to gain widespread use in political science. The one notable exception is the work on events data collection (which has already been well-reviewed in Schrodт 2006). In the conclusion (and Supplementary Material), we direct readers to further sources to learn more about the methods introduced in this article and methods that we regretfully exclude. We also caution the reader that the space allocated to methods in this article is not necessarily in proportion to their use among political scientists. Rather, some methods—such as supervised learning methods—are relatively conceptually simple and require less validation. Other methods, such as unsupervised classification methods, require much more explanation and greater space for validation.

## 2 Four Principles of Automated Text Analysis

This section presents our four principles of automated content analysis methods (summarized in Table 1). We provide a brief introduction to each principle here. Throughout our exploration of the methods, we will revisit these principles and see that they offer a useful guide for using and evaluating quantitative methods for content analysis.

### 2.1 Principle 1: All Quantitative Models of Language Are Wrong—But Some Are Useful

The data generation process for any text (including this one) is a mystery—even to linguists. If any one sentence has complicated dependency structure, its meaning could change drastically with

**Table 1** Four principles of quantitative text analysis

- 
- (1) All quantitative models of language are wrong—but some are useful.
  - (2) Quantitative methods for text amplify resources and augment humans.
  - (3) There is no globally best method for automated text analysis.
  - (4) Validate, Validate, Validate.
-

the inclusion of new words, and the sentence context could drastically change its meaning. Consider, for example, the classic joke “Time flies like an arrow. Fruit flies like a banana.” This joke relies on the extreme complexity of parsing the sentence, using the change in our understanding of “flies like” as the sentence progresses—from metaphorical to literal.

The complexity of language implies that all methods *necessarily* fail to provide an accurate account of the data-generating process used to produce texts. Automated content analysis methods use insightful, but wrong, models of political text to help researchers make inferences from their data. The goal of building text models is therefore different than model building to make causal inferences—the primary goal of most political science models. The usual advice for causal inference model building is that it is essential to include all the relevant features of the data-generating process—either in covariates or in the model structure (e.g., Greene 2007). But the same advice does not hold for quantitative text analysis methods. Including more realistic features into quantitative models does not necessarily translate into an improved method, and reducing the assumptions used may not imply more productive analyses. Rather, subtleties of applying the methods to any one data set mean that models that are less sophisticated in the use of language may provide more useful analysis of texts.

That all automated methods are based on incorrect models of language also implies that the models should be evaluated based on their ability to perform some useful social scientific task. As we explain below, emphasis in evaluations should be placed on helping researchers to assign documents into predetermined categories, discover new and useful categorization schemes for texts, or in measuring theoretically relevant quantities from large collections of text. Alternative model evaluations that rely on model fit or predicting the content of new texts can select substantively inferior models (see Chang et al. 2009).

## **2.2 Principle 2: Quantitative Methods Augment Humans, Not Replace Them**

Automated content analysis methods have demonstrated performance across a variety of substantive problems. These methods will not, however, eliminate the need for careful thought by researchers nor remove the necessity of reading texts. Indeed a deep understanding of the texts is one of the key advantages of the social scientist in applying automated methods. We will see below that researchers still guide the process, make modeling decisions, and interpret the output of the models. All these require the close reading of texts and thoughtful analysis by the researcher.

Rather than replace humans, computers *amplify* human abilities. The most productive line of inquiry, therefore, is not in identifying how automated methods can obviate the need for researchers to read their text. Rather, the most productive line of inquiry is to identify the best way to use both humans and automated methods for analyzing texts.

## **2.3 Principle 3: There Is No Globally Best Method for Automated Text Analysis**

Different data sets and different research questions often lead to different quantities of interest. This is particularly true with text models. Sometimes, the goal is to identify the words that distinguish the language of groups (Laver, Benoit, and Garry 2003; Monroe, Colaresi, and Quinn 2008). In other research projects, the quantity of interest is the proportion of documents that fall within a predetermined set of categories (Hopkins and King 2010). And in yet other research projects, scholars may want to use quantitative methods to discover a new way to organize texts (e.g., Grimmer and King 2011) or to locate actors in a spatial model (Laver, Benoit, and Garry 2003; Monroe and Maeda 2004; Slapin and Proksch 2008).

Each of the research questions imply different *models*, or families of models, to be used for analysis, and different methods of validation. Therefore, much of the debate across approaches for automated text analysis is misguided. We show below that much of the across-approach debate can be resolved simply by acknowledging that there are different research questions and designs that imply different types of models will be useful. Rather than debating across approaches to text analysis, we think one of the most important lines of inquiry is identifying effective ways to combine previously distinct methods for quantitative text analysis.

There is also substantial variation within families of models. Perhaps unsurprisingly, the same model will perform well on some data sets, but will perform poorly when applied to other data. Therefore, establishing one method to always use for one task is almost guaranteed to be impossible. Instead scholars will need to carefully think and apply different methods to generate useful and acceptable estimates for their problem.

#### 2.4 Principle 4: Validate, Validate, Validate

Automated text analysis methods can substantially reduce the costs and time of analyzing massive collections of political texts. When applied to any one problem, however, the output of the models may be misleading or simply wrong. Therefore, it is incumbent upon the researcher to *validate* their use of automated text analysis. This validation can take several forms, many of which we describe below. When categories are known in a calculation problem, scholars must demonstrate that the supervised methods are able to reliably replicate human coding. Validation for unsupervised methods is less direct. To validate the output of an unsupervised method, scholars must combine experimental, substantive, and statistical evidence to demonstrate that the measures are as conceptually valid as measures from an equivalent supervised model. Similar to unsupervised methods, validating ideological scalings requires numerous and substance-based evaluations (Budge and Pennings 2007; Slapin and Proksch 2008).

What should be avoided, then, is the blind use of any method without a validation step. This is why we discourage the use of many commercial tools for quantitative text analysis. These programs simply provide the researcher with output. It is often difficult, and sometimes impossible, to validate the output. But more damning for these methods is that once a problem is identified, it is exceedingly difficult to change how the analysis is conducted. Certainly, these methods can be validated and provide conceptually valid and useful output. But without a broader set of tools available, it is difficult to know if the methods included in the commercially available software are optimal for the particular problem at hand.

Before applying these principles, texts are needed. The next section describes common sources for data and characteristics of data that make automated methods more likely to succeed.

### 3 Acquiring Text

Political scientists have applied automated content analysis across a diverse set of texts. This includes archives of media data (Young and Soroka 2011); floor speeches in legislatures from across the world (Quinn et al. 2010); presidential, legislator, and party statements (Grimmer 2010); proposed legislation and bills (Adler and Wilkerson 2011); committee hearings (Jones, Wilkerson, and Baumgartner 2009); treaties (Spirling 2012); political science papers; and many other political texts.

This explosion is partly due to the rapid move to store and distribute documents in electronic text databases. Automated methods require texts stored in a plain text format—UTF-8 for Latin characters, for example. The easiest way to acquire text in this form is from online databases of articles. Lexis Nexis and ProQuest, for example, facilitate batch downloads of files, and J STOR provides already processed texts of academic articles. More sources are being added all the time; for example, the U.S. House of Representatives recently launched a new Web site dedicated to the distribution of all current House Resolutions under discussion in Xstensible Markup Language (XML) format, an easily accessible and widely used method for storing structured text.

Slightly more difficult to access are text data stored on Web sites, but automated scraping methods make acquiring these data easier (Jackman 2006). And even when scraping data is not an option—perhaps due to Web site restrictions—online platforms that distribute tasks across workers, such as Amazon’s Mechanical Turk, can make acquiring data efficient (Berinsky, Huber, and Lenz 2012). The most difficult to acquire are texts found in archives or yet-to-be-scanned books. But preparing these texts for analysis can be straightforward—using a high-quality scanner and Optical Character Recognition software, it is possible to convert archival

materials into computer readable texts (see e.g., the data collection process in Eggers and Hainmueller 2009).

Texts abound in politics and automated methods are potentially useful across most political texts. But some texts better fit the assumptions of the automated content methods described here. These methods tend to work better when the text is focused—either on the expression of one idea for classification or the expression of policy positions for scaling methods. In some instances, discarding text not related to the primary quantity of interest can actually improve the performance of automated clustering methods. Automated methods also rely on having a sufficient number of words to use reliably. This makes lengthier texts—newspapers or party platforms—often much easier to analyze than shorter statements, such as open-ended survey responses. For shorter texts, accompanying information (or an extremely large volume of texts) is often necessary for classification or scaling methods to perform reliably.

With texts in hand, the next section shows how to move words on the page to numbers analyzed statistically.

#### 4 Reducing Complexity: From Words to Numbers

Language is complex. But not all of language’s complexity is necessary to effectively analyze texts. In this section, we present a recipe for transforming text into quantitative data. Each of the steps are designed to retain information that will be used by the automated methods, while discarding information that will likely be unhelpful, ancillary, or too complex for use in a statistical model. The recipe is easy to apply, with several pieces of freely available software able to apply the sequence of steps we describe here. We present a recipe for representing text quantitatively, but emphasize that any one of the steps that we present here can—and oftentimes should—be modified. More important than following any individual recipe is to think carefully about the particular problem at hand, test different approaches, and validate the results.

Throughout, we will refer to the unit of analysis as a *text* or *document*, but it could apply to any unit of text: a tweet, Facebook status, spoken utterance, press briefing, sentence, or paragraph. We refer to the population of texts to be analyzed as the *corpus* and a collection of these as *corpora*. Some methods will work better on different types of tasks or documents of different lengths, but most methods begin in the same way with a series of preprocessing steps to reduce the awe inspiring diversity of language to a manageable set of features.

The most consequential, and shocking, step we will take is to discard the order in which words occur in documents (Jurafsky and Martin 2009). We will assume documents are a *bag of words*, where order does not inform our analyses. While it is easy to construct sample sentences where word order fundamentally changes the nature of the sentence, empirically these sentences are rare. A simple list of words, which we call *unigrams*, is often sufficient to convey the general meaning of a text. If this assumption is unpalatable, we can retain some word order by including *bigrams* (word pairs) or *trigrams* (word triples) into our analysis (Jurafsky and Martin 2009). This allows us to distinguish, for example, the “White House” from the color and the domicile. In practice, for common tasks like measuring sentiment, topic modeling, or search, *n-grams* do little to enhance performance (Manning, Raghavan, and Schütze 2008; Hopkins and King 2010).

After discarding word order, we simplify the vocabulary with *stemming*. Stemming removes the ends of words to reduce the total number of unique words in the data set, or reduce the *dimensionality* of text. Stemming reduces the complexity by mapping words that refer to the same basic concept to a single root. For example, family, families, families’, and familial all become *famili*. Stemming is actually an approximation to a linguistic concept called *lemmatization*, which seeks to reduce words to their base forms (Manning, Raghavan, and Schütze 2008; Jurafsky and Martin 2009). The critical difference is that a lemmatizer uses context and dictionaries to help discover (for example) that good is the base form of better and best. The stemmer is a much cruder algorithm, but considerably faster. The performance does not seem to matter for most applications, so the majority of applied research uses stemming. There are numerous stemming algorithms available that vary in the extent and frequency of word truncation.

But the Porter stemming algorithm (Porter 1980) is commonly employed because of its moderate approach to word simplification.

In addition to discarding word order, we also typically discard punctuation, capitalization, very common words (often we remove “stop” words, or function words that do not convey meaning but primarily serve grammatical functions), and very uncommon words (words that appear only once or twice in the corpus and thus are unlikely to be discriminating). We typically remove words which appear in less than 1% and more than 99% of documents in the corpus, although these choices need to be made contingent both on the diversity of the vocabulary, average length of the document, and the size of the corpus (Quinn et al. 2010; Hopkins and King 2010).

The result of the preprocessing steps is that each document  $i$  ( $i = 1, \dots, N$ ) is represented as a vector that counts the number of times each of the  $M$  unique words occur,  $W_i = (W_{i1}, W_{i2}, \dots, W_{iM})$ . Each  $W_{im}$  counts the number of times the  $m$ -th word occurs in the  $i$ -th document. The collection of count vectors into a matrix is often called the *document-term matrix*. For a moderate volume of documents without a particularly specialized vocabulary, this matrix will have between three thousand and five thousand *features* or terms and will contain mostly zeroes (a condition we call *sparsity*).

These steps seem to result in a shocking reduction of information, leaving many to conclude that there will be too little information to extract anything meaningful from the texts. But consistently across applications, scholars have shown that this simple representation of text is sufficient to infer substantively interesting properties of texts (Hopkins and King 2010).

#### 4.1 Alternative Methods for Representing Text

The recipe described above is one way to represent text as data. This can, and should, be varied as needed for specific applications. For example, in one of the first examples of quantitative text analysis, Mosteller and Wallace (1963) sought to infer the authorship of the unattributed Federalist Papers. Since they were interested specifically in the style of the documents and not their content, they used only the counts of function words or *stopwords*. Thus, their entire analysis relied on information that we customarily discard. Some other common strategies include: (1) using an indicator that a word occurs in a document, rather than a count (Pang, Lee, and Vaithyanathan 2002; Hopkins and King 2010); (2) including some commonly used stopwords such as gendered pronouns (Monroe, Colaresi, and Quinn 2008); (3) a subset of features (either by automated feature selection or a lower dimensional projection) (Hofmann 1999); and (4) weighting words by their rarity in the document set (often called *tf-idf* or term frequency by inverse document frequency weighting) (Manning, Raghavan, and Schütze 2008).

Although these are fundamentally similar ways of describing the same basic feature set (the list of unordered unigrams), sometimes the problem calls for a completely different approach. Spirling (2012), for example, analyzes treaties between the Native Americans and the U.S. government. In this case, discarding word order would mask information on land negotiations. In order to preserve word order information, Spirling (2012) uses *sub-string kernels*. The sub-string portion means that each feature is a small sequence of letters (e.g., five) that can span multiple words. Since using this feature space would be unimaginably large, Spirling (2012) uses technology from Lodhi et al. (2002) to calculate only the distance between the documents in this feature space using only sub-strings that occur in both documents. Thus, scaling performed on the resulting matrix of distances is able to account for word order.

### 5 Classifying Documents into Known Categories

Assigning texts to categories is the most common use of content analysis methods in political science. For example, researchers may ask if campaigns issue positive or negative advertisements (Anscombe and Iyengar 1995), if legislation is about the environment or some other issue area (Adler and Wilkerson 2011), if international statements are belligerent or peaceful (Schrodt 2000), or if local news coverage is positive or negative (Eshbaugh-Soha 2010). In each instance, the goal is

to infer either the category of each document, the overall distribution of documents across categories, or both.

Human-based methods for making these inferences are both time and resource intensive. Even after coding rules are developed and coders are trained, manual methods require that coders read each individual text. Automated methods can mitigate the cost of assigning documents to categories, by limiting the amount of classification humans perform. We characterize two broad groups of methods for reducing the costs of classification. *Dictionary* methods use the relative frequency of *key words* to measure the presence of each category in texts. Supervised learning methods replicate the familiar manual coding task, but with a machine. First, human coders are used to classify a subset of documents into a predetermined categorization scheme. Then, this *training set* is used to train an automated method, which then classifies the remaining documents.

### 5.1 Dictionary Methods

We begin with dictionary methods, perhaps the most intuitive and easy to apply automated method (Stone et al. 1966). Dictionaries use the rate at which key words appear in a text to classify documents into categories or to measure the extent to which documents belong to particular categories. For example, suppose the goal is to measure the *tone* in newspaper articles (e.g., Eshbaugh-Soha 2010): whether articles convey information positively or negatively. Dictionary methods use a list of words with attached tone scores and the relative rate at which words occur to measure a document's tone. A *dictionary* to measure tone is a list of words that are either dichotomously classified as positive or negative or contain more continuous measures of their content. Formally, each word  $m$  ( $m = 1, \dots, M$ ) will have associated score  $s_m$ . For the simplest measures,  $s_m = -1$  if the word is associated with a negative tone and  $s_m = 1$  if associated with a positive tone. If  $N_i = \sum_{m=1}^M W_{im}$  words are used in document  $i$ , then dictionary methods can measure the tone for any document  $t_i$  as,

$$t_i = \sum_{m=1}^M \frac{s_m W_{im}}{N_i}.$$

Scholars often use  $t_i$  as an approximately continuous measure of document tone, but it also can be used to classify documents into tone categories if a decision rule is assumed along with the dictionary method. Perhaps the simplest coding rule would assign all documents with  $t_i > 0$  to a positive tone category and  $t_i < 0$  to a negative tone.

Tone is just one type of analysis a dictionary method can perform. The general idea of dictionaries make them relatively easy and cheap to apply across a variety of problems: identify words that separate categories and measure how often those words occur in texts (for some recent examples that use dictionaries to measure a variety of concepts, see Kellstedt 2000; Laver and Garry 2000; Burden and Sanberg 2003; Young and Soroka 2011). Finding the separating words is also relatively easy. There are a variety of widely used off-the-shelf dictionaries that provide key words for a variety of categories (e.g., Bradley and Lang 1999; Hart 2000; Pennebaker, Francis, and Booth 2001; Turney and Littman 2003). And if scholars have documents already coded into categories, dictionaries can be produced using existing methods. Monroe, Colaresi, and Quinn (2008) describe a variety of methods that measure how well words separate already identified categories of interest (see also Taddy 2010 and Diermeier et al. 2011). Any one of these methods could be used to produce dictionary-like scores of words, which could then be applied in other contexts to classify documents.

For dictionary methods to work well, the scores attached to words must closely align with how the words are used in a particular context. If a dictionary is developed for a specific application, then this assumption should be easy to justify. But when dictionaries are created in one substantive area and then applied to another, serious errors can occur. Perhaps the clearest example of this is shown in Loughran and McDonald (2011). Loughran and McDonald (2011) critique the increasingly common use of off-the-shelf dictionaries to measure the tone of statutorily required corporate earning reports in the accounting literature. They point out that many words that have a negative

connotation in other contexts, like *tax*, *cost*, *crude* (oil), or *cancer*, may have a positive connotation in earning reports. For example, a health-care company may mention *cancer* often and oil companies are likely to discuss *crude* extensively. And words that are not identified as negative in off-the-shelf dictionaries may have quite negative connotation in earning reports (e.g., *unanticipated*).

Dictionaries, therefore, should be used with substantial caution, or at least coupled with explicit validation. When applying dictionaries, scholars should directly establish that word lists created in other contexts are applicable to a particular domain, or create a problem-specific dictionary. In either instance, scholars must validate their results. But measures from dictionaries are rarely validated. Rather, standard practice in using dictionaries is to assume the measures created from a dictionary are correct and then apply them to the problem. This is due, in part, to the exceptional difficulties in validating dictionaries. Dictionaries are commonly used to establish granular scales of a particular kind of sentiment, such as tone. Although this is useful for applications, humans are unable to produce the same granular measures reliably (Krosnick 1999). The result is that it is essentially impossible to derive gold-standard evaluations of dictionaries based on human coding of documents.

The consequence of domain specificity and lack of validation is that most analyses based on dictionaries are built on shaky foundations. Yes, dictionaries are able to produce measures that are claimed to be about tone or emotion, but the actual properties of these measures—and how they relate to the concepts they are attempting to measure—are essentially a mystery. Therefore, for scholars to effectively use dictionary methods in their future work, advances in the validation of dictionary methods must be made. We suggest two possible ways to improve validation of dictionary methods. First, the classification problem could be simplified. If scholars use dictionaries to code documents into binary categories (e.g., positive or negative tone), then validation based on human gold standards and the methods we describe in Section 5.2.3 is straightforward. Second, scholars could treat measures from dictionaries similar to how validations from unsupervised methods are conducted (see Section 6.4). This would force scholars to establish that their measures of underlying concepts satisfy several different standards of validity.

## 5.2 Supervised Learning Methods

Dictionary methods require scholars to identify words that separate classes beforehand. This can lead to inefficiencies in applying dictionaries to real data problems—particularly if dictionaries are applied outside of the domain for which they were originally developed. Supervised learning methods provide a useful alternative method for assigning documents to predetermined categories. The idea of supervised learning is simple: human coders categorize a set of documents by hand. The algorithm then “learns” how to sort the documents into categories using the training set and words: the algorithm uses characteristics of the documents to place the documents into the categories.

This approach to classification has two major advantages over dictionary methods. First, it is necessarily domain specific and therefore avoids the problems of applying dictionaries outside of their intended area of use. Applying supervised learning methods requires scholars to develop coding rules for the particular quantities of interest. The model is then trained using a sample of documents from the corpus that is to be classified. This also forces scholars to develop coherent definitions of concepts for particular applications, which leads to clarity in what researchers are measuring and studying. Second, supervised learning methods are much easier to validate, with clear statistics that summarize model performance.

Supervised methods for text classification is a massive—and rapidly expanding—area of research (see, e.g., the excellent software provided in Jurka et al. 2012). But all supervised learning methods share three basic steps: (1) construct a training set; (2) apply the supervised learning method—learning the relationship between features and categories in the training set, then using this to infer the labels in the test set; and (3) validate the model output and classify the remaining documents.

We outline each of the steps here—how to construct a training set, training and applying statistical models, and how to validate and assess the performance of the method.

### 5.2.1 Constructing a training set

The most important step in applying a supervised learning algorithm is constructing a reliable training set, because no statistical model can repair a poorly constructed training set, and well-coded documents can hide faults in simple models. We divide the construction of the training set into two components: (1) creating and executing a coding scheme and (2) sampling documents.

*Creating a Coding Scheme:* Ambiguities in language, limited attention of coders, and nuanced concepts make the reliable classification of documents difficult—even for expert coders. To address this difficulty, best practice is to iteratively develop coding schemes. Initially, a concise codebook is written to guide coders, who then apply the codebook to an initial set of documents. When using the codebook, particularly at first, coders are likely to identify ambiguities in the coding scheme or overlooked categories. This subsequently leads to a revision of the codebook, which then needs to be applied to a new set of documents to ensure that the ambiguities have been sufficiently addressed. Only after coders apply the coding scheme to documents without noticing ambiguities is a “final” scheme ready to be applied to the data set.

Creating coding schemes is a rich literature, with contributions across social science fields. For more on the schemes—including how to assess coder agreement and practical guides to scheme creation—see Krippendorff (2004); Neuendorf (2002); Weber (1990) and the documentation available in the *ReadMe* software (Hopkins et al. 2010).

*Selecting Documents:* Ideally, training sets should be representative of the corpus. Supervised learning methods use the relationship between the features in the training set to classify the remaining documents in the test set. In fact, almost all classification methods implicitly assume that the training set is a random sample from the population of documents to be coded (Hand 2006). Given this assumption, it is not surprising that best performance comes from random sampling to obtain a representative sample of documents—either through simple random sampling or from a more complicated stratified design. This may seem obvious, but presents particular difficulty when all the data are not available at the time of coding: either because it will be produced in the future or because it has yet to be digitized.

Training sets also need enough documents to apply supervised methods accurately. Hopkins and King (2010) offer five hundred as a rule of thumb with one hundred documents probably being enough. This is generally useful guidance, though it can be dangerous to apply a strict rule when selecting the number of documents. The number necessary will depend upon the specific application of interest. For example, as the number of categories in a coding scheme increases, the number of documents needed in the training set also increases. Supervised methods need enough information to learn the relationship between words and documents in *each* category of a coding scheme.<sup>1</sup>

### 5.2.2 Applying a supervised learning model

After hand classification is complete, the hand-labeled documents are used to train the supervised learning methods to learn about the test set—either classifying the individual documents into categories or measuring the proportion of documents in each category. The methods to do this classification are diverse, though they share a common structure that usefully unifies the methods (Hastie, Tibshirani, and Friedman 2001).

To see this common structure, suppose there are  $N_{\text{train}}$  documents ( $i = 1, \dots, N_{\text{train}}$ ) in our training set and each has been coded into one-of- $K$  categories ( $k = 1, \dots, K$ ). Each document  $i$ 's category is represented by  $Y_i \in \{C_1, C_2, \dots, C_K\}$  and the entire training set is represented as  $\mathbf{Y}_{\text{train}} = (Y_1, \dots, Y_{N_{\text{train}}})$ . Recall from Section 4 that each document  $i$ 's features are contained in an  $M$  length vector  $\mathbf{W}_i$ , which we collect in the  $N_{\text{train}} \times M$  matrix  $\mathbf{W}_{\text{train}}$ . Each supervised learning

---

<sup>1</sup>Determining the number of documents necessary to code documents can be challenging, but this need not be a blind search: the validation schemes that we use in Section 5.2.3 can be applied to determine the return on coding more documents.

algorithm assumes that there is some (unobserved) function that describes the relationship between the words and the labels,

$$\mathbf{Y}_{\text{train}} = f(\mathbf{W}_{\text{train}}).$$

Each algorithm attempts to learn this relationship—estimating the function  $f$  with  $\hat{f}$ .  $\hat{f}$  is then used to infer properties of the test set,  $\hat{\mathbf{Y}}_{\text{test}}$ —either each document’s category or the overall distribution of categories—using the test set’s words  $\mathbf{W}_{\text{test}}$ ,

$$\hat{\mathbf{Y}}_{\text{test}} = \hat{f}(\mathbf{W}_{\text{test}}).$$

We now overview three methods for inferring the relationship between words and categories.

*Individual Methods:* To provide a sense of how individual classification algorithms work, we present in detail a canonical example—the Naïve Bayes classifier (Maron and Kuhns 1960). The model has a simple, but powerful, approach to learning the relationship between words and categories. The training set is used to learn about the distribution of words for documents from category  $k$ . This distribution is then used to classify each of the documents in the test set. To perform this classification, Naïve Bayes begins with Bayes’s rule. The goal is to infer the probability that document  $i$  belongs to category  $k$  given word profile  $\mathbf{W}_i$ . Applying Bayes’s rule,

$$p(C_k|\mathbf{W}_i) \propto p(C_k)p(\mathbf{W}_i|C_k) \quad (1)$$

where we drop  $p(\mathbf{W}_i)$  from the denominator since we know that it is a constant across the different categories. [Equation \(1\)](#) shows that to estimate  $p(C_k|\mathbf{W}_i)$ , we need estimates of  $p(C_k)$  and  $p(\mathbf{W}_i|C_k)$ . If the documents in the training set are representative of the corpus, then the maximum likelihood estimate of  $p(C_k)$  is straightforward:

$$\hat{p}(C_k) = \frac{\text{Number Train Docs in Category } k}{N_{\text{train}}}$$

or the proportion of documents from the training set in each category. Estimating  $p(\mathbf{W}_i|C_k)$  is more challenging, because of the large number of words used to represent each text. For even moderately sized texts this implies that any one vector of word counts  $\mathbf{W}_i$  will only appear in a corpus once—implying that  $\hat{p}(\mathbf{W}_i|C_k) = 0$  for all  $i$  observations in the test set and all  $k$  categories. Therefore, without additional assumptions, the model is useless for classification. To make a useful estimation of  $p(\mathbf{W}_i|C_k)$  possible, we introduce the “naïve” assumption in the model: the model assumes that, given a document’s category, the words are generated independently,

$$p(\mathbf{W}_i|C_k) = \prod_{i=1}^M p(W_{im}|C_k).$$

Of course, this assumption must be wrong: the use of words is highly correlated any data set. But even though the assumption is wrong, the model is still able to capture information in texts useful for classification. Using this assumption, estimation of  $p(W_{im}|C_k)$  is straightforward,

$$\hat{p}(W_{im} = j|C_k) = \frac{\text{Number Train Docs in Category } k \text{ and With Word } m \text{ Used } j \text{ times}}{\text{Number Train Docs in Category } k}.$$

This simplified model still presents challenges, because some word-specific counts never occur in the data set. The common solution is to add a small amount to each probability, which is usually justified using a Bayesian Dirichlet-Multinomial model (Jurafsky and Martin 2009). Using the estimated components of the right-hand side of [equation \(1\)](#), Naïve Bayes then assigns each document in the test set to the document with the highest probability. Therefore, the estimated classifier for Naïve Bayes,  $\hat{f}$  is,

$$\hat{f}(\mathbf{W}_i) = \arg \max_k \left[ \hat{p}(C_k) \prod_{i=1}^M \hat{p}(W_{im}|C_k) \right]$$

The Naïve Bayes's classifier conforms neatly to our first principle: although the model is clearly wrong—of course features are not conditionally independent—it has proven to be a useful classifier for a diverse set of tasks (Hastie, Tibshirani, and Friedman 2001). But the Naïve Bayes's classifier is just one example of a very large and diverse literature, including Random Forests (Breiman 2001), Support Vector Machines (Venables and Ripley 2002), and neural networks (Bishop 1995).

*Ensembles:* On their own, individual methods for classification can provide accurate and reliable classifications. But it is also straightforward to combine classification models to produce a classifier which is superior to any of the individual classifiers. Heuristically, as long as the classifiers are accurate and diverse, combining the classifiers will improve accuracy (Jurafsky and Martin 2009). But ensembles are also useful for other reasons, including: increased out-of-sample stability and the ability to capture complex functional forms with relatively simple classifiers (Dietterich 2000; Hillard, Purpura, and Wilkerson 2008). Schemes for developing ensembles are diverse. For example, *Super-learning* uses cross-validation to assign weights to methods proportional to their out-of-sample accuracy (van der Laan, Polley, and Hubbard 2007). Additional methods include bagging—repeatedly drawing a sample with replacement of the training data and classifying the out of sample cases—and boosting—sequential training of classifiers increasing weight on misclassified cases (Hastie, Tibshirani, and Friedman 2001).

*Measuring Proportions:* A different way to improve the results is to change the quantity of interest. For many social science applications, only the proportion of documents in a category is needed—not the categories of each individual document. Shifting focus to estimating proportions,  $P(C)$ , can lead to substantial improvements in accuracy—even if the documents are not randomly sampled from the corpus (Hopkins and King 2010). The result is that *ReadMe* can provide reliable estimates of proportions across many domains and applications. To introduce *ReadMe*, we first modify the recipe described in Section 4, including an indicator of whether a word occurred in a document, rather than counts of the words. Using this representation, define a test-set-specific probability distribution over all possible documents,  $p(W_{\text{test}})$ . Without further assumptions, the data-generating process for the test set can be written as,

$$p(W_{\text{test}}) = p(W_{\text{test}}|C_{\text{test}})p(C_{\text{test}}) \quad (2)$$

where  $p(W_{\text{test}}|C_{\text{test}})$  is the distribution of documents in the test set conditional on categories and  $p(C_{\text{test}})$  is the proportion of documents in each class in the test set—the quantity of interest. The most important insight is that solving for  $p(C_{\text{test}})$  is simple if  $p(W_{\text{test}})$  and  $p(W_{\text{test}}|C_{\text{test}})$  are known. Of course, learning either quantities is the most challenging components of supervised learning: both quantities are high dimensional. One solution—used in Naïve Bayes—is to assume the words are generated independently. Hopkins and King (2010) avoid this assumption by employing matrix smoothing: this preserves higher order correlations between words, while also ensuring that the estimated probabilities are useful for learning from the training set. With this approach,  $\hat{p}(W_{\text{test}})$  can be estimated without coding any documents—it is inferred directly from the test set. Inferring  $\hat{p}(W_{\text{test}}|C_{\text{test}})$  requires labeled documents—which are unavailable for the test set. But if we assume that the conditional distributions are identical in the training and test sets, then we can substitute  $\hat{p}(W_{\text{test}}|C_{\text{test}})$  with  $\hat{p}(W_{\text{train}}|C_{\text{train}})$ . Heuristically,  $\hat{f}$ , used to estimate the proportion of documents in each category, is given by

$$\hat{f}(W_{\text{test}}) = (\hat{p}(W_{\text{train}}|C_{\text{train}})' \hat{p}(W_{\text{train}}|C_{\text{train}}))^{-1} \hat{p}(W_{\text{train}}|C_{\text{train}})' \hat{p}(W_{\text{test}}).$$

The move to proportions in Hopkins and King (2010) pays substantial dividends for inference—weakening assumptions used in other methods and reducing the potential bias systemic in other methods. But the *ReadMe* algorithm is not an option for all users, particularly those who have few documents or need to stratify the results by some external covariate. If there are a large number of documents (in excess of 100 at least) for each value of the covariate, the algorithm can simply be rerun on each strata, but otherwise it is necessary to return to individual classification.

**Table 2** Confusion matrix: comparing human and supervised coding

		Training data		
		Restrained	Activist	Neutral
Machine	Restrained	111	31	28
	Activist	10	17	0
	Neutral	26	9	68

### 5.2.3 Validation

Supervised methods are designed to automate the hand coding of documents into categories or measuring the proportion of documents in categories. If a method is performing well, it will directly replicate the hand coding. If it performs poorly, it will fail to replicate the coding—instead introducing serious errors. This clear objective implies a clear standard for evaluation: comparing the output of machine coding to the output of hand coding. The ideal validation procedure would divide the data into three subsets. Initial model fitting would be performed on the training set. Once a final model is chosen, a second set of hand-coded documents—the validation set—would be used to assess the performance of the model. The final model would then be applied to the test to complete the classification.

This approach to validation is difficult to apply in most settings. But *cross-validation* can be used to replicate this ideal procedure (Efron and Gong 1983; Hastie, Tibshirani, and Friedman 2001). In  $V$ -fold cross-validation, the training set is randomly partitioned into  $V$  ( $v = 1, \dots, V$ ) groups. The model's performance is assessed on each of the groups, ensuring all predictions are made on data out of sample. For each group  $v$ , the model is trained on the  $V - 1$  other groups, then applied to the  $V$ -th group to assess performance. Cross-validation is extremely powerful—it avoids overfitting by focusing on out-of-sample prediction and selects the best model for the underlying data from a set of candidate models (this is known as the Oracle property) (van der Vaart, Dudoit, and van der Laan 2006).<sup>2</sup>

### 5.2.4 Applying supervised learning methods: Russian military discourse

In order to demonstrate the use and validation of supervised learning methods, we adapt an example from Stewart and Zhukov (2009). To test a broader argument, Stewart and Zhukov (2009) compare the stances on foreign policy activism that civilian and military elites articulate in their public statements. They collect a corpus of 7920 Russian language public statements by political and military elites made between 1998 and 2008. Then following close reading of many of the documents, they develop a codebook to describe coding rules so that human coders could classify statements as having a restrained, activist, or neutral position on the Russian use of force.

Stewart and Zhukov (2009) then randomly sample and code three hundred documents. This seemingly low number is due to an additional constraint: finding and paying Russian-speaking coders substantially raised the costs of coding additional documents. With the set of human codings from Stewart and Zhukov (2009), we fit a *Random Forest* model to first learn the relationship between words and classes and then apply this relationship to classify the remaining documents.

To assess how well the Random Forest algorithm was able to replicate human coders, we performed a ten-fold cross-validation using the training data. This facilitates a direct comparison of machine and human classifications.

To summarize the model performance, Table 2 presents a *confusion matrix*. The rows of Table 2 are the out-of-sample codes from the Random Forest algorithm, the columns are the human-produced codes, and each cell entry counts the number of documents that received the

<sup>2</sup>These properties only apply when all steps (including selection of relevant predictors) are handled within the training set of the cross-validation and not within the full-labeled data. See Section 7.10.2 of Hastie, Tibshirani, and Friedman (2001) for more information on the proper application of cross-validation.

**Table 3** Document classifications by Elite type (proportion in parentheses)

		<i>Military</i>	<i>Political</i>
<i>Training set</i>	Restrained	27 (0.36)	119 (0.53)
	Activist	25 (0.34)	32 (0.14)
	Neutral	22 (0.30)	74 (0.33)
<i>Test set</i>	Restrained	870 (0.41)	3550 (0.62)
	Activist	500 (0.24)	260 (0.04)
	Neutral	749 (0.35)	1960 (0.34)

corresponding Random Forest and human codes. So, for example, the top-left cell counts 111 statements that the supervised method and the human agreed were both restrained. The goal of the supervised learning method is to place as many documents as possible in the on-diagonal cells—or replicate the human coding.

We use three statistics to summarize the confusion matrix. The first, and most widely used is *accuracy*—the proportion of correctly classified documents. Computing we find an overall accuracy of,

$$\text{Accuracy} = \frac{\text{No. Doc. On Diagonal}}{\text{Total No. Doc.}} = 0.65.$$

This is not an extremely accurate classifier, though it is comparable to accuracy levels for complex categories found in other studies (Hopkins and King 2010). It is also common to measure *precision* for a category—given that the machine guesses category  $k$ , what is the probability that the machine made the right guess. Operationally, this is estimated as number of documents correctly classified into category  $k$ , divided by the total number of documents the machine classifies as category  $k$ . For restrained foreign policy positions, the precision is 0.65. The final statistic is *recall* for category  $k$ —given that a human coder labels a document as belonging to category  $k$ , what is the chance the machine identifies the document. This is calculated by taking the number of correctly classified category  $k$  documents divided by the number of human coded documents in category  $k$ . For the restrained category, the recall is 0.75. The differences between the precision and recall exemplify why the different statistics are useful. The recall rate is higher than the precision here because the Random Forest algorithm guesses *too often* that a document is restrained. The result is that it labels a large portion of the human coder’s restrained positions correctly. But it also includes several documents that humans label differently.

Depending on the application, scholars may conclude that the supervised method is able to sufficiently replicate human coders. Or, additional steps can be taken to improve accuracy, including: applying other methods, using ensembles of methods, or switching the quantity of interest to proportions.<sup>3</sup> Given the limited space for this demonstration, after validation we move forward and apply the Random Forest classifier to the full set of civilian and political elite statements. The results are presented in Table 3. Stewart and Zhukov (2009) use similar results to show that Russian military elites are actually *more* activist in considering the use of force than their political counterparts, in contrast to the conventional wisdom (Gelpi and Feaver 2002).

## 6 Discovering Categories and Topics

Supervised and dictionary methods assume a well-defined set of categories. In some instances this poses no real challenge: researchers have a set of categories in mind before collecting texts, either from prior scholarship or a set of hypotheses that form the core of a research project. In other instances, however, the set of categories may be difficult to derive beforehand. For example scholars may struggle to identify the relevant topics of discussion in Senate floor speeches in 1887 or the subject of the daily briefings on the Falklands War.

<sup>3</sup>We provide suggestions for improving supervised learning performance in the Supplementary Material.

This difficulty is due in part to the massive number of potential organizations of even a small number of texts. Consider, for example, the number of ways to partition a set of objects—divide objects into a set of mutually exclusive and exhaustive groups. For any one collection of texts, the set of all possible partitions contains all possible substantively interesting ways to organize the texts (along with many uninteresting ways to organize texts). But enumerating these partitions is impossible. For even moderately sized data sets, say one hundred documents, the number of potential partitions is much larger than the estimated number of atoms in the universe.

*Unsupervised* learning methods are a class of methods that learn underlying features of text without explicitly imposing categories of interest. Rather than requiring users to condition on known categories beforehand—supervising the methods—unsupervised learning methods use modeling assumptions and properties of the texts to estimate a set of categories and simultaneously assign documents (or parts of documents) to those categories.

Unsupervised methods are valuable because they can identify organizations of text that are theoretically useful, but perhaps understudied or previously unknown. We divide unsupervised categorization methods into two broad categories (Grimmer and King 2011). The most widely used are FAC methods: methods that return a single *clustering* of the input objects. FAC methods are useful, but are *necessarily* model dependent.

Completely resolving the necessary model dependence in unsupervised models is impossible. But two strategies may prove useful at including additional information to make the models more flexible. The first strategy generalizes FAC models, using recently developed statistical models to incorporate context-specific structure into the analysis through a model. Including this information often leads to more interesting clusterings, but necessarily relies on small variations of similar models. The second strategy, CAC, allows researchers to efficiently search over millions of potential categorization schemes to identify interesting or useful organizations of the text. This embraces unsupervised learning methods as a method for generating new categorization schemes, but requires extensive additional analysis to classify all texts into categories.

Regardless of the strategy used to create clusterings, we should still view the output of the clustering methods with skepticism. All the clustering methods are based on incorrect models of language and *a priori* it is hard to know if any one method will provide substantively useful clusterings. Before using a clustering, validating the clustering is essential for demonstrating that the output of an unsupervised method is useful.

The need to validate clusterings does not negate the value of unsupervised methods, nor does it lead to them becoming a special case of supervised learning methods (as suggested in Hillard, Purpura, and Wilkerson 2008). As we show in Section 6.4, validations are done *conditional* on the clustering produced: without first seeing the clustering, assessing validity is impossible. The new organization scheme and documents classified according to those categories is the contribution of the method, not the time difference between applying supervised and unsupervised methods.

In fact, a recent debate in political science casts unsupervised and supervised as competitor methods (e.g., Hillard, Purpura, and Wilkerson 2008; Quinn et al. 2010). This debate is misplaced: supervised and unsupervised methods are different models with different objectives. If there are predetermined categories and documents that need to be placed in those categories, then use a supervised learning method. Using an unsupervised learning method to accomplish this task is at best frustrating—particularly if the predetermined categories are intended to capture subtle differences in tone or sentiment. If, however, researchers approach a problem without a predetermined categorization scheme, unsupervised methods can be useful. But supervised methods will never contribute a new coding scheme.

Far from competitors, supervised and unsupervised methods are most productively viewed as complementary methods, particularly for new projects or recently collected data sets. The categories of interest in a new project or a new corpus are usually unclear or could benefit from extensive exploration of the data. In this case, unsupervised methods provide insights into classifications that would be difficult (or impossible) to obtain without guidance. Once the unsupervised method is fit, we show below how supervised learning methods can be used to validate or generalize the findings.

## 6.1 FAC

We begin with FAC. This class of methods appears in an extensive literature across numerous fields, where they are used to estimate categories from data (for a review, see [Manning, Raghavan, and Schütze 2008](#) and [Jain, Murty, and Flynn 1999](#)). We consider two broad classes of FAC models: single membership and mixed membership models.

### 6.1.1 Single membership models

Single membership clustering models estimate a *clustering*: a partition of documents into mutually exclusive and exhaustive groups. Each group of documents in a clustering is a *cluster*, which represents an estimate of a *category*.<sup>4</sup> Across models,  $C_i$  will represent each document's cluster assignment and  $\mathbf{C} = (C_1, C_2, \dots, C_N)$  will represent a *partition* (clustering) of documents.

The FAC literature is *massive*, but each algorithm has three components: (1) a definition of document similarity or distance; (2) an objective function that operationalizes an *ideal* clustering; and (3) an optimization algorithm.

To build intuition, we introduce in detail the K-Means algorithm—perhaps the most widely used FAC method ([MacQueen 1967](#)). The goal of the K-means algorithm is to identify a partition of the documents that minimizes the squared Euclidean distance within clusters. To obtain this goal, the algorithm produces two quantities of interest: (1) a partition of the documents into  $K$  clusters ( $k = 1, \dots, K$ ) and (2)  $K$  cluster centers  $\mu_k$ . We now describe the three components of the K-means algorithm.

*Distance:* Standard  $K$ -means assumes the distance of a document  $W_i$  from a cluster center  $\mu_k$  is given by squared Euclidean distance,

$$d(W_i, \mu_k) = \sum_{m=1}^M (W_{im} - \mu_{km})^2.$$

Many other distance metrics are possible, and each will lead to different clusterings. Further, different *weights* can be attached within a distance metric. For example, scholars commonly use tf-idf weights within a Euclidean distance metric.

*Objective Function:* Heuristically, a “good” clustering under K-means is a partition where every document is close to its cluster center. Formally, this can be represented with the objective function,

$$\begin{aligned} f(\boldsymbol{\mu}, \mathbf{C}, \mathbf{W}) &= \sum_{i=1}^N \sum_{k=1}^K I(C_i = k) d(W_i, \mu_k) \\ &= \sum_{i=1}^N \sum_{k=1}^K I(C_i = k) \left( \sum_{m=1}^M (W_{im} - \mu_{km})^2 \right), \end{aligned} \quad (3)$$

where  $I(C_i = k)$  is an indicator function, equal to 1 if its argument is true. [Equation \(3\)](#) measures the quality of a clustering and set of cluster centers: the sum of all document's distance from their corresponding cluster centers.

*Optimization Method:* For any one data set, a distance metric and objective function identify an optimal partition of the data,  $\mathbf{C}^*$  and cluster centers  $\boldsymbol{\mu}^*$ . But directly identifying this optimum is extremely difficult—the K-means’ objective function is multimodal and non-continuous. Therefore,

<sup>4</sup>Our review of clustering algorithms is necessarily limited. For example, we make no mention of the distinction between hierarchical and nonhierarchical clustering algorithms ([Manning, Raghavan, and Schütze 2008](#)). We avoid this distinction between hierarchical clustering algorithms and clustering algorithms that estimate several partitions. As a result, there is a direct mapping from the output of hierarchical clustering algorithms to the partitions we describe here. We also avoid the distinction between soft and hard clustering. Our experience with soft clustering is that these examples usually assign most of a document to essentially one category, resulting in essentially a hard clustering. We do consider mixed membership models below, another form of soft clustering.

K-means, like many other FAC algorithms, employs an approximate and iterative optimization method. The standard K-means optimization algorithm proceeds in two steps. To begin, suppose that we have an initialized set of cluster centers  $\mu^{t-1}$ . The first step updates each document's assigned cluster to the closest cluster center,

$$C_i^t = \arg \min_k \sum_{m=1}^M (W_{im} - \mu_{km})^2.$$

Using the new cluster assignments,  $C^t$ , each cluster center  $\mu_k$  is updated by setting it equal to the *average* document assigned to the cluster

$$\mu_k^t = \frac{\sum_{i=1}^N I(C_i^t = k) W_i}{\sum_{i=1}^N I(C_i^t = k)}.$$

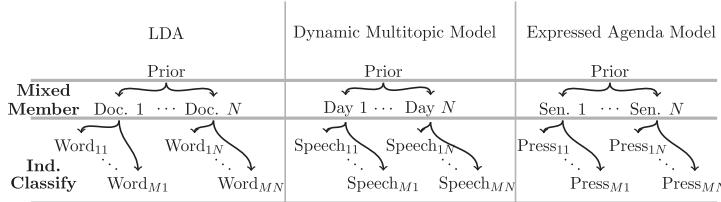
The steps are repeated until the change in objective function falls below a predetermined threshold. This algorithm for optimization—closely related to the Expectation–Maximization algorithm—only guarantees convergence to a local minimum (Dempster, Laird, and Rubin 1977; Bishop 2006) and therefore several random restarts are often necessary to find the global optimum. The K-means algorithm is just one of many potential algorithms for generating partitions of data. The literature is replete with alternative clustering methods. Some of the models vary the distance metric (Manning, Raghavan, and Schütze 2008), suggest different objective functions (Ng, Jordan, and Weiss 2001), and even numerous different optimization procedures (Frey and Dueck 2007). Some of the models do this explicitly with an algorithmic focus, while others vary the three components of the model implicitly in a statistical model for generating new clusterings. See Grimmer and King (2011) and Manning, Raghavan, and Schütze (2008) for more details on the diverse set of methods. While each of the clustering methods are well specified and based on careful derivation, each method relies on a set of assumptions that result in varying performance in different contexts. Unfortunately, the vast literature on clustering algorithms provides little guidance on when specific measures of similarity, objective functions, or optimization algorithms are likely to lead to useful or insightful partitions of data. Indeed, it appears that providing this guidance without specific knowledge of the data set may be impossible (Grimmer and King 2011). Therefore, we think any new clustering methods developed in other fields should be imported into political science with caution. Regardless of the claims made, substance-based validations, as outlined in Section 6.4, are necessary to establish the utility of an unsupervised method.

### 6.1.2 Mixed membership models

One way to improve the output of single-membership models is to include additional and problem-specific structure. *Topic models* have recently been proposed as one method for including this structure (Blei, Ng, and Jordan 2003). Topic models are a broad class of Bayesian generative models that encode problem-specific structure into an estimation of categories (see Blei 2012 for a review).

Topic models share two broad characteristics. The first is a definition of a topic. Statistically, a *topic* is a probability mass function over words. For a topic  $k$  ( $k = 1, \dots, K$ ), we represent this probability distribution over words with an  $M \times 1$  vector  $\theta_k$  where  $\theta_{mk}$  describes the probability the  $k$ -th topic uses the  $m$ -th word. Substantively, topics are distinct concepts. In congressional speech, one topic may convey attention to America's involvement in Afghanistan, with a high probability attached to words like *troop*, *war*, *taliban*, and *Afghanistan*. A second topic may discuss the health-care debate, regularly using words like *health*, *care*, *reform*, and *insurance*. To estimate a topic, the models use the co-occurrence of words across documents.

Second, the models share a basic hierarchical structure. The top of the model contains a prior which borrows information across units. In the middle of the structure is a *mixed membership*



**Fig. 2** A common structure across seemingly unrelated text models: in each case, the mixed member allows the scholar to estimate the parameter of particular substantive interest. The prior facilitates the borrowing of information across units, and the individual elements classified form the observed data.

level: this measures how a unit of interest allocates its attention to the estimated topics. And at the bottom of the hierarchy a word or document is assigned to a *single* topic.

The first and most widely used topic model, latent Dirichlet allocation (LDA), is one example of how this structure is used in a statistical model (Blei, Ng, and Jordan 2003). The left-hand side of Fig. 2 provides a nontechnical overview of LDA's data generation process.

LDA assumes that each *document* is a mixture of topics. For each document,  $i$  represent the proportion of the document dedicated to topic  $k$  as  $\pi_{ik}$  and collect the proportions across topics to be  $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{iK})$ . We will suppose that each document's proportions are drawn from a common Dirichlet prior,

$$\boldsymbol{\pi}_i \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

where  $\boldsymbol{\alpha}$  represent the Dirichlet distribution's shape parameters.

Within each document, the words are drawn according to the distribution of topics. Suppose that a document contains  $N_i$  total words ( $j = 1, \dots, N_i$ ). LDA assumes that a two-step process generates each word. To obtain the  $j$ -th word in the  $i$ -th document, the first step is to draw its topic  $\tau_{ij}$ ,

$$\tau_{ij} \sim \text{Multinomial}(1, \boldsymbol{\pi}_i).$$

Conditional on topic assignment, the actual word is drawn: if the  $j$ -th word in the  $i$ -th document is assigned to the  $k$ -th topic, then we draw from the corresponding topic,

$$W_{ij} \sim \text{Multinomial}(1, \theta_k).$$

*Topic Models in Political Science:* Political scientists have extended topic models so that the parameters correspond to politically relevant quantities of interest. The first topic model in political science is the dynamic multitopic model (Quinn et al. 2010), used to model the daily attention to topics in Senate floor speeches (center panel of Fig. 2). Following Quinn et al. (2010) is the *expressed agenda model* (Grimmer 2010), which measures the attention senators allocate to press releases (right-hand panel of Fig. 2). There are numerous novel features of both models that are missed when considered abstractly, but we will see that considering the models abstractly leads to a useful realization about the common elements of all three models. The dynamic multitopic model in Quinn et al. (2010) supposes that each day in the Senate is a mixture of attention to topics: days are at the mixed membership level. Each speech is then assigned to a single topic, analogous to assigning individual words to topics in LDA. And finally, a dynamic prior is used to make more efficient inferences about the proportion of each day's speeches allocated to each topic. The center panel of Fig. 2 shows the similarity of this model with LDA. The expressed agenda model in Grimmer (2010) demonstrates yet another way to exploit the same structure. The expressed agenda model is designed to measure how authors divide attention over topics and is applied to model how senators explain work to constituents in press releases. The key assumption is that each senator divides her attention over the set of topics: senators are at the mixed membership level in the model. Then, conditional on the senator's mixture of attention to topics, the topic of each press release is drawn.

The right-hand panel of Fig. 2 shows the similarity with LDA and the dynamic topic model. Like days in the dynamic topic model and documents in LDA, senators in the expressed agenda model are allowed to be members of several topics. And analogous to speeches in the dynamic topic model and words in LDA, each press release is assigned to a single topic and the prior is used to make more efficient inferences.

The similarity in structure across the three models in Fig. 2 demonstrates how statistical topic models can facilitate the production of substantively interesting and new models. Situating the models in a data-generating process makes including additional structure straightforward. For example, we could combine the dynamic topic model and the expressed agenda model to model how senators' attention to topics vary over time. More generally, we could include other problem-specific information, like the order of speeches made in the Senate in any one day, the sequence of ideas exchanged during a press conference, or covariates about the authors of texts. But requiring political scientists to tune each model to their task is a daunting task. The result is that only a few models are likely to be used for analyzing texts. Further, while the statistical structure allows for the inclusion of problem-specific information, the model is still limited to a single family of probability distributions and therefore a specific set of assumptions. CAC methods provide a way to explore these different assumptions.

## 6.2 CAC

Identifying the assumptions that will lead to a useful clustering of documents in a particular data set beforehand is difficult. After the fact, however, it is much easier to assess whether an organization of texts is useful within a particular context. With this idea in mind, CAC is a method for efficiently searching over a large space of clusterings (Grimmer and King 2011). To start, a diverse set of FAC methods is applied to a data set. The different methods vary the definition of similarity, objective functions, and optimization algorithms to provide diverse ways to organize the documents. Then, Grimmer and King (2011) show how to embed the partitions into a two-dimensional space such that two clusterings are close in the space if they organize the documents in similar ways. Using this space, Grimmer and King (2011) introduce a method for exploring it, easily searching over the included methods and millions of other clusterings that are the result of combinations of similar organizations of the data.

CAC has three broad uses. First, the method provides a way to identify new or understudied concepts—even in already extensively analyzed texts. The diverse set of partitions included ensures that researchers will encounter new ways of organizing data. Second, CAC provides an accelerated way to explore new collections of texts. CAC ensures that these explorations are not limited to only one organization of the data. Third, CAC provides a way to evaluate the originality of new clustering methods. If a clustering method is truly new, then it should occupy a distinct part of the space—at least for some collections. If it is useful, then the distinct clusterings should lead to new or useful organizations of the documents that other methods miss.

But CAC does place substantial burden on users to identify clusterings that are useful. In one respect, this is a virtue of the method—it is limited only by human creativity. But, it does imply that the method will only perform as well as the person conducting the exploration and interrogation of the space.

## 6.3 Setting the Number of Clusters

Both CAC and FAC methods require setting the number of clusters in a model. For K-means the number of clusters— $K$ —has to be set, for mixed membership models the number of topics must be chosen, and for CAC the number of clusters in the final clustering must be determined. Determining the number of clusters is one of the most difficult questions in unsupervised learning. Some methods attempt to eliminate this decision and estimate the number of features (Frey and Dueck 2007; Wallach et al. 2010), but recent studies show that the estimated number of clusters is strongly model dependent (Wallach et al. 2010). We also cannot turn to fit statistics, as Chang et al. (2009) show

that there is often a negative relationship between the best-fitting model and the substantive information provided.

When setting the number of clusters, we caution in general that *you can't get something for nothing*. Models that estimate the number of clusters are heavily model dependent (Wallach et al. 2010). Nonparametric models, such as the Dirichlet process prior, still make model-based decisions on the number of clusters or topics to include. But the choice has been reparameterized as a hyper prior (Wallach et al. 2010) or as a tuning parameter in an algorithm (Frey and Dueck 2007).

Rather than statistical fit, model selection should be recast as a problem of measuring *substantive fit*. Unsupervised methods for content analysis *reduce* the information in large text collections substantially. Measures of statistical fit measure how well the models fit by comparing the estimated parameters to the actual data. But this relies on the assumption that the goal is to model well the representation of texts after preprocessing. It is not. The preprocessed texts represent a substantial simplification of the documents. The goal is revelation of substantively interesting information.

We think a productive line of inquiry will replace the use of the preprocessed texts with carefully elicited evaluations based on the substance of the model. Quinn et al. (2010) provide one method for performing this model selection. At the first stage of the process, candidate models are fit varying the number of clusters or topics. At the second stage, human judgment is used to select a final model, assessing the *quality* of the clusters. This can be done approximately—assessing whether the clusters group together documents that are distinct from other clusters and internally consistent. Or an explicit search across models based on elicited subject expert evaluations can be employed, using measures developed in Chang et al. (2009) or Grimmer and King (2011).

#### **6.4 Validating Unsupervised Methods: How Legislators Present Their Work to Constituents**

As Quinn et al. (2010) observe, unsupervised methods shift the user burden from determining categories before analysis to validating model output afterward. The post-fit validations necessary can be extensive, but the methods are still *useful* because they suggest new, or at least understudied, ways to organize the data. As an example of this process, we perform a validation of senators' expressed priorities introduced in Grimmer (2012). To obtain the measures of expressed priorities, Grimmer (2012) applies a version of the expressed agenda model to over 64,000 Senate press releases issued from 2005 to 2007.

Applying the unsupervised model leads to an understudied organization of senators, based on how they present their work to constituents. Grimmer (2012) shows that the primary variation underlying senators' expressed priorities is how senators balance position taking and credit claiming in their press releases. Some senators allocate substantial attention to articulating positions, others allocate much more attention to credit claiming, and still others adopt a more evenly balanced presentational style. Grimmer (2012) shows that this spectrum is theoretically interesting: it approximates spectra suggested in earlier qualitative work (Fenno 1978; Yiannakis 1982), and predicted in formal theoretic models of Congress and Congressional elections (Weingast, Shepsle, and Johnsen 1981; Ashworth and Bueno de Mesquita 2006). And Grimmer (2012) shows that where senators fall on this spectrum has real consequences for representation.

To make these inferences, however, requires extensive validation of both the estimated topics and expressed priorities. We overview some of those validations now—of both the topics and the expressed priorities. Before proceeding, we caution that the validations performed here are only a subset of the evaluations any researcher would need to perform before using measures from unsupervised methods in their own work. For a more comprehensive review of validity and topic models, see Quinn et al. (2010).

##### **6.4.1 Validating topics**

As a prelude to validating the topic output, the topics must be labeled: it must be determined *what* each topic measures. Table 4 provides three examples of the topics from Grimmer (2012), based on the broader forty-four topic model. One method for labeling is reading: sampling ten to fifteen documents assigned to a topic and inferring the commonality across the press releases. Examples of

**Table 4** An example of topic labeling

Description	Discriminating words
Iraq War	Iraq, iraqi, troop, war, sectarian
Honorary	Honor, prayer, remember, fund, tribute
Fire Department Grants	Firefight, homeland, afgp, award, equipment

the labels are posted in the first column of **Table 4**. Statistical methods are also used. While many have been proposed, they share the same intuition: identify words that are highly predictive of documents belonging to a particular topic. Examples of the *discriminating* words are also in **Table 4**

The topics in **Table 4** exemplify the utility of unsupervised learning. The first row of **Table 4** identifies a topic about the Iraq war—one of the most salient debates during the time the press releases were issued. But the model also identifies a topic of press releases that *honor* constituents—commemorating a national holiday or a tribute to a deceased constituent. The final row of **Table 4** is a topic about claiming credit for grants allocated to fire departments through the Assistance to Firefighter Grant Program (AFGP). This is a prominent type of credit claiming—bureaucratic agencies creating opportunities for legislators—often missed from standard models of legislative speech.

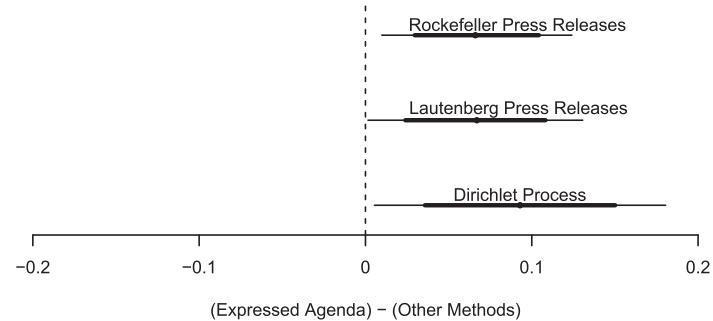
With the topic labels in hand for all forty-four topics, we now describe two methods for validating the topic labels. The first evaluation assesses *semantic validity* (Quinn et al. 2010): the extent to which our topics identify coherent groups of press releases that are internally homogeneous, yet distinctive from other topics. The second evaluation assesses *predictive validity* (Quinn et al. 2010): measuring how well variation in topic usage corresponds with expected events.

*Semantic Validation via Experiments:* To assess *semantic validity*, Grimmer and King (2011) introduce a method based on experimentally elicited human input. Using this method allows direct comparison of the quality of clusterings, but it is unable to provide an absolute measure of semantic validity. To assess the semantic validity of the topics, we compare them to two clusterings produced by Senate press secretaries and a clustering from a state-of-the-art nonparametric clustering method (Blei and Jordan 2006). Following Grimmer and King (2011), we first sample pairs of documents assigned to the same and different clusters. Research assistants were then asked to evaluate the pairs of documents on a three-point scale, rating a pair of press releases as (1) unrelated, (2) loosely related, or (3) closely related. We then average over the human evaluations to create the measure of cluster quality: the average evaluation of press releases assigned to the same cluster and less the average evaluation of press releases assigned to different clusters. For a method  $i$  call this  $\text{Cluster Quality}_i$ . We then compare alternative partitions to the clustering used in Grimmer (2012) to compare relative semantic validity, Cluster Quality<sub>ExpressedAgenda</sub> – Cluster Quality<sub>Alt.Method</sub>. The differences are plotted in **Fig. 3**, with each point representing the mean difference and the thick and thin lines constituting 80% and 95% credible intervals for the difference, respectively.

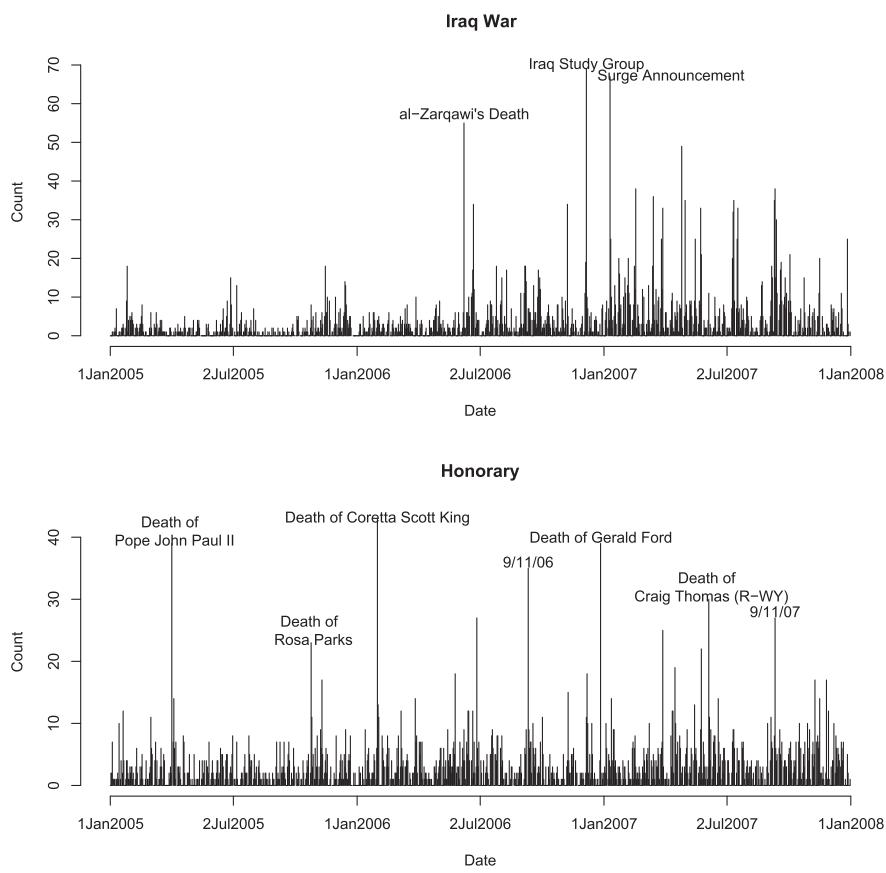
**Figure 3** shows that the expressed agenda model applied to the press releases produces a higher quality clustering—more semantically valid—than any of the comparison methods on the same press releases. This includes clusterings from the nonparametric method and clusterings produced by two Senate press secretaries.

*Predictive Validity:* Quinn et al. (2010) argue that if topics are valid, then external events should explain sudden increases in attention to a topic (Grimmer 2010 performs a similar validation). **Figure 4** plots the number of press releases issued each day about the Iraq war (top plot) and honorary press releases (bottom plot), with the date plotted on the vertical axis.

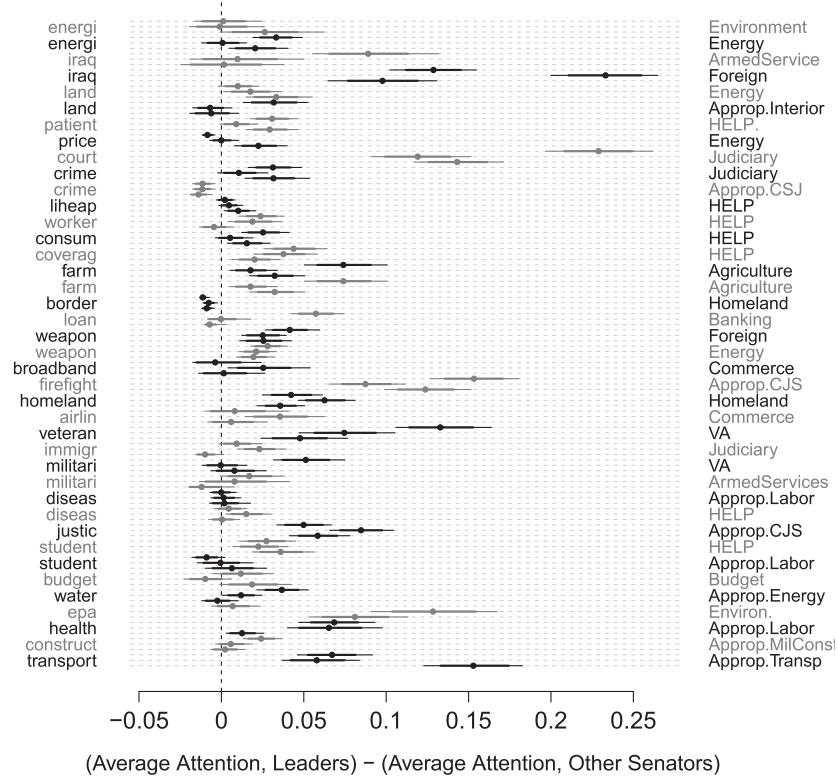
Both plots in **Fig. 4** show that external events predict spikes in attention to press releases. When major events happen surrounding the war—such as the release of the Iraq study group report—more press releases are issued about Iraq. Likewise, when major figures die, such as Pope John Paul II, more honorary press releases are issued.



**Fig. 3** Semantic validity of topics. This figure shows that the model applied in Grimmer (2012) provides higher quality clusters than press secretaries grouping press releases on senators' Web sites and state-of-the-art nonparametric topic models.



**Fig. 4** Predictive validity of topics.



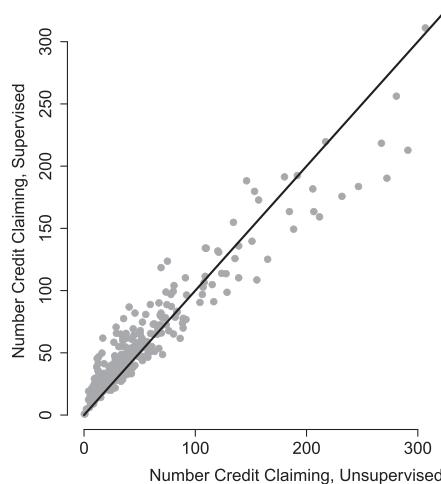
**Fig. 5** Predictive validity of expressed priorities. This figure compares the attention that Senate committee leaders—chairs or ranking members—dedicate to topics under their committee jurisdictions to the attention allocated by the rest of the Senate. The solid black dots represent the expected difference, the thick lines are 80% credible intervals, and the thin lines are 95% intervals. Along the left-hand vertical axis, the topics are listed, and, on the right-hand side, the corresponding committee names are listed. As this figure clearly illustrates, committee leaders allocate substantially more attention to issues under their jurisdiction than other members of Congress.

Downloaded from <http://pan.oxfordjournals.org/> at London School of economics on July 7, 2013

#### 6.4.2 Validating expressed priorities

The expressed agenda model also produces a measure of senators' expressed priorities: how senators divide their attention over the topics. This too requires rigorous validation—a demonstration that the estimated priorities measure the concept claimed. We apply two tests to the expressed priorities from Grimmer (2012). First, we show that they satisfy one test of *predictive validity*. Second, we show that the measures have convergent validity with supervised learning methods created after observing the categorization scheme from the unsupervised method.

*Predictive Validity:* Grimmer (2010) argues that the leaders of Senate committees—committee chairs and ranking members—should allocate more attention to issues that fall under the jurisdiction of their committee than other senators. This straightforward expectation provides a test of the predictive validity of the estimated priorities. Figure 5 carries out this comparison. In Fig. 5, committee leaders' average attention dedicated to an issue under their committee's jurisdiction is compared with the average attention among the other ninety-eight senators for forty committee topic pairs, for each year in the analysis. The left-hand vertical axis denotes the topics that were



**Fig. 6** Convergent validity of unsupervised methods with supervised methods.

used for the comparison and the right-hand vertical axis contains an abbreviated committee or appropriations subcommittee name. The solid dot represents the expected difference between committee leaders and the rest of the Senate; the thick lines are 80% and 95% credible intervals, respectively. If committee leaders discuss issues related to their committee more often, then the estimates should be to the right of the vertical dotted line at zero.

Figure 5 shows that committee leaders allocate more attention to issues under their committee's jurisdiction than the average senator. For almost every topic committee pair in each year, leaders of Senate committees allocate substantially more attention to issues under their jurisdiction than other senators. This suggests that the measures are, at least by this test, predictively valid.

*Convergent Validity:* Even after several validations, skepticism can remain about measures from an unsupervised method. This skepticism is perhaps most acute in political science, where unsupervised methods have a colored past (Armstrong 1967). This skepticism would be allayed—at least partially—if we were confident that an unsupervised measure was just as valid as the equivalent measure from a supervised method. Here, we provide one way to gain this confidence by using supervised methods to validate the output of unsupervised methods. For simplicity and space, we perform this validation on a subset of the full expressed priorities used in Grimmer (2012). Grimmer (2012) shows that several of the topics identify press releases that are claiming credit for money directed to the district (Mayhew 1974) (the differences across topics capture differences in the type of money claimed). Grimmer (2012) aggregates across categories to create an estimate of the number of press releases senators issue claiming credit for money in their state. We validate this measure using ReadMe. First, we developed a codebook that contained a category for claiming credit for money, along with other categories to improve our performance (Mayhew 1974). We then asked a research assistant to classify five hundred press releases according to our scheme. Then, we used those five hundred press releases and ReadMe to measure the number of credit-claiming press releases issued.<sup>5</sup> Figure 6 shows the strong correlation between the supervised and unsupervised methods. In it, each senator's estimated number of credit-claiming press releases are plotted against the estimate from the expressed agenda model. The black line in the plot is the 45° line: if the two estimates were equivalent, the points would fall along this line.

<sup>5</sup>ReadMe estimates proportions, so the number of press releases in each category can be retrieved by multiplying by the total number of press releases.

Figure 6 shows clearly that the gray points group around the  $45^\circ$  line, demonstrating that the unsupervised method is essentially equivalent to the supervised method. This is true across all senators, resulting in a correlation of 0.96. The unsupervised estimate of credit-claiming behavior is essentially *equivalent* to the estimate from ReadMe. This implies that all the confidence we would attach to using the estimates from the supervised method can also be attached to the unsupervised method. We caution that this validation *does not* obviate the need for unsupervised methods. The validation using the supervised method is possible *only after the unsupervised method suggests a classification scheme*. It does, however, provide one direct test to ensure that the output from an unsupervised method is just as valid, reliable, and useful as the categorization schemes from supervised methods. It also serves to underscore the point that future work based on a categorization scheme developed by an unsupervised method may well use supervised classification to extend and generalize that point.

## 7 Measuring Latent Features in Texts: Scaling Political Actors

One of the most promising applications of automated content analysis methods is to locate political actors in ideological space. Estimating locations using existing data is often difficult and sometimes impossible. Roll call votes are regularly used to scale legislators (Poole and Rosenthal 1997; Clinton, Jackman, and Rivers 2004), but outside the U.S. Congress roll call votes are less reliable (Spirling and McLean 2007). And other political actors—presidents, bureaucrats, and political candidates—do not cast votes. Other methods for scaling political actors have been developed (e.g., Gerber and Lewis 2004; Bonica 2011), but they rely on particular disclosure institutions that are often absent in other democracies.

But nearly all political actors speak. A method that could use this text to place actors in a political space would facilitate testing some of the most important theories of politics. We describe two methods for scaling political actors using texts. One method, based on Laver, Benoit, and Garry (2003), is a supervised method—analogous to dictionary methods—to situate actors in space based on their words. A second method is an unsupervised method for locating actors in space (Monroe and Maeda 2004; Slapin and Proksch 2008).

The scaling literature holds great promise for testing spatial theories of politics. Recognizing this, several recent papers have offered important technical contributions that improve the methods used to perform the scalings (Martin and Vanberg 2007; Lowe 2008; Lowe et al. 2011). These papers are important, but we think that the scaling literature would benefit from a clearer articulation of its goals. Recent papers have implicitly equated the goal of scaling methods as replicating expert opinion (Benoit, Laver, and Mikhaylov 2009; Mikhaylov, Laver, and Benoit 2010) or well-validated scalings made using nontext data (Beauchamp 2011). Certainly plausibility of measures is important, but if the goal is to replicate expert opinion, or already existent scalings, then text methods are unnecessary. Simple extrapolation from the experts or existing scaling would suffice.

One clear goal for the scaling literature could be *prediction* of political events. Beauchamp (2011), for example, shows that the output from text scaling methods can be used to predict votes in Congress. More generally, text scalings should be able to predict legislative coalitions throughout the policy creation process. Or when applied to campaigns, text scalings should be able to predict endorsements and campaign donations.

Improving the validation of scales will help improve current models, which rely on the strong assumption of *ideological dominance* in speech. Both supervised and unsupervised scaling methods rely on the strong assumption that actors' ideological leanings determine what is discussed in texts. This assumption is often useful. For example, Beauchamp (2011) shows that this works well in Senate floor speeches, and we replicate an example from Slapin and Proksch (2008) that shows that the model works well with German political platforms. But in other political speech, this may not be true—we show below that the ideological dominance assumption appears to not hold in Senate press releases, where senators regularly engage in nonideological credit claiming.

Scaling methods will have more even performance across texts if they are accompanied with methods that separate ideological and non-ideological statements. Some of this separation is now done manually. For example, it is recommended in Slapin and Proksch (2008). But more nuanced methods for separating ideological content remain an important subject of future research.

### 7.1 Supervised Methods for Scaling

Laver, Benoit, and Garry (2003) represent a true breakthrough in placing political actors in a policy space. Rather than rely on difficult to replicate and hard to validate manual coding or dictionary methods, Laver, Benoit, and Garry (2003) introduced a fully automated method for scaling political actors: *wordscores*.

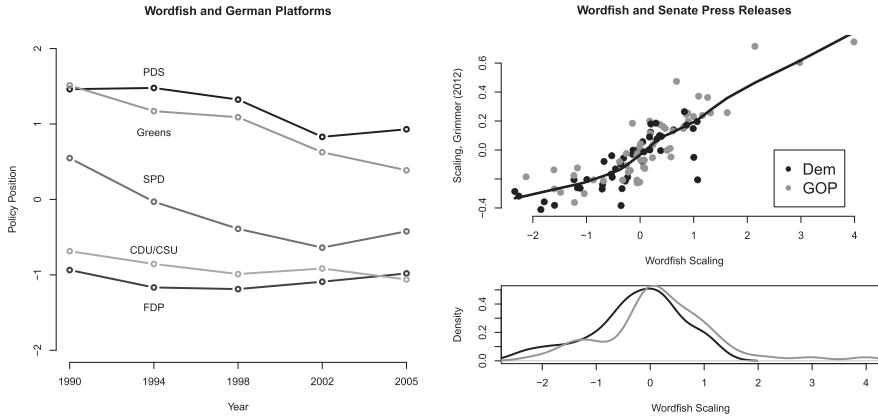
*Wordscores* is a special case of the dictionary methods that we presented in Section 5.1. The first step is the selection of *reference* texts that define the political positions in the space. In the simplest example, we may select two texts to define the liberal and conservative ends of the spectrum. If we wanted to scale U.S. senators based on their speeches, for example, we may define as a reference text all the speeches from a very liberal senator, like Ron Wyden (D-OR) or Barbara Boxer (D-CA), and a very conservative senator, like Tom Coburn (R-OK) or Jim DeMint (R-OK). The reference (training) texts are then used to generate a *score* for each word. The score measures the relative rate each word is used in the reference texts. This creates a measure of how well the word separates liberal and conservative members—one measure of whether a word is *liberal* or *conservative*. The word scores are then used to scale the remaining texts. Laver, Benoit, and Garry (2003) calls these the *virgin* texts, but in supervised learning we would call these texts the *test set*. To scale the documents using the word scores, first Laver, Benoit, and Garry (2003) calculate the relative rate words are used in each of the test documents. The position of the texts is then determined by taking the weighted average of the word scores of the words in a text, where the weights are given by the rate at which the words are used.

Wordscores is rich and generalizable to multiple dimensions and to include several reference texts. But facets of wordscores constrain the method and make it difficult to recommend for general use (see Lowe 2008 for an extended critique). By defining “liberal” and “conservative” using *only* reference texts, Laver, Benoit, and Garry (2003) conflate ideological language with stylistic differences across authors and impose the ideological dominance assumption on the texts. The result is that every use of wordscores will depend strongly on the reference texts that are used, in part because of stylistic differences across authors and in part because the reference texts will discuss non-ideological content. Careful preprocessing of texts, to remove words that are likely only stylistic, can mitigate part of this problem. Beauchamp (2011), for instance, shows that results are significantly improved by removing technical language, which coincides more with party power than with ideology. But no amount of preprocessing can completely eliminate it. The explicit adoption of a supervised learning approach might limit the influence of style substantially. Unfortunately, this also requires a substantial increase in effort and time, which makes it application unwieldy.

### 7.2 Unsupervised Methods for Scaling

Rather than rely on reference texts for scaling documents, unsupervised scaling methods *discover* words that distinguish locations on a political spectrum. First Monroe and Maeda (2004), then later Slapin and Proksch (2008), introduce statistical models based on *item response theory* (IRT) to automatically estimate the spatial location of the parties. Politicians are assumed to reside in a low-dimensional political space, which is represented by the parameter  $\theta_i$  for politician  $i$ . A politician’s (or party’s) position in this space is assumed to affect the rate at which words are used in texts. Using this assumption and text data, unsupervised scaling methods estimate the underlying positions of political actors.

Slapin and Proksch (2008) develop their method, *wordfish*, as a Poisson-IRT model. Specifically, Slapin and Proksch (2008) assume that each word  $j$  from individual  $i$ ,  $W_{ij}$  is drawn from a Poisson distribution with rate  $\lambda_{ij}$ ,  $W_{ij} \sim \text{Poisson}(\lambda_{ij})$ .  $\lambda_{ij}$  is modeled as a function of individual  $i$ ’s



**Fig. 7** Wordfish algorithm: performance varies across context.

loquaciousness ( $\alpha_i$ ), the frequency word  $j$  is used ( $\psi_j$ ), the extent to which a word discriminates the underlying ideological space ( $\beta_j$ ), and the politician's underlying position ( $\theta_i$ ),

$$\lambda_{ij} = \exp(\alpha_i + \psi_j + \beta_j \times \theta_i).$$

The next section applies this model to political texts from two different contexts, demonstrating conditions when the model is able to reliably retrieve underlying policy positions.

### 7.3 Applying Unsupervised Methods to Political Texts

Both the strength and limitations of IRT methods for scaling are the lack of supervision. When the model works well it provides reliable estimates of political actors' spatial locations with little resource investment. But the lack of supervision and the use of an IRT model implies that the model will seize upon the primary variation in language across actors. This might be ideological. Or, the differences across actors may be about their focus on policy or pork, the style in which the essays were written, or the tone of the statements. Because the model does not include supervision explicitly, it is difficult to guarantee that the output of the model will reliably identify the revealed ideological locations of political actors. It is worth emphasizing that this *is not* a shortcoming of wordfish. In fact, we will show below that non-ideological locations that wordfish identifies are quite useful. But this does suggest that one should not assume that wordfish output measures an ideological location without careful validation.

When the ideological dominance assumption fits the data the model can reliably retrieve valid ideological dimensions from political texts. Take, for example, the left-hand plot of Fig. 7. This replicates a plot in Fig. 1 of Slapin and Proksch (2008), who apply the wordfish algorithm to German party platforms. As Slapin and Proksch (2008) show, the estimates in the left-hand plot separate the German parties and replicate expert assessments of German party movement over time.

But when ideological dominance assumption fails to fit the data, wordfish fails to retrieve underlying policy dimensions. The right-hand plot applies the wordfish algorithm to Senate press-release data introduced in Grimmer (2012). The bottom right-hand plot in Fig. 7 is a density of the Democrats (black line) and Republican (gray line) positions from the wordfish algorithm. The model clearly fails to separate Democrat and Republican senators—a necessity for any valid scaling in the now polarized Senate.

The wordfish scaling is meaningful substantively, but it does not correspond to standard policy space. The top plot shows that the wordfish algorithm reproduces the spectrum Grimmer (2012)

identified using the expressed agenda model—how senators balance position taking and credit claiming in press releases. This plot presents the scaling from Grimmer (2012) against the scaling from wordfish on the horizontal axis and the black line is a lowess curve (Cleveland 1979). The relationship between the two measures is extremely strong—correlating at 0.86. Clear evidence that wordfish has identified this interesting—though nonideological—spectrum in the Senate press releases.

This exemplifies when scaling methods are likely to recover an ideological position. When political actors are engaging in heavily ideological speech—as in German party platforms—unsupervised methods appear to retrieve reliable position estimates. But when political actors can avoid ideological speech—as in Senate press releases—scaling methods retrieve some other, nonideological scaling. Therefore when applying scaling algorithms, careful validation is needed to confirm that the intended space has been identified. And an essential future area of future research will simultaneously isolate ideological statements and then employ those ideological statements to scale political actors.

## 8 Text as Data in Political Science

Automated content analysis methods provide a wide range of tools to measure diverse quantities of interest. This ranges from classifying documents—either into existing or yet to be determined categories—or scaling political actors into policy space. We emphasize that any one method's performance will be context specific. And because text analysis methods are necessarily *incorrect* models of language, the output always necessitates careful validation. For supervised classification methods, this requires demonstrating that the classification from machines replicates hand coding. For unsupervised classification and scaling methods, this requires validating that the measures produced correspond with the concepts claimed.

The automated content literature extends well beyond the methods discussed in this article. Textbooks produced for other fields provide excellent overviews of methods not discussed here, including natural language processing tools. (Manning, Raghavan, and Schütze 2008; Jurafsky and Martin 2009). We also recommend political science papers that make use of other methods not profiled here (e.g., Schrot 2000).

While there are many possible paths for this research to advance along, we identify three of the most important here.

*New Texts Need New Methods:* Perhaps the most obvious future research pursuit will be the development of new statistical models for text. Indeed, this is already actively underway within political science, complementing long-standing literatures in computer science, statistics, and machine learning. New text data in political science will necessitate the development of new methods. But as methodologists develop problem-specific tools, they should also think generally about their methods. Identifying commonalities will allow scholars to share creative solutions to common problems.

*Uncertainty in Automated Content Methods:* Measuring uncertainty in automated text methods remains one of the most important challenges. One of the greatest strengths of the quantitative treatment of text as data is the ability to estimate uncertainty in measurements. And there has been progress in measuring uncertainty, particularly in supervised classification methods. Hopkins and King (2010) show how *simulation-extrapolation* (SIMEX) can allow for uncertainty in the categories human coders place training documents. Similarly, Benoit, Laver, and Mikhaylov (2009) use SIMEX to include error in text-based scales into generalized linear models. But solutions across models are needed. This may take the form of characterizing full posterior distributions for Bayesian statistical models, determining fast and reliable computational methods for algorithmic models, or methods for including uncertainty generated when including humans in the analysis process.

*New Frontiers: New Texts and New Questions:* Beyond methodological innovation, there are vast stacks of texts that can now be analyzed efficiently using automated text analysis. From political theory, to law, to survey research, scholars stand to learn much from the application of automated text analysis methods to their domain of interest. Political scientists may also use texts to

accomplish tasks beyond those that we have highlighted here. Part of the utility of these texts is that they will provide new data to test long-standing theories. But the new texts can also suggest new ideas, concepts, and processes previously undiscovered.

The vast array of potential applications captures well the promise of automated text analysis and its potential pitfalls. The promise is that the methods will make possible inferences that were previously impossible. If political scientists can effectively use large collections of texts in their inferences, then many substantively important questions are likely to be answered. The pitfalls are that applying these methods will take careful thought and reasoning. Applying any one of the methods described here without careful thought will likely lead to few answers and a great deal of frustration.

This essay provides some guidance to avoid these pitfalls. If scholars recognize the limitations of statistical text models and demonstrate the validity of their measurements, automated methods will reach their promise and revolutionize fields of study within political science.

### Funding

Brandon Stewart gratefully acknowledges a Graduate Research Fellowship from the National Science Foundation.

### References

- Adler, E. Scott, and John Wilkerson. 2011. *The Congressional bills project*. <http://www.congressionalbills.org>.
- Ansolabehere, Stephen, and Shanto Iyengar. 1995. *Going negative: How political advertisements shrink and polarize the electorate*. New York, NY: Simon & Schuster.
- Armstrong, J. S. 1967. Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. *The American Statistician* 21(1):17–21.
- Ashworth, Scott, and Scott Bueno de Mesquita. 2006. Delivering the goods: Legislative particularism in different electoral and institutional settings. *Journal of Politics* 68(1):168–79.
- Beauchamp, Nick. 2011. Using text to scale legislatures with uninformative voting. New York University Mimeo.
- Benoit, K., M. Laver, and S. Mikhaylov. 2009. Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science* 53(2):495–513.
- Berinsky, Adam, Greg Huber, and Gabriel Lenz. 2012. Using mechanical turk as a subject recruitment tool for experimental research. *Political Analysis* 20:351–68.
- Bishop, Christopher. 1995. *Neural networks for pattern recognition*. Gloucestershire, UK: Clarendon Press.
- . 2006. *Pattern recognition and machine learning*. New York, NY: Springer.
- Blei, David. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84.
- Blei, David, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning and Research* 3:993–1022.
- Blei, David, and Michael Jordan. 2006. Variational inference for dirichlet process mixtures. *Journal of Bayesian Analysis* 1(1):121–44.
- Bonica, Adam. 2011. Estimating ideological positions of candidates and contributions from campaign finance records. Stanford University Mimeo.
- Bradley, M. M., and P. J. Lang. 1999. Affective Norms for English Words (ANEW): Stimuli, instruction, manual and affective ratings. University of Florida Mimeo.
- Breiman, L. 2001. Random Forests. *Machine Learning* 45:5–32.
- Budge, Ian, and Paul Pennings. 2007. Do they work? Validating computerised word frequency estimates against policy series. *Electoral Studies* 26:121–29.
- Burden, Barry, and Joseph Sanberg. 2003. Budget rhetoric in presidential campaigns from 1952 to 2000. *Political Behavior* 25(2):97–118.
- Chang, Jonathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, 288–96. Cambridge, MA: The MIT Press.
- Cleveland, William S. 1979. Robust locally weighted regression and scatterplots. *Journal of the American Statistical Association* 74(368):829–36.
- Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review* 98(02):355–70.
- Dempster, Arthur, Nathan Laird, and Donald Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39(1):1–38.
- Diermeier, Daniel, Jean-Francois Godbout, Bei Yu, and Stefan Kaufmann. 2011. Language and ideology in Congress. *British Journal of Political Science* 42(1):31–55.
- Dietterich, T. 2000. Ensemble methods in machine learning. *Multiple Classifier Systems* 1–15.

- Efron, Bradley, and Gail Gong. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician* 37(1):36–48.
- Eggers, Andy, and Jens Hainmueller. 2009. MPs for sale? Returns to office in postwar British politics. *American Political Science Review* 103(04):513–33.
- Eshbaugh-Soha, Matthew. 2010. The tone of local presidential news coverage. *Political Communication* 27(2):121–40.
- Fenno, Richard. 1978. *Home style: House members in their districts*. Boston, MA: Addison Wesley.
- Frey, Brendan, and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science* 315(5814):972–6.
- Gelpi, C., and P. D. Feaver. 2002. Speak softly and carry a big stick? Veterans in the political elite and the American use of force. *American Political Science Review* 96(4):779–94.
- Gerber, Elisabeth, and Jeff Lewis. 2004. Beyond the median: Voter preferences, district heterogeneity, and political representation. *Journal of Political Economy* 112(6):1364–83.
- Greene, William. 2007. *Econometric analysis*. 6th ed. Upper Saddle River, NJ: Prentice Hall.
- Grimmer, Justin. 2010. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis* 18(1):1–35.
- . Forthcoming 2012. Appropriators not position takers: The distorting effects of electoral incentives on congressional representation. *American Journal of Political Science*.
- Grimmer, Justin, and Gary King. 2011. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences* 108(7):2643–50.
- Hand, David J. 2006. Classifier technology and the illusion of progress. *Statistical Science* 21(1):1–15.
- Hart, R. P. 2000. *Diction 5.0: The text analysis program*. Thousand Oaks, CA: Sage-Scolari.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The elements of statistical learning*. New York, NY: Springer.
- Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2008. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics* 4(4):31–46.
- Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 50–7.
- Hopkins, Daniel, and Gary King. 2010. Extracting systematic social science meaning from text. *American Journal of Political Science* 54(1):229–47.
- Hopkins, Daniel, Gary King, Matthew Knowles, and Steven Melendez. 2010. *ReadMe: Software for automated content analysis*. <http://gking.harvard.edu/readme>.
- Jackman, Simon. 2006. Data from Web into R. *The Political Methodologist* 14(2):11–6.
- Jain, A. K., M. N. Murty, and P. J. Flynn. 1999. Data clustering: A review. *ACM Computing Surveys* 31(3):264–323.
- Jones, Bryan, John Wilkerson, and Frank Baumgartner. 2009. *The policy agendas project*. <http://www.policyagendas.org>.
- Jurafsky, Dan, and James Martin. 2009. *Speech and natural language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.
- Jurka, Timothy P., Loren Collingwood, Amber Boydston, Emiliano Grossman, and Wouter van Atteveldt. 2012. RTextTools: Automatic text classification via supervised learning. <http://cran.r-project.org/web/packages/RTextTools/index.html>.
- Kellstedt, Paul. 2000. Media framing and the dynamics of racial policy preferences. *American Journal of Political Science* 44(2):245–60.
- Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology*. New York: Sage.
- Krosnick, Jon. 1999. Survey research. *Annual Review of Psychology* 50(1):537–67.
- Laver, Michael, and John Garry. 2000. Estimating policy positions from political texts. *American Journal of Political Science* 44(3):619–34.
- Laver, Michael, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review* 97(02):311–31.
- Lodhi, H., C. Saunders, J. Shawe-Taylor, N. Christianini, and C. Watkins. 2002. Text classifications using string kernels. *Journal of Machine Learning Research* 2:419–44.
- Loughran, Tim, and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66(1):35–65.
- Lowe, Will. 2008. Understanding wordscores. *Political Analysis* 16(4):356–71.
- Lowe, Will, Ken Benoit, Slava Mihaylov, and M. Laver. 2011. Scaling policy preferences from coded political texts. *Legislative Studies Quarterly* 36(1):123–55.
- MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1: 281–97. London, UK: Cambridge University Press.
- Manning, Christopher, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press.
- Maron, M. E., and J. L. Kuhns. 1960. On relevance, probabilistic indexing, and information retrieval. *Journal of the Association for Computing Machinery* 7(3):216–44.
- Martin, Lanny, and Georg Vanberg. 2007. A robust transformation procedure for interpreting political text. *Political Analysis* 16(1):93–100.
- Mayhew, David. 1974. *Congress: The electoral connection*. New Haven, CT: Yale University Press.
- Mikhaylov, S., M. Laver, and K. Benoit. 2010. Coder reliability and misclassification in the human coding of party manifestos. 66th MPSA annual national conference, Palmer House Hilton Hotel and Towers.

- Monroe, Burt, and Ko Maeda. 2004. Talk's cheap: Text-based estimation of rhetorical ideal points. Paper presented at the 21st annual summer meeting of the Society of Political Methodology.
- Monroe, Burt, Michael Colaresi, and Kevin Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16(4):372.
- Mosteller, F., and D. L. Wallace. 1963. Inference in an authorship problem. *Journal of the American Statistical Association* 58:275–309.
- Neuendorf, K. A. 2002. *The content analysis guidebook*. Thousand Oaks, CA: Sage Publications, Inc.
- Ng, Andrew, Michael Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems 14: Proceeding of the 2001 conference*, eds. T. Dietterich, S. Becker, and Z. Gharamani, 849–56. Cambridge, MA: The MIT Press.
- Pang, B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing* 10:79–86.
- Pennebaker, James, Martha Francis, and Roger Booth. 2001. *Linguistic inquiry and word count: LIWC 2001*. Mahwah, NJ: Erlbaum Publishers.
- Poole, Keith, and Howard Rosenthal. 1997. *Congress: A political-economic history of roll call voting*. Oxford, UK: Oxford University Press.
- Porter, Martin. 1980. An algorithm for suffix stripping. *Program* 14(3):130–37.
- Quinn, Kevin. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54(1):209–28.
- Schrodt, Philip. 2000. Pattern recognition of international crises using Hidden Markov Models. In *Political complexity: Nonlinear models of politics*, ed. Diana Richards, 296–328. Ann Arbor, MI: University of Michigan Press.
- Schrodt, Philip A. 2006. Twenty years of the Kansas event data system project. *Political Methodologist* 14(1):2–6.
- Slapin, Jonathan, and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science* 52(3):705–22.
- Spirling, Arthur. 2012. US treaty-making with American Indians. *American Journal of Political Science* 56(1):84–97.
- Spirling, Arthur, and Iain McLean. 2007. UK OC OK? Interpreting optimal classification scores for the UK House of Commons. *Political Analysis* 15(1):85–96.
- Stewart, Brandon M., and Yuri M. Zhukov. 2009. Use of force and civil–military relations in Russia: An automated content analysis. *Small Wars & Insurgencies* 20:319–43.
- Stone, Phillip, Dexter Dunphy, Marshall Smith, and Daniel Ogilvie. 1966. *The general inquirer: A computer approach to content analysis*. Cambridge, MA: The MIT Press.
- Taddy, Matthew A. 2010. Inverse regression for analysis of sentiment in text. *Arxiv preprint arXiv:1012.2098*.
- Turney, P., and M. L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)* 21(4):315–46.
- van der Laan, Mark, Eric Polley, and Alan Hubbard. 2007. Super learner. *Statistical Applications in Genetics and Molecular Biology* 6(1):1544–6115.
- van der Vaart, A. W., S. Dudoit, and M. J. van der Laan. 2006. Oracle inequalities for multifold cross validation. *Statistics and Decisions* 24(3):351–71.
- Venables, W. N., and B. D. Ripley. 2002. *Modern applied statistics with S*. 4th ed. New York: Springer.
- Wallach, Hanna, Lee Dicker, Shane Jensen, and Katherine Heller. 2010. An alternative prior for nonparametric Bayesian Clustering. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)* 9: 892–99.
- Weber, Robert P. 1990. *Basic content analysis*. Newbury Park, CA: Sage University Paper Series on Quantitative Applications in the Social Sciences.
- Weingast, Barry, Kenneth Shepsle, and Christopher Johnsen. 1981. The political economy of benefits and costs: A neoclassical approach to distributive politics. *The Journal of Political Economy* 89(4):642.
- Yiannakis, Diana Evans. 1982. House members' communication styles: Newsletter and press releases. *The Journal of Politics* 44(4):1049–71.
- Young, Lori, and Stuart Soroka. 2011. Affective news: The automated coding of sentiment in political texts. *Political Communication* 29(2):205–31.

# CHAPTER 1

## History

*Empirical inquiries into the meanings of communications date back to theological studies in the late 1600s, when the Church found the printing of nonreligious materials to be a threat to its authority. Such inquiries have since mushroomed, moving into numerous areas and becoming the backbone of communication research. This chapter discusses several stages in the history of content analysis: quantitative studies of the press; propaganda analysis during World War II; social scientific uses of the technique in studies of political symbols, historical documents, anthropological data, and psychotherapeutic exchanges; computer text analysis and the new media; and qualitative challenges to content analysis.*

### 1.1 Some Precursors

---

Content analysis entails a systematic reading of a body of texts, images, and symbolic matter, not necessary from an author's or user's perspective. Although the term *content analysis* did not appear in English until 1941 (Waples & Berelson, 1941, p. 2; cited in Berelson & Lazarsfeld, 1948), the systematic analysis of text can be traced back to inquisitorial pursuits by the Church in the 17th century. Religions have always been captivated by the written word, so it is not surprising that the first known dissertations about newspapers were defended in 1690, 1695, and 1699 by individuals pursuing academic degrees in theology. After the advent of the printing press, the Church became worried about the spread of printed matter of a nonreligious nature, and so it dealt with newspaper content in moralizing terms (Groth, 1948, p. 26). Surprisingly, in spite of the rhetorical tradition of ancient Greece, which was normative and oral in orientation, the 17th century contributed very little to the methodology of content analysis.

Probably the first well-documented quantitative analyses of printed matter occurred in 18th-century Sweden. According to Dovring's (1954–1955; see also

Krippendorff & Bock, 2009, Chapter 1.1) account, these analyses were undertaken as the result of the publication of the *Songs of Zion*, a collection of 90 hymns of unknown authorship. The collection had passed the Royal Swedish censor, but soon after its publication it was blamed for undermining the orthodox clergy of the Swedish state church. When the collection became popular, it was said to be “contagious” and was accused of aiding a dissenting group. Outstanding in this case is the fact that literary scholars of good reputation participated in the controversy, which crystallized around the question of whether the songs harbored dangerous ideas and, if so, how. Scholars on one side made a list of the religious symbols in the songs and became alarmed. Those on the other side, however, found the very same symbols in established songbooks and so discounted the claimed difference. Then some scholars noted that the symbols in the songs occurred in different contexts and had acquired meanings that were different from those taught in the official church. A debate arose about whether the meanings should be interpreted literally or metaphorically. The interpretations came to be compared with the results of a German study of the outlawed Moravian Brethren, a religious sect whose members later emigrated to the United States. This process—of revising a method in response to criticism—continued until it became clear to both sides in the debate how the symbols in the *Songs of Zion* differed from the symbols used in the official songbooks and how this (in the end political) phenomenon could be explained. The controversy generated many ideas that are now part of content analysis and stimulated debates about methodology that continue today.

In 1903, Eugen Löbl published in German an elaborate classification scheme for analyzing the “inner structure of content” according to the social functions that newspapers perform. His book, which became well-known in journalistic circles, contributed to the idea of *Publizistik*, or newspaper science, and foreshadowed functionalism, but it did not stimulate empirical investigations.

At the first meeting of the German Sociological Society in 1910, Max Weber (1911; see also Krippendorff & Bock, 2009, Chapter 1.2) proposed a large-scale content analysis of the press, but for a variety of reasons the research never got off the ground. During the same period, Andrei Markov (1913), who was working on a theory of chains of symbols, published a statistical analysis of a sample of Pushkin’s novel in verse, *Eugene Onegin*. These inquiries were discovered only recently or influenced the content analysis literature only indirectly. For example, Weber is celebrated as one of the great sociologists, but his advocacy of the use of content analysis as a method for understanding the mass media is relatively unknown. And Markov’s probability theories entered the content analysis literature only through Shannon’s mathematical theory of communication (see Shannon & Weaver, 1949), which influenced Osgood’s (1959) contingency analysis and cloze procedure.

## Quantitative Newspaper Analysis 1.2

---

The beginning of the 20th century saw a visible increase in the mass production of newsprint. In the United States, the boom in newspapers created mass markets and interest in public opinion. Journalism schools emerged, leading to demands for

ethical standards and for empirical inquiries into the phenomenon of the newspaper. These demands, plus a somewhat simplistic notion of scientific objectivity, were met by what was then called *quantitative newspaper analysis*.

Probably the first quantitative newspaper analysis, published in 1893, asked the rhetorical question, “Do newspapers now give the news?” (Speed, 1893). Its author showed how, between 1881 and 1893, New York newspapers had dropped their coverage of religious, scientific, and literary matters in favor of gossip, sports, and scandals. In a similar but far more simplistic study published in 1910, Mathews attempted to reveal the overwhelming space that one New York daily newspaper devoted to “demoralizing,” “unwholesome,” and “trivial” matters as opposed to “worthwhile” news items. By simply measuring the column inches that newspapers devoted to particular subject matters, journalists in the early 20th century attempted to reveal “the truth about newspapers” (Street, 1909). Some believed that they had found a way of showing that the profit motive was the cause of “cheap yellow journalism” (Wilcox, 1900); others became convinced that they had established “the influence of newspaper presentations on the growth of crime and other antisocial activity” (Fenton, 1910). At least one concluded that a “quarter century survey of the press content shows demand for facts” (White, 1924).

Quantitative newspaper analysis seemingly provided the needed scientific ground for journalistic arguments. The respect for numbers has a long history, and facts that could be quantified were considered irrefutable. In a footnote, Berelson and Lazarsfeld (1948) quote from a source published more than 200 years ago:

Perhaps the spirit of the battle over ratification is best reflected in the creed ironically attributed to each of the contending parties by its opponents. The recipe for an Anti-Federalist essay which indicates in a very concise way the class-bias that actuated the opponents of the Constitution, ran in this manner: “wellborn, nine times—Aristocracy, eighteen times—Liberty of the Press, thirteen times repeated—Liberty of Conscience, once—Negro Slavery, once mentioned—Trial by Jury, seven times—Great men, six times repeated—Mr. Wilson, forty times . . .—put them together and dish them up at pleasure. (p. 9; quoted from *New Hampshire Spy*, November 30, 1787)

Quantitative newspaper analysis led to the development of many valuable ideas, however. In 1912, Tenney (see also Krippendorff & Bock, 2009, Chapter 1.4) made a far-reaching proposal for a large-scale and continuous survey of press content to establish a system of bookkeeping of the “social weather” “comparable in accuracy to the statistics of the U.S. Weather Bureau” (p. 896). He demonstrated what he had in mind with an analysis of a few New York newspapers for different ethnic groups, but his proposal exceeded the scope of what was then feasible. Quantitative newspaper analysis culminated in sociologist Malcolm M. Willey’s 1926 book *The Country Newspaper*. In this model study, Willey traced the emergence of Connecticut country weeklies, examining circulation figures, changes in subject matter, and the social role these papers acquired in competition with large city dailies.

When other mass media became prominent, researchers extended the approach first used in newspaper analysis—measuring volumes of coverage in various subject matter categories—initially to radio (Albig, 1938) and later to movies and television.

Content analysis in subject matter categories continues today and is applied to a wide variety of printed matter, such as textbooks, comic strips, speeches, and print advertising.

## Early Content Analysis 1.3

---

The second phase in the intellectual growth of content analysis, which took place in the 1930s and 1940s, involved at least four factors:

- During the period following the 1929 economic crisis, numerous social and political problems emerged in the United States. Many Americans believed that the mass media were at least partially to blame for such problems as yellow journalism, rising crime rates, and the breakdown of cultural values.
- New and increasingly powerful electronic media of communication, first radio and later television, challenged the cultural hegemony of the newspapers. Researchers could not continue to treat these new media as extensions of newspapers, because they differed from the print media in important ways. For example, users of radio and television did not have to be able to read.
- Major political challenges to democracy were linked to the new mass media. For example, the rise of fascism was seen as nourished by the as-yet little-known properties of radio.
- Perhaps most important, this period saw the emergence of the behavioral and social sciences as well as increasing public acceptance of the theoretical propositions and empirical methods of inquiry associated with them.

In the 1930s, sociologists started to make extensive use of survey research and polling. The experience they gained in analyzing public opinion gave rise to the first serious consideration of methodological problems of content analysis, published by Woodward in a 1934 article titled “Quantitative Newspaper Analysis as a Technique of Opinion Research.” From writings about public opinion, interest in social stereotypes (Lippmann, 1922) entered the analysis of communications in various forms. Questions of representations were raised, with researchers examining topics such as how Negroes were presented in the Philadelphia press (Simpson, 1934); how U.S. textbooks described wars in which the United States had taken part, compared with textbooks published in countries that were former U.S. enemies (Walworth, 1938); and how nationalism was expressed in children’s books published in the United States, Great Britain, and other European countries (Martin, 1936).

One of the most important concepts that emerged in psychology during this time was the concept of “attitude.” It added evaluative dimensions to content analysis, such as “pro-con” or “favorable-unfavorable,” that had escaped the rough subject matter categories of quantitative newspaper analysis. Attitude measures redefined journalistic standards of fairness and balance and opened the door to the systematic assessment of bias. Among the explicit standards developed, Janis and Fadner’s (1943/1965) “coefficient of imbalance” deserves mention. Psychological experiments in rumor transmission led Allport and Faden to study newspaper content from an

entirely new perspective. In their 1940 article “The Psychology of Newspapers: Five Tentative Laws,” they attempted to account for the changes that information undergoes as it travels through an institution and finally appears on the printed page.

The interest in political symbols added another feature to the analysis of public messages. McDiarmid (1937), for example, examined 30 U.S. presidential inaugural addresses for symbols of national identity, of historical significance, of government, and of fact and expectations. Most important, Lasswell (1938), viewing public communications within his psychoanalytical theory of politics, classified symbols into such categories as “self” and “others” and forms of “indulgence” and “deprivation.” His symbol analysis led to his “World Attention Survey,” in which he compared trends in the frequencies with which prestige newspapers in several countries used national symbols (Lasswell, 1941; see also Krippendorff & Bock, 2009, Chapter 5.3).

Researchers in several disciplines examined the trends in scholarship, as reflected in the topics that representative journals published. Rainoff’s (1929) Russian study regarding physics was probably the first of this kind, but the most thorough analyses were conducted in the field of sociology (Becker, 1930, 1932; Shanas, 1945) and later in journalism (Tannenbaum & Greenberg, 1961).

Several factors influenced the transition from quantitative newspaper analysis, which was largely journalism driven, to content analysis:

- Eminent social scientists became involved in these debates and asked new kinds of questions.
- The concepts these social scientists developed were theoretically motivated, operationally defined, and fairly specific, and interest in stereotypes, styles, symbols, values, and propaganda devices began to replace interest in subject matter categories.
- Analysts began to employ new statistical tools borrowed from other disciplines, especially from survey research but also from experimental psychology.
- Content analysis data became part of larger research efforts (e.g., Lazarsfeld, Berelson, & Gaudet, 1948), and so content analysis no longer stood apart from other methods of inquiry.

The first concise presentation of these conceptual and methodological developments under the new umbrella term *content analysis* appeared in a 1948 mimeographed text titled *The Analysis of Communication Content*, authored by Berelson and Lazarsfeld, which was later published as Berelson’s *Content Analysis in Communications Research* (1952). This first systematic presentation codified the field for years to come.

## 1.4 Propaganda Analysis

---

Berelson described content analysis as the use of mass communications as data for testing scientific hypotheses and for evaluating journalistic practices. Yet the most important and large-scale challenge that content analysis faced came during World

War II, when it was employed in efforts to extract information from propaganda. Before the war, researchers analyzed texts in order to identify “propagandists,” to point fingers at individuals who were attempting to influence others through devious means. Fears concerning such influence had several origins. Propaganda was used extensively during World War I (Lasswell, 1927), and the years between the two world wars witnessed the effective use of propaganda by antidemocratic demagogues in Europe. In addition, Americans tend to have deep-seated negative attitudes toward religious fanatics, and the lack of knowledge concerning what the extensive use of the new mass media (radio, film, and television) could do to people raised concerns as well. According to the Institute for Propaganda Analysis (1937), propagandists reveal themselves through their use of tricks such as “name-calling,” employing “glittering generalities,” “plain folks” identifications, “card stacking,” “bandwagon” devices, and so on. Such devices could be identified easily in many religious and political speeches, even in academic lectures, and this approach to propaganda analysis led to a kind of witch-hunt for propagandists in the United States. Theories concerning subliminal messages, especially in advertising, raised widespread suspicion as well.

In the 1940s, as U.S. attention became increasingly devoted to the war effort, the identification of propagandists was no longer an issue. Nor were researchers particularly interested in revealing the power of the mass media of communication to mold public opinion; rather, military and political intelligence were needed. In this climate, two centers devoted to propaganda analysis emerged. Harold D. Lasswell and his associates, having written on political symbolism, worked with the Experimental Division for the Study of Wartime Communications at the U.S. Library of Congress, and Hans Speier, who had organized a research project on totalitarian communication at the New School for Social Research in New York, assembled a research team at the Foreign Broadcast Intelligence Service of the U.S. Federal Communications Commission (FCC). The Library of Congress group focused on analyzing newspapers and wire services from abroad and addressed basic issues of sampling, measurement problems, and the reliability and validity of content categories, continuing the tradition of early quantitative analysis of mass communications (Lasswell, Leites, & Associates, 1965).

The FCC group analyzed primarily domestic enemy broadcasts and surrounding conditions to understand and predict events within Nazi Germany and the other Axis countries, and to estimate the effects of Allied military actions on the war mood of enemy populations. The pressures of day-to-day reporting left the analysts little time to formalize their methods, and Berelson (1952) thus had little to say about the accomplishments of the FCC group. After the war, however, Alexander L. George worked through the volumes of reports that resulted from these wartime efforts to describe methods that had evolved in the process and to validate the inferences the researchers had made by comparing them with documentary evidence now available from Nazi archives. These efforts resulted in his book *Propaganda Analysis* (1959a; see also Krippendorff & Bock, 2009, Chapter 1.5), which made major contributions to the conceptualization of the aims and processes of content analysis.

The assumptions that propagandists are rational, in the sense that they follow their own propaganda theories in their choice of communications, and that the meanings of propagandists' communications may differ for different people reoriented the FCC analysts from a concept of "content as shared" (Berelson would later say "manifest") to conditions that could explain the motivations of particular communicators and the interests they might serve. The notion of "preparatory propaganda" became an especially useful key for the analysts in their effort to infer the intents of broadcasts with political content. In order to ensure popular support for planned military actions, the Axis leaders had to inform, emotionally arouse, and otherwise prepare their countrymen and -women to accept those actions; the FCC analysts discovered that they could learn a great deal about the enemy's intended actions by recognizing such preparatory efforts in the domestic press and broadcasts. They were able to predict several major military and political campaigns and to assess Nazi elites' perceptions of their situation, political changes within the Nazi governing group, and shifts in relations among Axis countries. Among the more outstanding predictions that British analysts were able to make was the date of deployment of German V weapons against Great Britain. The analysts monitored the speeches delivered by Nazi propagandist Joseph Goebbels and inferred from the content of those speeches what had interfered with the weapons' production and when. They then used this information to predict the launch date of the weapons, and their prediction was accurate within a few weeks.

Several lessons were learned from these applications of content analysis, including the following:

- Content is not inherent to communications. People typically differ in how they read texts. The intentions of the senders of broadcast messages may have little to do with how audience members hear those messages. Temporal orderings, individuals' needs and expectations, individuals' preferred discourses, and the social situations into which messages enter are all important in explaining what communications come to mean. Interpretations on which all communicators readily agree are rare, and such interpretations are usually relatively insignificant.
- Content analysts must predict or infer phenomena that they cannot observe at the time of their research. The inability to observe phenomena of interest tends to be the primary motivation for using content analysis. Whether the analyzed source has reasons to hide what the analyst desires to know (as in the case of an enemy during wartime or the case of someone needing to impress) or the phenomena of interest are inaccessible in principle (e.g., an individual's attitudes or state of mind, or historical events) or just plain difficult to assess otherwise (such as what certain mass-media audiences could learn from watching TV), the analyst seeks answers to questions that go outside a text. To be sure, the questions that a content analyst seeks to answer are the analyst's questions, and as such they are potentially at odds with whether others could answer them and how. Quantitative newspaper analysts made inferences without acknowledging their own conceptual contributions to

what they thought they found but actually inferred. Content is not the whole issue; rather, the issue is what can be legitimately inferred from available texts.

- In order to interpret given texts or make sense of the messages intercepted or gathered, content analysts need elaborate models of the systems in which those communications occur (or occurred). The propaganda analysts working during World War II constructed such models more or less explicitly. Whereas earlier content analysts had viewed mass-produced messages as inherently meaningful and analyzable unit by unit, the propaganda analysts succeeded only when they viewed the messages they analyzed in the context of the lives of the diverse people presumed to use those messages.
- For analysts seeking specific political information, quantitative indicators are extremely insensitive and shallow. Even where large amounts of quantitative data are available, as required for statistical analyses, these tend not to lead to the “most obvious” conclusions that political experts would draw from qualitative interpretations of textual data. Qualitative analyses can be systematic, reliable, and valid as well.

Convinced that content analysis does not need to be inferior to unsystematic explorations of communications, numerous writers in the postwar years, such as Kracauer (1947, 1952–1953) and George (1959a), challenged content analysts’ simplistic reliance on counting qualitative data. Smythe (1954) called this reliance on counting an “immaturity of science” in which objectivity is confused with quantification. However, the proponents of the quantitative approach largely ignored the criticism. In his 1949 essay “Why Be Quantitative?” Lasswell (1949/1965b) continued to insist on the quantification of symbols as the sole basis of scientific insights. His approach to propaganda analysis produced several working papers but very few tangible results compared with the work of the FCC group of scholars. Today, quantification continues, although perhaps no longer exclusively.

## Content Analysis Generalized 1.5

---

After World War II, and perhaps as the result of the first integrated picture of content analysis provided by Berelson (1952), the use of content analysis spread to numerous disciplines. This is not to say that content analysis emigrated from mass communication. In fact, the very “massiveness” of available communications continued to attract scholars who looked at the mass media from new perspectives. For example, Lasswell (1941) realized his earlier idea of a “world attention survey” in a large-scale study of political symbols in French, German, British, Russian, and U.S. elite press editorials and key policy speeches. He wanted to test the hypothesis that a “world revolution” had been in steady progress for some time (Lasswell, Lerner, & Pool, 1952). Gerbner and his colleagues pursued Gerbner’s (1969) proposal to develop “cultural indicators” by analyzing, for almost two decades, one week of fictional television programming per year, mainly to establish “violence profiles” for different networks, to trace trends, and to see how various groups (such as

women, children, and the aged) were portrayed on U.S. television (see, e.g., Gerbner, Gross, Signorielli, Morgan, & Jackson-Beeck, 1979).

Psychologists began to use content analysis in four primary areas. The first was the inference of motivational, mental, or personality characteristics through the analysis of verbal records. This application started with Allport's (1942) treatise on the use of personal documents, Baldwin's (1942) application of "personal structure analysis" to cognitive structure, and White's (1947) value studies. These studies legitimated the use of written material, personal documents, and individual accounts of observed phenomena as an addition to the then-dominant experimental methods. A second application was the use of verbal data gathered in the form of answers to open-ended interview questions, focus group conversations, and verbal responses to various tests, including the construction of Thematic Apperception Test (TAT) stories. In the context of TAT stories, content analysis acquired the status of a supplementary technique. As such, it allowed researchers to utilize data that they could gather without imposing too much structure on subjects and to validate findings they had obtained through different techniques. Psychological researchers' third application of content analysis concerned processes of communication in which content is an integral part. For example, in his "interaction process analysis" of small group behavior, Bales (1950) used verbal exchanges as data through which to examine group processes. The fourth application took the form of the generalization of measures of meaning over a wide range of situations and cultures (which derived from individualist notions of meaning or content). Osgood (1974a, 1974b) and his students found numerous applications for Osgood, Suci, and Tannenbaum's (1957) semantic differential scales and conducted worldwide comparisons of cultural commonalities and differences.

Anthropologists, who started using content analysis techniques in their studies of myths, folktales, and riddles, have made many contributions to content analysis, including the componential analysis of kinship terminology (Goodenough, 1972). Ethnography emerged in anthropology, and although ethnographers often interact with their informants in ways that content analysts cannot interact with authors or readers, after ethnographers gather their field notes they start to rely heavily on methods that are similar to those that content analysts use.

Historians are naturally inclined to look for systematic ways to analyze historical documents, and they soon embraced content analysis as a suitable technique, especially where data are numerous and statistical accounts seem helpful. Social scientists also recognized the usefulness of educational materials, which had long been the focus of research. Such materials are a rich source of data on processes of reading (Flesch, 1948, 1951) as well as on a society's larger political, attitudinal, and value trends. In addition, literary scholars began to apply the newly available techniques of content analysis to the problem of identifying the authors of unsigned documents.

On one hand, this proliferation of the use of content analysis across disciplines resulted in a loss of focus: Everything seemed to be content analyzable, and every analysis of symbolic phenomena became a content analysis. On the other hand, this trend also broadened the scope of the technique to embrace what may well be the essence of human behavior: talk, conversation, and mediated communication.

In 1955, responding to increasing interest in the subject, the Social Science Research Council's Committee on Linguistics and Psychology sponsored a conference on content analysis. The participants came from such disciplines as psychology, political science, literature, history, anthropology, and linguistics. Their contributions to the conference were published in a volume titled *Trends in Content Analysis*, edited by Ithiel de Sola Pool (1959a). Despite obvious divergence among the contributors in their interests and approaches, Pool (1959a, p. 2) observed, there was considerable and often surprising convergence among them in two areas: They exhibited (a) a shift from analyzing the "content" of communications to drawing inferences about the antecedent conditions of communications and (b) an accompanying shift from measuring volumes of subject matter to counting simple frequencies of symbols, and then to relying on contingencies (co-occurrences).

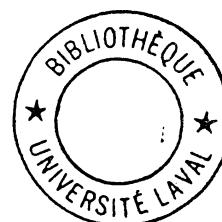
## Computer Text Analysis

### 1.6

The late 1950s witnessed considerable interest among researchers in mechanical translation, mechanical abstracting, and information retrieval systems. Computer languages suitable for literal data processing emerged, and scholarly journals started to devote attention to computer applications in psychology, the humanities, and the social sciences. The large volumes of written documents to be processed in content analysis and the repetitiveness of the coding involved made the computer a natural but also a difficult ally of the content analyst.

The development of software for literal (as opposed to numerical) data processing stimulated new areas of exploration, such as information retrieval, information systems, computational stylistics (Sedelow & Sedelow, 1966), computational linguistics, word processing technology, and computational content analysis. New software also revolutionized tedious literary work, such as indexing and the creation of concordances. Probably the first computer-aided content analysis was reported by Sebeok and Zeps (1958), who made use of simple information retrieval routines to analyze some 4,000 Cheremis folktales. In a Rand Corporation paper titled *Automatic Content Analysis*, Hays (1960) explored the possibility of designing a computer system for analyzing political documents. Unaware of both these developments, Stone and Bales, who were engaged in a study of themes in face-to-face interacting groups, designed and programmed the initial version of the General Inquirer system. This culminated in a groundbreaking book by Stone, Dunphy, Smith, and Ogilvie (1966) in which they presented an advanced version of this system and demonstrated its application in numerous areas, ranging from political science to advertising and from psychotherapy to literary analysis.

The use of computers in content analysis was also stimulated by developments in other fields. Scholars in psychology became interested in simulating human cognition (Abelson, 1963; Schank & Abelson, 1977). Newell and Simon (1963) developed a computer approach to (human) problem solving. Linguistics researchers developed numerous approaches to syntactic analysis and semantic interpretation of linguistic expressions. Researchers in the field of artificial intelligence focused on designing machines that could understand natural language (with very little success).



In 1967, the Annenberg School of Communications (which later became the Annenberg School for Communication) sponsored a major conference on content analysis. Discussions there focused on many areas—the difficulties of recording nonverbal (visual, vocal, and musical) communications, the need for standardized categories, the problems involved in drawing inferences, the roles of theories and analytical constructs, what developments content analysts could expect in the near future—but the subject of the use of computers in content analysis permeated much of the conference. Stone et al.'s (1966) book on the General Inquirer had just been published, and it had created considerable hope among content analysts. The contributions to the 1967 conference are summarized in a 1969 volume edited by Gerbner, Holsti, Krippendorff, Paisley, and Stone, the publication of which coincided with Holsti's (1969) survey of the field.

In 1974, participants in the Workshop on Content Analysis in the Social Sciences, held in Pisa, Italy, saw the development of suitable algorithms for computer content analysis as the only obstacle to better content analyses (Stone, 1975). Since that time, computational approaches have moved in numerous directions. One has been the development of customizable content analysis packages, of which the General Inquirer was the most important precursor. Attempts to apply the General Inquirer system to German texts revealed that software's English-language biases and led to more general versions of General Inquirers, such as TextPack. The basic ingredient of the General Inquirer and TextPack is a dictionary of relevant words. In the 1980s, Sedelow (1989) proposed the idea of using a thesaurus instead, as a thesaurus might be more accurate than a dictionary in reflecting "society's collective associative memory" (p. 4; see also Sedelow & Sedelow, 1986). In the 1990s, George Miller initiated a major research effort to chart the meanings of words using a computer-traceable network called WordNet (see Miller et al., 1993). In the 1980s, some authors observed that the enthusiasm associated with large systems that had appeared in the 1960s was fading (see Namenwirth & Weber, 1987), but today the development of text analysis software is proliferating, fueled largely by the historically unprecedented volumes of electronic and digital texts available for content analysis. More recently, Diefenbach (2001) reviewed the history of content analysis by focusing on four specific areas: mass communication research, political science, psychology, and literature.

Naturally, many researchers have compared computer-based content analyses with human-based content analyses. For example, Schnurr, Rosenberg, and Ozman (1992, 1993) compared the Thematic Apperception Test (Murray, 1943) with a computer content analysis of open-ended free speech and found the low agreement between the two to be discouraging. Zeldow and McAdams (1993) challenged Schnurr et al.'s conclusion, however. Nacos et al. (1991) compared humans' coding of political news coverage with data from Fan's (1988) computer-coded approach to the same coverage and found satisfactory correlations between the two. Nacos et al. came to the conclusion that content analysts can best use computers in their research by thinking of them as aids, not as replacements for the highly developed human capabilities of reading, transcribing, and translating written matter. As one might expect, today scholars hold many different opinions regarding the future of the use of computer-based content analysis.

Another development that has influenced how content analysts employ computers in their work is the increasingly common use of word processing software, which provides users with such features as spell-checkers, word- or phrase-finding and -replacing operations, and even readability indices. Although not intended for this purpose, ordinary word processing software makes it possible for a researcher to perform basic word counts and KWIC (keyword in context) analyses, albeit laboriously.

Word processing software is inherently interactive; it is driven by the user's reading of the textual material, not fixed. In the absence of computational theories of text interpretation, content analysts have found the symbiosis of the human ability to understand and interpret written documents and the computer's ability to scan large volumes of text systematically and reliably increasingly attractive. In such collaborations, human coders are no longer used as text-level content analysts; rather, they serve as translators of text or sections of text into categories that emerge during reading and then into a data language (that preserves relevant meanings), which enables various computational algorithms (that cannot respond to meanings) to do housekeeping and summarizing chores. This has given rise to a new class of software designed for computer-aided qualitative text analysis, of which NVivo and ATLAS.ti are two examples. Such interactive-hermeneutic text analysis software is becoming increasingly accessible, especially to students.

The most important stimulus in the development of computational content analysis, however, has been the growing availability of text in digital form. It is very costly to enter handwritten documents, such as transcripts of audio recordings of interviews, focus group protocols, minutes of business meetings, and political speeches, into a computer. Scanners have vastly improved in recent years, but they are still too unreliable to be used without additional manual editing. In the 1970s, data consortia emerged through which social scientists could share costly data, but the operations of these consortia were marred by a lack of standards and the usually highly specialized nature of the data. Then, in 1977, DeWeese proposed and took the remarkable step of bypassing the costly transcription process by feeding the typesetting tapes of a Detroit newspaper directly into a computer to conduct an analysis of the paper's content the day after it was published. Since that time, word processing software has come to be an integral part of the internal operations of virtually all social organizations; personnel create texts digitally before they appear on paper, use electronic mail systems, and surf the internet to download materials relevant to their work.

Today, a fantastic amount of raw textual data is being generated daily in digital form, representing almost every topic of interest to social scientists. Electronic full-text databases, to which all major U.S. newspapers, many social science and legal journals, and many corporations contribute all of the materials they publish, are growing exponentially and have become easily available and inexpensive to use online. Add to this the volume of electronic publications, the research potential of the internet, data available from online multiuser discussions (MUDs) and newsgroups, and online survey systems, which may well replace focus groups and interviews in certain empirical domains, and it is clear that the landscape of how society presents itself has been altered drastically. With more and more people interested in

this wealth of digital data, there is a corresponding demand for increasingly powerful search engines, suitable computational tools, text base managing software, encryption systems, devices for monitoring electronic data flows, and translation software, all of which will eventually benefit the development of computer-aided content analysis. The current culture of computation is moving content analysis into a promising future.

## 1.7 Qualitative Approaches

---

Perhaps in response to the now dated “quantitative newspaper analysis” of more than a century ago or as a form of compensation for the sometimes shallow results reported by the content analysts of 60 years ago, a variety of research approaches have begun to emerge that call themselves *qualitative*. I question the validity and usefulness of the distinction between quantitative and qualitative content analyses. Ultimately, all reading of texts is qualitative, even when certain characteristics of a text are later converted into numbers. The fact that computers process great volumes of text in a very short time and represent these volumes in ways someone can understand does not remove the qualitative nature of the texts being analyzed and the algorithms used to process them: On the most basic level, computers recognize zeros and ones and change them as instructed, proceeding one step at a time. Nevertheless, proponents of qualitative approaches to content analysis offer alternative protocols for exploring texts systematically.

*Discourse analysis* is one such approach. Generally, *discourse* is defined as text above the level of sentences. Discourse analysts tend to focus on how particular phenomena are represented. For example, Van Dijk (1991) studied manifestations of racism in the press: how minorities appear, how ethnic conflicts are described, and how stereotypes permeate given accounts, for example, in advertisements during sports events (Wonsek, 1992). Other discourse analysts have examined how television news programs and other TV shows in the United States manifest a particular ideological vision of the U.S. economy (Jensen, 2006), the components of “age markers” in the humorous context of the TV series *The Golden Girls* (Harwood & Giles, 1992), and the portrayal of the peace movement in news editorials during the Gulf War (Hackett & Zhao, 1994).

Researchers who conduct *social constructivist analyses* focus on discourse as well, but less to criticize (mis)representations than to understand how reality comes to be constituted in human interactions and in language, including written text (Gergen, 1985). Such analysts may address how emotions are conceptualized (Averill, 1985) or how facts are constructed (Fleck, 1935/1979; Latour & Woolgar, 1986), or they may explore changing notions of self (Gergen, 1991) or of sexuality (Katz, 1995).

*Rhetorical analysis*, in contrast, focuses on how messages are delivered, and with what (intended or actual) effects. Researchers who take this approach rely on the identification of structural elements, tropes, styles of argumentation, speech acts, and the like; Kathleen Hall Jamieson’s book *Packaging the Presidency* (1984) is an example of such an analysis. Efforts to study negotiations (Harris, 1996), what works and what doesn’t, might be described as rhetorical analyses as well.

*Ethnographic content analysis*, an approach advocated by Altheide (1987), does not avoid quantification but encourages content analysis accounts to emerge from readings of texts. This approach works with categories as well as with narrative descriptions but focuses on situations, settings, styles, images, meanings, and nuances presumed to be recognizable by the human actors/speakers involved.

*Conversation analysis* is another approach that is considered to be qualitative. The researcher performing such an analysis tends to start with the recording of verbal interactions in natural settings and aims at analyzing the transcripts as records of conversational moves toward a collaborative construction of conversations. This tradition is indebted to the work of Harvey Sacks, who studied numerous interactive phenomena, including the collaboration among communicators in the telling of jokes (Sacks, 1974). Goodwin (1977, 1981) extended conversation analysis by incorporating video data in his groundbreaking study of turn taking.

Qualitative approaches to content analysis have their roots in literary theory, the social sciences (symbolic interactionism, ethnomethodology), and critical scholarship (Marxist approaches, British cultural studies, feminist theory). Sometimes they are given the label *interpretive*. They share the following characteristics:

- They require a close reading of relatively small amounts of textual matter.
- They involve the rearticulation (interpretation) of given texts into new (analytical, deconstructive, emancipatory, or critical) narratives accepted within particular scholarly communities that are sometimes opposed to positivist traditions of inquiry.
- The analysts acknowledge working within hermeneutic circles in which their own socially or culturally conditioned understandings constitutively participate. (For this reason, I refer to these approaches as *interactive-hermeneutic*, a description that speaks to the process of engaging in systematic interpretations of text.)

One could summarize and say that content analysis has evolved into a repertoire of methods of research that promise to yield inferences from all kinds of verbal, pictorial, symbolic, and communication data. Beyond the technique's initially journalistic roots, the past century has witnessed the migration of content analysis into various fields and the clarification of many methodological issues. After a short period of stagnation in the 1970s, content analysis is today growing exponentially, largely due to the widespread use of computers for all kinds of text processing. As of February 2011, an internet search for “*content analysis*” using the Google search engine found 1,650,000 documents. By comparison, “*survey research*” turned up 275,000 hits and “*psychological testing*,” 894,000. Since the first casual mention of “*content analysis*” in 1941—that is, seventy years ago and with a frequency of one—the public interest in the body of content analysis research has clearly grown to an astonishing extent.

# Chapter 1

## AN INTRODUCTION TO TEXT MINING

Charu C. Aggarwal

*IBM T. J. Watson Research Center  
Yorktown Heights, NY*  
charu@us.ibm.com

ChengXiang Zhai

*University of Illinois at Urbana-Champaign  
Urbana, IL*  
czhai@cs.uiuc.edu

### Abstract

The problem of text mining has gained increasing attention in recent years because of the large amounts of text data, which are created in a variety of social network, web, and other information-centric applications. Unstructured data is the easiest form of data which can be created in any application scenario. As a result, there has been a tremendous need to design methods and algorithms which can effectively process a wide variety of text applications. This book will provide an overview of the different methods and algorithms which are common in the text domain, with a particular focus on mining methods.

### 1. Introduction

Data mining is a field which has seen rapid advances in recent years [8] because of the immense advances in hardware and software technology which has lead to the availability of different kinds of data. This is particularly true for the case of text data, where the development of hardware and software platforms for the web and social networks has enabled the rapid creation of large repositories of different kinds of data. In particular, the web is a technological enabler which encourages the

creation of a large amount of text content by different users in a form which is easy to store and process. The increasing amounts of text data available from different applications has created a need for advances in algorithmic design which can learn interesting patterns from the data in a dynamic and scalable way.

While structured data is generally managed with a database system, text data is typically managed via a search engine due to the lack of structures [5]. A search engine enables a user to find useful information from a collection conveniently with a keyword query, and how to improve the effectiveness and efficiency of a search engine has been a central research topic in the field of information retrieval [13, 3], where many related topics to search such as text clustering, text categorization, summarization, and recommender systems are also studied [12, 9, 7].

However, research in information retrieval has traditionally focused more on facilitating information access [13] rather than analyzing information to discover patterns, which is the primary goal of text mining. The goal of information access is to connect the right information with the right users at the right time with less emphasis on processing or transformation of text information. Text mining can be regarded as going beyond information access to further help users analyze and digest information and facilitate decision making. There are also many applications of text mining where the primary goal is to analyze and discover any interesting patterns, including trends and outliers, in text data, and the notion of a query is not essential or even relevant.

Technically, mining techniques focus on the primary models, algorithms and applications about what one can learn from different kinds of text data. Some examples of such questions are as follows:

- What are the primary supervised and unsupervised models for learning from text data? How are traditional clustering and classification problems different for text data, as compared to the traditional database literature?
- What are the useful tools and techniques used for mining text data? Which are the useful mathematical techniques which one should know, and which are repeatedly used in the context of different kinds of text data?
- What are the key application domains in which such mining techniques are used, and how are they effectively applied?

A number of key characteristics distinguish text data from other forms of data such as relational or quantitative data. This naturally affects the

mining techniques which can be used for such data. The most important characteristic of text data is that it is *sparse* and *high dimensional*. For example, a given corpus may be drawn from a lexicon of about 100,000 words, but a given text document may contain only a few hundred words. Thus, a corpus of text documents can be represented as a *sparse term-document matrix* of size  $n \times d$ , when  $n$  is the number of documents, and  $d$  is the size of the lexicon vocabulary. The  $(i, j)$ th entry of this matrix is the (normalized) frequency of the  $j$ th word in the lexicon in document  $i$ . The large size and the sparsity of the matrix has immediate implications for a number of data analytical techniques such as dimensionality reduction. In such cases, the methods for reduction should be specifically designed while taking this characteristic of text data into account. The variation in word frequencies and document lengths also lead to a number of issues involving document representation and normalization, which are critical for text mining.

Furthermore, text data can be analyzed at different levels of representation. For example, text data can easily be treated as a bag-of-words, or it can be treated as a string of words. However, in most applications, it would be desirable to represent text information *semantically* so that more meaningful analysis and mining can be done. For example, representing text data at the level of named entities such as people, organizations, and locations, and their relations may enable discovery of more interesting patterns than representing text as a bag of words. Unfortunately, the state of the art methods in natural language processing are still not robust enough to work well in unrestricted text domains to generate accurate semantic representation of text. Thus most text mining approaches currently still rely on the more shallow word-based representations, especially the bag-of-wrods approach, which, while losing the positioning information in the words, is generally much simpler to deal with from an algorithmic point of view than the string-based approach. In special domains (e.g., biomedical domain) and for special mining tasks (e.g., extraction of knowledge from the Web), natural language processing techniques, especially information extraction, are also playing an important role in obtaining a semantically more meaningful representation of text.

Recently, there has been rapid growth of text data in the context of different web-based applications such as social media, which often occur in the context of multimedia or other heterogeneous data domains. Therefore, a number of techniques have recently been designed for the *joint mining* of text data in the context of these different kinds of data domains. For example, the Web contains text and image data which are often intimately connected to each other and these links can be used

to improve the learning process from one domain to another. Similarly, cross-lingual linkages between documents of different languages can also be used in order to transfer knowledge from one language domain to another. This is closely related to the problem of transfer learning [11].

The rest of this chapter is organized as follows. The next section will discuss the different kinds of algorithms and applications for text mining. We will also point out the specific chapters in which they are discussed in the book. Section 3 will discuss some interesting future research directions.

## 2. Algorithms for Text Mining

In this section, we will explore the key problems arising in the context of text mining. We will also present the organization of the different chapters of this book in the context of these different problems. We intentionally leave the definition of the concept "text mining" vague to broadly cover a large set of related topics and algorithms for text analysis, spanning many different communities, including natural language processing, information retrieval, data mining, machine learning, and many application domains such as the World Wide Web and Biomedical Science. We have also intentionally allowed (sometimes significant) overlaps between chapters to allow each chapter to be relatively self contained, thus useful as a standing-alone chapter for learning about a specific topic.

**Information Extraction from Text Data:** Information Extraction is one of the key problems of text mining, which serves as a starting point for many text mining algorithms. For example, extraction of entities and their relations from text can reveal more meaningful semantic information in text data than a simple bag-of-words representation, and is generally needed to support inferences about knowledge buried in text data. Chapter 2 provides an survey of key problems in Information Extraction and the major algorithms for extracting entities and relations from text data.

**Text Summarization:** Another common function needed in many text mining applications is to summarize the text documents in order to obtain a brief overview of a large text document or a set of documents on a topic. Summarization techniques generally fall into two categories. In extractive summarization, a summary consists of information units extracted from the original text; in contrast, in abstractive summarization, a summary may contain "synthesized" information units that may not necessarily occur in the text documents. Most existing summarization methods are extractive, and in Chapter 3, we give a brief survey of these

commonly used summarization methods.

**Unsupervised Learning Methods from Text Data:** Unsupervised learning methods do not require any training data, thus can be applied to any text data without requiring any manual effort. The two main unsupervised learning methods commonly used in the context of text data are *clustering* and *topic modeling*. The problem of clustering is that of segmenting a corpus of documents into partitions, each corresponding to a topical cluster. The problems of clustering and topic modeling are closely related. In topic modeling we use a probabilistic model in order to determine a *soft* clustering, in which each document has a membership probability of the cluster, as opposed to a hard segmentation of the documents. Topic models can be considered as the process of clustering with a generative probabilistic model. Each *topic* can be considered a probability distribution over words, with the representative words having the highest probability. Each document can be expressed as a probabilistic combination of these different topics. Thus, a topic can be considered to be analogous to a cluster, and the membership of a document to a cluster is probabilistic in nature. This also leads to a more elegant cluster membership representation in cases in which the document is known to contain distinct topics. In the case of hard clustering, it is sometimes challenging to assign a document to a single cluster in such cases. Furthermore, topic modeling relates elegantly to the dimension reduction problem, where each topic provides a conceptual dimension, and the documents may be represented as a linear probabilistic combination of these different topics. Thus, topic-modeling provides an extremely general framework, which relates to both the clustering and dimension reduction problems. In chapter 4, we study the problem of clustering, while topic modeling is covered in two chapters (Chapters 5 and 8). In Chapter 5, we discuss topic modeling from the perspective of dimension reduction since the discovered topics can serve as a low-dimensional space representation of text data, where semantically related words can “match” each other, which is hard to achieve with bag-of-words representation. In chapter 8, topic modeling is discussed as a general probabilistic model for text mining.

**LSI and Dimensionality Reduction for Text Mining:** The problem of dimensionality reduction is widely studied in the database literature as a method for representing the underlying data in compressed format for indexing and retrieval [10]. A variation of dimensionality reduction which is commonly used for text data is known as *latent semantic indexing* [6]. One of the interesting characteristics of latent semantic indexing is that it brings out the key semantic aspects of the text data, which makes it more suitable for a variety of mining applications. For ex-

ample, the noise effects of synonymy and polysemy are reduced because of the use of such dimensionality reduction techniques. Another family of dimension reduction techniques are probabilistic topic models, notably PLSA, LDA, and their variants; they perform dimension reduction in a probabilistic way with potentially more meaningful topic representations based on word distributions. In chapter 5, we will discuss a variety of LSI and dimensionality reduction techniques for text data, and their use in a variety of mining applications.

**Supervised Learning Methods for Text Data:** Supervised learning methods are general machine learning methods that can exploit training data (i.e., pairs of input data points and the corresponding desired output) to learn a classifier or regression function that can be used to compute predictions on unseen new data. Since a wide range of application problems can be cast as a classification problem (that can be solved using supervised learning), the problem of supervised learning is sometimes also referred to as classification. Most of the traditional methods for text mining in the machine learning literature have been extended to solve problems of text mining. These include methods such as rule-based classifier, decision trees, nearest neighbor classifiers, maximum-margin classifiers, and probabilistic classifiers. In Chapter 6, we will study machine learning methods for automated text categorization, a major application area of supervised learning in text mining. A more general discussion of supervised learning methods is given in Chapter 8. A special class of techniques in supervised learning to address the issue of lack of training data, called *transfer learning*, are covered in Chapter 7.

**Transfer Learning with Text Data:** The afore-mentioned example of cross-lingual mining provides a case where the attributes of the text collection may be heterogeneous. Clearly, the feature representations in the different languages are heterogeneous, and it can often provide useful to transfer knowledge from one domain to another, especially when there is paucity of data in one domain. For example, labeled English documents are copious and easy to find. On the other hand, it is much harder to obtain labeled Chinese documents. The problem of transfer learning attempts to *transfer* the learned knowledge from one domain to another. Some other scenarios in which this arises is the case where we have a mixture of text and multimedia data. This is often the case in many web-based and social media applications such as *Flickr*, *Youtube* or other multimedia sharing sites. In such cases, it may be desirable to transfer the learned knowledge from one domain to another with the use of cross-media transfer. Chapter 7 provides a detailed survey of such learning techniques.

**Probabilistic Techniques for Text Mining:** A variety of probabilistic methods, particularly unsupervised topic models such as PLSA and LDA and supervised learning methods such as conditional random fields are used frequently in the context of text mining algorithms. Since such methods are used frequently in a wide variety of contexts, it is useful to create an organized survey which describes the different tools and techniques that are used in this context. In Chapter 8, we introduce the basics of the common probabilistic models and methods which are often used in the context of text mining. The material in this chapter is also relevant to many of the clustering, dimensionality reduction, topic modeling and classification techniques discussed in Chapters 4, 5, 6 and 7.

**Mining Text Streams:** Many recent applications on the web create massive streams of text data. In particular web applications such as social networks which allow the simultaneous input of text from a wide variety of users can result in a continuous stream of large volumes of text data. Similarly, news streams such as *Reuters* or aggregators such as *Google news* create large volumes of streams which can be mined continuously. Such text data are more challenging to mine, because they need to be processed in the context of a one-pass constraint [1]. The one-pass constraint essentially means that it may sometimes be difficult to store the data offline for processing, and it is necessary to perform the mining tasks continuously, as the data comes in. This makes algorithmic design a much more challenging task. In chapter 9, we study the common techniques which are often used in the context of a variety of text mining tasks.

**Cross-Lingual Mining of Text Data:** With the proliferation of web-based and other information retrieval applications to other applications, it has become particularly useful to apply mining tasks in different languages, or use the knowledge or corpora in one language to another. For example, in cross-language mining, it may be desirable to cluster a group of documents in different languages, so that documents from different languages but similar semantic topics may be placed in the same cluster. Such cross-lingual applications are extremely rich, because they can often be used to leverage knowledge from one data domain into another. In chapter 10, we will study methods for cross-lingual mining of text data, covering techniques such as machine translation, cross-lingual information retrieval, and analysis of comparable and parallel corpora.

**Text Mining in Multimedia Networks:** Text often occurs in the context of many multimedia sharing sites such as *Flickr* or *Youtube*. A natural question arises as to whether we can enrich the underlying mining process by simultaneously using the data from other domains

together with the text collection. This is also related to the problem of transfer learning, which was discussed earlier. In chapter 11, a detailed survey will be provided on mining other multimedia data together with text collections.

**Text Mining in Social Media:** One of the most common sources of text on the web is the presence of social media, which allows human actors to express themselves quickly and freely in the context of a wide range of subjects [2]. Social media is now exploited widely by commercial sites for influencing users and targeted marketing. The process of mining text in social media requires the special ability to mine dynamic data which often contains poor and non-standard vocabulary. Furthermore, the text may occur in the context of linked social networks. Such links can be used in order to improve the quality of the underlying mining process. For example, methods that use both link and content [4] are widely known to provide much more effective results which use only content or links. Chapter 12 provides a detailed survey of text mining methods in social media.

**Opinion Mining from Text Data:** A considerable amount of text on web sites occurs in the context of product reviews or opinions of different users. Mining such opinionated text data to reveal and summarize the opinions about a topic has widespread applications, such as in supporting consumers for optimizing decisions and business intelligence. spam opinions which are not useful and simply add noise to the mining process. Chapter 13 provides a detailed survey of models and methods for opinion mining and sentiment analysis.

**Text Mining from Biomedical Data:** Text mining techniques play an important role in both enabling biomedical researchers to effectively and efficiently access the knowledge buried in large amounts of literature and supplementing the mining of other biomedical data such as genome sequences, gene expression data, and protein structures to facilitate and speed up biomedical discovery. As a result, a great deal of research work has been done in adapting and extending standard text mining methods to the biomedical domain, such as recognition of various biomedical entities and their relations, text summarization, and question answering. Chapter 14 provides a detailed survey of the models and methods used for biomedical text mining.

### 3. Future Directions

The rapid growth of online textual data creates an urgent need for powerful text mining techniques. As an interdisciplinary field, text data mining spans multiple research communities, especially data mining,

natural language processing, information retrieval, and machine learning with applications in many different areas, and has attracted much attention recently. Many models and algorithms have been developed for various text mining tasks have been developed as we discussed above and will be surveyed in the rest of this book.

Looking forward, we see the following general future directions that are promising:

- **Scalable and robust methods for natural language understanding:** Understanding text information is fundamental to text mining. While the current approaches mostly rely on bag of words representation, it is clearly desirable to go beyond such a simple representation. Information extraction techniques provide one step forward toward semantic representation, but the current information extraction methods mostly rely on supervised learning and generally only work well when sufficient training data are available, restricting its applications. It is thus important to develop effective and robust information extraction and other natural language processing methods that can scale to multiple domains.
- **Domain adaptation and transfer learning:** Many text mining tasks rely on supervised learning, whose effectiveness highly depends on the amount of training data available. Unfortunately, it is generally labor-intensive to create large amounts of training data. Domain adaptation and transfer learning methods can alleviate this problem by attempting to exploit training data that might be available in a related domain or for a related task. However, the current approaches still have many limitations and are generally inadequate when there is no or little training data in the target domain. Further development of more effective domain adaptation and transfer learning methods is necessary for more effective text mining.
- **Contextual analysis of text data:** Text data is generally associated with a lot of context information such as authors, sources, and time, or more complicated information networks associated with text data. In many applications, it is important to consider the context as well as user preferences in text mining. It is thus important to further extend existing text mining approaches to further incorporate context and information networks for more powerful text analysis.
- **Parallel text mining:** In many applications of text mining, the amount of text data is huge and is likely increasing over time,

thus it is infeasible to store the data in one machine, making it necessary to develop parallel text mining algorithms that can run on a cluster of computers to perform text mining tasks in parallel. In particular, how to parallelize all kinds of text mining algorithms, including both unsupervised and supervised learning methods is a major future challenge. This direction is clearly related to cloud computing and data-intensive computing, which are growing fields themselves.

## References

- [1] C. Aggarwal. *Data Streams: Models and Algorithms*, Springer, 2007.
- [2] C. Aggarwal. *Social Network Data Analytics*, Springer, 2011.
- [3] R. A. Baeza-Yates, B. A. Ribeiro-Neto, *Modern Information Retrieval - the concepts and technology behind search, Second edition*, Pearson Education Ltd., Harlow, England, 2011.
- [4] S. Chakrabarti, B. Dom, P. Indyk. Enhanced Hypertext Categorization using Hyperlinks, *ACM SIGMOD Conference*, 1998.
- [5] W. B. Croft, D. Metzler, T. Strohmaier, *Search Engines - Information Retrieval in Practice*, Pearson Education, 2009.
- [6] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, R. Harshman. Indexing by Latent Semantic Analysis. *JASIS*, 41(6), pp. 391–407, 1990.
- [7] D. A. Grossman, O. Frieder, *Information Retrieval: Algorithms and Heuristics (The Kluwer International Series on Information Retrieval)*, Springer-Verlag New York, Inc, 2004.
- [8] J. Han, M. Kamber. *Data Mining: Concepts and Techniques*, 2nd Edition, Morgan Kaufmann, 2005.
- [9] C. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
- [10] I. T. Jolliffe. Principal Component Analysis. *Springer*, 2002.
- [11] S. J. Pan, Q. Yang. A Survey on Transfer Learning, *IEEE Transactions on Knowledge and Data Engineering*, 22(10): pp 1345–1359, Oct. 2010.
- [12] G. Salton. *An Introduction to Modern Information Retrieval*, McGraw Hill, 1983.
- [13] K. Sparck Jones P. Willett (ed.). *Readings in Information Retrieval*, Morgan Kaufmann Publishers Inc, 1997.

---

# Computational Text Analysis for Social Science: Model Assumptions and Complexity

---

Brendan O'Connor\* David Bamman† Noah A. Smith†\*

\*Machine Learning Department

†Language Technologies Institute

Carnegie Mellon University

{brenocon, dbamman, nasmith}@cs.cmu.edu

December 8, 2011

Second Workshop on Computational Social Science  
and Wisdom of the Crowds (NIPS 2011)

## Abstract

Across many disciplines, interest is increasing in the use of computational text analysis in the service of social science questions. We survey the spectrum of current methods, which lie on two dimensions: (1) computational and statistical model complexity; and (2) domain assumptions. This comparative perspective suggests directions of research to better align new methods with the goals of social scientists.

## 1 Use cases for computational text analysis in the social sciences

The use of computational methods to explore research questions in the social sciences and humanities has boomed over the past several years, as the volume of data capturing human communication (including text, audio, video, etc.) has risen to match the ambitious goal of understanding the behaviors of people and society [1]. Automated content analysis of text, which draws on techniques developed in natural language processing, information retrieval, text mining, and machine learning, should be properly understood as a class of quantitative social science methodologies. Employed techniques range from simple analysis of comparative word frequencies to more complex hierarchical admixture models. As this nascent field grows, it is important to clearly present and characterize the assumptions of techniques currently in use, so that new practitioners can be better informed as to the range of available models.

To illustrate the breadth of current applications, we list a sampling of substantive questions and studies that have developed or applied computational text analysis to address them.

- Political Science: How do U.S. Senate speeches reflect agendas and attention? How are Senate institutions changing [27]? What are the agendas expressed in Senators' press releases [28]? Do U.S. Supreme Court oral arguments predict justices' voting behavior [29]? Does social media reflect public political opinion, or forecast elections [12, 30]? What determines international conflict and cooperation [31, 32, 33]? How much did racial attitudes affect voting in the 2008 U.S. presidential election [34]?
- Economics: How does sentiment in the media affect the stock market [2, 3]? Does sentiment in social media associate with stocks [4, 5, 6]? Do a company's SEC filings predict aspects of stock performance [7, 8]? What determines a customer's trust in an online merchant [9]? How can we measure macroeconomic variables with search queries and social media text [10, 11, 12]? How can we forecast consumer demand for movies [13, 14]?
- Psychology: How does a person's mental and affective state manifest in their language [15]? Are diurnal and seasonal mood cycles cross-cultural [16]?

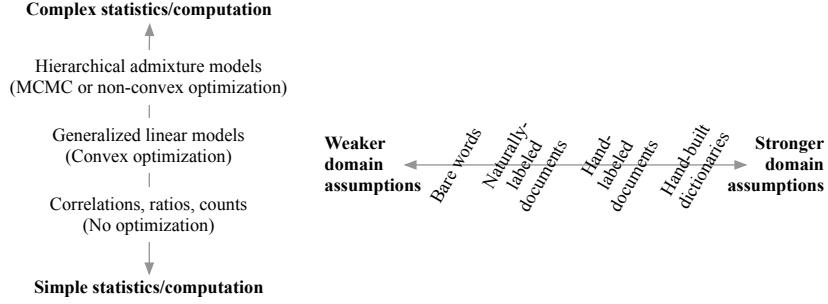


Figure 1: Schematic of model complexity versus domain assumptions for various computational text analysis methods. Statistical models are listed with their respective inference/training algorithms; computational expense increases with model expressiveness.

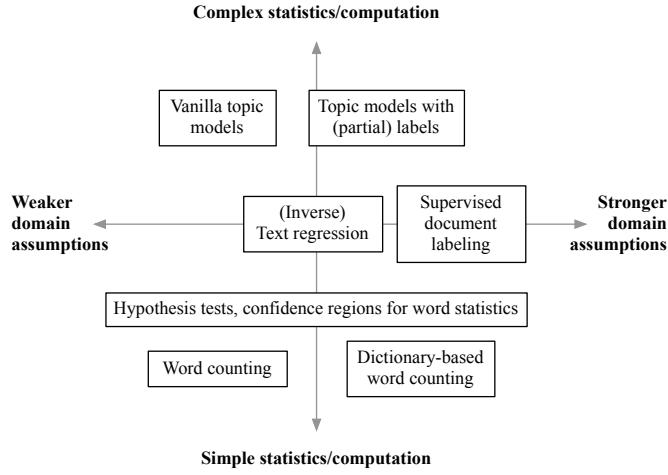


Figure 2: Typical methods used in computational text analysis. Compare to Table 1 in [27].

- Scientometrics/Bibliometrics: What are influential topics within a scientific community? What determines a paper's citations [35, 36, 37, 38]?
- Sociolinguistics: What are the geographic determinants of lexical variation in social media [39]? Demographic determinants [12, 40]?
- Public Health: How can search queries and social media help measure levels of the flu and other public health issues [41, 42, 43]?
- History: How did modern English legal institutions develop over the 17th to 20th centuries [17]? When did concepts of religion, secularism, and social institutions develop over two millennia of Latin literature [18]? What can topical labels in Diderot's 18th Encyclopédie reveal about contemporary ways of thought [19]? Who were the true authors of a particular piece of historical text [20]? The last deserves mention as a classic (1964) work that analyzed pseudonymous Federalist papers and answered long-standing questions about their authorship—one of the earliest instances of automated, statistical *stylometry* and automated text analysis for social science in general.
- Literature: What are the textual allusions in Classical Latin poetry [21] and the synoptic gospels [22]? How do demographic determinants of fictional characters affect their language use [23]? Who is the true author of a work of literature [24]? Roberto Busa's work in digitizing and lemmatizing the complete works of Thomas Aquinas, begun in 1949, also deserves mention as one of the earliest efforts at creating a machine-readable *annotated corpus* [25, 26].

This list is incomplete, both in the works cited and in the range of areas that have used or could use these methods. These techniques are still in their infancy: while several of the works above thoroughly address substantive questions (and, in a few cases, there exist lines of work published in high-quality social science journals), most tend to focus on developing new methodologies or data sources. There are also more exploratory analyses not aimed at specific research questions [44, 45]. In most cases, automated text analysis functions as a tool for *discovery* and *measurement* of prevalent attitudes, concepts, or events in textual data.

## 2 Classes of methods

In most cases, the analysis is restricted to the frequencies of words or short phrases ( $n$ -grams) in documents and corpora.<sup>1</sup> Even so, there is still a rich variety of methods, with two important axes of variation: statistical model assumptions and domain assumptions (Figure 1).

**Domain assumptions** refer to how much knowledge of the substantive issue in question is used in the analysis. A purely exploratory, “bare words” analysis only considers the words of documents; for example, examining the most common words in a corpus, or latent topics extracted automatically from them. Next, non-textual metadata about the documents is almost always used; for example, tracking word frequencies by year of book publication [45], or spatial location of a microblogger [39]. We call these *naturally-labeled* documents; typically the labels take the form of continuous, discrete, or ordinal variables associated with documents or segments of them. In contrast, *manually-labeled* documents may be created in order to better understand particular quantities of interest; for example, annotating news articles with information about the events they describe [32]. Creating and evaluating the codebook (that carefully defines the semantics of the annotations a coder will produce) can be a laborious and iterative process, but is essential to understand the problem and create a psychometrically reliable coding standard [46]. Finally, another source of domain information can take the form of *dictionaries*: lists of terms of interest to the analysis, such as names, emotional words, topic-specific keywords, etc. Ideally, they may be custom-built for the problem (if done manually, a labor-intensive task similar to coding documents), or they may be reused or adapted from already-existing dictionaries (e.g., Freebase [47] for names or LIWC [15] for affect, though see [48]’s critical comments on the naïve use of affect dictionaries). Useful information can be revealed with just a handful of terms; for example, [34] analyzes the Google search query frequencies of one highly charged racial epithet as a proxy for racial attitudes.

The second dimension is **computational and statistical complexity**.<sup>2</sup> The simplest techniques count words. A typical analysis is to compare word frequencies between groups (e.g., listing the most common words per speaker in a debate). Note that any type of comparison requires some form of natural labels for text; generally, metadata is what links text to interesting substantive questions. In this case, the metadata is speaker identity, but time, social group membership, and others have also been considered—as in the common analysis of plotting a word’s frequency over time.

Frequency ratios and correlations with response variables can be seen as parameters or hypothesis tests for simple two-variable models between text frequencies and response/metadata variables; in the case of a categorical metadata variable, words’ conditional probabilities  $p(x | y)$  correspond to parameters of the naïve Bayes model of text. A hallmark of this class is that they are computationally straightforward to calculate; typically they involve a single pass through the data to compute counts, sums, and other quantities.

Another popular set of techniques sits on the other side of the computational spectrum: hierarchical admixture models, specifically LDA-style topic models [49]. Here, documents undergo dimension reduction by being modeled as mixtures of multinomials, where each component is a distribution over words—called a topic. The output of a topic model can be used for exploratory analysis, or post-hoc compared across observed variables. With some work, these models can also be usefully customized for a variety of applications; typically, an important change is to incorporate the natural labels and structure of the domain. (Models that can incorporate reasonably generic types of labels, and in substantially different ways, include SLDA, DMR, and PLDA [50, 51, 36].)

---

<sup>1</sup>A few interesting exceptions: [18, 9, 31].

<sup>2</sup>We use the term “complexity” informally, not intending to imply any of its technical senses.

Another class of techniques is generalized linear models, and specifically regularized linear and logistic regression [52]. In *text regression*, the response variable is modeled as a conditional distribution given a linear combination of text features,  $p(y | x)$ . This model has often been used for the task of *text categorization*, to predict a document’s category according to a training set of prespecified labeled documents; research has shown that linear models are state-of-the-art for this problem ([53] §14.7-8). A researcher can manually label documents and train a classifier to aid in the analysis of a large document collection; but these models can also be used for natural labels, to directly model a response variable of interest with text. We prefer the term *regression* for both discrete and continuous response variables, to emphasize these models’ connections to the extensively developed statistical literature in GLMs and applied regression analysis [54, 55, 56].

An alternative is *inverse text regression*, where  $p(x | y)$  is modeled as a multinomial logistic regression over the vocabulary, using the document labels [57] (or possibly latent variables [58]) as features. This direction of conditioning is more like naïve Bayes and (labeled) topic models in that it grounds out as multinomials over the vocabulary, but with linear parameterization of the multinomials, using additive effects instead of mixture memberships to select word probabilities.

### 3 Considerations

Which method to use completely depends on the goals and needs of the analysis: all three can be used for descriptive analysis and prediction. One consideration is the usual tradeoff between **simplicity** and **expressiveness**. Frequencies and correlations are easily computed and replicable; regressions require more computation, though often have unique solutions and off-the-shelf solvers; while topic models use fitting procedures that are more expensive (MCMC), or less flexible (variational inference), and may be less stable in that different runs can produce different results. This is part of the tradeoff of their greater expressive power. Regressions have the same level of expressiveness as word frequencies, but control for covariation through additive effects, where a word’s coefficient explains the specific effect of that word when controlling for other words and other covariates. ([37] illustrates how this can make a difference for analysis.)

We should note that all the methods described in the previous section assign vectors of weights across the vocabulary, giving words associations to non-textual document-level variables, and are therefore fundamentally **interpretable**, because the researcher can inspect words’ numeric weights. Word correlations and regressions associate words to observed document label variables, while topic models associate words to hidden topic variables. (Per-word association weights are individual correlations, regression coefficients, or conditional topic probabilities, respectively.) In all these cases, a way to summarize a particular document-level variable, then, is to look at the top-weighted words for that vector – e.g., the top 10 words with highest probability under a topic, or highest coefficient for a label class, or highest correlation/frequency. An analyst can then view the corpus through the lens of these top-words lists and their associated variables. This level of interpretability is a major advantage over black-box non-linear methods like kernel methods (e.g. kernelized SVMs) or neural networks, especially given that linear methods often have similar predictive performance.

A third consideration is what sort of the relationship between text and **observed variables** the researcher is interested in. If there are few observed variables, then topic models can still be used for purely exploratory analysis. However, since many of the substantive questions researchers are interested in typically involve conditioning on observed variables to make comparisons (whether the observed variables are natural or hand-labeled), it is useful to allow the model to tie relevant textual features to the variables in question.

For some problems, like analyzing Congressional floor speeches, topics correspond quite well to the substantive issues under consideration [27]. But for other problems, they can work less well. As one example, we have observed several cases where SLDA (an LDA variant that models a document-level variable through a GLM regression on topic proportions [50]) has similar [59] or worse [39, 57] predictive performance than regularized text regression. For the problem of predicting U.S. users’ locations from their microblog text [39], we observed that Lasso regression selected a small number of words to have non-zero coefficients (e.g., “taco” to indicate the West Coast). We believe that in SLDA, the impact of these sparse cues was blunted from their incorporation into broad topics, since the model had to explain not just the response, but also the entirety of all the text. Sometimes the relationship between text and the document variable is better explained by individual words alone.

The extremes of individual word frequencies versus broad topic proportions are only two points in the space of possible text representations; it remains an interesting open question how to design models that can reliably abstract beyond individual words in service of social science analysis.

## Acknowledgments

This research was supported by the NSF through grants IIS-0915187 and IIS-1054319 and by Google through the Worldly Knowledge Project at CMU. The authors thank the anonymous reviewers for helpful comments.

## References

- [1] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszl Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915):721–723, February 2009.
- [2] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.
- [3] Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, and James Allan. Mining of concurrent text and time series. In *Proceedings of KDD Workshop on Text Mining*, pages 37–44, 2000.
- [4] Eric Gilbert and Karrie Karahalios. Widespread worry and the stock market. In *Proceedings of the International Conference on Weblogs and Social Media*, 2010.
- [5] Sanjiv R. Das and Mike Y. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, September 2007.
- [6] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *1010.3003*, October 2010.
- [7] Shimon Kogan, Dmitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 272280, 2009.
- [8] Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance (forthcoming)*, 2011.
- [9] Nikolay Archak, Anindya Ghose, and Panagiotis Ipeirotis. Deriving the pricing power of product features by mining consumer reviews. *Management Science*, page mnscl110, 2011.
- [10] Nikolaos Askitas and Klaus F. Zimmermann. Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2):107–120, April 2009.
- [11] Matthew E. Kahn and Matthew J. Kotchen. Environmental concern and the business cycle: The chilling effect of recession. <http://www.nber.org/papers/w16241>, July 2010.
- [12] Brendan O’Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *International AAAI Conference on Weblogs and Social Media, Washington, DC*, 2010.
- [13] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. *1003.5699*, March 2010.
- [14] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 293296, 2010.
- [15] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 2009.
- [16] Scott A. Golder and Michael W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333:1878–1881, September 2011.

- [17] Dan Cohen, Frederick Gibbs, Tim Hitchcock, Geoffrey Rockwell, Jorg Sander, Robert Shoemaker, Stefan Sinclair, Sean Takats, William J. Turkel, Cyril Briquet, Jamie McLaughlin, Milena Radzikowska, John Simpson, and Kirsten C. Uszkalo. Data mining with criminal intent. Final white paper, 2011.
- [18] David Bamman and Gregory Crane. Measuring historical word sense variation. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries*, page 110, 2011.
- [19] Russell Horton, Robert Morrissey, Mark Olsen, Glenn Roe, and Robert Voyer. Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopédie. *Digital Humanities Quarterly*, 3(2), 2009.
- [20] Frederick Mosteller and David Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, 1964.
- [21] David Bamman and Gregory Crane. The logic and discovery of textual allusion. In *Proceedings of the 2008 LREC Workshop on Language Technology for Cultural Heritage Data*, 2008.
- [22] John Lee. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 472–479, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- [23] Shlomo Argamon, Charles Cooney, Russell Horton, Mark Olsen, Sterling Stein, and Robert Voyer. Gender, race, and nationality in black drama, 1950-2006: Mining differences in language use in authors and their characters. *Digital Humanities Quarterly*, 3(2), 2009.
- [24] David I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998.
- [25] Roberto Busa. The annals of humanities computing: The index thomisticus. *Language Resources and Evaluation*, 14:83–90, 1980.
- [26] Roberto Busa. *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SI*. Frommann-Holzboog, Stuttgart-Bad Cannstatt, 1974–1980.
- [27] Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209228, 2010.
- [28] Justin Grimmer. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1, 2010.
- [29] Ryan C. Black, Sarah A. Treul, Timothy R. Johnson, and Jerry Goldman. Emotions, oral arguments, and Supreme Court decision making. *The Journal of Politics*, 73(2):572–581, April 2011.
- [30] Panagiotis T. Metaxas, Eni Mustafaraj, and Daniel Gayo-Avello. How (Not) to predict elections. Boston, MA, 2011.
- [31] Philip A. Schrodt, Shannon G. Davis, and Judith L. Weddle. KEDS – a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561 –587, December 1994.
- [32] Gary King and Will Lowe. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57(3):617–642, July 2003.
- [33] Stephen M. Shellman. Coding disaggregated intrastate conflict: machine processing the behavior of substate actors over time and space. *Political Analysis*, 16(4):464, 2008.
- [34] Seth Stephens-Davidowitz. The effects of racial animus on voting: Evidence using Google search data. Job market paper, downloaded from <http://www.people.fas.harvard.edu/~sstephen/papers/RacialAnimusAndVotingSethStephensDavidowitz.pdf>, November 2011.
- [35] Sean M. Gerrish and David M. Blei. A language-based approach to measuring scholarly impact. In *Proceedings of ICML Workshop on Computational Social Science*, 2010.

- [36] Daniel Ramage, Christopher D. Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 457465, 2011.
- [37] Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. Predicting a scientific community's response to an article. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011.
- [38] Steven Bethard and Dan Jurafsky. Who should I cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, page 609618, 2010.
- [39] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, page 12771287, 2010.
- [40] Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of ACL*, 2011.
- [41] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, February 2009.
- [42] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. 2010.
- [43] Michael J. Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *Proceedings of ICWSM*, 2011.
- [44] Peter S. Dodds and Christopher M. Danforth. Measuring the happiness of Large-Scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, page 116, 2009.
- [45] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176 –182, January 2011.
- [46] Klaus Krippendorff. *Content analysis: an introduction to its methodology*. Sage Publications, Inc, 2004.
- [47] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, Vancouver, Canada, 2008. ACM.
- [48] Justin Grimmer and Brandon M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. <http://www.stanford.edu/~jgrimmer/tad2.pdf>, 2011.
- [49] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:9931022, 2003.
- [50] David M. Blei and Jon D. McAuliffe. Supervised topic models. *arXiv:1003.0783*, March 2010.
- [51] David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*, page 411418, 2008.
- [52] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, June 2009.
- [53] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 1st edition, July 2008.
- [54] Sanford Weisberg. *Applied linear regression*. John Wiley and Sons, 2005.
- [55] Alan Agresti. *Categorical data analysis*. John Wiley and Sons, 2002.
- [56] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multi-level/Hierarchical Models*. Cambridge University Press, 1 edition, December 2006.
- [57] Matthew A. Taddy. Inverse regression for analysis of sentiment in text. *arXiv:1012.2098*, December 2010.
- [58] Jacob Eisenstein, Ahmed Ahmed, and Eric P. Xing. Sparse additive generative models of text. *Proceedings of ICML*, 2011.

- [59] Sean M. Gerrish and David M. Blei. Predicting legislative roll calls from text. In *Proceedings of ICML*, 2011.

## Introduction to the Special Issue: The Statistical Analysis of Political Text

**Burt L. Monroe**

*Department of Political Science, Pennsylvania State University, University Park, PA 16802*  
*e-mail: burtmonroe@psu.edu (corresponding author)*

**Philip A. Schrodt**

*Department of Political Science, University of Kansas, Lawrence, KS 66045*

Text is arguably the most pervasive—and certainly the most persistent—artifact of political behavior. Extensive collections of texts with clearly recognizable political—as distinct from religious—content go back as far as 2500 BCE in the case of Mesopotamia and 1300 BCE for China, and 2400-year-old political discussions dating back to the likes of Plato, Aristotle, and Thucydides are common fare even in the introductory study of political thought. Political tracts were among the earliest productions following the introduction of low-cost printing in Europe—fueling more than a few revolutions and social upheavals—and continuous printed records of legislative debates, such as the British parliament’s *Hansard* and precursors tracing to 1802, cover centuries of political discussion.

The possibility that the analysis of texts could provide insights into the political processes also has a long pedigree. The Italian humanist Lorenzo Valla’s careful philological analysis of the reputed *Donation of Constantine* in 1439 convincingly demonstrated, using purely textual methods, that the document was a medieval forgery that must have postdated the Emperor Constantine I by at least four centuries. Moving toward our own day, the first modern, theoretically driven content analysis project was Harold Lasswell’s Wartime Communications Project just prior to the outbreak of World War II (Janowitz 1968), and subsequently, content analysis became a standard analytical tool in the West, particularly for the analysis of “enemy” communications, first Nazi and later Communist.

Given the tedious nature of human coding, researchers recognized early on that systematic textual analysis might be suited to automation. Harvard’s *General Inquirer* (Stone et al. 1966) emerged as the first widely used computer program for automated content analysis, and re-written from the original IBM language PL/I into Java, it persists to this day.<sup>1</sup> However, most of the development of automated tools within the context of the social sciences occurred in Europe (see Alexa and Zuell 1999; Popping 2000), though automated natural language processing in general continued apace in computer science (e.g., Salton 1989; Smith 1990), and occasionally, these efforts would move into the political realm (e.g., ARPA 1993).

<sup>1</sup><http://www.wjh.harvard.edu/~inquirer/>; accessed August 4, 2008.

© The Author 2009. Published by Oxford University Press on behalf of the Society for Political Methodology.  
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Although limited computer capacity was a problem in some of the early applications—any reasonably sized corpus of text will contain thousands of distinct words—by the 1980s, the primary bottleneck to the practical application of text analysis was the availability of machine-readable input. The manual entry of texts was at least as costly and time consuming as simply coding them directly from paper or microfilm, particularly when large-scale and relatively diffuse sources such as debates and news reports were concerned, and consequently, human coding remained the norm. This problem gradually began to abate in the late 1980s with the availability of large-scale online textual databases such as Lexis-Nexis and then, in the late 1990s, the floodgates opened as the World Wide Web expanded.

The Web revolutionized the availability of texts, providing material that was accessible, reasonably standardized in the form of HTML files, and, critically for academic researchers, free of charge. By the first decade of the 21st century, vast quantities of politically relevant texts were available, ranging from the rants of political bloggers to official campaign statements to parliamentary debates to day-to-day news reports. Although such sources cannot be immediately analyzed—text still needs to be extracted from individual web sites using various forms of “web scraping” (Schrenk 2007)—tools for downloading and filtering web content have become increasingly common, user-friendly, and generalized, and once a system has been developed for a particular site, the marginal cost of acquiring additional data is usually close to zero.

As a consequence of these developments, automated content analysis in political science has experienced considerable growth in recent years. The pivotal application in the political science mainstream was Benoit and Laver’s *Wordscores* (Benoit and Laver 2003; Laver et al. 2003), which attracted considerable attention and is discussed in several of the papers in this volume. Another early contribution was the discussion of latent semantic analysis in this journal by Simon and Xenos (2004). The 2006 American Political Science Association meeting featured a workshop titled “Automated Content Analysis and Computer Annotation” organized by Stephen Purpura, which attracted extensive participation, and the 2007 Midwest Political Science Association meeting featured two standing-room-only panels on systematic textual analysis. The topic has taken off in the methodological community as well. Following on a series of single workshops and papers given at the annual summer meetings (Monroe and Maeda 2004; Quinn et al. 2006; Schrodt 2003), the 2008 meetings featured four graduate student posters, from four different institutions, on text analysis topics (Goodrich 2008 [NYU]; Grimmer 2008 [Harvard]; Pemstein 2008 [Illinois]; Sagarzazu 2008 [Houston]). Based on this dramatic growth in a topic that barely existed 5 years ago, we suggested this special issue of *Political Analysis*.

The APSA workshop had focused on two topics, and ours has been on the first: fully automated methods. In our call for papers, we defined the scope of this issue as

... techniques where most of the data-generation is fully automated, as distinct from computer-aided mark-up systems. This does not mean the entire system has to be fully automated—a system might, for example, use lists of words or phrases or parsing rules that were derived by human coders—but the final text processing needs to be completely automated. As a rule-of-thumb, we consider a system fully automated if the marginal cost of analyzing additional texts goes to zero as the size of the corpus being analyzed increases, and the coding is completely replicable given a set of software, dictionaries, and so forth.

The other topic of the workshop—computer-assisted text mark-up—has also attracted considerable attention and is the focus of a recent issue of the *Journal of Information Technology and Politics* (Cardie and Wilkerson 2008).

The automated methods discussed in this volume generally derive from two literatures. The first is classical content analysis (Holsti 1969; Krippendorff 2004; Neuendorf 2001;

Roberts 1997; Weber 1990), updated with the use of machine-readable texts and automated coding. The second is the very large literature from computer science and computational linguistics on natural language processing in general (see, e.g., Jurafsky and Martin 2008; Manning and Schütze 2002). All these methods are directed toward specific applications in the study of politics, such as determining ideological position from texts, coding political interactions, and identifying the content of political conflict.

Methodologically, there is a practical trade-off between an emphasis on *statistical* modeling and *language* modeling. Variants of the “bag-of-words” methods are concerned with the frequencies of words or *n*-grams (*n*-word phrases), without concern for syntax. This limits considerably the amount of information that can be extracted from an individual phrase or sentence, but allows inferential models of large corpora to be built on assumptions of count or discrete choice processes, and can easily be applied in multiple languages. Conversely, some problems require far more attention to syntax. For example, we may want to infer from a news report not just the identity of two countries involved in a military engagement, but *who* attacked *whom*. This requires modeling of the rules of a language—for example, how it distinguishes subjects from objects—in a more sophisticated way than is possible with bag-of-words. This trade-off can be observed across the issue, with relatively more emphasis on statistics in the first papers and more emphasis on language in the last.

Lowe takes on the most widely used algorithm, the aforementioned *WordScores*. Although acknowledging the flexibility and computational tractability of the method, Lowe notes at least two major problems with it: scaling issues and the absence of an underlying statistical model. Lowe goes on to demonstrate that *WordScores* is essentially equivalent to an older and well-understood scaling method, correspondence analysis but that methods based on item response theory are reasonable estimators of “ideal points” under a broader range of conditions.

Monroe, Colaresi, and Quinn discuss a variety of different approaches to the problem of *feature selection* and *feature evaluation*. They use a variety of techniques, with increasingly specific underlying assumptions about the data generation process, to examine the lexical differences between Democrats and Republicans in the United States Senate. Many techniques in broad use for such purposes miss the mark by poor information accounting from (generally unspecified) underlying models and from overfitting. They demonstrate that techniques of regularization and shrinkage, accomplished through the use of Bayesian priors, provide more useful and substantively meaningful results.

Bailey and Schonhardt-Bailey demonstrate the theoretical traction that automated analysis can afford. Substantively, they are concerned with the dramatic shift in United States monetary policy, the “Volcker Revolution,” which occurred in 1979. Methodologically, they are interested in demonstrating the existence of deliberative persuasion as it occurred in meetings of the Federal Open Market Committee. To this end, they deploy clustering and scaling methods available in a commercial software package, ALCESTE (a contrast with the “roll-your-own” approaches evident in the other papers of the issue), to demonstrate dynamic change in the structure of monetary policy debate over several years.

Beigman Klebanov, Diermeier, and Beigman investigate the concept of *lexical cohesion*. In this piece, we see an effort to explore and exploit semantic relationships among words in a text. They build on the hierarchical conceptual database *WordNet* (Miller 1990) and prior work by Beigman Klebanov (2006) to use *WordNet* to develop a measure of “semantic relatedness” and, in turn, cohesion. They apply the technique to evaluate the ideological cohesion in speeches by Margaret Thatcher, contrasting their technique with conventional approaches based on the study of rhetoric.

The article by Shellman deals with the issue of automated event data coding, with a particular focus on the degree to which automated coding allows the creation of substantially

more detailed data sets than are found in the older, human-coded approaches. Shellman's PCS system, which he has applied to several southeast Asia conflicts, provides a much higher level of detail on internal actors and can also use a far greater number of source texts, while still maintaining the relatively low costs, high speed, and reliability of automated coding.

The final contribution, by van Atteveldt, Kleinnijenhuis, and Ruigrok, is distinctive in several respects. First, it is the only article in our collection working in a language other than English (Dutch, in this instance), though we should note that with little or no modification, all the methods discussed in this issue should work in any human language. The article is the sharpest departure from a bag-of-words approach, applying a more complex *semantic network analysis* based on a syntactic analysis and pattern matching. Perhaps most critically—and fittingly for the final article in our collection—van Atteveldt et al. systematically demonstrate that when data generated by automated methods is used in hypothesis testing, it yields results statistically indistinguishable from those generated by manually coded data.

The techniques presented in this issue by no means exhaust the available methods, and the source texts explored in these studies are only a tiny fraction of those available. Our hope is that readers interested in these methods will use the approaches discussed here—as well as those referenced tangentially in many of the articles—as a jumping-off point for further research and development. The past 10 years have seen a dramatic expansion of work on the automated analysis of political texts, but our sense is that we have only begun to tap the potential in this field. We present this issue as the departure lounge, not the baggage claim, and hope that it will inspire additional innovative work in the future.

## References

- Advanced Research Projects Agency (ARPA). 1993. *Proceedings of the Fifth Message Understanding Conference (MUC-5)*. Los Altos, CA: Morgan Kaufmann.
- Alexa, Melina, and Cornelia Zuell. 1999. *A review of software for text analysis*. Mannheim, Germany: Zentrum für Umfragen, Methoden und Analysen.
- Beigman Klebanov, Beata. 2006. "Measuring semantic relatedness using people and WordNet." *Proceedings of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp 13–7. New York, NY: Association for Computational Linguistics.
- Benoit, Kenneth, and M. Laver. 2003. Estimating Irish party positions using computer wordscore: The 2002 elections. *Irish Political Studies* 17:97–107.
- Bond, Doug, J. Craig Jenkins, Charles L. Taylor, and Kurt Schock. 1997. Mapping mass political conflict and civil society: The automated development of event data. *Journal of Conflict Resolution* 41:553–79.
- Cardie, Claire, and John Wilkerson. 2008. Special issue: Text annotation for political science. *Journal of Information Technology and Politics* 5:1–6.
- Goodrich, Melanie. 2008. "A coding methodology for open-ended survey questions. Poster presented to the Political Methodology Society, Ann Arbor, MI, July 10–12, 2008.
- Grimmer, Justin. 2008. "Expanding the study of political representation: Measuring and explaining representatives' expressed agenda. Poster presented to the Political Methodology Society, Ann Arbor, MI, July 10–12, 2008.
- Holsti, Ole R. 1969. *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.
- Janowitz, Morris. 1968. Harold D. Lasswell's contribution to content analysis. *Public Opinion Quarterly* 32:646–53.
- Jurafsky, Daniel, and James H. Martin. 2008. *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics*. 2nd ed. Upper Saddle River, NJ: Prentice-Hall.
- Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology*. 2nd ed. Thousand Oaks, CA: Sage.
- Laver, Michael, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review* 97:311–31.

- Manning, Christopher D., and Hinrich Schütze. 2002. *Foundations of statistical natural language processing*. 5th ed. Cambridge, MA: MIT Press.
- Miller, George. 1990. "WordNet: An on-line lexical database." *International Journal of Lexicography* 3:235–312.
- Monroe, Burt L., and Ko Maeda. 2004. "Talk's cheap: Text-based ideal point estimation. Paper presented to the Political Methodology Society, Palo Alto, July 29–31, 2004.
- Neuendorf, Kimberly A. 2001. *The content analysis guidebook*. New York: Sage.
- Pemstein, Daniel. 2008. "Predicting strategic roll calls with legislative text. Poster presented to the Political Methodology Society, Ann Arbor, MI, July 10–12, 2008.
- Popping, Roel. 2000. *Computer-assisted text analysis*. New York: Sage.
- Quinn, Kevin M., Burt L. Monroe, Michael P. Colaresi, Michael H. Crespin, and Dragomir Radev. 2006. "An automated method of topic-coding legislative speech over time with application to the 105<sup>th</sup>–108<sup>th</sup> United States Senate. Paper presented to the Political Methodology Society, Davis, CA, July 20–22, 2006.
- Roberts, Carl W. 1997. Text analysis for the social sciences: Methods for drawing inferences from texts and transcripts. Mahwah, NJ: Lawrence Erlbaum.
- Sagarzazu, Inaki. 2008. "Look who's talking: Analyzing the Dynamics of Political Discourse. Poster presented to the Political Methodology Society, Ann Arbor, MI, July 10–12, 2008.
- Salton, Gerard. 1989. *Automatic Text Processing*. Reading, MA: Addison-Wesley.
- Schrenk, Michael. 2007. *Webbots, spiders, and screen scrapers*. San Francisco, CA: No Starch Press.
- Schrodt, Philip. 2003. "Analyzing text using statistical methods. Workshop presented to the Political Methodology Society, Minneapolis, MN, July 17–19, 2003.
- Simon, Adam F., and Michael Xenos. 2004. "Dimensional reduction of word-frequency data as a substitute for intersubjective content analysis." *Political Analysis* 12:63–75.
- Smith, Peter D. 1990. *An introduction to text processing*. Cambridge, MA: MIT Press.
- Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The general inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.
- Weber, Robert P. 1990. *Basic content analysis*. 2nd ed. Newbury Park, CA: Sage Publications.



ANNUAL REVIEWS **Further**

Click here to view this article's online features:

- Download figures as PPT slides
- Navigate linked references
- Download citations
- Explore related articles
- Search keywords

# Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges

John Wilkerson and Andreu Casas

Department of Political Science, University of Washington, Seattle, Washington 98195;  
email: jwilker@uw.edu

Annu. Rev. Polit. Sci. 2017. 20:529–44

The *Annual Review of Political Science* is online at  
[polisci.annualreviews.org](http://polisci.annualreviews.org)

<https://doi.org/10.1146/annurev-polisci-052615-025542>

Copyright © 2017 by Annual Reviews.  
All rights reserved

## Keywords

text as data, automatic coding, machine learning, computational social sciences

## Abstract

Text has always been an important data source in political science. What has changed in recent years is the feasibility of investigating large amounts of text quantitatively. The internet provides political scientists with more data than their mentors could have imagined, and the research community is providing accessible text analysis software packages, along with training and support. As a result, text-as-data research is becoming mainstream in political science. Scholars are tapping new data sources, they are employing more diverse methods, and they are becoming critical consumers of findings based on those methods. In this article, we first describe the four stages of a typical text-as-data project. We then review recent political science applications and explore one important methodological challenge—topic model instability—in greater detail.

## INTRODUCTION

Words are an integral part of politics. Officials and citizens use words to express opinions, make proposals, and defend their actions. Laws and regulations are also largely codified in words. Political scientists have always been interested in words, but a revolution has occurred that is creating unprecedented research opportunities (Cardie & Wilkerson 2008, Monroe & Schrodт 2008, Alvarez 2016). The internet is providing an avalanche of data related to politics. For example, all departments and agencies of the US federal government must now post their public records on the internet, and many other governments adhere to similar practices. Most major newspapers offer online access to their archives. Project Gutenberg and Google Books offer free access to the complete texts of millions of books. Social media sites such as Twitter and Facebook encourage researchers to use their data. The Internet Archive offers archival information about millions of government web pages dating back to 1996.

The research community has responded to this surfeit of data by developing accessible open-source text analysis libraries in R, Python, and other programming languages (e.g., Munzert et al. 2014). The combination of so many untapped research opportunities and accessible tools and training makes this an excellent time for specialists in all areas to invest in text. Legislative scholars can now systematically investigate floor speeches, constituent communications, revisions to laws and regulations, and much more. International relations scholars can systematically compare final treaties or agreements to hundreds of proposals made at earlier stages. Political theorists can explore political thought by searching across centuries of published works. This newfound ability to computationally investigate text (as well as many other innovative data sources, such as images and sound) will transform political science research as scholars become more adept at exploiting the available opportunities.

Because not all readers may be familiar with text-as-data research, we first provide an overview of the four stages of a typical project. This overview highlights key considerations for potential projects and provides context for appreciating recent developments and methodological challenges. We then review recent political science applications and explore one important methodological challenge—topic model instability—in greater detail.

## FOUR STAGES OF A TEXT AS DATA PROJECT

Text-as-data methods expand research opportunities for political scientists in two ways. First, they leverage the power of computing to make ambitious data collection tasks feasible. Second, they offer a growing number of options for analyzing large volumes of text quantitatively. A typical text-as-data project proceeds through four stages. Text must be obtained, converted to quantitative data, analyzed, and validated.

### Obtaining Text

The first stage of a project usually entails downloading digitized content. For many projects, this is now a fairly minor step. However, it is probably wise to investigate what will be required before committing to a project. Some sources make it easy for researchers to get exactly what they need, whereas extracting relevant information from other sources can be difficult and time consuming. Application user interfaces (APIs) enable users to “request” selected content from an underlying structured database using a single line of code. APIs are ideal when they include options that serve the needs of a project. Examples include the multiple APIs offered by the *New York Times* (e.g.,

Article Search API, Congress API), the Sunlight Foundation (e.g., Open States, Capitol Words), and prominent social media sites (e.g., Twitter, Facebook).

If an API is not available, the next best option in terms of ease of use is obtaining documents that are similarly formatted. Identical formatting makes it possible to write a single script to extract more specific content from many documents at once, such as the thousands of congressional bill texts available through the Government Printing Office. Almost all documents contain hidden formatting language that may also be helpful for systematically extracting more specific content. The look and feel of a web page come from embedded HTML or XML tags. These tags may do little more than format the visible text, but they can be used to isolate desired content [see, e.g., the @unitedstates project (<https://theunitedstates.io/>)]. Other types of documents (.doc, .docx, .txt, .pdf) also contain hidden formatting that may provide unique markers to facilitate splitting. Even text formatting can be helpful. In transcripts of Federal Reserve Board meetings, only the speaker's name is printed in all capital letters ("MS. YELLEN") and can be used to easily split transcripts by speaker statement.

The most challenging text extraction or "scraping" projects are those that draw content from diverse sources. For example, extracting the same content from many different candidate websites is challenging because each website has a different structure. One option is to write multiple scripts. The OpenStates project (<https://openstates.org>) recruited volunteer programmers to write scores of scripts to extract information about legislative bills for different state government websites. For less ambitious projects, crowdsourcing may be more practical. Sites such as Mechanical Turk and Crowdflower farm out small tasks to thousands of workers around the world. For a small fee (often a few cents), these workers will (for example) copy and paste website content. Another option is to collect simpler metrics at the source, such as counts of keywords, a common approach of many "big data" projects (Carneiro & Mylonakis 2009, Leskovec et al. 2009, Schmidt 2015).

### From Text to Data

The content of each document must then be converted to quantitative data. Frequently, the objective is to create a term–document or term–frequency matrix where each row is a document and each column is a feature found in at least one of those documents.<sup>1</sup> Thus, at this stage researchers need to decide on the appropriate unit of analysis. For example, US presidential State of the Union addresses (SOUs) are lengthy and cover many different subjects; a project that examines SOU policy topics will probably be improved by splitting each address into more focused paragraphs or sentences.

The next step is to specify which features within each document will be used in the quantitative analysis. The starting point is usually to treat every unique word as a separate feature. Researchers then exclude document content that is thought to be irrelevant to the analysis and potentially misleading. Standard options include removing punctuation, common words (stopwords), very infrequent words (sparse terms), and word suffixes (stemming). However, each of these actions deserves careful consideration. For example, standard stopwords such as "can't" and "cannot" might be relevant features for a study of presidential address tone. The next step may be to create features beyond the basic bag of words. One common practice is to include word pairs (bigrams) as additional features. But the possibilities are truly endless. Instead of treating synonyms as separate

<sup>1</sup>The cell values indicate whether a feature is present (0,1) in a term–document matrix, or how often it is found (0,N) in a term–frequency matrix.

words, researchers might combine them into a single feature. They might also assign more weight to features that are thought to be especially informative, or create new features from outside information. Roberts et al. (2016) find that incorporating information about whether a blog has liberal or conservative leanings helps to predict its topics.

### Quantitative Analysis of Text

Simple metrics can be very useful and have the added virtues of transparency and replicability. Eggers & Spirling (2017) study parliamentary dynamics by examining frequencies of specific word usage across time. Casas et al. (2016) use lists of positive and negative words to study how the media portrays protestors. However, today much of the focus (some would say hype) is on statistical machine learning methods. Scholars continue to debate, water-cooler style, the differences between machine learning and statistics. We are certainly not going to settle that debate, but we do think that the distinction can help to highlight general differences in approach. Political scientists are accustomed to using statistical methods to test theories. They choose the best model for the data (ordinary least squares, logistic regression, etc.) before testing model specifications that include a limited number of theoretically derived input (independent) variables. The focus is typically on the coefficients or parameters for the input variables—e.g., other things equal, are women significantly more likely to identify as Democrats than men? Whether the model accurately predicts the partisan identification of each voter is usually of secondary concern.

In machine learning research, the focus is usually on the outputs rather than the inputs. Instead of asking whether women are more likely to identify as Democrats, a more typical objective would be to predict state-level political opinion using Twitter (Beauchamp 2017). This focus on outputs leads researchers to be more concerned with prediction accuracy and less concerned with explanation. Beauchamp reports the features most associated with pro-Obama and pro-Romney poll shifts but does not try to explain why (for example) the most important predictor for Obama support is “75” and the most important for Romney is “cia.” The focus on prediction also encourages more experimentation with different algorithms and features (Domingos 2015). We review some of the most relevant machine learning applications later in this article.

### Evaluating Performance

Validation is a critical component of every text-as-data project (Saldana 2009, Grimmer & Stewart 2013). For some methods validation is straightforward. Supervised machine learning results are validated by comparing an algorithm’s predictions to pre-existing “gold standard” results. These may be documents labeled by human annotators, but there are many other possibilities. The gold standard for Beauchamp (2017) are state-level public opinion polls. In computer science, researchers frequently take advantage of online ratings and reviews to train and validate algorithms capturing sentiment. To guard against overfitting, researchers typically train the algorithm on one set of labeled examples before testing accuracy using a different, held-out, set.<sup>2</sup> Whether the gold standard labels validly capture the phenomenon of interest is a separate (and important) question. For other methods, where no gold standard is available, validation is typically multifaceted. For unsupervised machine learning methods, scholars have delved into specific examples within

<sup>2</sup>Repeating this process several times, using different training and testing sets, and then aggregating the validation results ( $N$ -fold cross-validation), is an even better approach (Kohavi 1995, Arlot 2010).

topics to show that the topics make sense; demonstrated that different algorithms produce similar clusters; and established that variations in topic emphasis across time or venues correlate with real-world events (Blei & Lafferty 2009, Quinn et al. 2010, Grimmer & King 2011, Roberts et al. 2014).

## RECENT DEVELOPMENTS IN POLITICAL SCIENCE

The purpose of this section is to provide a sense of the research opportunities available for political scientists. We make no attempt to be comprehensive but instead focus on four general research objectives. Two (classification and scaling) will be familiar to many readers (Grimmer & Stewart 2013). The other two (text reuse and semantics) have received less attention to date.

### Classification

Classification is a popular objective of text-as-data projects. Unsupervised machine learning methods [e.g., K-means, principal components analysis (PCA), latent Dirichlet allocation (LDA)] compare the similarity of documents based on co-occurring features. Despite their name, unsupervised methods require a lot of input from the user, who must (among other things) specify the number of topics in advance and interpret their meaning. In one of the earliest applications by political scientists, Quinn et al. (2010) used an unsupervised learner to classify Senate speeches by policy topic. They then validated their results by showing that their topics were similar to those developed using more time-consuming methods. Bousaills & Coan (2016) and Farrell (2016) use topic modeling to investigate climate change “skepticism” in reports and communications by think tanks and interest groups. Grimmer & King (2011) demonstrate how unsupervised methods can lead to new discoveries. They find that congressional press releases cluster in ways that match Mayhew’s (1974) typology of constituent advertising, position taking, and credit claiming, but they also observe an additional cluster they label “partisan taunting” (see also Grimmer 2013). Roberts et al. (2014) show how incorporating additional information about documents (beyond the bag of words) into topic models can aid in interpretation of open-ended survey responses.

Whereas unsupervised methods are often used for discovery, supervised learning methods are primarily used as a labor-saving device. For example, Workman (2015) and Collingwood & Wilkerson (2011) use supervised methods to apply a well-established Policy Agendas topic-coding system to new research domains (federal regulations and congressional bills). Boydston et al. (2016) are currently labeling thousands of newspaper articles for issue frame with the long-term goal of developing a supervised learner that can predict frames in other articles. The fact that supervised methods often require thousands of training examples makes them a nonstarter for many researchers and projects. However, there are often creative ways to reduce the effort required. Examining 250,000 Enron emails, Drutman & Hopkins (2013) use simple identification techniques to first exclude the 99% that were not political in nature. Crowdsourcing is also frequently used to build training sets in computer science. When a project does not require individual document labels, ReadMe is a supervised method that reliably predicts class proportions using a much smaller number of training examples (Hopkins & King 2010). King et al. (2013) use ReadMe to classify millions of social media posts by topic in a study of government censorship in China. Ceron et al. (2014) use it to study citizens’ policy preferences in Italy and France.

Sentiment analysis is another important area of classification research where supervised and unsupervised methods are often used. The objective is to classify text ordinally (from negative to positive, for example) rather than categorically. Because businesses care about how consumers are responding to their products online, sentiment analysis is a well-funded area of research in

computer science. As a result, political scientists can take advantage of many pre-existing training corpora for a wide variety of research domains.<sup>3</sup>

### Scaling

Some of the earliest applications of automated text analysis in political science focused on using speeches and manifestos to locate European political parties in continuous ideological space (Laver et al. 2003, Lowe 2008, Slapin & Proksh 2008). Subsequent research has extended this by employing new methods and investigating new domains. In a pathbreaking study, Benoit et al. (2016) show that crowdsourcing can be a viable, even preferable, alternative to expert-based approaches to locating parties on policy dimensions. Kluver (2009) uses statements by interest groups and EU regulators to estimate ideological positions and gauge influence. Diermeier et al. (2012) test several different approaches to estimating legislator ideology from statements in the *Congressional Record* (see also Lauderdale & Herzog 2016). Barbera (2015) uses Twitter data and information about posters' followers to estimate the ideological positions of politicians, parties, and individual citizens. Lauderdale & Clark (2015) combine past votes with topic modeling of judicial opinions to critique single-dimensional scaling of justices and to develop separate estimates of judicial ideology for different issue areas.

### Text Reuse

Text reuse, as the name implies, is about discovering instances of similar language usage. The distinctive feature of text reuse algorithms is that they explicitly value word sequencing in judging document similarity. Political scientists have recently employed them to trace the origins of policy proposals in legislation (Wilkerson et al. 2015), to study the influence of interest groups in state legislatures (Hertel-Fernandez & Kashin 2015),<sup>4</sup> and to study party messaging strategies (Jansa et al. 2015). Other possibilities yet to be exploited by political scientists include studying the diffusion of political memes and contagion effects in new and old media (Leskovec et al. 2009, Smith et al. 2013). Different algorithms also support different types of analyses. Global alignment approaches (e.g., Needleman-Wunsch 1970) measure the overall similarity of documents whereas local alignment approaches (e.g., Smith & Waterman 1981) identify and score shared word sequences within documents. Thus, in a study of lawmaking or treaty negotiations, a global alignment approach might be used to see how much the entire proposal changes as it moves from one stage of the process to the next, whereas a local alignment approach could be used to investigate the fates of more specific provisions or proposals.

### Natural Language Processing

Social network analysis often employs text to investigate relationships among actors (Ward et al. 2011). Natural language processing (NLP) makes it possible to go beyond simply establishing connections to investigating the state of relationships—moving from “whom?” to “who did what to whom?” (Van Atteveldt et al. 2016). For example, political event data analysis draws on media reports to systematically monitor interactions between international actors. Instead of simply counting the number of times two actors are mentioned in reports, event data analysis incorporates

<sup>3</sup>Examples include <http://www.cs.cornell.edu/home/llee/data/> and [http://mpqa.cs.pitt.edu/corpora/mpqa\\_corpus/](http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/).

<sup>4</sup>See also the Legislative Influence Detector project (<https://dssg.uchicago.edu/lid/>).

syntax (sentence structure) and semantics (word meaning) to systematically track whether a relationship is improving or worsening and (possibly) to attribute credit or blame for developments.

Early event data research relied on human annotators to develop dictionaries of named entities and actions (Schrodt & Gerner 1994, Gerner et al. 2014). More recent research seeks to dramatically expand the scope of this research by taking advantage of extensive NLP resources developed by computer scientists and linguists (Leetaru & Schrodt 2013; see Ward et al. 2013 for an overview). For example, the Stanford Parser and the Stanford Named Entity Recognizer can be used to automatically extract specific parts of speech from documents and to tag different references to the same entity (e.g., USA, America, United States). Other valuable resources such as Wordnet can be used to identify synonyms for similar actions or sentiments. Denny et al. (2015) demonstrate how NLP methods can be used to systematically isolate the substantive provisions in legislation that typically includes lots of irrelevant “boilerplate” language. The creative possibilities are extensive, and Bird et al. (2009) provide an excellent primer on available NLP resources.

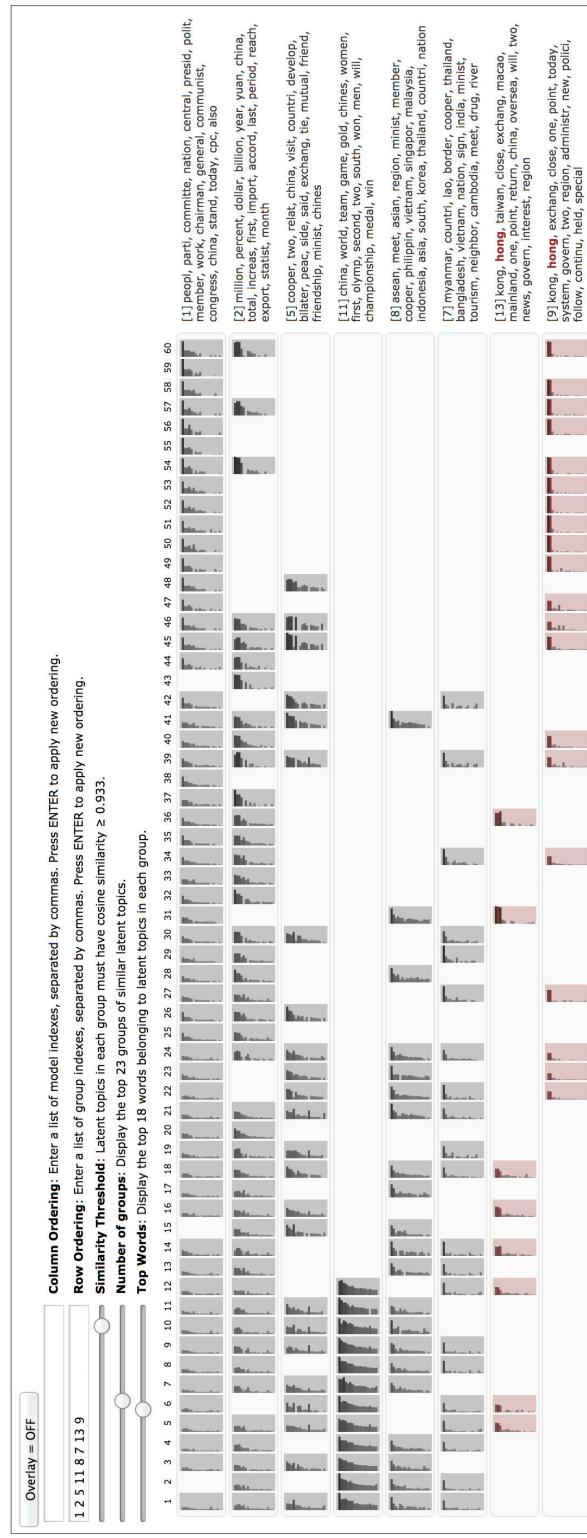
## TOPIC MODEL INSTABILITY AND A CALL FOR GREATER ATTENTION TO ROBUSTNESS IN TEXT-AS-DATA RESEARCH

In this final section, we shift from providing an overview of the field to delving into one contemporary challenge in more detail. Unsupervised machine learning methods (topic models) are very popular in political science in part because they classify documents without the extensive labeling efforts often required for supervised learning methods. The common practice has been to report and validate a single topic model after comparing results for several different models that vary by the number of topics specified by the researcher. This choice is usually based on the researcher’s subjective judgment about which model’s clusters best reflect the substantive goals of the project.

Above, we noted that the absence of a gold standard makes validation more challenging for these methods. A second challenge is model instability. Chuang et al. (2015, figure 1) estimate the same structural topic model 50 times to find that only two of 25 topics persist across all of the estimations. This can happen because different estimations can converge at different local maxima (Roberts et al. 2015). In a second experiment, the same authors find that manipulating just one feature of a structural topic model also leads to very different results (**Figure 1**). Many machine learning packages remove rarely used words by default to reduce processing time and avoid overfitting. In **Figure 1**, varying only this feature leads to important differences in terms of the topics that emerge from different estimations of the same model.

A number of recent studies have proposed different ways to assess and respond to topic-model instability (Grimmer & King 2011, Schmidt 2012, Boyd-Graber et al. 2014, Roberts et al. 2014). However, the focus, as far as we are aware, continues to be on selecting and validating a single best model. In conventional statistical studies, researchers try to demonstrate that their results are robust by reporting results for multiple model specifications. A study examining gender and voting will test and report several combinations of theoretically derived independent variables to demonstrate that the central findings persist. Supervised machine learning analyses also commonly address robustness by basing results on the consensus prediction of an ensemble rather than a single algorithm. Grimmer & King (2011) propose comparing topic model results for different algorithms but do not incorporate those differences into their findings.

Robustness can be evaluated with respect to methods, parameters, features, and data partitions. No study can consider all permutations, but we do think that political scientists using text-as-data methods should explicitly address robustness in their results. Do the central findings stand up to reasonable variations in modeling choices? Where topic models are concerned, one option is to move away from the current convention of reporting results for a single model.



**Figure 1**

Impact of a feature on topic stability (from Chuang et al. 2015). Each of the 50 columns is a 25-topic latent Dirichlet allocation model where the only difference is the threshold used to exclude sparse terms. Each row is a topic. The shaded cells indicate when a model includes the topic. Chuang et al. assume that two models share the same topic if the cosine similarity of the topic terms is greater than 0.9. Darker shades indicate higher similarity.

## Exploring the Topics of Legislators' Floor Speeches

In this section, we illustrate how topic robustness can inform a study of congressional floor speeches.<sup>5</sup> Members of the US House of Representatives gave almost 10,000 “one-minute” floor speeches during the 113th Congress (2013–2014). These speeches are given before ordinary business and are primarily intended for public consumption (Schneider 2015; <https://www.fas.org/sgp/crs/misc/RL30135.pdf>). A quick review indicates that their subjects are often quite diverse. Some honor constituent accomplishments (such as a state basketball championship), whereas others address political and legislative issues. However, to our knowledge, no one has systematically investigated what members talk about in these speeches. What topics are covered and which are the most common? Do Republicans and Democrats tend to talk about the same issues or emphasize different ones?

To examine these questions, we first used the Sunlight Foundation’s Capitol Words API to download all member statements from the *Congressional Record* of the 113th Congress. We then removed statements that did not begin with the opening phrase of a one-minute speech: “Mr. Speaker, I rise today . . . .” This produced a corpus of 5,346 one-minute speeches given by 179 Democrats and 4,358 given by 213 Republicans. We converted the words in each speech to lower case and removed punctuation, stopwords, word stems, and words of two characters or fewer. Finally, we constructed a term–document matrix where each row is a one-minute speech and each column is a vector indicating whether a feature/word is present in a given speech.

The next step was to estimate a series of latent Dirichlet allocation (LDA) models where the number of topics ( $k$ ) ranges from 10 to 90 in five topic increments (Blei et al. 2003). These 17 models yield 850 topics ( $10 + 15 + 20 \dots + 90$ ). To determine which topics were robust, we first calculated cosine similarity<sup>6</sup> for all topic pairs (resulting in 722,500 similarity scores) and then used the Spectral Clustering algorithm to group the 850 topics based on cosine similarity. The Spectral Clustering algorithm does this by maximizing average intra-cluster cosine similarity for a given number of clusters  $c$ . The substance of a given cluster can then be investigated by examining the most predictive words (“top terms”) in each cluster.

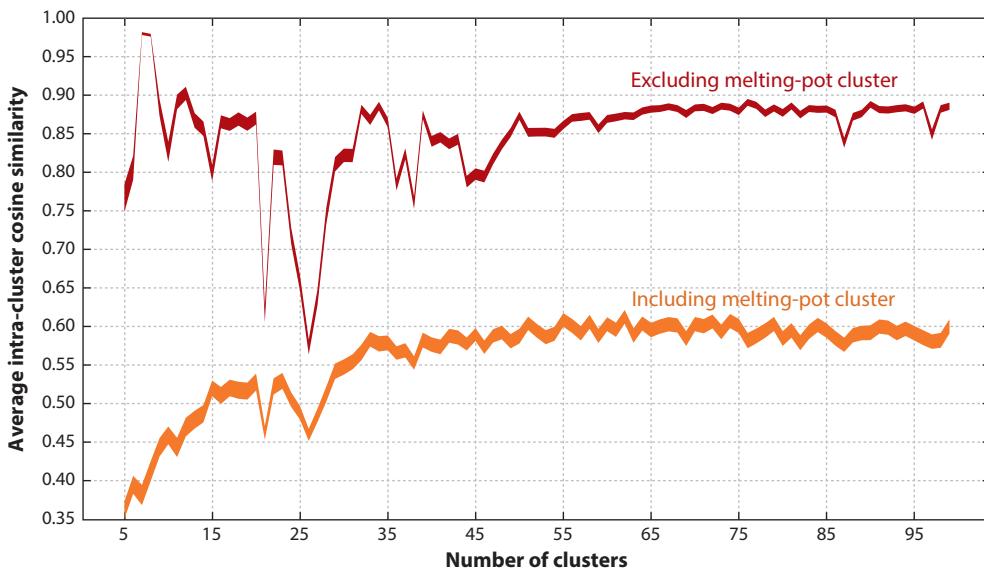
Ten thousand speeches by 435 lawmakers should cover a diverse set of topics. On the other hand, dividing the speeches into too many clusters may complicate the analysis without improving the overall fit (average intra-cluster similarity) of the model to the data. **Figure 2** displays how fit improves as the number of speech clusters ( $c$ ) is varied from five to 100. More clusters improve overall fit until approximately 50 clusters. The average similarity of the clusters excluding the largest catch-all “melting pot” cluster is also quite volatile until approximately 50 clusters. We therefore base our analysis on the robust topics from a 50-cluster model.

After clustering the 850 topics from 17 models into 50 clusters (see **Figure 3**), we grouped some of the clusters into what we will call metatopics. For example, the education metatopic includes three clusters about more specific aspects of education. All of the topics in which we were unable to discern a consistent theme were excluded by assignment to one “unclear” metatopic. Thus, the results presented are based on 697 of the original 850 topics from 16 of the 17 original topic models. In **Figure 4**, the education metatopic, for example, includes 37 topics found in 14 different topic models. In our view, the figure underscores the drawbacks of presenting results

<sup>5</sup>Supplemental and replication materials for this section can be found in the **Supplemental Materials** section of the Annual Reviews website (<http://www.annualreviews.org/db/suppl>) and at [https://github.com/CasAndreu/wilkerson\\_casas\\_2016\\_TAD](https://github.com/CasAndreu/wilkerson_casas_2016_TAD). These materials include a Python module, r1da, to apply the robust latent Dirichlet allocation models used here (<https://github.com/CasAndreu/r1da>).

<sup>6</sup>For each possible pair of topics,  $\cos(\theta) = \frac{a \cdot b}{\|a\| \|b\|}$ , where  $a$  and  $b$  are vectors of counts recording topic-word assignments in the final estimation iteration.





**Figure 2**

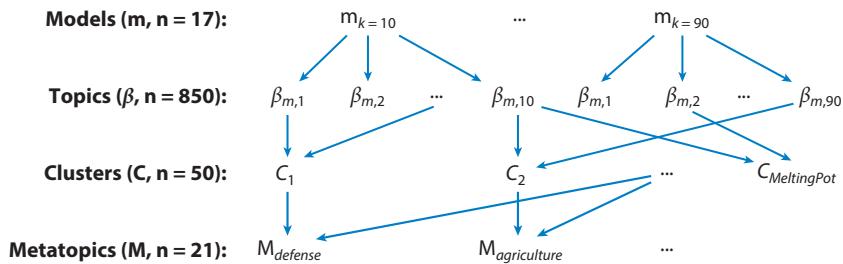
Number of clusters and average intra-cluster similarity. The lower line indicates that more clusters improve overall fit until approximately 50 clusters. The upper line indicates that the average similarity of the clusters excluding the largest catch-all “melting pot” cluster is also quite volatile until approximately 50 clusters.

based on a single model. Topics that are common to many models are often missing from any one of them.

### Topic Attention in One-Minute Speeches

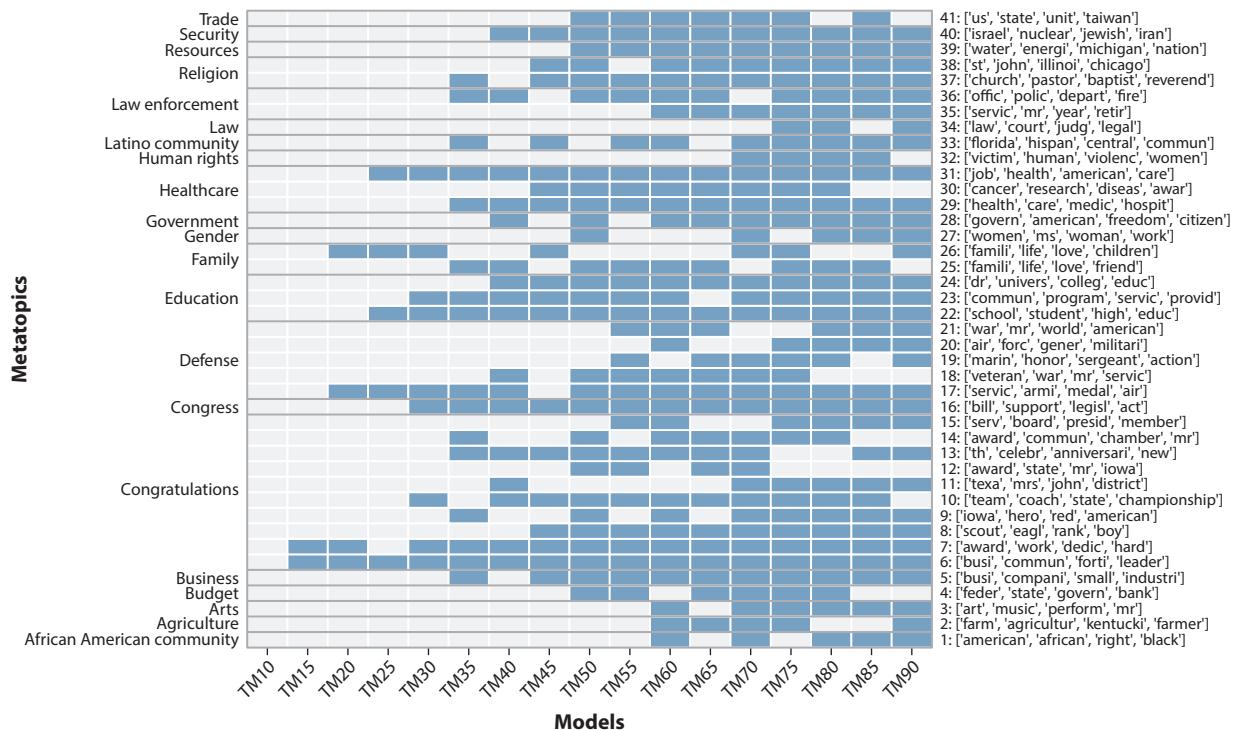
In an LDA model, the topics exist before the documents (see Blei et al. 2003, pp. 996–97). Each document is assumed to be about each topic with some positive probability. To study speech attention, we must first label individual speeches for primary topic. We assume each speech is about its most probable topic. Thus, we classify 9,704 speeches for each of 16 topic models. We then report results for only those topics from each model that are part of the 21 metatopics.

**Figure 5** displays those results. For example, for education, the consensus of the different topic models is that Democrats gave more speeches about education than Republicans did. It is reassuring



**Figure 3**

Workflow of moving from 17 topic models to 21 metatopics.



**Figure 4**

The 21 metatopics of a 50-cluster model. Each column is one of the 17 LDA topic models (ranging from 10 to 90 topics), and each row is a topic cluster. The 21 substantive metatopics are listed on the left. The shaded cells indicate where the topics in each cluster or metatopic originate.

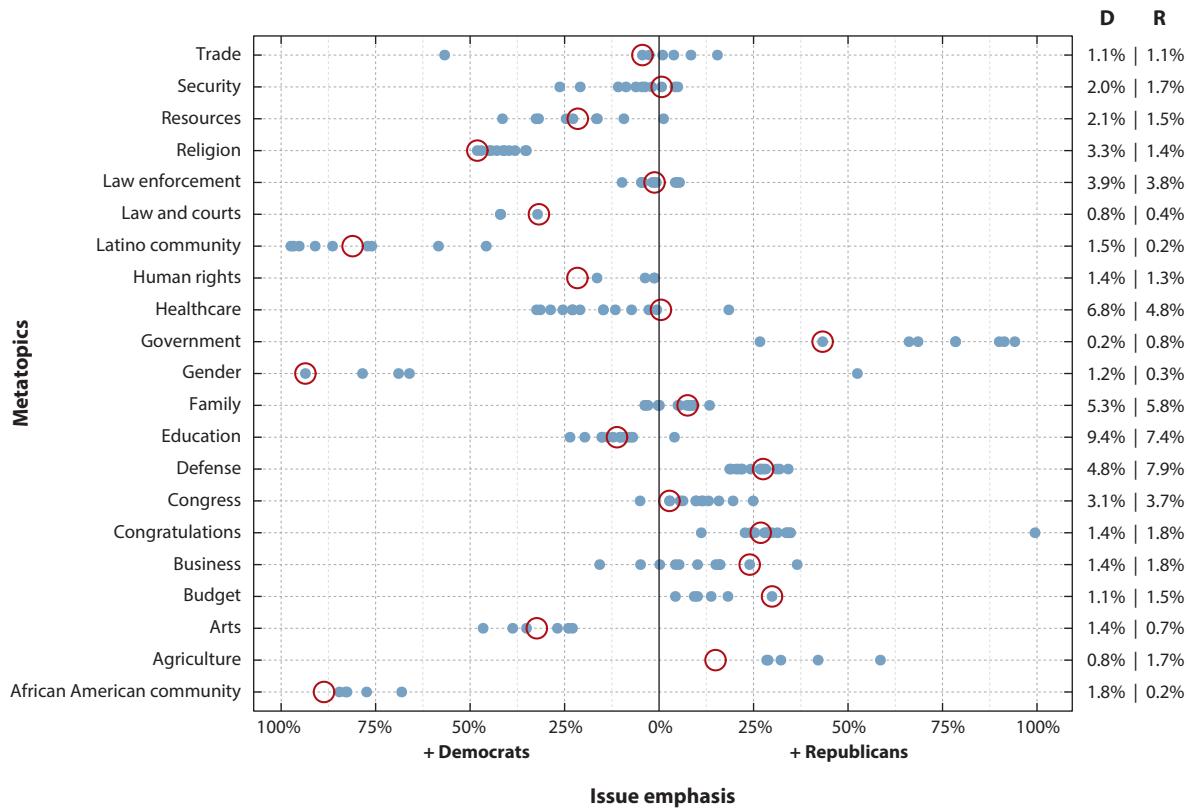
that the models generally agree concerning partisan emphasis for most of the metatopics. However, there is often considerable disagreement regarding the size of the difference.

The average amount of attention (across the models) given to different metatopics (such as education) by Republicans and Democrats is displayed on the far right. The results indicate that Republicans were most likely to give “Congrats” speeches (17%), followed by speeches about “Defense” (8%), “Education” (7%), and “Family” (6%). Democrats also gave lots of congratulatory speeches (9%), but were as likely to give speeches about education (9%), followed by health care (7%) and family (5%).

## Validation

We think that similarity of estimates of topic emphasis across different models is an important type of validation. **Figure 5** should inspire confidence in the robustness of general differences in speech topic emphasis but less confidence in the amount of difference in many cases.

Our results generally support Petrocik's (1996) “issue ownership” argument. The main exception seems to be agriculture. According to Petrocik, Democrats own the issue of agriculture, whereas most of the models of our analysis indicate that Republicans own it. We therefore took a closer look at who was giving speeches about agriculture and found a strong correlation between the proportion of a member's speeches that were about agriculture and the number of district



**Figure 5**

A robust examination of issue emphasis in one-minute speeches. Each row is one of the 21 metatopics. Each dot is a result for one topic model. The average amount of attention (across the models) given to different metatopics by Republicans and Democrats is displayed on the far right. The red circles display relative topic attention for a single ( $k = 50$ ) topic model.

workers employed in the agriculture, forestry, fishing, hunting, and mining industries (Pearson's  $r = 0.4$ ).<sup>7</sup> Thus, it seems likely that there has been a transfer of ownership on this issue since Petrocik's article was published 20 years ago.

## DISCUSSION

Computerized text analysis is transforming political science research because scholars now have the ability to explore massive amounts of politically relevant text using increasingly sophisticated tools. These developments have already produced important advances in research methods (Hopkins & King 2010, Benoit et al. 2016), opened the door to new research questions (Wilkerson et al. 2015), and altered current understandings (Lauderdale et al. 2015). We have argued that researchers do not need to be computer programmers or statistical methodologists to use text-as-data methods in their research. They do need to be attentive to the same concerns about validity and reliability that apply to all methods.

### Supplemental Material

<sup>7</sup>See the appendix (Supplemental Materials; <http://www.annualreviews.org/db/suppl>) for more details.

The other area where political scientists are making important advances is in assessing the quality of findings. Recent studies are examining the sensitivity of findings to alternative feature specifications (Spirling & Denny 2017) and proposing new approaches to explicitly incorporate information about reliability into research findings (Grimmer et al. 2016). Unsupervised learning methods (such as topic models) are among the more popular methods used in political science. A central attraction for many researchers is that they do not require labeled training sets. To be sure, supervised learning methods have their own limitations. However, the absence of any gold standard makes choosing and validating a model even more challenging for unsupervised methods.

Scholars have recently proposed new ways of selecting among alternative topic models, for example by examining the cohesiveness and distinctiveness of the topic words (Roberts et al. 2014). A robust approach to reporting topic-model results takes advantage of the information provided by alternative specifications. This approach has its own limits, but in our view it is informative and transparent and adheres to current conventions that lead researchers to explicitly address robustness in statistical studies.

## DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

The authors thank Jeffrey Arnold, Noah Smith, Jason Chuang, and an anonymous reviewer for helpful feedback.

## LITERATURE CITED

- Alvarez RM, ed. 2016. *Computational Social Science: Discovery and Prediction. Analytical Methods for Social Research*. New York: Cambridge Univ. Press
- Arlot C. 2010. A survey of cross-validation procedures for model selection. *Stat. Surv.* 4:40–79
- Barbera P. 2015. Birds of the same feather tweet together. Bayesian ideal point estimation using Twitter data. *Polit. Anal.* 23(1):76–91
- Beauchamp N. 2017. Predicting and interpolating state-level polls using Twitter textual data. *Am. J. Polit. Sci.* In press. doi: 10.1111/ajps.12274
- Benoit K, Conway D, Lauderdale B, Laver M, Mikhaylov S. 2016. Crowd-sourced text analysis: reproducible and agile production of political data. *Am. Polit. Sci. Rev.* 110(2):278–95
- Bird S, Klein E, Loper E. 2009. *Natural Language Processing with Python—Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media
- Blei D, Lafferty J. 2009. Topic models. In *Text Mining: Classification, Clustering, and Applications*, ed. AN Srivastava, M Sahami, pp. 71–94. Data Mining and Knowledge Discovery Ser. Boca Raton, FL: Chapman & Hall/CRC
- Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022
- Boussalis C, Coan TG. 2016. Text-mining the signals of climate change doubt. *Glob. Environ. Change* 36:89–100
- Boyd-Graber J, Mimno D, Newman D. 2014. Care and feeding of topic models: problems, diagnostics, and improvements. In *Handbook of Mixed Membership Models and Their Applications*, pp. 3–34. Boca Raton, FL: CRC Press

- Boydston A, Butters R, Card D, Gross J, Resnik P, Smith N. 2016. *Under what conditions does media framing influence public opinion on immigration?* Presented at Annu. Meet. Midwest Polit. Sci. Assoc., Chicago, IL, Apr. 7–9
- Cardie C, Wilkerson J. 2008. Text annotation for political science research. *J. Inf. Technol. Polit.* 5(1):1–6
- Carneiro HA, Mylonakis E. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin. Infect. Dis.* 49(10):1557–64
- Casas A, Davesa F, Congosto M. 2016. The media coverage of a connective action: the interaction between the 15-M Movement and the mass media. *Rev. Espan. Investig. Sociol.* 155:73–96
- Ceron A, Curini L, Iacus SM, Porro G. 2014. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media Soc.* 16(2):340–58
- Chang J, Boyd-Graber J, Wang C, Gerrish S, Blei DM. 2009. Reading tea leaves: how humans interpret topic models. In *Advances in Neural Information Processing Systems*, ed. Y Bengio, D Schuurmans, J Lafferty, CKI Williams, A Culotta, pp. 288–96. Cambridge, MA: MIT Press
- Chuang J, Roberts M, Stewart B, Weiss R, Tingley D, et al. 2015. TopicCheck: interactive alignment for assessing topic model stability. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pp. 175–84. Denver, CO: Assoc. Comput. Linguist.
- Chuang J, Wilkerson JD, Weiss R, Tingley D, Stewart BM, et al. 2014. *Computer-assisted content analysis: topic models for exploring multiple subjective interpretations*. Presented at Advances in Neural Information Processing Systems Workshop on Human-Propelled Machine Learning, Montreal, Dec. 8–13
- Collingwood L, Wilkerson J. 2011. Tradeoffs in accuracy and efficiency in supervised learning methods. *J. Inf. Technol. Polit.* 4:1–28
- Denny MJ, O'Connor B, Wallach H. 2015. *A little bit of NLP goes a long way: finding meaning in legislative texts with phrase extraction*. Presented at Annu. Meet. Midwest Polit. Sci. Assoc., 73rd, Apr. 16–19
- Denny MJ, Spirling A. 2017. *Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it*. Unpublished manuscript, Dep. Polit. Sci., Stanford Univ and Inst. Quant. Soc. Sci., Harvard Univ. <https://ssrn.com/abstract=2849145>
- Diermeier D, Yu B, Kaufmann S, Godbout JE. 2012. Language and ideology in Congress. *Br. J. Polit. Sci.* 42(1):31–55
- Domingos P. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books
- Drutman L, Hopkins DJ. 2013. The inside view: using the Enron email archive to understand corporate political attention. *Legis. Stud. Q.* 38(1):5–30
- Eggers A, Spirling A. 2017. The shadow cabinet in Westminster systems: modeling opposition agenda setting in the House of Commons, 1832–1915. *Br. J. Polit. Sci.* In press
- Farrell J. 2016. Corporate funding and ideological polarization about climate change. *PNAS* 113(1):92–97
- Gerner DJ, Schrod PA, Francisco RA, Weddle JL. 2014. Machine coding of event data using regional and international sources. *Int. Stud. Q.* 38(1):91
- Grimmer J. 2013. Appropriators not position takers: the distorting effects of electoral incentives on congressional representation. *Am. J. Polit. Sci.* 57(3):624–42
- Grimmer J, King G. 2011. General purpose computer-assisted clustering and conceptualization. *PNAS* 108(7):2643–50
- Grimmer J, King G, Superti C. 2016. *The unreliability of measures of intercoder reliability, and what to do about it*. Unpublished manuscript, Dep. Polit. Sci., Stanford Univ. <http://web.stanford.edu/~jgrimmer/Handbib.pdf>
- Grimmer J, Stewart BM. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* 21(3):267–97
- Hertel-Fernandez A, Kashin K. 2015. *Capturing business power across the states with text reuse*. Presented at Annu. Meet. Midwest Polit. Sci. Assoc., Chicago, IL, Apr. 16–19
- Hopkins DJ, King G. 2010. A method of automated nonparametric content analysis for social science. *Am. J. Polit. Sci.* 54(1):229–47
- Huang A. 2008. Similarity measures for text document clustering. In *Proc. Sixth New Zealand Computer Science Research Student Conference*, pp. 49–56. Christchurch, New Zealand: NZCSRSC

- Jansa J, Hansen E, Gray V. 2015. *Copy and paste lawmaking: the diffusion of policy language across American state legislatures*. Work. Pap., Dep. Polit. Sci., Univ. North Carolina, Chapel Hill
- Jockers ML. 2014. *Text Analysis with R for Students of Literature*. New York: Springer
- King G, Pan J, Roberts ME. 2013. How censorship in China allows government criticism but silences collective expression. *Am. Polit. Sci. Rev.* 107(2):326–43
- Kluver H. 2009. Measuring interest group influence using quantitative text analysis. *Eur. Union Polit.* 10(4):535–49
- Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. Int. Joint Conf. Artificial Intelligence*, pp. 1137–43. San Francisco: Morgan Kaufmann
- Lauderdale BE, Clark TS. 2014. Scaling politically meaningful dimensions using texts and votes. *Am. J. Polit. Sci.* 58(3):754–71
- Lauderdale BE, Herzog A. 2016. Measuring political positions from legislative speech. *Polit. Anal.* 26:374–94
- Laver M, Benoit K, Garry J. 2003. Extracting policy positions from political texts using words as data. *Am. Polit. Sci. Rev.* 2:311–31
- Leetaru K, Schrodt P. 2013. *GDELT: global data on events, location, and tone, 1979–2012*. Presented at International Studies Association Annu. Conv., San Francisco, CA, Apr.
- Leskovec J, Backstrom L, Kleinberg J. 2009. *Memetracking and the dynamics of the news cycle*. Presented at ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), Paris, June
- Lowe W. 2008. Understanding Wordscores. *Polit. Anal.* 16(4):356–71
- Mayhew DR. 1974. *Congress: The Electoral Connection*. New Haven, CT: Yale Univ. Press
- Monroe BL, Schrodt PA. 2008. Introduction to the special issue: the statistical analysis of political text. *Polit. Anal.* 16(4):351–55
- Munzert S, Rubba C, Meissner P, Nyhuis D. 2014. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. Hoboken, NJ/Chichester, UK: Wiley & Sons
- Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48(3):443–53
- Petrocik JR. 1996. Issue ownership in presidential elections, with a 1980 case study. *Am. J. Polit. Sci.* 40(3):825–50
- Quinn KM, Monroe BL, Colaresi M, Crespin MH, Radev DR. 2010. How to analyze political attention with minimal assumptions and costs. *Am. J. Polit. Sci.* 54(1):209–28
- Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, et al. 2014. Structural topic models for open-ended survey responses: structural topic models for survey responses. *Am. J. Polit. Sci.* 58(4):1064–82
- Roberts M, Stewart B, Tingley D. 2016. Navigating the local modes of big data: the case of topic models. In *Computational Social Sciences*, ed. RM Alvarez, pp. 51–97. New York: Cambridge Univ. Press
- Saldana J. 2009. *The Coding Manual for Qualitative Researchers*. Los Angeles: Sage
- Schmidt BM. 2012. Words alone: dismantling topic models in the humanities. *J. Digit. Humanit.* (2)1. <http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/>
- Schmidt B. 2015. Is it fair to rate professors online? *New York Times*, Dec. 16, Sec. Room for Debate
- Schneider J. 2015. *One-minute speeches: current house practices*. Congr. Res. Serv. Rep. 7-5700, 1–7
- Schrodt PA, Gerner DJ. 1994. Validity assessment of a machine-coded event data set for the Middle East, 1982–92. *Am. J. Polit. Sci.* 38(3):825
- Slapin JB, Proksch S-O. 2008. A scaling model for estimating time-series party positions from texts. *Am. J. Polit. Sci.* 52(3):705–22
- Smith DA, Cordell R, Dillon EM. 2013. Infectious texts: modeling text reuse in nineteenth-century newspapers. In *Proc. IEEE Int. Conf. Big Data*, pp. 86–94. Santa Clara, CA: Inst. Electrical and Electronics Engineers
- Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147(1):195–97
- Van Atteveldt W, Shenhav SR, Fogel-Dror Y. 2017. Clause analysis: using syntactic information to automatically extract source, subject, and predicate from texts with an application to the 2008–2009 Gaza War. *Polit. Anal.* In press

- Wallach H, Dicker L, Jensen S. 2010. An alternative prior for nonparametric Bayesian clustering. In *Proc. Thirteenth International Conference on Artificial Intelligence and Statistics, May 13–15, 2010, Chia Laguna Resort, Sardinia, Italy*, ed. YW Teh, M Titterington, 9:892–99. <http://www.jmlr.org/proceedings/papers/v9/>
- Ward M, Beger A, Josh C, Dickenson M, Dorff C, Radford B. 2013. Comparing GDELT and ICEWS event data. *Analysis* 21:267–97
- Ward M, Stovel K, Sacks A. 2011. Network analysis and political science. *Annu. Rev. Polit. Sci.* 14:245–64
- Wilkerson J, Smith D, Stramp N. 2015. Tracing the flow of policy ideas in legislatures: a text reuse approach. *Am. J. Polit. Sci.* 59(4):943–56
- Workman S. 2015. *The Dynamics of Bureaucracy in the US Government: How Congress and Federal Agencies Process Information and Solve Problems*. Cambridge, UK: Cambridge Univ. Press

---

# 1

# Defining Content Analysis

---

## An Introduction

---

Content analysis is one of the most popular and rapidly expanding techniques for quantitative research. Advances in computer applications and in digital media have made the organized study of messages quicker and easier . . . but not automatically better. This book explores the current options for quantitative analyses of messages.

*Content analysis* may be briefly defined as the systematic, objective, quantitative analysis of message characteristics. It includes both human-coded analyses and computer-aided text analysis (CATA). Its applications can include the careful examination of face-to-face human interactions; the analysis of character portrayals in media venues ranging from novels to online videos; the computer-driven analysis of word usage in news media and political speeches, advertising, and blogs; the examination of interactive content such as video gaming and social media exchanges; and so much more.

Content analysis has been applied to many areas of inquiry. It has been used to investigate naturally occurring language (Markel, 1998), newspaper coverage of the greenhouse effect (Miller, Boone, & Fowler, 1992), letters to the editor (Perrin & Vaisey, 2008), and how characters of different genders are shown on TV (Greenberg, 1980). It has been used in such highly specific studies as those analyzing Turkish elementary school math books (Özgeldi & Esen, 2010), greenway plans in northwest Indiana (Floress et al., 2009), questions asked by patients and companions in physician–patient interactions (Eggly et al., 2006), web page hits and Google Group threadedness for living and dead public intellectuals (Danowski & Park, 2009), the emotional tone of social networking comments (i.e., sentiment analysis; Thelwall, Wilkinson, & Uppal, 2010), the linguistic substance of the writings of a 19th-century explorer leading up to his suicide (Baddeley, Daniel, & Pennebaker, 2011), and the substance of Canadian winery web sites (Zhu, Basil, & Hunter, 2009).

Content analyses have resulted in eclectic and often surprising findings. A study analyzing Hollywood actresses' facial features predicted good economic

times from the prevalence of neonate (babylke) features among top movie stars (Pettijohn & Tesser, 1999). Johnson (1987) analyzed Porky Pig's vocalics from a clinical speech therapy standpoint, finding stuttering in 11.6% to 51.4% of words uttered (per cartoon), with certain behaviors statistically associated with the stuttering (e.g., eye blinks, grimaces). Hirdes, Woods, and Badzinski (2009) examined the prevalence of persuasive appeals associated with a wide range of types of "Jesus merchandise." Atkinson and Herro (2010) discovered that *The New York Times* mentioned tennis star Andre Agassi's age much more often when he was atypically young or atypically old for competitive tennis. And Wansink and Wansink (2010) measured the food-to-head ratio in 52 *Last Supper* paintings produced over a millennium, finding that the relative sizes of the main dish, bread, and plates have all increased linearly and significantly over the past thousand years. Chapter 9 presents an overview of some of the major areas of study—the main "contexts" of content analysis research—but the above examples show that the range of applications is limited only by the researcher's imagination.

Content-analytic measures may be combined with other types of measurement, as in Pian, Khoo, and Chang's (2014) study of users' attention to an online health discussion forum. They used an eye-tracking system to first identify text segments that users' attention was focused on (via eye fixations) and then used content analysis to identify the types of information attended to. Himelboim, McCreery, and Smith (2013) combined network analysis and content analyses to examine exposure to cross-ideological political views on Twitter. They mapped the Twitter networks of 10 controversial political topics, identifying user clusters (groups of highly connected individuals) and content analyzed messages for political orientation, finding that Twitter users were unlikely to be exposed to cross-ideological content from the user clusters they followed; the within-cluster content was likely to be quite homogeneous. Content-analytic data may be more broadly combined with survey or experimental data about message sources or receivers as well. Chapter 2 elaborates on this "integrative" approach to content analysis.

This book will explore the expansion and variety of the techniques of content analysis. In this chapter, we will follow the development of a full definition of content analysis—how one attempts to ensure objectivity, how the scientific method provides a means of achieving systematic study, and how the various scientific criteria (e.g., validity, reliability) are met. Furthermore, standards are established, extending the expectations of readers who may hold a view of content analysis as necessarily simplistic.

## The Growing Popularity of Content Analysis

---

The repertoire of techniques that make up the methodology of content analysis has been growing in range and usage. In the field of mass communication research, content analysis has been the fastest-growing technique

over the past 40 years or so (Yale & Gilly, 1988). Riffe and Freitag (1997) noted a nearly sixfold increase in the number of content analyses published in *Journalism & Mass Communication Quarterly* over a 24-year period—from 6.3% of all articles in 1971 to 34.8% in 1995, making this journal one of the primary outlets for content analyses of mass media. Kamhawi and Weaver (2003) studied articles in 10 major mass communication journals for the period 1980 through 1999, finding content analysis to be the second-most popular method reported, after surveys (30% and 33% of all studies, respectively). Freimuth, Massett, and Meltzer (2006) examined the first 10 years of *The Journal of Health Communication*, finding that a fifth of all quantitative studies presented in the journal were content analyses. Manganello and Blake (2010) looked at the frequency and types of content analyses in the interdisciplinary health literature between 1985 and 2005, finding a steady increase in the number of studies of health-related media messages over the period.

One great expansion in analysis capability has been the rapid advancement in computer-aided text analysis (CATA) software (see Chapter 5 of this volume), with a corresponding proliferation of online archives and databases (Evans, 1996; Gottschalk & Bechtel, 2008; see also Chapter 7 of this volume). There has never been such ready access to archived electronic messages, and it has never been easier to perform at least basic analyses with computer-based speed and precision. Further, scholars and practitioners alike have begun to merge the traditions of content analysis, especially CATA, with such expanding fields of endeavor as natural language processing (bringing to bear some of the capabilities of machine learning of language to the analysis of text and even images; Indurkhy & Damerau, 2010), computational linguistics, text mining of “big data,” message-centric applications of social media metrics, and sentiment analysis (or opinion mining; Pang & Lee, 2008); see also Chapter 5 of this volume. While content analysis, with its traditions extending back nearly a century, might be considered the grandparent of all “message analytics,” it has been stretched and adapted to the changing times.

Content analysis has a long history of use in communication, journalism, sociology, psychology, and business. And content analysis is being used with increasing frequency by a growing array of researchers. White and Marsh (2006) demonstrate the method’s growing acceptance in library and information science. Expansions in medical fields, such as nursing, psychiatry, and pediatrics (Neuendorf, 2009), and in political science (Monroe & Schrot, 2008) have been noted. The importance of the method to gender studies was recognized in two special issues of the interdisciplinary journal *Sex Roles* in 2010 and 2011 (Rudy, Popova, & Linz, 2010, 2011).

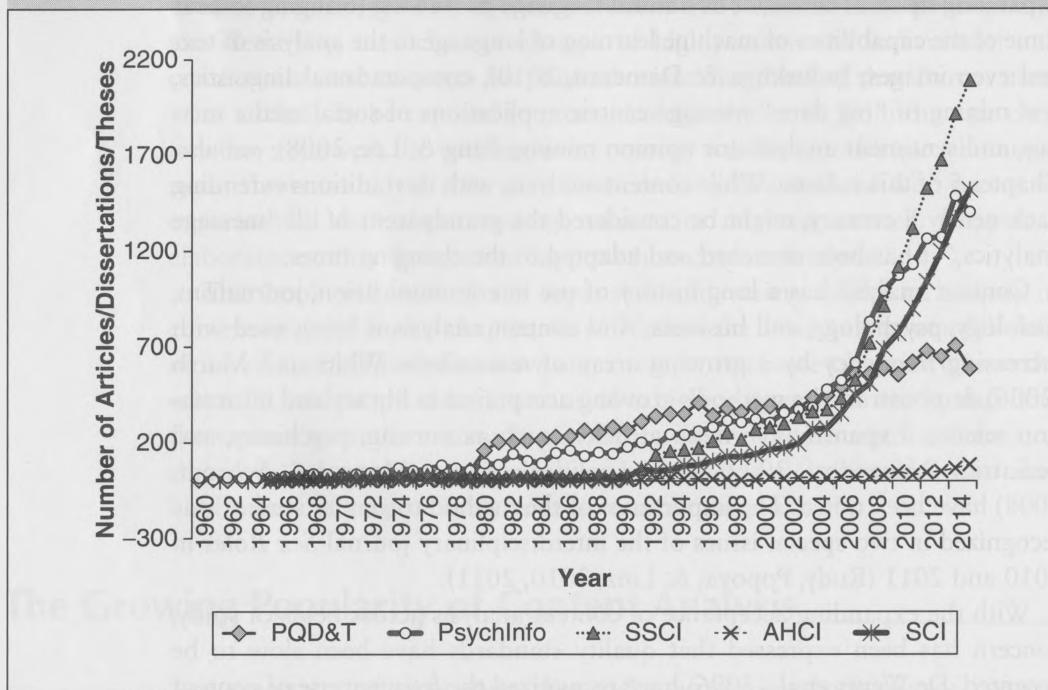
With the expanding acceptance of content analysis across fields of study, concern has been expressed that quality standards have been slow to be accepted. De Wever et al. (2006) have recognized the frequent use of content analysis to analyze transcripts of asynchronous, computer-mediated discussion groups in formal education settings, while noting that “standards are

not yet established" (p. 6). And Strijbos et al. (2006) have pointed out methodological deficiencies in the application of content analysis to computer-supported collaborative learning.

The explosion of content analysis in various areas of scholarship is demonstrated in Figure 1.1. Here, we may see the growth of content analysis as a research technique over a period of 50+ years, from 1960 through 2014. To produce this summary, five scholarly indexes were searched for dissertations, theses, and research articles containing the term *content analysis* in titles, subjects, or abstracts: ProQuest Dissertations and Theses (PQD&T), PsychInfo, Social Science Citation Index (SSCI), Arts and Humanities Citation Index (AHCI), and Science Citation Index (SCI).<sup>1</sup>

The graphed lines should be viewed cautiously and interpreted as the outcome of simple searches for a term in publications available since 1960, without contextual information about how the term has been used by the researchers. That is, a number of studies labeled "content analyses" are actually qualitative text analyses or other studies that do not fit the definition of content analysis assumed in this book. Further, a portion of the articles counted by the Science Citation Index are actually "content analyses" of chemical compounds; however, a perusal of the searches indicates that no more than 10% of contemporary SCI articles are of this type. Second, the

**Figure 1.1** Timeline of Content Analysis Publications by Year



SOURCE: ProQuest Dissertations and Theses (PQD&T), PsychInfo, Social Science Citation Index (SSCI), Arts and Humanities Citation Index (AHCI), and Science Citation Index (SCI).

indexes overlap in their coverage. For example, a number of psychology journals are indexed in both PsychInfo and the Social Science Citation Index. Third, it should be noted that some of the growth in content analysis applications is surely due to the expansion in the number of journals indexed (via new journals and the addition of cross-listings).

Taking these caveats into account, the evidence is still clear: Never has content analysis received more attention in the research literature than at present. And never has content analysis been embraced by more disciplines.<sup>2</sup> Only the arts and humanities have remained relatively aloof to quantitative content analysis techniques.

## The Myths of Content Analysis

There have been evident certain misconceptions about the methods of content analysis: Conducting a content analysis is by nature simplistic and substantially easier than conducting other types of research, content analysis is anything a scholar says it is, and anyone can do it without much training or forethought. It has also been widely assumed that there is little interest in or reason to use content analysis for commercial or other nonacademic research. Unfortunately, these preconceptions have occasionally been reinforced by academic journals that may fail to hold content analyses to the same standards of methodological rigor as they do other social and behavioral science methods, such as surveys, experiments, and participant observation studies. Based on over 30 years of involvement in over 200 content analyses, I would like to dispel common myths about this method before providing a full working definition.

### Myth 1: Content Analysis Is Limited to Simple Analyses

**Truth:** Content analysis may be as simple—or as complex—as the researcher determines it to be. It is not necessarily more limited than a survey, experiment, or other type of study.

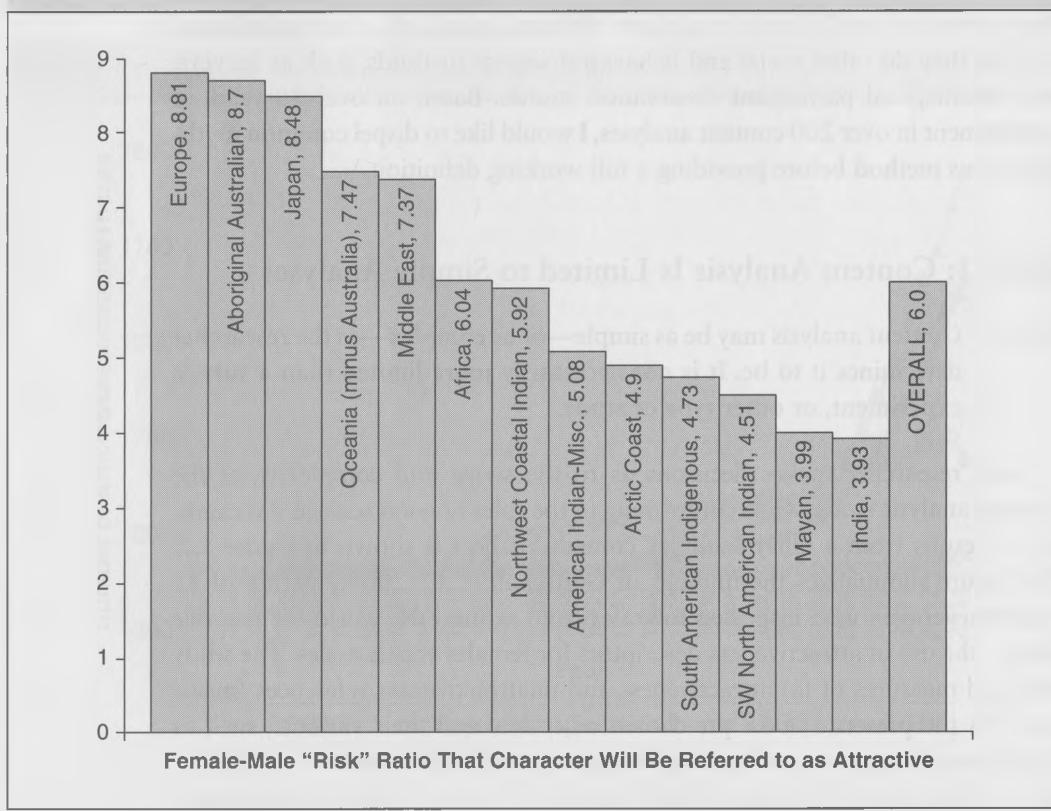
Each researcher makes decisions as to the scope and complexity of the content-analytic study, while conforming to the rules of good science. An example of results from a fairly “simple” content analysis is shown in Figure 1.2. This figure summarizes the findings of Gottschall et al. (2008), a team of 31 coauthors/coders who inspected folktales from around the world for just one thing—the use of attractiveness descriptors for females versus males. The study included measures of (a) attractiveness and unattractiveness references (measured via the presence of 58 pre-chosen adjectives and their variants, such as pretty/prettier/prettiest and ugly/uglier/ugliest) and (b) the gender of the character to whom each reference applied (measured via use of personal pronouns). Additionally, (c) a rough measure of how many characters in each tale were female and male was executed via electronic word searches for pronouns so

that attractiveness references could be expressed as proportional to the number of characters of that gender. So just three measures were developed for this study. The coder training task was relatively simple, and acceptable intercoder reliability was achieved, even with 31 coders.

Although using an elegantly simple coding scheme, the researchers chose an ambitiously large sample for its application: 90 volumes of traditional folktales from 13 regions around the world. In total, 8.17 million words in 16,541 single-spaced pages were analyzed.

Figure 1.2 shows the main findings—the female-to-male ratio of “risk” that a character will be referred to with attractiveness terminology. These figures take into account the rough numbers of females and males in the tales. Thus, we see that stories from European folktales show the greatest “gender bias”—a female character in these tales is 8.81 times more likely to be referred to as attractive/unattractive than is a male.<sup>3</sup> Overall, female characters are 6.0 times more likely to be referred to with regard to attractiveness than are males. And there is no region of the world that seems to generate folktales with gender parity, or with male predominance, when it comes to attractiveness references (Gottschall et al., 2008).

**Figure 1.2** Female–Male Attractiveness Emphasis in World Folktales



SOURCE: Adapted from Gottschall et al. (2008).

Even with such a limited content analysis scheme, broad claims might be made from the findings. The researchers indicate that the consistency of the results across cultures and world regions “strongly support[s] the evolutionary prediction that greater emphasis on female physical attractiveness will be the rule across human culture areas” and that “the main elements of the beauty myth are no myths” (Gottschall et al., 2008, p. 185).

Near the complex end of a simple-to-complex continuum of content analyses might be an ambitious master’s thesis (Smith, 1999) that examined the gender role portrayals of women in popular films from the 1930s, 1940s, and 1990s. The sampling was extremely problematic, given that no valid lists (i.e., sampling frames) of top box office hits are available for years prior to 1939. For many years after that date, all that are available are lists of the top *five* films. The researcher made the analysis even more complex by measuring 18 variables for each film and 97 variables for each primary or secondary character in each film (the complete coding scheme may be found at *The Content Analysis Guidebook Online, CAGO*). Some of the variables were untried in content analysis. For example, psychologist Eysenck’s (1990) measures of extraversion (e.g., sociable, assertive, sensation-seeking), typically measured on individuals by self-report questionnaire, were applied to film characters, with not always successful results. One hypothesis, that female portrayals will become less stereotypic over time, resulted in the measurement and analysis of 27 different dependent variables. With four active coders, the study took six months to complete.

The multifaceted results reflected the complexity and breadth of the study. The results included such wide-ranging points as these:

- Across the decades (1930s, 1940s, 1990s), there were several significant trends indicating a decrease in stereotypical portrayals of women in films.
- The average body shape for women varied across the decades at a near-significant level, indicating a trend toward a thinner body shape.
- Screen women who exhibited more traditional sex-role stereotyping experienced more negative life events.
- Female characters who exhibited more male sex-role traits and experienced negative life events tended to appear in films that were more successful at the box office.
- Screen women were portrayed somewhat more traditionally in films with greater female creative control (i.e., in direction, writing, producing, or editing; Smith, 1999).

## Myth 2: Anyone Can Do Content Analysis; It Doesn’t Take Any Special Preparation

**Truth:** Indeed, anyone can do it—but only with at least some training and with substantial research planning.

Despite the popularity of content analysis, rigorous methodological standards have not always been evident, notably with regard to issues of validity and reliability (Lombard, Snyder-Duch, & Bracken, 2002; Neuendorf, 2009, 2011; Pasadeos et al., 1995). Even contemporary reviews of content analyses find important standards lacking in many published studies. For example, an analysis of 133 health media content analyses failed to find a single instance of full reliability assessment and reportage (Neuendorf, 2009), with 38% of studies including no reliability assessment whatsoever. This figure is comparable to the 31% found by Lombard et al. (2002) in their review of content analyses in the field of communication. Coder training is an essential part of all human-coded content analyses, yet meta-analytic reviews of content analyses have revealed deficiencies in this regard—an analysis of 59 content analyses on the information content of advertising noted that “many authors give no information on whether or how coders were trained” (Abernethy & Franke, 1996, p. 5), and an analysis of 132 content analyses in the field of consumer behavior/marketing found 48% of studies failing to report any information about coder training (Kolbe & Burnett, 1991). Other deficiencies identified by Kolbe and Burnett included a lack of research questions or hypotheses (39% of studies), poor sampling (80% were convenience samples), and nonindependence of coders (over 50% of studies).

In order for content analysis to enjoy the same rigor as other research methods, those engaged in such analysis need to take serious stock of their own training and abilities. Just as no researcher would attempt to execute a true experiment without having studied some widely accepted text on the topic, the content analyst should be guided by one or more accepted reference texts on the methodology (see Neuendorf, 2011). And, as will become apparent in the chapters that follow, the planning stage of a content analysis may take substantial time and effort.

While the individual who designs a content analysis must have some special knowledge and preparation, a central notion in the methodology of content analysis is that all individuals are potentially useful “human coders” (i.e., people who make judgments about variables as applied to each message unit). The coding scheme must be so objective and so reliable that, once they are trained, coders from varied backgrounds and with different orientations will generally agree in its application (Neuendorf, 2009).

Clearly, however, each coder must be proficient in the language(s) of the message pool. This may require some special training for coders. To analyze natural speech, coders may need to be trained in the nuances of a given dialect. Before coding television or film content, coders may have to learn about production techniques and other aspects of visual communication. To code print advertising, coders may need to learn a bit about graphic design. All of this is in addition to training *with the coding scheme*, which is a necessary step for all coders.

For analyses that do not use human coders (i.e., those that use CATA), the burden rests squarely on the researcher to establish complete and carefully researched dictionaries or other protocols. Because the step of making sure

coders can understand and reliably apply a scheme is missing, the researcher needs to execute additional checks. Chapter 5 presents some notions on how this might be done.

### Myth 3: The Term *Content Analysis* Applies to All Examinations of Messages

**Truth:** The term does not apply to every analysis of messages—only those investigations that meet a particular definition. Calling an investigation a content analysis does *not* make it so.

There are many forms of analysis—from frivolous to seminal—that may be applied to the human production of messages. Content analysis is only one type, a technique presented by this book as systematic and quantitative. Even in the scholarly literature, some contestation exists as to what may be called a content analysis. On a number of occasions, the term has been applied erroneously (e.g., Council on Interracial Books for Children, 1977; DeJong & Atkin, 1995; Goble, 1997; Hicks, 1992; Thompson, 1996), and at times, studies that warrant the term do not use it (e.g., Bales, 1950; Fairhurst, Rogers, & Sarr, 1987; Thorson, 1989).

The term “qualitative content analysis” has been applied in some fields to a range of nonquantitative analyses of messages (Altheide, 1996; Mayring, 2000; Schreier, 2012; Zhang & Wildemuth, 2009). Altheide and Schneider (2013) present “ethnographic content analysis,” a blend of objective content analysis and participant observation that is intended to reveal “how a researcher interacts with documentary materials” (p. 5; see also Gormly, 2004). Fink and Gantz (1996) delineate between “interpretive” and “critical” analyses, the former embracing a qualitative/holistic method, and the latter resting on value judgments derived from ideological theory. In this book, the working definition of content analysis assumes a quantitative approach. Quantitative analyses typically rely on the soundness of *a priori* measurement instruments; qualitative and critical analyses usually rely on the expertise of an expert scholar. In quantitative content analysis, the empirical process is independent of the particular scholar; in qualitative or critical message analyses, it is not.

That said, it should be noted that the dividing line between quantitative and qualitative might be viewed as “a rather thin and discreet line. . . . Even the most sophisticated piece of quantitative research remains dependent on natural language (words), while most qualitative studies do contain some kind of quantitative information (numbers)” (Schedler & Mudde, 2010, pp. 418–419; see, for example, Weisburd, 2009).

Further, we might consider applying the labels of *quantitative* and *qualitative* separately to the phenomenon under investigation and to the analytical strategies used to describe or summarize the phenomenon. Often, the core task of quantitative measures is to put numerical values, either counts or

amounts, to *qualities* of a phenomenon (e.g., Fukkink & Hermanns, 2009). Indeed, in survey and experimental research we accept *quantitative* self-report measures of such human *qualities* as state depression, extraversion, and communication apprehension. Similarly, in content analysis, we have seen *quantitative* measures of such *qualities* as the framing of a news item or the emotional tone of a political speech. That is, the phenomenon under investigation, or the constructs being examined, might be very *qualitative* in nature, and the analyses applied might be indisputably *quantitative*. The reverse is also possible, in which *quantitative* events might be interpreted in a *qualitative* fashion. Here, the focus will be on the analytical strategies employed and their underlying assumptions.

A complete review of all the types of qualitative message analyses that may complement quantitative content analysis is beyond the scope of this volume. But the reader should become aware of some of the main options for such analyses of messages (Lindlof & Taylor, 2011).

An important methodological source for qualitative content analysis of mediated messages is Altheide's (1996) canonical text (see also Altheide & Schneider, 2013). At its core, the method relies on identifying thematic patterns in a text (i.e., message or set of messages). The themes are not imposed upon the text from outside (e.g., via a theoretically informed coding mechanism or past studies) or *a priori*, but they emerge as the researcher undertakes a close reading of a text. Once themes are identified, the analyst looks for thematic patterns in the text.

Another useful source is Hijmans's (1996) typology of "qualitative content analyses" applied to media content. She presents accurate descriptions of some of the main qualitative analytic methods that have been applied to messages. Based on descriptions by Hijmans (pp. 103–104) and by Gunter (2000), they are as follows.

### *Rhetorical Analysis*

For this historically revered technique, properties of the text (both words and images) are crucial. The analyst engages in a reconstruction of characteristics of text or image or both, such as the message's construction, form, metaphors, argumentation structure, and choices. The emphasis is not so much on *what* the message says as on *how* the message is presented. The message is viewed not as an aesthetic object, but as an artistically structured instrument for communication and persuasion, with consideration given to the interaction among text, source, and audience. The analysis involves breaking the text down into parts; by understanding how the different parts operate, the analyst develops insights into the overall persuasive strategies used. There is an assumption that the researcher is a competent rhetorician. This technique has a *very* long history, with its principal origins among the Greek philosophers (Aristotle, 1991), and is the legitimate forebearer of many of today's academic disciplines. Rhetorical analysis has been widely

applied to news content, political speech, advertising, and many other forms of communication (McCroskey, 2005).

### *Narrative Analysis*

Informed by narrative theory, the goal of narrative analysis is to understand relationships between a text and social reality (Altman, 2008). Through all forms of communication, humans tell stories, and narrative is regarded as a basic and universal mode of verbal expression (Smith, 2000). Via narrative analysis, the scholar can unpack individual experiences and representations in stories and plots (Franzosi, 1998; Riessman, 2008). This technique involves a description of formal narrative structure. Attention focuses on characters—their difficulties, choices, conflicts, complications, and developments. The analysis involves reconstruction of the composition of the narrative. The assumption is that the researcher is a competent reader of narratives. One of the most complex and interesting applications of this technique is Propp's exhaustive analysis of Russian fairy tales (Propp, 1968), which establishes common character roles (e.g., hero, helper, villain, dispatcher), an identifiable linear sequence of elements in the narrative (e.g., initial situation, absention, interdiction), and particular functions in the narrative (e.g., disguise, pursuit, transfiguration, punishment).

### *Discourse Analysis*

This process engages in characteristics of manifest language and word use—description of topics in media texts—through consistency and connection of words to theme analysis of content and the establishment of central terms. The technique aims at typifying media representations (e.g., communicator motives, ideology). The focus is on the researcher as competent language user. Gunter (2000) identifies van Dijk's *Racism and the Press*, published in 1991, as a clear example of a large-scale discourse analysis. According to Gunter, van Dijk analyzes the “semantic macrostructures,” or the overall characteristics of meanings, with regard to ethnic minorities in the news media (p. 88), concluding that minority groups are depicted as problematic.

Discourse analysis has been a popular method for analyzing public communication, with analyses ranging from the macroscopic to the very microscopic. Duncan (1996) examined the 1992 New Zealand National Kindergarten Teachers' Collective Employment Contract Negotiations and identified two discourses—“Children First” and “For the Sake of the Children.” Both discourses were evident in arguments used by each side in the labor negotiations, in arguments *for* teacher pay and benefits by the teachers' representatives and in arguments *against* such expenditures by employers and government representatives. Duncan's article presents numerous direct quotes from the negotiations to support her point of view. Typical of this method,

she points out that her analysis “is *one* reading of the texts, and that there will be numerous other readings possible” (p. 161).

### *Structuralist or Semiotic Analysis*

The focus here is on deep meanings of messages. The technique aims at discovering deep structures, latent meanings, and the signifying process through signs, codes, and binary oppositions. The assumption is that the researcher is a competent member of the culture. Structural semiotic analysis is informed by a theory of signs (Peirce, 1931/1958). According to semiotics, meaning is not only an outcome of a relationship between signifier and signified but also of the relationships between signs in thinking and language (Saussure, 1974). The aim of semiotic analysis is to identify linguistic structures (e.g., rules of language and culture) that organize relationships between signs in a communication process (Eco, 1976; Hodge & Kress, 1988; Saussure, 1974).

Semiotics has been a valuable technique for examining cultural artifacts. Christian Metz’s (1974) classic text, *A Semiotics of the Cinema*, applies the wide range of semiotic techniques to narrative film. He provides a syntagmatic analysis (i.e., one that examines relationships between segments [syntagms] in the text of the film) for the French film, *Adieu Philippine*, indicating the *structure* of the film in shots, scenes, sequences, and the like. He also offers a detailed semiotic analysis of the self-reflexive “mirror construction” of Federico Fellini’s semiautobiographical film, *8-1/2*.

### *Interpretative Analysis*

The focus of this technique is on the formation of theory from the observation of messages and the coding of those messages. With its roots in social scientific inquiry, it involves theoretical sampling; analytical categories; cumulative, comparative analysis; and the formulation of types or conceptual categories. The methodology is clearly spelled out, but it differs from scientific inquiry in its wholly qualitative nature and its cumulative process, whereby the analyst is in a constant state of discovery and revision. The researcher is assumed to be a competent observer.

Many of the systems of analysis developed by such interpretative methods are empirical and detailed and in fact are more precise and challenging than most content analyses (e.g., Berger, 1998, 2014). With only minor adjustment, many are appropriate for use in content analysis as well.

In addition to these qualitative message analysis types reviewed by Hijmans (1996), several others deserve mention.

### *Conversation Analysis*

Conversation analysis is a technique for analyzing naturally occurring conversations, used by social scientists in the disciplines of psychology,

communication, and sociology (Sudnow, 1972). The procedure has been described as a “rigorously empirical approach which avoids premature theory construction and employs inductive methods . . . to tease out and describe the way in which ordinary speakers use and rely on conversational skills and strategies” (Kottler & Swartz, 1993, pp. 103–104). Most typically, it relies on transcribed conversations. The technique generally falls within the rubric of ethnomethodology, scholarly study in which the precise and appropriate methods emerge from within the process of study, with the clearly subjective involvement of the investigator. Examples of its applications have included an analysis of doctor–patient interaction (Manning & Ray, 2000) and an in-depth analysis of a notorious interview of Vice President George Bush by television reporter Dan Rather as they jockeyed for position in order to control the flow of a “turbulent” interview (Nofsinger, 1988/1989).

### Critical Analysis

Critical analysis, often conducted in a tradition of cultural studies, has been a widely used method for the analysis of media messages (Newcomb, 1987). Critical analysis is informed by critical theory and Marxist criticism of capitalism and neoliberalism. The aim of critical theory in the study of communication is to identify structures of power that maintain social differences between classes, genders, and races (Habermas, 1981, 1987). One of the foundational principles of critical theory of the “Frankfurt School” has been to search for practical solutions to the problem of human emancipation and “liberate human beings” from the cultural, political, and economic conditions that enslave humans and undermine true democracy (Horkheimer, 1982; Horkheimer & Adorno, 1972).

The area of film studies provides a good example of a fully developed, theoretically sound literature that primarily uses the tools of critical analysis (e.g., Cooper, 2010; Lyman, 1997). For example, Strong’s (1996) essay about how Native Americans are “imaged” in two mid-1990s media forms—Disney Studio’s *Pocahontas* and Paramount’s *The Indian in the Cupboard*—is influenced heavily by her own roles as mother, musician—American raised during a period when “playing Indian” was a childhood rite of passage—and anthropologist long interested in White America’s representations of Native Americans. She acknowledges these various roles and perspectives, provides precise details to back her assertions (including many lines and song lyrics from the movies), and gives summative statements that bring the details into line with cultural frameworks. For example, she concludes that “Disney has created a New Age Pocahontas to embody our millennial dreams for wholeness and harmony, while banishing our nightmares of savagery without and emptiness within” (p. 416).

### *Normative Analysis*

Some analyses are explicitly normative or prescriptive (e.g., Legg, 1996). For example, a guide to *Stereotypes, Distortions and Omissions in U.S. History Textbooks: A Content Analysis Instrument for Detecting Racism and Sexism* (Council on Interracial Books for Children, 1977), compiled by 32 educators and consultants, provides checklists for history textbook coverage of African Americans, Asian Americans, Chicanos, Native Americans, Puerto Ricans, and women. For each group, an instrument is presented with criteria for parents and teachers to use when examining children's history texts. For instance, in the Native American checklist, the following criteria are included:

The myth of "discovery" is blatantly Eurocentric. . . . War and violence were not characteristic of Native nations. . . . The Citizenship Act of 1924 was not a benevolent action . . . and the BIA [Bureau of Indian Affairs] is a corrupt and inefficient bureaucracy controlling the affairs of one million people. (pp. 84–85)

The guide is certainly well intended and a powerful tool for social change. Its prescriptive approach, however, does not fit most definitions of content analysis.

Similarly, in their article, "Evaluation Criteria and Indicators of Quality for Internet Resources," Wilkinson, Bennet, and Oliver (1997) offer a list of 125 questions to ask about a web site. Their goal is to pinpoint characteristics that indicate accuracy of information, ease of use, and aesthetic qualities of Internet material. The work is a normative prescription for a "good" web site. Although they call their proposal a content analysis, it does not meet the definition given in this book.

### *Computers and Qualitative Message Analysis*

In recent decades, computer adjuncts have been developed to support the tasks of these various qualitative methods. NVivo, a qualitative counterpart to quantitative CATA programs, is used to provide detailed markup, retrieval, and description of textual and related documents (Bazeley & Jackson, 2013). It is based on the organization of coded text via a system of concept nodes, grouped hierarchically in a tree structure, which is displayed by the program. Because qualitative methods emphasize researchers being the "research instrument" for data collection and data analysis, the qualitative uses of NVivo are usually in the form of managing data and assisting qualitative coding and memoing. It is unlike quantitative analyses in which researchers construct or use built-in algorithms to mine textual data. While NVivo has added quantitative supplements to its repertoire over the years, its core utility remains in support of qualitative methods (Bazeley & Jackson, 2013).

An example may be seen in a study by Creed, DeJordy, and Lok (2010), who used NVivo to assist in their narrative analysis of in-depth interview responses by 10 gay, lesbian, bisexual, and transgender ministers serving in two mainline Protestant denominations in the United States. They used an inductive narrative analysis, moving “iteratively between the data, the emerging themes, and existing theory in several phases” (p. 1342). Through these techniques, they developed a model of “identity work” for the ministers, with eight first-level constructs (e.g., healing and accepting, challenging orthodoxy from within) that merged into three second-level microprocesses (e.g., identity reconciliation work). In this and similar studies, computer applications such as NVivo bring coherence to what otherwise would be a daunting—if not impossible—task of making sense of complex message content.

## Myth 4: Content Analysis Is for Academic Use Only

**Truth:** Not so.

Certainly, the majority of content analyses have been conducted by academics for scholarly purposes. However, there has been growing interest among commercial researchers and communication practitioners in particular applications of content analysis. Whitney, Wartella, and Kunkel (2009) have provided a thorough consideration of reasons why governmental agencies, media institutions, issue advocates, and the general public can find utility in content analysis. Content analysis is often used in applied, nonacademic situations. For example, law firms have hired academics to conduct content analyses of news coverage of their high-profile clients, to be used as evidence in conjunction with change-of-venue motions (i.e., excessive and negative coverage may warrant moving a court case to another city in order to obtain a fair trial; McCarty, 2001) or to establish particular patterns of news coverage that may refute plaintiff claims of information availability.

In response to criticisms, a southern daily newspaper hired a journalism scholar to systematically document coverage of the local African American community (Riffe, Lacy, & Fico, 2014). In 2009, the U.S. Secret Service National Threat Assessment Center (NTAC) engaged the expertise of The National Academies and the committee of experts it convened to explore the utility of a variety of message-focused methods—including content analysis—for the prediction of threat outcomes.

As part of a legal settlement with the ACLU to address poor police–civilian relations that culminated in three days of civil unrest in Cincinnati, Ohio, the city of Cincinnati funded a RAND Corporation study of traffic stops that had been recorded via vehicle-mounted cameras. Dixon et al. (2008) used communication accommodation theory (CAT) as a template for the analysis of the “dashcam” footage. With random sampling stratified by the combination of officer/driver race(s), the study detected that (a) Black drivers were more

likely to experience extensive policing (i.e., longer stops); (b) the communication quality of White drivers was more positive (i.e., accommodating) than that of Black drivers (although statistical controls indicated that some of this was due to the greater length of the stops for Black drivers); and (c) officers' communication was more positive (i.e., more accommodating) when the officer and driver were of the same race. The findings have clear implications for communication skills training for police officers and for community intervention programs that might ease police–civilian tensions.

Internal corporate research initiatives sometimes include content analyses. The marketing research unit of a large-city newspaper systematically compared its own coverage of regional issues with that provided by local television news. Organizational communication consultants often include a content analysis of recorded messages (e.g., emails, memos) in their audit of the communication patterns within the organization. Rittenhouse Rankings, an investor-relations firm, has used content analysis of annual CEO letters to effectively predict the following year's stock prices for 100 top companies (Blumenthal, 2013). And the clinical diagnostic tools of criterion-based content analysis (e.g., PCAD) have been used in nonacademic settings by psychologists and legal professionals (Gottschalk & Bechtel, 2008).

Increasingly, methods of content analysis are included by marketing research and public opinion firms as part of their template of research offerings, ranging from coding of open-ended responses on surveys to analyses of news coverage. Some firms even specialize in custom content analyses, such as Talkhouse LLC, which has supplied its CATPAC III software to General Motors suppliers for the monitoring of the impact of GM Super Bowl ads. And Social Science Automation offers software and analyses with its Profiler Plus Text Coding Platform; its services have been engaged by both government and private-sector clients.

## A Six-Part Definition of Content Analysis

---

This book assumes that content analysis is conducted within the scientific method but with certain additional characteristics that place it in a unique position as a primary message-centric methodology.

### Box 1.1 Defining Content Analysis

Some of the main players in the development of quantitative message analysis present their points of view:

Berelson (1952, p. 18): Content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication.

Stone et al. (1966, p. 5, with credit given to Dr. Ole Holsti): Content analysis is any research technique for making inferences by systematically and objectively identifying specified characteristics within text.

Carney (1971, p. 52): The general purpose technique for posing questions to a “communication” in order to get findings which can be substantiated. . . . [T]he “communication” can be anything: A novel, some paintings, a movie, or a musical score—the technique is applicable to all alike and *not* only to analysis of literary materials.

Kassarjian (1977, p. 9): [After reviewing definitions to date, t]hese researchers and others agree that the distinguishing characteristics of content analysis are that it must be *objective, systematic, and quantitative*.

Weber (1990, p. 9): Content analysis is a research method that uses a set of procedures to make valid inferences from text.

Berger (1998, p. 23): Content analysis . . . is a research technique that is based on measuring the amount of something (violence, negative portrayals of women, or whatever) in a representative sampling of some mass-mediated popular art form.

Smith (2000, p. 314): Content analysis is a technique used to extract desired information from a body of material (usually verbal) by systematically and objectively identifying specified characteristics of the material . . . [thereby] yielding unbiased results that can be reproduced by other qualified investigators. Content analysis differs from clinical *interpretation*, which is more holistic and provisional, and for which specific criteria are not made explicit in advance.

Ahuvia (2001, p. 139): “Content analysis” will be used as a . . . general term for methodologies that code text into categories and then count the frequencies of occurrences within each category.

Krippendorff (2013, p. 24): Content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use.

Riffe, Lacy, & Fico (2014, p. 19): Quantitative content analysis is the systematic and replicable examination of symbols of communication, which have been assigned numeric values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods, to describe the communication, draw inferences about its meaning, or infer from the communication to its context, both of production and consumption.

Babbie (2013, p. 330): The study of recorded human communications.

**This book:** Content analysis is a summarizing, quantitative analysis of messages that follows the standards of the scientific method (including attention to objectivity-intersubjectivity, a priori design, reliability, validity, generalizability, replicability, and hypothesis testing based on theory) and is not limited as to the types of variables that may be measured or the context in which the messages are created or presented.

Box 1.1 presents some alternative definitions of content analysis for the sake of comparison. More details on this book’s definition are presented in the discussion that follows.

## 1. Content Analysis as Following the Standards of the Scientific Method

Perhaps the most distinctive characteristic that differentiates content analysis from other, more qualitative or interpretive message analyses is the attempt to meet the standards of the scientific method (Bird, 1998; Klee, 1997); by most definitions, it fits the positivism paradigm of social research (Gunter, 2000).<sup>4</sup> The goal of the scientific method is generalizable knowledge, with the concomitant functions of description, prediction, explanation, and control (Hanna, 1969; Kaplan, 1964).

A commitment to the scientific method includes attending to such criteria as the following:

### *Objectivity–Intersubjectivity*

A major goal of any scientific investigation is to provide a description or explanation of a phenomenon in a way that avoids the biases of the investigator. Thus, objectivity is desirable. However, as the classic work *The Social Construction of Reality* (Berger & Luckman, 1966) points out, there is no such thing as true objectivity—“knowledge” and “facts” are what are socially agreed upon. According to this view, all human inquiry is inherently subjective, but still we must strive for consistency among inquiries. We do not ask “Is it true?” but rather “Do we agree it is true?” Scholars sometimes refer to this standard as *intersubjectivity* (Babbie, 1986, p. 27).

### *An A Priori Design*

Although an a priori (i.e., before the fact) design is actually a part of the task of meeting the requirement of objectivity–intersubjectivity, it is given its own listing here to provide emphasis. Too often, a so-called content analysis report describes a study in which variables were chosen and “measured” *after* all the messages were observed. This wholly inductive approach violates the guidelines of scientific endeavor. All decisions on variables, their measurement, and coding rules must be made before the final measurement process begins. In the case of human coding, the code-book and coding form must be constructed in advance. In the case of computer coding in CATA, the dictionary or other coding protocol should be established *a priori*.

However, the self-limiting nature of this “normal science” approach should be mentioned. As Kuhn’s (1970) seminal work on paradigms has pointed out, deduction based on past research, theories, and bodies of evidence within the current popular paradigm does not foster innovation. Content analysis has a bit of this disadvantage, with the insistence that coding schemes be developed *a priori*. Still, creativity and innovation can thrive within the method. As described in Chapter 4, a lot of exploratory work can and should be done

before a final coding scheme is “set in stone.” The entire process may be viewed as a combination of induction and deduction.

### ***Reliability***

Reliability has been defined as the extent to which a measuring procedure yields the same results on repeated trials (Carmines & Zeller, 1979). When human coders are used in content analysis, this translates to *intercoder reliability*, or level of agreement among two or more coders. In content analysis, reliability is paramount. Without acceptable levels of reliability, content analysis measures are meaningless. Chapter 6 addresses this important issue in detail.

### ***Validity***

Validity refers to the extent to which an empirical measure adequately reflects what humans agree on as the real meaning of a concept (Babbie, 2013, p. 151). Generally, it is addressed with the question “Are we really measuring what we want to measure?” Although in content analysis the researcher is the boss, making final decisions on what concepts to measure and how to measure them, there are a number of good guidelines available for assessing and improving validity (Carmines & Zeller, 1979). Chapter 5 gives a more detailed discussion.

### ***Generalizability***

The generalizability of findings is the extent to which they may be applied to other cases, usually to a larger set that is the defined population from which a study’s sample has been drawn. After completing a poll of 300 city residents, the researchers obviously hope to generalize their findings to all residents of the city. Likewise, in a study of 800 personal ads in newspapers, Kolt (1996) generalized his findings to all personal ads in U.S. newspapers in general. He was in a good position to do so because he (a) randomly selected U.S. daily newspapers, (b) randomly selected dates for specific issues to analyze, and then (c) systematically random sampled personal ads in each issue. In Chapter 3, the options for selecting representative samples from populations will be presented.

### ***Replicability***

The replication of a study is a safeguard against overgeneralizing the findings of one particular research endeavor. Replication involves repeating a study with different cases or in a different context, checking to see if similar results are obtained each time (Babbie, 2013, p. 7). Whenever possible, research reports should provide enough information about the methods and

protocols so that others are free to conduct replications. Throughout this book, the assumption is made that *full reportage* of methods is optimal, for both academic and commercial research.

As Hogenraad and McKenzie (1999) caution, content analyses are sometimes at a unique disadvantage with regard to replication. Certain messages are historically situated, and repeated samplings are not possible, as with their study of political speeches leading up to the formation of the European Union. They propose an alternative—*bootstrap replication*—which compares and pools multiple random subsamples of the original data set.

### *Hypothesis Testing Based on Theory*

The scientific method is generally considered to be hypothetico-deductive. That is, from theory, one or more hypotheses (conjectural statements or predictions about the relationship among variables) are derived. Each hypothesis is tested deductively: Measurements are made for each of the variables, and relationships among them are examined statistically to see if the predicted relationship holds true. If so, the hypothesis is supported and lends further support to the theory from which it was derived. If not, the hypothesis fails to receive support, and the theory is called into question to some extent. Ultimately, theory may be revised in the face of nonconfirming evidence. If existing theory is not strong enough to warrant a prediction, a sort of fallback position is to offer one or more research questions. A research question poses a query about possible relationships among variables. In the deductive scientific model, hypotheses and research questions are both posed *before* data are collected. Chapter 4 presents examples of hypotheses and research questions appropriate to content analysis.

## **2. The Message as the Unit of Analysis, the Unit of Data Collection, or Both**

The unit in a research study is the individual “thing” that is the subject of study—what or whom is studied. Frequently, it is useful to distinguish between the *unit of data collection* (sometimes referred to as the *unit of observation*; Babbie, 2013) and the *unit of analysis*, although in many studies, these two things are the same. The unit of data collection is the element on which each variable is measured. The unit of analysis is the element on which data are analyzed and for which findings are reported.

In most social and behavioral science investigations, the individual *person* is both the unit of data collection and the unit of analysis. For example, when a survey of city residents is conducted to measure opinions toward the president and the mayor, let’s say, the unit of data collection is the individual respondent—the person. That is, telephone interviews may be conducted, and normally, each person responds alone. The variables (e.g., attitude toward the

president, attitude toward the mayor, gender, age) are measured on each unit. The unit of analysis is also typically the individual person. That is, in the data set, each respondent's answers will constitute one line of data, and statistical analyses will be conducted on the data set, with  $n$  equaling the number of people responding. When "average rating of confidence in the president" is reported as 6.8 on a 0-to-10 scale, that's the mean based on  $n$  respondents.

Sometimes, the unit of data collection and the unit of analysis are not the same. For example, a study of marital discord may record interactions between married partners. The unit of data collection may be the "turn" in verbal interaction: Each time an individual speaks, the tone and substance of his or her turn may be coded. However, the ultimate goal of the study may be to compare the interactions of those couples who have received intervention counseling and those who have not. Thus, the unit of *analysis* may be the dyad, pooling information about all turns and interactions for each married pair.

In content analysis, the unit of data collection or the unit of analysis—or both—must be a *message unit*. Quite simply, there must be communication *content* as a primary subject of the investigation for the study to be deemed a content analysis. In the marital-discord example just described, the unit of data collection is a message unit (an interaction turn), and the unit of analysis is not. It may be called a content analysis. Chapter 3 provides more examples of unitizing.

### 3. Content Analysis as Quantitative

The goal of any quantitative analysis is to produce *counts* of key categories and measurements of the *amounts* of other variables (Fink, 2009). For both counts and amounts, there is a numerical process. A quantitative content analysis has as its goal a numerically based summary of a chosen message set. It is neither a gestalt impression nor a fully detailed description of a message or message set.

There is often confusion between what is considered *quantitative* and what is considered *empirical*. Empirical observations are those based on real, apprehendable phenomena. Accordingly, both quantitative and qualitative investigations may be empirical. What, then, is not empirical? Efforts to describe theory and conditions without making observations of events, behaviors, and other "real" aspects of the world, such as abstract theorizing, many portions of the discipline of philosophy, and (perhaps surprisingly) certain types of scholarship in mathematics (which is, of course, quite quantitative in focus) might be considered nonempirical. Much of the social and behavioral science literature is based on empirical work, which may be quantitative or qualitative.

As noted earlier, we may distinguish between the quantitative or qualitative nature of the analysis and the quantitative or qualitative attributes of the phenomenon under examination. Clearly, *qualities* of a message are routinely

subject to *quantification* (Smith, 2000). Very often, a study that might be characterized as “qualitative” is actually quite quantitative—the phenomenon being studied is what is qualitative in nature. Farrell, Wallis, and Evans (2007) conducted individual and focus group interviews concerning attitudes toward nursing programs and, as they put it, “analyzed the qualitative data using a standardized codebook and content analysis” (p. 267). And in a study of lower-level service workers’ commentaries on the experience of part-time work, Walsh (2007) collected open-ended survey responses, and the “qualitative comments were analysed with respect to [23 discrete] categories and themes and were decomposed in relation to their frequency of occurrence” (p. 163). In these cases, quantitative analyses are applied to what the researchers quite properly view as qualitative information.

It should be made clear at the outset that this book takes the viewpoint that critical and qualitative analyses that are empirical are typically extremely useful to the content analyst. They have the potential to provide a highly valid source of detailed or “deep” information about a text. (Note that the term *text* is a preferred term in many critical analyses and denotes not just written text but also any other message type that is considered in its entirety. For example, the text of a film includes its dialog, its visuals, production techniques, music, characterizations, and anything else of meaning presented in the film.) The empiricism of a careful and detailed critical analysis is one of its prime strengths and may produce such a lucid interpretation of the text as to provide us with a completely new encounter with the text. Such an analysis may bring us into the world of the text (e.g., into what is called the *diegesis* of a film, “the sum of a film’s denotation: the narration itself, but also the fictional space and time dimensions implied in and by the narrative, and consequently the characters, the landscapes, the events, and other narrative elements” [Metz, 1974, p. 98]). It may illuminate the intentions of the source of the text, or it may allow us to view the text through the eyes of others who may experience the text (e.g., as in providing an understanding of a child’s view of a favorite TV program, something that may be essential to a full appreciation of child-centric content).

When approaching a text—a message or message set—the researcher needs to evaluate his or her needs and the outcomes possible from both quantitative (i.e., content analysis) and nonquantitative analyses. For example, to identify and interpret pacifist markers in the film *Saving Private Ryan*, a critical analysis, perhaps with a Marxist approach, is in order. To establish the prevalence of violent acts in top-grossing films of the 2000s, a content analysis is more appropriate. The content analysis uses a broader brush and is typically more generalizable. As such, it is also typically less in-depth and less detailed.

As noted above, a concerted pairing of quantitative content analysis with qualitative or critical message analysis has obvious advantages, given the complementary goals of each (Hardy, Harley, & Phillips, 2004; Neuendorf, 2004; Stepchenkova, Kirilenko, & Morrison, 2009). This outlook coincides nicely with the view presented by Gray and Densten (1998): “Quantitative

and qualitative research may be viewed as different ways of examining the same research problem" (p. 420). This *triangulation* of methods "strengthens the researcher's claims for the validity of the conclusions drawn where mutual confirmation of results can be demonstrated" (p. 420).<sup>5</sup> Such triangulation is unfortunately relatively rare (e.g., Hymans, 2010; Pinto & McKay, 2006; Southall et al., 2008) and not always embraced by a particular discipline. Indeed, Phelan and Shearer (2009) described their analyses as "bastardised" in that they supplemented traditional discourse analysis with some quantification.

One study combined quantitative content analysis and semiotic analysis to assess gender portrayals in drug advertisements in an Irish medical publication (Curry & O'Brien, 2006). Another examined storytelling in Taiwanese and European American families, combining ethnographic fieldwork with content-analytic coding of audio and video recordings of naturally occurring talk in the home (Miller et al., 1997). In another example, Kumar (2005) combined quantitative content analysis of news coverage of the Abu Ghraib incident with qualitative historical contextual analysis that helped explain the dynamics of the political and media interactions relevant to the case. (See also Lieberman et al., 2009, for a "fusion" of quantitative experimental research and critical message analyses.)

#### 4. Content Analysis as Summarizing

As noted in the previous point, a content analysis summarizes rather than reports all details concerning a message set. This is consistent with a *nomothetic* approach to scientific investigations (i.e., seeking to generate generalizable conclusions from an aggregate of cases), rather than an *idiographic* approach (i.e., focusing on a full and precise conclusion about a particular case, as in a case study). An idiographic study seeks to fully describe a single artifact or case from a phenomenological perspective and to connect the unique aspects of the case with more general truths or principles. A nomothetic study hopes to identify generalizable findings, usually from multiple cases, and demands "specific and well-defined questions that in order to answer them it is desirable to adopt standardized criteria having known . . . characteristics" (Te'eni, 1998). Idiographic study implies conclusions that are unique, nongeneralizable, subjective, rich, and well-grounded; nomothetic study implies conclusions that are broadly based, generalizable, objective, summarizing, and inflexible.

The goal of some message analyses, not deemed to be quantitative content analyses, is a type of microdocumenting. Historians have contributed a number of examples of very precise, fully explicated analyses that rely on original textual sources. Because these analyses are based on texts, we might be tempted to call them content analyses. But some of them display an obvious attempt to report all possible details across a wide variety of units of data

collection rather than to summarize information for a chosen unit of data collection or analysis. One example is Kohn's (1973) book on Russia during World War I, in which he professes to attempt "an *exhaustive* inquiry into the vital statistics of Russia" (p. 3), ultimately to assess the economic and noneconomic consequences of the war on Russian society. The work is largely a reportage of numerical facts taken from a variety of textual sources. Another example, the book *Plantation Slaves of Trinidad, 1783–1816*, brings the reader into the daily lives of those Caribbean slaves during the nation's slave period of that time (John, 1988). Aggregate figures on slave mortality and childbearing are presented side by side with drawings of slave life on the Trinidad plantations. Also typical of a qualitative analysis of text, Creed, DeJordy, and Lok (2010) present "exemplars from the data" as their findings—these are extended verbatim quotes from in-depth interviews, with no summarization.

Hesse-Biber, Dupuis, and Kinder (1997) used the qualitative analysis computer program, HyperRESEARCH, to identify, index (which they term *code*), and search a broad mix of photographs, text samples, audio segments, and video segments. The emphasis was on cataloging discrete exemplars of desired content in a manner that made their retrieval and comparison easy. For example, after indexing is complete, the researchers might query the program to produce all examples that have been tagged "expression of self-esteem" (p. 7). These cases may be examined and cross-indexed according to other characteristics, but the responsibility for making sense of these interwoven networks of similarities rests with the analyst, and there is no goal of providing a summary of the complexities of the text.

In contrast, the quantitative content analysis summarizes characteristics across a set of messages. For example, in a study of television news coverage of Belgian automobile crashes, Beullens, Roe, and Van den Bulck (2008) provided a neat summary for all 2005 news broadcasts dealing with traffic accidents from the top two television channels. They found that the most prominent "contributing factors" mentioned were weather (11.8%), alcohol use (7.1%), and speeding (6.4%). Further, 48% of stories were framed as human interest, while 47% were framed as responsibility-oriented. Throughout their findings, the results *summarized* the state of news reporting across the sample of 297 stories.

## 5. Content Analysis as Applicable to All Contexts

The term *content analysis* is not reserved for studies of mass media or for any other type of message content or context. As long as other pertinent characteristics apply (e.g., quantitative, summarizing), the study of any type of message pool may be deemed a content analysis. The messages may be mediated—that is, having some message reproduction or transmittal device interposed between source and receiver. Or they may be nonmediated—that is, experienced face to

face. Although not attempting to create an exhaustive typology of communication purposes and context, the sections to follow give some examples of the range of applications of the techniques of content analysis.

### *Individual Messaging*

Some analyses examine the creation of messages by a single individual, with the typical goal of making some inference to that source (Chapter 2 will provide further discussion regarding limits to the ability to make inferences from content analysis findings).

In psychology, there is a growing use of content analysis of naturally produced text and speech as a type of psychometric instrument (Gottschalk, 1995; Gottschalk & Bechtel, 2008; Horowitz, 1998; Tully, 1998). This technique analyzes statements made by an individual to diagnose psychological disorders and tendencies, to measure psychological traits of the source, or to assess the credibility of the source (Doris, 1994). Nearly all these efforts stem from the work of Philip Stone (Stone et al., 1966) in the Harvard Department of Social Relations. His “General Inquirer” computer program was the first to apply content-analytic techniques to free-speech words (see “Milestones in Content Analysis History” at *The Content Analysis Guidebook Online*, CAGO). Rosenberg and others (e.g., Rosenberg & Tucker, 1979) applied the computer technique to the language of schizophrenics, with the goal of better diagnosis. In an example of a further refinement of such procedures, Broehl and McGee (1981) analyzed the writings of historical figures—three British lieutenants serving during the Indian Mutiny of 1957 to 1958—and on this basis developed psychological profiles for the officers. Even the Watergate tapes have been studied using content analysis to gain insights into the underlying psychological motives of the individuals involved (Weintraub & Plant, as cited in Broehl & McGee, 1981, p. 288).

Others in the field of psychology have continued to develop computer analyses that produce diagnoses from written or spoken text. For example, Gottschalk, Stein, and Shapiro (1997) compared results from standard psychometric tests, such as the MMPI (Minnesota Multiphasic Personality Inventory), with content analysis results from a CATA analysis of transcripts of five-minute speeches. Their study of 25 new psychiatric outpatients found strong construct validity—the speech analyses were highly correlated with corresponding questionnaire outcomes. They point out the potential value in being able to use ordinary spoken or written material for an initial, rapid diagnostic appraisal that can easily remain unobtrusive (i.e., the individual does not have to submit to a lengthy questionnaire administration; p. 427). The content analysis scheme used—the 16-part Gottschalk-Gleser Content Analysis Scales—became a software program (PCAD) developed and validated over a period of many years.

Another application of content analysis to the individual as message generator is the common method of coding responses to open-ended questionnaire

items and in-depth interviews (Gray & Densten, 1998). For example, Farrow et al. (2009) coded open-ended responses in a survey of Irish coroners' attitudes toward suicide. Although the first steps in this process usually include a qualitative review of the message pool and the development of an emergent coding scheme based on what's represented in the pool, it must be remembered that the true content analysis portion is the subsequent, careful application of the a priori coding scheme to the message pool.

In the fields of linguistics, history, and literature, some attempts have been made at analyzing individual authors or other sources. In recent decades, CATA analyses have been conducted either to describe a source's style, to verify a questionable source, or to identify an unknown source (Floud, 1977; Olsen, 1993). For example, Elliott and Valenza's (1996) "Shakespeare Clinic" has developed computer tests for Shakespeare authorship, and Martindale and McKenzie (1995) used CATA to confirm James Madison's authorship of *The Federalist*.

Content analysis may be applied to nonverbal communication of the individual as well. Magai et al. (2006) used a facial affect coding scheme to measure emotional experience in a study of age-related differences in experience and expressed affect and emotion regulatory skills. They utilized the Maximally Discriminative Facial Movement Coding System (MAX), introduced by Izard (1979). Another popular system, the Facial Action Coding System (FACS; Ekman & Friesen, 1978; Ekman, Friesen, & Hager, 2002) is a rich system for human coding of facial "action units," marked by very manifest motions such as "nostril wings widen and raise" or "inner and/or central portion of brow lowers slightly," which are intended to link up with overall expressions of emotion (although FACS does not ask the coder to make such judgments).

### *Interpersonal and Group Messaging*

This book assumes a definition of interpersonal communication that acknowledges the *intent* of the messaging to reach and be understood by a *particular individual*. This may occur face to face, or it may be mediated, as in the cases of telephoning, emailing, or social media messaging. It may occur in a dyad or a small group.

To study face-to-face group processes, Bales (1950) developed a content analysis scheme that calls for the coding of each communication act. A verbal act is "usually the simple subject-predicate combination," whereas a nonverbal act is "the smallest overt segment of behavior that has 'meaning' to others in the group" (Bales et al., 1951, p. 462). Each act is coded into one of 12 categories: (a) shows solidarity, (b) shows tension release, (c) agrees, (d) gives suggestion, (e) gives opinion, (f) gives orientation, (g) shows antagonism, (h) shows tension, (i) disagrees, (j) asks for suggestion, (k) asks for opinion, or (l) asks for orientation. Bales's scheme has been widely used and elaborated on (Bales & Cohen, 1979) and has also been adapted for analyzing human interaction in mass media content (Greenberg, 1980; Neuendorf & Abelman, 1987).

**Box 1.2** Analyzing Communication in Crisis**Perpetrator and Negotiator Interpersonal Exchanges**

Most standoffs between police and perpetrators are resolved nonviolently. An analysis of 137 crisis hostage incidents handled by the New York City Police Department revealed that in 91% of the cases, neither hostages nor hostage takers were killed (Rogan & Hammer, 1995, p. 554). Nonetheless, those crisis situations that end violently—such as the 1993 Branch Davidian conflagration in Waco, Texas—focus attention on the need to better understand the negotiation process. There is interest among scholars and police professionals alike in studying the communication content of negotiations in crisis situations so that outcomes may be predicted and negative outcomes prevented.

Rogan and Hammer (1995) had such a goal for their content analysis of audio recordings of three authentic crisis negotiations obtained from the FBI training academy. They looked at message affect—a combination of message valence and language intensity—across eight phases of each negotiation process. The unit of data collection was the uninterrupted talking turn. Each turn was coded by human coders for positive–negative valence and for Donohue's (1991) five correlates of language intensity: (a) obscure words, (b) general metaphors, (c) profanity and sex, (d) death statements, and (e) expanded qualifiers. The analysis was highly systematic and achieved good reliability (i.e., agreement between independent coders).

Total “message affect” scores were calculated for perpetrator and negotiator for each of the eight time periods in each negotiation. In all three situations, the negotiator’s message profile remained positive throughout, whereas the perpetrator’s score became more strongly negative during Periods 2 and 3. Eventually, between Periods 6 and 8, the perpetrator’s message affect shifted to a positive valence, approaching that of the negotiator. In the one successful negotiation studied, the perpetrator’s scores remained high and positive; in the two unsuccessful incidents (one culminating in the perpetrator’s suicide), the perpetrator’s scores began an unrelenting slide to intense negativity at Period 6 or 7.

The researchers point out certain limitations of the study—primarily, that the analysis was limited to message affect, with no consideration of other characteristics of the communicators, no examination of substantive or relational communication content, and so on. Nevertheless, just based on message affect, the results are striking. By looking at the charted message affect scores, you can visualize the process of negotiation success or failure. Although currently not useful for real-time application to ongoing crisis situations, this content analysis technique shows promise for the development of such applications. And researching past negotiation successes and failures provides practitioners insight into the dynamics of the process. As Rogan and Hammer (1995) note, “Ultimately, such insight could enable a negotiator to more effectively control a perpetrator’s level of emotional arousal, such that a negotiator could take actions to reduce a perpetrator’s highly negative and intense emotionality in an effort to negate potentially violent behavior” (p. 571), perhaps the ultimate useful application of the technique of content analysis.

**Box 1.3 The Variety of Content Analysis****Religious TV—Tapping Message Characteristics,  
Ranging From Communicator Style to Dollar Signs**

In the 1980s, religious broadcasting reached a peak of popularity with the rapid growth of “televangelism” (Frankl, 1987). Concerned with a growing perception of religious broadcasting as invasive and inordinately focused on fund-raising, the organization of Roman Catholic broadcasters, UNDA-USA, commissioned a set of content analyses. During the mid-1980s, researchers at Cleveland State University conducted an extensive five-part project. All the components of the project were quantitative content analyses, and they drew on a wide array of theories and research perspectives.

A set of 81 episodes of religious programs provided the content to be analyzed. These were three randomly sampled episodes for each of the top religious television or cable programs, as determined by an index of availability in a random sample of 40 U.S. towns and cities. These programs ranged from talk format shows, such as *The 700 Club*, to televangelist programs like *Jim Bakker* to narrative forms, such as the soap opera *Another Life* and the children’s stop-motion animated “daily lesson” program, *Davey and Goliath*. Different teams of coders were trained for the five types of analysis:

1. The demography of religious television

With the unit of data collection and analysis the *individual character* (real or fictional), a dozen demographic variables were assessed (based on previous content analyses of TV characters, such as Greenberg [1980] and Gerbner et al. [1980]), including social age (child, adolescent, young adult, mature adult, elderly), occupation, and religious affiliation. An example of the results was the finding that 47% of the characters were mature adults, with 37% being young adults. Children constituted only 7% of the sample, with the elderly at only 5% (Abelman & Neuendorf, 1984a).

2. Themes and topics on religious television

Here, the unit of data collection was a *period of time*: the five-minute interval. At the end of each five-minute period, a checklist coding form was completed by the coder, with 60 measures indicating simple presence or absence of a given social, political, or religious topic within all verbalizations in the period (pulling from existing analyses of religious communication, e.g., Hadden & Swann, 1981). Also, both explicit and implied appeals for money were recorded at the end of each five-minute period. Overall, \$328.13 was explicitly requested of the viewer per hour across the sample of religious programs (Abelman & Neuendorf, 1985a, 1985b).

3. Interaction analysis of religious television content

Using a scheme derived and adapted from Bales (1950), Borke (1969), and Greenberg (1980), interpersonal interactions among characters on religious television were examined. The unit of data collection was each

*verbal utterance* (act), which was coded as falling into one of 20 modes (e.g., offering information, seeking support, attacking, evading). The results suggested age and gender differences in interaction patterns; most interactions were male dominated, and the elderly were often shown as conflict-producing individuals who were the frequent targets of guidance from those who were younger (Neuendorf & Abelman, 1987).

#### 4. Communicator style of televangelists

Drawing on the considerable interpersonal communication literature on communicator style, notably the work of Robert Norton (1983), this aspect of the project targeted the 14 televangelists in the program sample and used as the unit of data collection *each verbal utterance within a monologue*. Each utterance was coded for a variety of characteristics, including mode (similar to the interaction coding scheme), vocal intensity, pace, and facial nonverbal intensity. Based on an overall intensity index, the top three “most intense” televangelists were James Robison, Robert Schuller, and Ernest Angley (Neuendorf & Abelman, 1986).

#### 5. Physical contact on religious television programming

Drawing on work in nonverbal communication (e.g., Knapp, 1978), this portion of the content analyses examined physical touch. The unit of data collection was *the instance of nonaccidental physical contact*. Characteristics of the initiator and recipient of the touching were tapped, as were type of touch (religious in nature, nonreligious), anatomical location of the touch, and the recipient’s reaction to the touch. A sample result was that there was a clear similarity with real-life touching along gender lines: Males were the primary initiators of physical contact, and it tended to be rather formal and ritualistic (i.e., a substantial portion of the contact was religious in nature, such as healing; Abelman & Neuendorf, 1984b).

### **Organizational Messaging**

Content analysis has been used less frequently for profiling messages within a defined organization than it has in other contexts (Tangpong, 2011). More often, messages within an organization have been scrutinized using more qualitative techniques (Stohl & Redding, 1987). Nevertheless, an assortment of content analyses in the organizational context have used a variety of techniques.

Organizational applications of content analysis have included the analysis of open-ended responses to employee surveys (DiSanza & Bullis, 1999), the word network analysis of voicemail (Rice & Danowski, 1991), the use of CATA to analyze levels of narcissism among CEOs of Fortune 100 corporations (Spangler et al., 2012), and the application of interpersonal interaction coding to manager–subordinate control patterns (Fairhurst et al., 1987). Developing a novel coding scheme, Larey and Paulus (1999) analyzed the

transcripts of brainstorming discussion groups of four individuals looking for unique ideas. They found that interactive groups were less successful in generating unique ideas than were “nominal,” noninteractive groups. Increasingly, content analysis has been used to identify patterns of communication from the organization to various publics or constituencies (e.g., Bravo et al., 2013), but these messages are more properly thought of as mass, rather than organizational, in nature.

### *Mass Messaging*

Mass messaging is the creation of messages that are *intended* for a relatively *large, undifferentiated audience*. These messages are most commonly mediated (e.g., via television, newspaper, radio, online), but they do not necessarily have to be, as in the case of a public speech.

Mass messages have been heavily studied by sociologists, social psychologists, communication scientists, marketing and advertising scholars, and others. Fully 34.8% of the mass communication articles published during 1995 in *Journalism & Mass Communication Quarterly*, one of the most prominent mass communication journals, were content analyses (Riffe & Freitag, 1997). The range of types of investigations is staggering, although some areas of study are much better represented in the content analysis literature than others; for instance, studies of journalistic coverage are common, whereas studies of films are relatively rare.

### *Applied Contexts*

In addition to the aforementioned means of dividing up message contexts, we might also consider such applied contexts as health communication, political communication, and social media, all of which transcend the distinctions of interpersonal, group, organizational, and mass communication. That is, content analyses within the health context might include analyses of doctor–patient interaction (interpersonal), the flow of email among hospital employees (organizational), and images of medical professionals on television (mass; Berlin Ray & Donohew, 1990). Yet all these varied studies would be informed by a clear grasp of the norms, values, behaviors, legal constraints, and business practices within the health care environment. Thus, special consideration of such applied contexts is useful. A number of these are presented in Chapter 9.

Some applications of content analysis may be highly practical. Rather than attempting to answer questions of theoretical importance, some analyses are aimed at building predictive power within a certain message arena. Box 1.2 highlights one such study. Rogan and Hammer (1995) applied a scheme to actual crisis negotiation incidents, such as hostage taking. Their findings offer insight into message patterns that may predict successful and unsuccessful resolutions to crisis incidents.

Another applied context is that of religious television. Box 1.3 describes a set of studies that took into consideration the special nature of religion on television during a time of critical discourse. A variety of communication and religious perspectives informed the analyses, ranging from interpersonal communication theories to practical considerations of religious mass media.

## 6. All Message Characteristics Are Available to Content Analyze

This book takes a broad view of what types of messages and message characteristics may be analyzed. A few clarifications on terminology are in order:

### *The Use of the Term “Content”*

As Smith (2000) points out, “The term ‘content’ in content analysis is something of a misnomer because verbal materials may be examined for content, for form (e.g., style, structure), function, or sequence of communications” (p. 314). Similarly, Morgan and Shanahan (2010, p. 351) note that the terminology “message system analysis,” used by scholar George Gerbner in the 1960s, was more inclusive than the term *content analysis*—“Gerbner specifically meant to point out that the entirety of a message system is what matters.” Thus, we should take a liberal view of the term *content* in “content analysis,” extending it to all message characteristics.

### *Manifest Versus Latent Content*

Early content analyses tended to concentrate on *manifest content*, the “elements that are physically present and countable” (Gray & Densten, 1998, p. 420). An alternative is to also consider the *latent content*, consisting of unobserved concept(s) that “cannot be measured directly but can be represented or measured by one or more . . . indicators” (Hair et al., 2010, p. 614). These two types of content might be seen as analogous to “surface” and “deep” structures of language and have their roots in Freud’s interpretations of dreams.<sup>6</sup> Other scholarship has compared manifest content to denotative meanings and latent content to connotative meanings (Ahuvia, 2001; Berelson, 1952; Eco, 1976).

Although the early definition of content analysis by Berelson (1952) indicated that it is ordinarily limited to manifest content only, many have attempted to measure the more subtle aspects of message meaning. As Ahuvia (2001, p. 141) notes, manifest and latent measures look at different aspects of the message. Manifest analysis examines obvious and straightforward aspects (e.g., Does the ad claim that the car has greater than 100 horsepower?), while latent analysis examines the subtler aspects (e.g., Does the ad position the car as powerful?).

Content analyses commonly include measures of manifest characteristics of messages, such as many of Baruh's (2009) measures applied to reality TV programming—whether a scene's setting was public or private, whether partial or full nudity was shown, and whether personal financial information was disclosed, for example.

The measurement of latent constructs is typically more problematic. At least two different approaches have been used. The first is a direct attempt to measure a latent characteristic via coder assessment. For example, Perrin's (2005) analysis of letters to the editor of major U.S. newspapers focused on an assessment of the degree of authoritarianism and antiauthoritarianism in the writings, finding an increase in both following the 9/11 attacks.

A second approach to the measurement of latent constructs in content analysis is to use multiple measures (often ones that are quite manifest) in concert, much as in survey and experimental research, standard self-report scales measuring global latent constructs (e.g., state depression) are comprised of multiple specific items. For example, in the Smith (1999) study, the latent construct, "sexism," was measured by 27 manifest variables that tapped "stereotypic images of women," extracted from a variety of theoretic works (largely from feminist literature) and critical, qualitative analyses of film (e.g., Haskell, 1987).

In the case of Ghose and Dou's (1998) study of Internet web sites, the latent variable, "interactivity," was represented by 23 manifest variables that are easily measurable, such as presence or absence of a key word search, electronic couponing, online contests, and downloading of software. Kinney (2006) used principal components analysis to group 35 manifest measures of word usage in news articles covering the charitable choice policy innovation aspect of the 1996 welfare law. The discovered "latent" themes were further interpreted by an independent panel of scholars. And, Van Gorp (2005, 2007) has approached content analysis of news coverage from the perspective that the construct of news "framing" can be considered a "latent message from the journalist" and that "sequences of manifest variables can represent" this latent construct (2005, pp. 487–488).

Scholars have empirically identified a tendency toward unreliability of human coding associated with measures of latent constructs (vs. manifest constructs; Carlyle, Slater, & Chakroff, 2008; Manganello et al., 2010), and some have questioned whether quantitative content analysis can even properly measure latent constructs (e.g., Ahuvia, 2001). In fact, early work by Berelson (1952) suggested that the focus of quantitative content analysis is manifest meaning, while qualitative content analysis is necessarily focused on latent meaning, a distinction that is further supported by Schreier (2012).

Potter and Levine-Donnerstein (1999) have delineated between two types of latent content—"pattern" and "projective." Pattern content "focuses on patterns in the content itself," while projective content "shifts the focus more onto coders' interpretations of the meaning of the content" (p. 259). An example of the former would be mode of dress of a female political candidate

(e.g., formal suit, soft feminine suit, dress, casual), which would be established by a coder examining combinations, or patterns, of types of clothing. An example of the latter would be the candidate's rhetorical style (e.g., exhortive, bureaucratic, emotional, informative), which would require the coder to access his or her own preexisting mental schema in order to make a judgment. According to Potter and Levine-Donnerstein, both types of latent content rely on "content cues and coder schema"—the distinction is which of the two is emphasized.

Gray and Densten (1998) promote the use of latent constructs as a way of integrating quantitative content analysis and qualitative message analysis. They used both methods to study *locus of control*, the broad latent concept from Rotter's internal/external locus of control construct: An individual holding a more external locus of control feels that his or her life events are the product of circumstances beyond his or her personal control (p. 426). Their findings indicate a surprising correspondence between quantitative and qualitative methods in the discovery of new locus-of-control dimensions reflected in a variety of very specific manifest indicators.

A number of researchers have criticized any dependence on the manifest-latent dichotomy, noting the often fuzzy distinction between the two (Potter & Levine-Donnerstein, 1999; Riffe, Lacy, & Fico, 2014; Shapiro & Markoff, 1997). It is perhaps more useful to think of a *continuum* from "highly manifest" to "highly latent" and to address issues of subtlety of measurement for those message aspects that are very latent and therefore a challenge in achieving objective and reliable measurement.

### ***Content/Substance Versus Form Characteristics***

Many scholars have differentiated between content and form elements of a message (Berelson, 1952; Huston & Wright, 1983; Naccarato & Neuendorf, 1998) or work of art (Tolhurst, 1985). Content attributes—sometimes more appropriately called *substance characteristics*—are those that may appear or exist in any medium. They are generally able to survive the translation from medium to medium. Form attributes—often called *formal features*, although there's usually nothing formal about them in the colloquial sense—are those that are relevant to the medium through which the message is sent. They are in a sense contributed by the particular medium or form of communication.

For example, the examination of self-disclosure by women to other women has been analyzed for movie characters (Capwell, 1997). The same measures of level and type of self-disclosure could be used to analyze naturally occurring discussions between real women, interactions between characters on TV programs or commercials, or relationship building between characters in novels. The measures are *content/substance* measures, applicable regardless of the medium. On the other hand, measurements of the type of camera shot (e.g., close-up vs. long shot) used when self-disclosure occurs in a film is a measure of *form*, or how the content is treated in a particular medium.

Even though the distinction between substance and form is an important one, the primary focus should not be on placing each variable in one category or the other. Some variables may be on the fine line between the two types, exhibiting characteristics of each. What's important is that both substance and form characteristics of messages ought to be considered for *every* content analysis conducted. Form characteristics are often extremely important mediators of the content elements. Huston and Wright (1983) have summarized how formal features of TV influence the cognitive processing of TV content, notably for children. This speaks once again to the importance of the content analyst becoming well versed in the norms and syntax of any medium he or she chooses to study.

### ***Text Analysis Versus Other Types of Content Analysis***

You'll notice that some of the classic definitions of content analysis shown in Box 1.1 apply the term *only* to analyses of text (i.e., written or transcribed words). The view presented in this book is not so limiting. Content analysis may be conducted on written text, transcribed speech, verbal interactions, visual images, characterizations, nonverbal behaviors, sound events, or any other message type. In this book, the term *content analysis* encompasses all such studies; the terms *text analysis* or *text content analysis* refer to the specific type of content analysis that focuses on written or transcribed words. Historically, content analyses did begin with examinations of written text. And text analysis remains a vibrant part of content analysis research, both human-coded analyses and increasingly popular computer-aided text analyses (Roberts, 1997b; Gottschalk & Bechtel, 2008). Those seeking more information on the historical trends in content analysis that saw expansion beyond the written word are advised to read "Milestones in Content Analysis History" at the CAGO.

## **Notes for Chapter 1**

---

1. It should be noted that terminologies for content analyses have become increasingly fluid. For example, the term *sentiment analysis*, a special form of computer-aided text analysis (Liu, 2010), appears for the first time in the PQD&T database in 2003. By 2011, it occurs regularly, but unfortunately is undetectable in searches for "content analysis," or even "text analysis."
2. Additionally, a search of Google Scholar revealed exponential growth in online articles that include the term *content analysis*, with an increase from approximately 6,000 sources dated 1997 to over 97,000 citations for 2015.
3. Although they note this as a limitation, Gottschall et al. (2008) made the choice to lump adjectives denoting *attractive* and *unattractive* together, forming an overall measure of "attractiveness references." They contend that this may actually understate the disproportion of female–male attractiveness emphasis. "When

these attributes are separated, 15% of male ‘hits’ are for adjectives associated with *unattractiveness*, compared with just 5% of female ‘hits’” (p. 184).

4. According to Gunter (2000), the “overriding objective” of the positivism paradigm is to “prove or disprove hypotheses and ultimately to establish universal laws of behaviour through the use of numerically defined and quantifiable measures analogous to those used by the natural sciences” (p. 4).
5. There is a difference between *triangulation*, which refers to the testing of the same hypotheses or research questions with different methodologies, and *mixed method* approaches, in which different research hypotheses or questions within a study are addressed using different methodologies.
6. According to Gregory (1987), “Freud’s approach to the interpretation of dreams was by way of the method of free association [from which Freud’s psychoanalysis procedures would evolve].... As in psychoanalysis proper, the subject is required to relax and allow his mind to wander freely from elements in the dream to related ideas, recollections, or emotional reactions which they may chance to suggest” (p. 274). The dream as reported was termed the *manifest content* by Freud, and the dream’s underlying thoughts and wishes Freud called the *latent content*.

# Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data

KENNETH BENOIT *London School of Economics and Trinity College*

DREW CONWAY *New York University*

BENJAMIN E. LAUDERDALE *London School of Economics and Political Science*

MICHAEL LAVER *New York University*

SLAVA MIKHAYLOV *University College London*

**E**mpirical social science often relies on data that are not observed in the field, but are transformed into quantitative variables by expert researchers who analyze and interpret qualitative raw sources. While generally considered the most valid way to produce data, this expert-driven process is inherently difficult to replicate or to assess on grounds of reliability. Using crowd-sourcing to distribute text for reading and interpretation by massive numbers of nonexperts, we generate results comparable to those using experts to read and interpret the same texts, but do so far more quickly and flexibly. Crucially, the data we collect can be reproduced and extended transparently, making crowd-sourced datasets intrinsically reproducible. This focuses researchers' attention on the fundamental scientific objective of specifying reliable and replicable methods for collecting the data needed, rather than on the content of any particular dataset. We also show that our approach works straightforwardly with different types of political text, written in different languages. While findings reported here concern text analysis, they have far-reaching implications for expert-generated data in the social sciences.

Political scientists have made great strides toward greater reproducibility of their findings since the publication of Gary King's influential article *Replication, Replication* (King 1995). It is now standard practice for good professional journals to insist that authors lodge their data and code in a prominent open access repository. This allows other scholars to replicate and extend published results by reanalyzing the data, rerunning and modifying the code. Replication of an *analysis*, however, sets a far weaker standard than reproducibility of the *data*, which is typically seen as a fundamental principle of the scientific method. Here, we propose a step towards a more comprehensive scientific replication standard in which the mandate is to replicate data production, not just data analysis. This shifts attention from specific datasets as the essential scientific objects of interest, to the *published and reproducible method by which the data were generated*.

We implement this more comprehensive replication standard for the rapidly expanding project of analyzing

the content of political texts. Traditionally, a lot of political data are generated by experts applying comprehensive classification schemes to raw sources in a process that, while in principle repeatable, is in practice too costly and time-consuming to reproduce. Widely used examples include<sup>1</sup> the *Polity* dataset, rating countries on a scale "ranging from -10 (hereditary monarchy) to +10 (consolidated democracy)"<sup>2</sup>; the *Comparative Parliamentary Democracy* data with indicators of the "number of inconclusive bargaining rounds" in government formation and "conflictual" government terminations<sup>3</sup>; the *Comparative Manifesto Project* (CMP), with coded summaries of party manifestos, notably a widely used left-right score<sup>4</sup>; and the *Policy Agendas Project*, which codes text from laws, court decisions, political speeches into topics and subtopics (Jones and Baumgartner 2013). In addition to the issue of reproducibility, the fixed nature of these schemes and the considerable infrastructure required to implement them discourages change and makes it harder to adapt them to specific needs, as the data are designed to fit general requirements rather than a particular research question.

Here, we demonstrate a method of crowd-sourced text annotation for generating political data that are both *reproducible* in the sense of allowing the data generating process to be quickly, inexpensively, and reliably repeated, and *agile* in the sense of being capable of flexible design according to the needs of a

<sup>1</sup> Other examples of coded data include expert judgments on party policy positions of party positions (Benoit and Laver 2006; Hooghe et al. 2010; Laver and Hunt 1992), and democracy scores from *Freedom House* and corruption rankings from *Transparency International*.

<sup>2</sup> <http://www.systemicpeace.org/polity/polity4.htm>

<sup>3</sup> [http://www.erdda.se/cpd/data\\_archive.html](http://www.erdda.se/cpd/data_archive.html)

<sup>4</sup> <https://manifesto-project.wzb.eu/>

specific research project. The notion of agile research is borrowed from recent approaches to software development, and incorporates not only the flexibility of design, but also the ability to iteratively test, deploy, verify, and, if necessary, redesign data generation through feedback in the production process. In what follows, we apply this method to a common measurement problem in political science: locating political parties on policy dimensions using text as data. Despite the lower expertise of crowd workers compared to experts, we show that properly deployed crowd-sourcing generates results indistinguishable from expert approaches. Given the millions of available workers online, crowd-sourced data collection can also be *repeated* as often as desired, quickly and with low cost. Furthermore, our approach is easily tailored to specific research needs, for specific contexts and time periods, in sharp contrast to large “canonical” data generation projects aimed at maximizing generality. For this reason, crowd-sourced data generation may represent a paradigm shift for data production and reproducibility in the social sciences. While, as a proof of concept, we apply our particular method for crowd-sourced data production to the analysis of political texts, the core problem of specifying a *reproducible data production process* extends to all subfields of political science.

In what follows, we first review the theory and practice of crowd-sourcing. We then deploy an experiment in content analysis designed to evaluate crowd-sourcing as a method for reliably and validly extracting meaning from political texts, in this case party manifestos. We compare expert and crowd-sourced analyses of the same texts, and assess external validity by comparing crowd-sourced estimates with those generated by completely independent expert surveys. In order to do this, we design a method for aggregating judgments about text units of varying complexity, by readers of varying quality,<sup>5</sup> into estimates of latent quantities of interest. To assess the external validity of our results, our core analysis uses crowd workers to estimate party positions on two widely used policy dimensions: “economic” policy (right-left) and “social” policy (liberal-conservative). We then use our method to generate “custom” data on a variable not available in canonical datasets, in this case party policies on immigration. Finally, to illustrate the general applicability of crowd-sourced text annotation in political science, we test the method in a multilingual and technical environment to show that crowd-sourced text analysis is effective for texts other than party manifestos and works well in different languages.

## HARVESTING THE WISDOM OF CROWDS

The intuition behind crowd-sourcing can be traced to Aristotle (Lyon and Pacuit 2013) and later Galton (1907), who noticed that the average of a large number of individual judgments by fair-goers of the weight

of an ox is close to the true answer and, importantly, closer to this than the typical individual judgment (for a general introduction see Surowiecki 2004). Crowd-sourcing is now understood to mean using the Internet to distribute a large package of small tasks to a large number of anonymous workers, located around the world and offered small financial rewards per task. The method is widely used for data-processing tasks such as image classification, video annotation, data entry, optical character recognition, translation, recommendation, and proofreading. Crowd-sourcing has emerged as a paradigm for applying human intelligence to problem-solving on a massive scale, especially for problems involving the nuances of language or other interpretative tasks where humans excel but machines perform poorly.

Increasingly, crowd-sourcing has also become a tool for social scientific research (Bohannon 2011). In sharp contrast to our own approach, most applications use crowds as a cheap alternative to traditional subjects for experimental studies (e.g., Horton et al. 2011; Lawson et al. 2010; Mason and Suri 2012; Paolacci et al. 2010). Using subjects in the crowd to populate experimental or survey panels raises obvious questions about external validity, addressed by studies in political science (Berinsky et al. 2012), economics (Horton et al. 2011) and general decision theory and behavior (Chandler et al. 2014; Goodman et al. 2013; Paolacci et al. 2010). Our method for using workers in the crowd to label *external* stimuli differs fundamentally from such applications. We do not care at all about whether our crowd workers represent any target population, as long as different workers, on average, make the same judgments when faced with the same information. In this sense our method, unlike online experiments and surveys, is a canonical use of crowd-sourcing as described by Galton.<sup>6</sup>

All data production by humans requires expertise, and several empirical studies have found that data created by domain experts can be matched, and sometimes improved at much lower cost, by aggregating judgments of nonexperts (Alonso and Baeza-Yates 2011; Alonso and Mizzaro 2009; Carpenter 2008; Hsueh et al. 2009; Ipeirotis et al. 2013; Snow et al. 2008). Provided crowd workers are not systematically biased in relation to the “true” value of the latent quantity of interest, and it is important to check for such bias, the central tendency of even erratic workers will converge on this true value as the number of workers increases. Because experts are axiomatically in short supply while members of the crowd are not, crowd-sourced solutions also offer a straightforward and *scalable* way to address reliability in a manner that expert solutions cannot. To improve confidence, simply employ more crowd workers. Because data production is broken down into many simple specific tasks, each performed by many different exchangeable workers, it tends to wash out biases that might affect a single worker, while also making it

<sup>5</sup> In what follows we use the term “reader” to cover a person, whether expert, crowd worker, or anyone else, who is evaluating a text unit for meaning.

<sup>6</sup> We are interested in the weight of the ox, not in how different people judge the weight of the ox.

possible to estimate and correct for worker-specific effects using the type of scaling model we employ below.

Crowd-sourced data generation inherently requires a method for aggregating many small pieces of information into valid measures of our quantities of interest.<sup>7</sup> Complex calibration models have been used to correct for worker errors on particular difficult tasks, but the most important lesson from this work is that increasing the number of workers reduces error (Snow et al. 2008). Addressing statistical issues of “redundant” coding, Sheng et al. (2008) and Ipeirotis et al. (2014) show that repeated coding can improve the quality of data as a function of the individual qualities and number of workers, particularly when workers are imperfect and labeling categories are “noisy.” Ideally, we would benchmark crowd workers against a “gold standard,” but such benchmarks are not always available, so scholars have turned to Bayesian scaling models borrowed from item-response theory (IRT), to aggregate information while simultaneously assessing worker quality (e.g., Carpenter 2008; Raykar et al. 2010). Welinder and Perona (2010) develop a classifier that integrates data difficulty and worker characteristics, while Welinder et al. (2010) develop a unifying model of the characteristics of both data and workers, such as competence, expertise, and bias. A similar approach is applied to rater evaluation in Cao et al. (2010) where, using a Bayesian hierarchical model, raters’ judgments are modeled as a function of a latent item trait, and rater characteristics such as bias, discrimination, and measurement error. We build on this work below, applying both a simple averaging method and a Bayesian scaling model that estimates latent policy positions while generating diagnostics on worker quality and sentence difficulty. We find that estimates generated by our more complex model match simple averaging very closely.

## A METHOD FOR REPLICABLE CODING OF POLITICAL TEXT

We apply our crowd-sourcing method to one of the most wide-ranging research programs in political science, the analysis of political text, and in particular text processing by *human* analysts that is designed to extract meaning systematically from some text corpus, and from this to generate valid and reliable data. This is related to, but quite distinct from, spectacular recent advances in *automated* text analysis that in theory scale up to unlimited volumes of political text (Grimmer and Stewart 2013). Many automated methods involve *supervised machine learning* and depend on labeled training data. Our method is directly relevant to this enterprise, offering a quick, effective, and, above all, *reproducible* way to generate labeled training data. Other, *unsupervised*, methods intrinsically require

*a posteriori* human interpretation that may be haphazard and is potentially biased.<sup>8</sup>

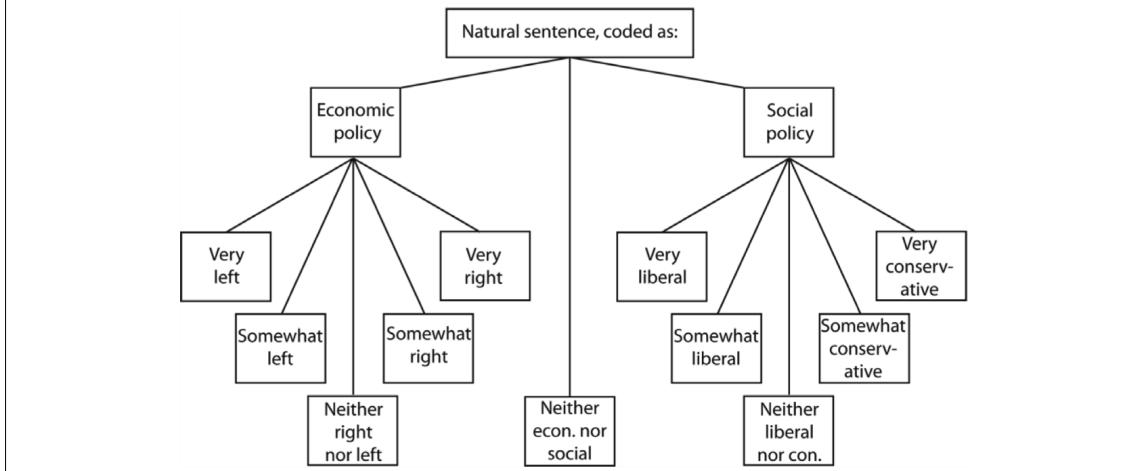
Our argument here speaks directly to more traditional content analysis within the social sciences, which is concerned with problems that automated text analysis cannot yet address. This involves the “reading” of text by real humans who interpret it for meaning. These interpretations, if systematic, may be classified and summarized using numbers, but the underlying human interpretation is fundamentally qualitative. Crudely, human analysts are employed to engage in *natural language processing* (NLP) which seeks to extract “meaning” embedded in the syntax of language, treating a text as more than a bag of words. NLP is another remarkable growth area, though it addresses a fundamentally difficult problem and fully automated NLP still has a long way to go. Traditional human experts in the field of inquiry are of course highly sophisticated natural language processors, finely tuned to particular contexts. The core problem is that they are in very short supply. This means that text processing by human experts simply does not scale to the huge volumes of text that are now available. This in turn generates an inherent difficulty in meeting the more comprehensive scientific replication standard to which we aspire. Crowd-sourced text analysis offers a compelling solution to this problem. Human workers in the crowd can be seen, perhaps rudely, as generic and very widely available “biological” natural language processors. Our task in this article is now clear. Design a system for employing generic workers in the crowd to analyze text for meaning in a way that is as reliable and valid as if we had used finely tuned experts to do the same job.

By far the best known research program in political science that relies on expert human readers is the long-running *Manifesto Project* (MP). This project has analyzed nearly 4,000 manifestos issued since 1945 by nearly 1,000 parties in more than 50 countries, using experts who are country specialists to label sentences in each text in their original languages. A single expert assigns every sentence in every manifesto to a single category in a 56-category scheme devised by the project in the mid-1980s (Budge et al. 1987; Budge et al. 2001; Klingemann et al. 1994; Klingemann et al. 2006; Laver and Budge 1992).<sup>9</sup> This has resulted in a widely used “canonical” dataset that, given the monumental coordinated effort of very many experts over 30 years, is unlikely ever to be recollected from scratch and in this sense is unlikely to be replicated. Despite low levels of interexpert reliability found in experiments using the MP’s coding scheme (Mikhaylov et al. 2012), a proposal to re-process the entire manifesto corpus many times, using many independent experts, is in practice a non-starter. Large canonical datasets such as this, therefore, tend not to satisfy the deeper standard of reproducible research that requires the transparent repeatability of data generation. This deeper replication standard can

<sup>7</sup> Of course aggregation issues are no less important when combining any multiple judgments, including those of experts. Procedures for aggregating nonexpert judgments may influence both the quality of data and convergence on some underlying “truth,” or trusted expert judgment. For an overview, see Quoc Viet Hung et al. (2013).

<sup>8</sup> This human interpretation can be reproduced by workers in the crowd, though this is not our focus in this article.

<sup>9</sup> <https://manifesto-project.wzb.eu/>

**FIGURE 1. Hierarchical Coding Scheme for Two Policy Domains with Ordinal Positioning**

however be satisfied with the crowd-sourced method we now describe.

### A simple coding scheme for economic and social policy

We assess the potential for crowd-sourced text analysis using an experiment in which we serve up an identical set of documents, and an identical set of text processing tasks, to both a small set of experts (political science faculty and graduate students) and a large and heterogeneous set of crowd workers located around the world. To do this, we need a simple scheme for labeling political text that can be used reliably by workers in the crowd. Our scheme first asks readers to classify each sentence in a document as referring to economic policy (left or right), to social policy (liberal or conservative), or to neither. Substantively, these two policy dimensions have been shown to offer an efficient representation of party positions in many countries.<sup>10</sup> They also correspond to dimensions covered by a series of expert surveys (Benoit and Laver 2006; Hooghe et al. 2010; Laver and Hunt 1992), allowing validation of estimates we derive against widely used independent estimates of the same quantities. If a sentence was classified as economic policy, we then ask readers to rate it on a five-point scale from very left to very right; those classified as social policy were rated on a five-point scale from liberal to conservative. Figure 1 shows this scheme.<sup>11</sup>

We did not use the MP's 56-category classification scheme, for two main reasons. The first is methodological: complexity of the MP scheme and uncertain boundaries between many of its categories were major

sources of unreliability when multiple experts applied this scheme to the same documents (Mikhaylov et al. 2012). The second is practical: it is impossible to write clear and precise instructions, to be understood reliably by a diverse, globally distributed, set of workers in the crowd, for using a detailed and complex 56-category scheme quintessentially designed for highly trained experts. This highlights an important trade-off. There may be data production tasks that cannot feasibly be explained in clear and simple terms, sophisticated instructions that can only be understood and implemented by highly trained experts. Sophisticated instructions are designed for a more limited pool of experts who can understand and implement them and, for this reason, imply less scalable and replicable data production. Such tasks may not be suitable for crowd-sourced data generation and may be more suited to traditional methods. The striking alternative now made available by crowd-sourcing is to break down complicated data production tasks into simple small jobs, as happens when complex consumer products are manufactured on factory production lines. Over and above the practical need to have simple instructions for crowd workers, furthermore, the scheme in Figure 1 is motivated by the observation that most scholars using manifesto data actually seek simple solutions, typically estimates of positions on a few general policy dimensions; they do not need estimates of these positions in a 56-dimensional space.

### Text corpus

While we extend this in work we discuss below, our baseline text corpus comprises 18,263 natural sentences from British Conservative, Labour and Liberal Democrat manifestos for the six general elections held between 1987 and 2010. These texts were chosen for two main reasons. First, for systematic external validation, there are diverse independent estimates of British

<sup>10</sup> See Chapter 5 of Benoit and Laver (2006) for an extensive empirical review of this for a wide range of contemporary democracies.

<sup>11</sup> Our instructions—fully detailed in the Online Appendix (Section 6)—were identical for both experts and nonexperts, defining the economic left-right and social liberal-conservative policy dimensions we estimate and providing examples of labeled sentences.

party positions for this period, from contemporary expert surveys (Benoit 2005, 2010; Laver 1998; Laver and Hunt 1992) as well as MP expert codings of the same texts. Second, there are well-documented substantive shifts in party positions during this period, notably the sharp shift of Labour towards the center between 1987 and 1997. The ability of crowd workers to pick up this move is a good test of external validity.

In designing the breakdown and presentation of the text processing tasks given to both experts and the crowd, we made a series of detailed operational decisions based on substantial testing and adaptation (reviewed in the Appendix). In summary, we used natural sentences as our fundamental text unit. Recognizing that most crowd workers dip into and out of our jobs and would not stay online to code entire documents, we served target sentences from the corpus in a random sequence, set in a two-sentence context on either side of the target sentence, without identifying the text from which the sentence was drawn. Our coding experiments showed that these decisions resulted in estimates that did not significantly differ from those generated by the classical approach of reading entire documents from beginning to end.

## SCALING DOCUMENT POLICY POSITIONS FROM CODED SENTENCES

Our aim is to estimate the policy positions of entire documents: not the code value of any single sentence, but some aggregation of these values into an estimate of each document's position on some meaningful policy scale while allowing for reader, sentence, and domain effects. One option is simple averaging: identify all economic scores assigned to sentences in a document by all readers, average these, and use this as an estimate of the economic policy position of a document. Mathematical and behavioral studies on aggregations of individual judgments imply that simpler methods often perform as well as more complicated ones, and often more robustly (e.g., Ariely et al. 2000; Clemen and Winkler 1999). Simple averaging of individual judgments is the benchmark when there is no additional information on the quality of individual coders (Armstrong 2001; Lyon and Pacuit 2013; Turner et al. 2014). However, this does not permit direct estimation of misclassification tendencies by readers who for example fail to identify economic or social policy "correctly," or of reader-specific effects in the use of positional scales.

An alternative is to model each sentence as containing information about the document, and then scale these using a measurement model. We propose a model based on item response theory (IRT), which accounts for both individual reader effects and the strong possibility that some sentences are intrinsically harder to interpret. This approach has antecedents in psychometric methods (e.g., Baker and Kim 2004; Fox 2010; Hambleton et al. 1991; Lord 1980), and has been used to aggregate crowd ratings (e.g., Ipeirotis et al. 2014;

Welinder et al. 2010; Welinder and Perona 2010; Whitehill et al. 2009).

We model each sentence,  $j$ , as a vector of parameters,  $\theta_{jd}$ , which corresponds to sentence attributes on each of four latent dimensions,  $d$ . In our application, these dimensions are latent *domain propensity* of a sentence to be labeled economic (1) and social (2) versus none; latent *position* of the sentence on economic (3) and social (4) dimensions. Individual readers  $i$  have potential *biases* in each of these dimensions, manifested when classifying sentences as "economic" or "social" and when assigning positions on economic and social policy scales. Finally, readers have four *sensitivities*, corresponding to their relative responsiveness to changes in the latent sentence attributes in each dimension. Thus, the latent coding of sentence  $j$  by reader  $i$  on dimension  $d$  is

$$\mu_{ijd}^* = \chi_{id} (\theta_{jd} + \psi_{id})$$

where the  $\chi_{id}$  indicate relative *responsiveness* of readers to changes in latent sentence attributes  $\theta_{jd}$ , and the  $\psi_{id}$  indicate relative *biases* towards labeling sentences as economic or social ( $d = 1, 2$ ), and rating economic and social sentences as right rather than left ( $d = 3, 4$ ).

We cannot observe readers' behavior on these dimensions directly. We therefore model their responses to the choice of label between economic, social and "neither" domains using a multinomial logit given  $\mu_{ij1}^*$  and  $\mu_{ij2}^*$ . We model their choice of scale position as an ordinal logit depending on  $\mu_{ij3}^*$  if they label the sentence as economic and on  $\mu_{ij4}^*$  if they label the sentence as social.<sup>12</sup> This results in the following model for the 11 possible combinations of labels and scales that a reader can give a sentence:<sup>13</sup>

$$p(\text{none}) = \left( \frac{1}{1 + \exp(\mu_{ij1}^*) + \exp(\mu_{ij2}^*)} \right),$$

$$p(\text{econ; scale}) = \left( \frac{\exp(\mu_{ij1}^*)}{1 + \exp(\mu_{ij1}^*) + \exp(\mu_{ij2}^*)} \right)$$

$$\times (\text{logit}^{-1}(\xi_{\text{scale}} - \mu_{ij3}^*) - \text{logit}^{-1}(\xi_{\text{scale}-1} - \mu_{ij3}^*)),$$

<sup>12</sup> By treating these as independent, and using the logit, we are assuming independence between the choices and between the social and economic dimensions (IIA). It is not possible to identify a more general model that relaxes these assumptions without asking additional questions of readers.

<sup>13</sup> Each policy domain has five *scale* points, and the model assumes proportional odds of being in each higher scale category in response to the sentence's latent policy positions  $\theta_3$  and  $\theta_4$  and the coder's sensitivities to this association. The cutpoints  $\xi$  for ordinal scale responses are constrained to be symmetric around zero and to have the same cutoffs in both social and economic dimensions, so that the latent scales are directly comparable to one another and to the raw scales. Thus,  $\xi_2 = \infty$ ,  $\xi_1 = -\xi_{-2}$ ,  $\xi_0 = -\xi_{-1}$ , and  $\xi_{-3} = -\infty$ .

$$p(soc; scale) = \left( \frac{\exp(\mu_{ij2}^*)}{1 + \exp(\mu_{ij1}^*) + \exp(\mu_{ij2}^*)} \right) \times (\text{logit}^{-1}(\xi_{scale} - \mu_{ij4}^*) - \text{logit}^{-1}(\xi_{scale-1} - \mu_{ij4}^*)).$$

The primary quantities of interest are not sentence level attributes,  $\theta_{jd}$ , but rather aggregates of these for entire documents, represented by the  $\bar{\theta}_{k,d}$  for each document  $k$  on each dimension  $d$ . Where  $\epsilon_{jd}$  are distributed normally with mean zero and standard deviation  $\sigma_d$ , we model these latent sentence level attributes  $\theta_{jd}$  hierarchically in terms of corresponding latent document level attributes:

$$\theta_{jd} = \bar{\theta}_{k(j),d} + \epsilon_{jd}.$$

As at the sentence level, two of these ( $d = 1, 2$ ) correspond to the overall frequency (importance) of economic and social dimensions relative to other topics, and the remaining two ( $d = 3, 4$ ) correspond to aggregate left-right positions of documents on economic and social dimensions.

This model enables us to generate estimates of not only our quantities of interest for the document-level policy positions, but also a variety of reader- and sentence-level diagnostics concerning reader agreement and the “difficulty” of domain and positional coding for individual sentences. Simulating from the posterior also makes it straightforward to estimate Bayesian credible intervals indicating our uncertainty over document-level policy estimates.<sup>14</sup>

Posterior means of the document level  $\bar{\theta}_{kd}$  correlate very highly with those produced by the simple averaging methods discussed earlier: 0.95 and above, as we report below. It is therefore possible to use averaging methods to summarize results in a simple and intuitive way that is also invariant to shifts in mean document scores that might be generated by adding new documents to the coded corpus. The value of our scaling model is to estimate reader and sentence fixed effects, and correct for these if necessary. While this model is adapted to our particular classification scheme, it is general in the sense that nearly all attempts to measure policy in specific documents will combine domain classification with positional coding.

## BENCHMARKING A CROWD OF EXPERTS

Our core objective is to compare estimates generated by workers in the crowd with analogous estimates generated by experts. Since readers of all types will likely disagree over the meaning of particular sentences, an important benchmark for our comparison of expert and crowd-sourced text coding concerns levels of disagreement between experts. The first stage of our empirical work therefore employed multiple

<sup>14</sup> We estimate the model by MCMC using the JAGS software, and provide the code, convergence diagnostics, and other details of our estimations in Section 2 of the Online Appendix.

(four to six)<sup>15</sup> experts to independently code each of the 18,263 sentences in our 18-document text corpus, using the scheme described above. The entire corpus was processed twice by our experts. First, sentences were served in their natural sequence in each manifesto, to mimic classical expert content analysis. Second, about a year later, sentences were processed in random order, to mimic the system we use for serving sentences to crowd workers. Sentences were uploaded to a custom-built, web-based platform that displayed sentences in context and made it easy for experts to process a sentence with a few mouse clicks. In all, we harvested over 123,000 expert evaluations of manifesto sentences, about seven per sentence. Table 1 provides details of the 18 texts, with statistics on the overall and mean numbers of evaluations, for both stages of expert processing as well as the crowd processing we report below.

## External validity of expert evaluations

Figure 2 plots two sets of estimates of positions of the 18 manifestos on economic and social policy: one generated by experts processing sentences in natural sequence (x axis); the other generated by completely independent expert surveys (y axis).<sup>16</sup> Linear regression lines summarizing these plots show that expert text processing predicts independent survey measures very well for economic policy ( $R = 0.91$ ), somewhat less well for the noisier dimension of social policy ( $R = 0.81$ ). To test whether coding sentences in their natural sequence affected results, our experts also processed the entire text corpus taking sentences in random order. Comparing estimates from sequential and random-order sentence processing, we found almost identical results, with correlations of 0.98 between scales.<sup>17</sup> Moving from “classical” expert content analysis to having experts process sentences served at random from anonymized texts makes no substantive difference to point estimates of manifesto positions. This reinforces our decision to use the much more scalable random sentence sequencing in the crowd-sourcing method we specify.

## Internal reliability of expert coding

**Agreement between experts.** As might be expected, agreement between our experts was far from perfect. Table 2 classifies each of the 5,444 sentences in the 1987 and 1997 manifestos, all of which were processed by the same six experts. It shows how many experts agreed the sentence referred to economic, or social, policy. If experts are in perfect agreement on the policy content of each sentence, either all six label each sentence as

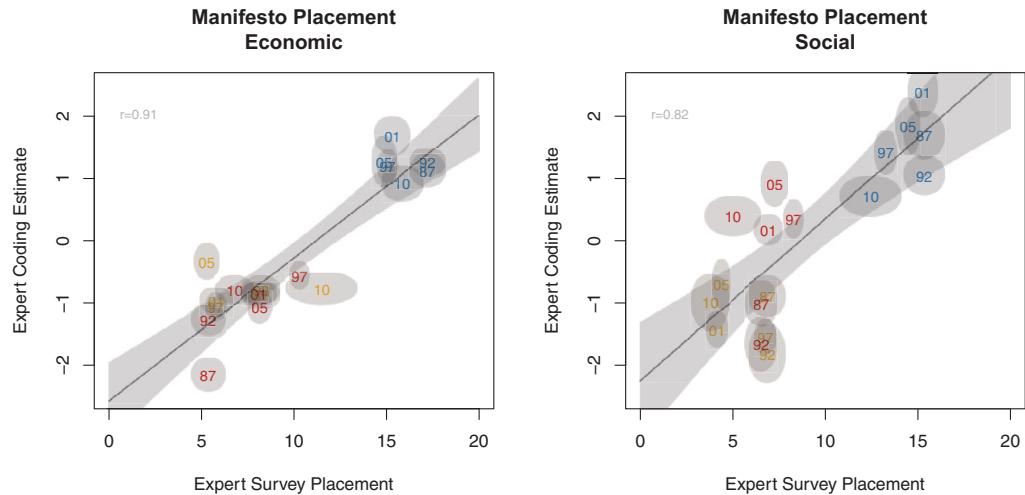
<sup>15</sup> Three of the authors of this article, plus three senior PhD students in Politics from New York University processed the six manifestos from 1987 and 1997. One author of this article and four NYU PhD students processed the other 12 manifestos.

<sup>16</sup> These were Laver and Hunt (1992); Laver (1998) for 1997; Benoit and Laver (2006) for 2001; Benoit (2005, 2010) for 2005 and 2010.

<sup>17</sup> Details provided in the Online Appendix, Section 5.

**TABLE 1. Texts and Sentences Coded: 18 British Party Manifestos**

Manifesto	Total Sentences in Manifesto	Mean Expert Evaluations: Natural Sequence	Mean Expert Evaluations: Random Sequence	Total Expert Evaluations	Mean Crowd Evaluations	Total Crowd Evaluations
Con 1987	1,015	6.0	2.4	7,920	44	36,594
LD 1987	878	6.0	2.3	6,795	22	24,842
Lab 1987	455	6.0	2.3	3,500	20	11,087
Con 1992	1,731	5.0	2.4	11,715	6	28,949
LD 1992	884	5.0	2.4	6,013	6	20,880
Lab 1992	661	5.0	2.3	4,449	6	23,328
Con 1997	1,171	6.0	2.3	9,107	20	11,136
LD 1997	873	6.0	2.4	6,847	20	5,627
Lab 1997	1,052	6.0	2.3	8,201	20	4,247
Con 2001	748	5.0	2.3	5,029	5	3,796
LD 2001	1,178	5.0	2.4	7,996	5	5,987
Lab 2001	1,752	5.0	2.4	11,861	5	8,856
Con 2005	414	5.0	2.3	2,793	5	2,128
LD 2005	821	4.1	2.3	4,841	5	4,173
Lab 2005	1,186	4.0	2.4	6,881	5	6,021
Con 2010	1,240	4.0	2.3	7,142	5	6,269
LD 2010	855	4.0	2.4	4,934	5	4,344
Lab 2010	1,349	4.0	2.3	7,768	5	6,843
Total	18,263	91,400	32,392	123,792		215,107

**FIGURE 2. British Party Positions on Economic and Social Policy 1987–2010**

Notes: Sequential expert text processing (vertical axis) and independent expert surveys (horizontal). Labour red, Conservatives blue, Liberal Democrats yellow, labeled by last two digits of year.

dealing with economic (or social) policy, or none do. The first data column of the table shows a total of 4,125 sentences which all experts agree have no social policy content. Of these, there are 1,193 sentences all experts also agree have no economic policy content, and 527 that all experts agree do have economic policy content. The experts disagree about the remaining 2,405 sen-

tences: some but not all experts label these as having economic policy content.

The shaded boxes show sentences for which the six experts were in unanimous agreement—on economic policy, social policy, or neither. There was unanimous expert agreement on about 35 percent of the labeled sentences. For about 65 percent of the sentences, there

**TABLE 2. Domain Classification Matrix for 1987 and 1997 Manifestos: Frequency with which Sentences were Assigned by Six Experts to Economic and Policy Domains**

Experts Assigning Economic Domain	Experts Assigning Social Policy Domain							Total
	0	1	2	3	4	5	6	
0	<b>1,193</b>	196	67	59	114	190	<b>170</b>	1,989
1	326	93	19	11	9	19	—	477
2	371	92	15	15	5	—	—	498
3	421	117	12	7	—	—	—	557
4	723	68	10	—	—	—	—	801
5	564	31	—	—	—	—	—	595
6	<b>527</b>	—	—	—	—	—	—	527
Total	4,125	597	123	92	128	209	170	5,444

Note: Shaded boxes: perfect agreement between experts.

**TABLE 3. Interexpert Scale Reliability Analysis for the Economic Policy, Generated by Aggregating All Expert Scores for Sentences Judged to have Economic Policy Content**

Item	N	Sign	Item-scale Correlation	Item-rest Correlation	Cronbach's Alpha
Expert 1	2,256	+	0.89	0.76	0.95
Expert 2	2,137	+	0.89	0.76	0.94
Expert 3	1,030	+	0.87	0.74	0.94
Expert 4	1,627	+	0.89	0.75	0.95
Expert 5	1,979	+	0.89	0.77	0.95
Expert 6	667	+	0.89	0.81	0.93
Overall k Policy Domain					0.95 0.93

was disagreement, even about the policy area, among trained experts of the type usually used to analyze political texts.

**Scale reliability.** Despite substantial disagreement among experts about individual sentences, we saw above that we can derive externally valid estimates of party policy positions if we aggregate the judgments of all experts on all sentences in a given document. This happens because, while each expert judgment on each sentence is a noisy realization of some underlying signal about policy content, the expert judgments taken as a whole scale nicely—in the sense that in aggregate they are all capturing information about the same underlying quantity. Table 3 shows this, reporting a scale and coding reliability analysis for economic policy positions of the 1987 and 1997 manifestos, derived by treating economic policy scores for each sentence allocated by each of the six expert coders as six sets of independent estimates of economic policy positions.

Despite the variance in expert coding of the policy domains as seen in Table 2, overall agreement as to the policy domain of sentences was 0.93 using Fleiss' kappa, a very high level of inter-rater agreement (as  $\kappa$

ranges from 0 to 1.0).<sup>18</sup> A far more important benchmark of reliability, however, focuses on the construction of the scale resulting from combining the coders' judgments, which is of more direct interest than the codes assigned to any particular fragment of text. *Scale* reliability, as measured by a Cronbach's alpha of 0.95, is "excellent" by any conventional standard.<sup>19</sup> We can therefore apply our model to aggregate the noisy information contained in the combined set of expert judgements at the sentence level to produce coherent estimates of policy positions at the document level. This is the essence of crowd-sourcing. It shows that our experts are really a small crowd.

<sup>18</sup> Expert agreement for the random order coding as to the precise scoring of positions within the policy domains had  $\kappa = 0.56$  for a polarity scale (left, neutral, right) and  $\kappa = 0.41$  for the full five-point scale. For position scoring agreement rates can be estimated only roughly, however, as sentences might have been assigned different policy domains by different raters, and therefore be placed using a different positional scale.

<sup>19</sup> Conventionally, an alpha of 0.70 is considered "acceptable." Nearly identical results for social policy are available in the Online Appendix (Section 1d). Note that we use Cronbach's alpha as a measure of scale reliability across readers, as opposed to a measure of inter-reader agreement (in which case we would have used Krippendorff's alpha).

## DEPLOYING CROWD-SOURCED TEXT CODING

### CrowdFlower: A crowd-sourcing platform with multiple channels

Many online platforms now distribute crowd-sourced microtasks (Human Intelligence Tasks or “HITs”) via the Internet. The best known is Amazon’s Mechanical Turk (MT), an online marketplace for serving HITs to workers in the crowd. Workers must often pass a pretask qualification test, and maintain a certain quality score from validated tasks that determines their status and qualification for future jobs. However, MT has for legal reasons become increasingly difficult to use for non-U.S. researchers and workers, with the result that a wide range of alternative crowd-sourcing channels has opened up. Rather than relying on a single crowd-sourcing channel, we used CrowdFlower, a service that consolidates access to dozens of channels.<sup>20</sup> CrowdFlower not only offers an interface for designing templates and uploading tasks that look the same on any channel but, crucially, also maintains a common training and qualification system for potential workers from any channel before they can qualify for tasks, as well as cross-channel quality control while tasks are being completed.

### Quality control

Excellent quality assurance is critical to all reliable and valid data production. Given the natural economic motivation of workers in the crowd to finish as many jobs in as short a time as possible, it is both tempting and easy for workers to submit bad or faked data. Workers who do this are called “spammers.” Given the open nature of the platform, it is vital to prevent them from participating in a job, using careful screening and quality control (e.g., Berinsky et al. 2014; Eickhoff and de Vries 2012; Kapelner and Chandler 2010; Nowak and Rger 2010.). Conway used coding experiments to assess three increasingly strict screening tests for workers in the crowd (Conway 2013).<sup>21</sup> Two findings directly inform our design. First, using a screening or qualification test *substantially* improves the quality of results; a well-designed test can screen out spammers and bad workers who otherwise tend to exploit the job. Second, once a suitable test is in place, increasing its difficulty *does not* improve results. It is vital to have a filter on the front end to keep out spammers and bad workers, but a tougher filter does not necessarily lead to better workers.

The primary quality control system used by CrowdFlower relies on completion of “gold” HITs: tasks with

unambiguous correct answers specified in advance.<sup>22</sup> Correctly performing “gold” tasks, which are both used in qualification tests and randomly sprinkled through the job, is used to monitor worker quality and block spammers and bad workers. We specified our own set of gold HITs as sentences for which there was unanimous expert agreement on both policy area (economic, social, or neither), and policy direction (left or right, liberal or conservative), and seeded each job with the recommended proportion of about 10% “gold” sentences. We therefore used “natural” gold sentences occurring in our text corpus, but could also have used “artificial” gold, manufactured to represent archetypical economic or social policy statements. We also used a special type of gold sentences called “screeners,” (Berinsky et al. 2014). These contained an exact instruction on how to label the sentence,<sup>23</sup> set in a natural two-sentence context, and are designed to ensure coders pay attention throughout the coding process.

Specifying gold sentences in this way, we implemented a two-stage process of quality control. First, workers were only allowed into the job if they correctly completed 8 out of 10 gold tasks in a qualification test.<sup>24</sup> Once workers are on the job and have seen at least four more gold sentences, they are given a “trust” score, which is simply the proportion of correctly labeled gold. If workers get too many gold HITs wrong, their trust level goes down. They are ejected from the job if their trust score falls below 0.8. The current trust score of a worker is recorded with each HIT, and can be used to weight the contribution of the relevant piece of information to some aggregate estimate. Our tests showed this weighting made no substantial difference, however, mainly because trust scores all tended to range in a tight interval around a mean of 0.84.<sup>25</sup> Many more potential HITs than we use here were rejected as “untrusted,” because the workers did not pass the qualification test, or because their trust score subsequently fell below the critical threshold. Workers are not paid for rejected HITs, giving them a strong incentive to perform tasks carefully, as they do not know which of these have been designated as gold for quality assurance. We have no hesitation in concluding that a system of thorough and continuous monitoring of worker quality is necessary for reliable and valid crowd sourced text analysis.

### Deployment

We set up an interface on CrowdFlower that was nearly identical to our custom-designed expert web

<sup>20</sup> See <http://www.crowdflower.com>.

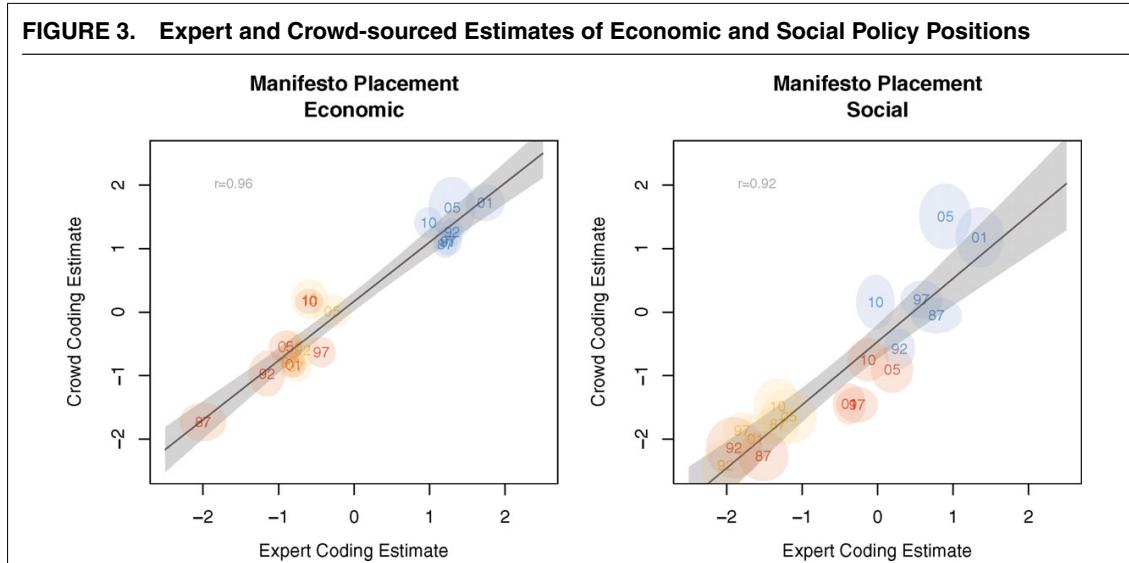
<sup>21</sup> There was a baseline test with no filter, a “low-threshold” filter where workers had to correctly code 4/6 sentences correctly, and a “high-threshold” filter that required 5/6 correct labels. A “correct” label means the sentence is labeled as having the same policy domain as that provided by a majority of expert coders. The intuition here is that tough tests also tend to scare away good workers.

<sup>22</sup> For CrowdFlower’s formal definition of gold, see <https://success.crowdflower.com/hc/en-us/articles/201855809-Guide-to-Test-Question-Data>.

<sup>23</sup> For example, “Please code this sentence as having economic policy content with a score of very right.”

<sup>24</sup> Workers giving wrong labels to gold questions are given a short explanation of why they are wrong.

<sup>25</sup> Our Online Appendix (Section 4) reports the distribution of trust scores from the complete set of crowd codings by country of the worker and channel, in addition to results that scale the manifesto aggregate policy scores by the trust scores of the workers.

**FIGURE 3. Expert and Crowd-sourced Estimates of Economic and Social Policy Positions**

system and deployed this in two stages. First, we oversampled all sentences in the 1987 and 1997 manifestos, because we wanted to determine the number of judgments per sentence needed to derive stable estimates of our quantities of interest. We served up sentences from the 1987 and 1997 manifestos until we obtained a minimum of 20 judgments per sentence. After analyzing the results to determine that our estimates of document scale positions converged on stable values once we had five judgments per sentence—in results we report below—we served the remaining manifestos until we reached five judgments per sentence. In all, we gathered 215,107 judgments by crowd workers of the 18,263 sentences in our 18 manifestos, employing a total of 1,488 different workers from 49 different countries. About 28 percent of these came from the United States, 15 percent from the United Kingdom, 11 percent from India, and 5 percent each from Spain, Estonia, and Germany. The average worker processed about 145 sentences; most processed between 10 and 70 sentences, 44 workers processed over 1,000 sentences, and four processed over 5,000.<sup>26</sup>

### CROWD-SOURCED ESTIMATES OF PARTY POLICY POSITIONS

**Figure 3** plots crowd-sourced estimates of the economic and social policy positions of British party manifestos against estimates generated from analogous ex-

pert text processing.<sup>27</sup> The very high correlations of aggregate policy measures generated by crowd workers and experts suggest both are measuring the same latent quantities. Substantively, **Figure 3** also shows that crowd workers identified the sharp rightwards shift of Labour between 1987 and 1997 on both economic and social policy, a shift identified by expert text processing and independent expert surveys. The standard errors of crowd-sourced estimates are higher for social than for economic policy, reflecting both the smaller number of manifesto sentences devoted to social policy and higher coder disagreement over the application of this policy domain.<sup>28</sup> Nonetheless **Figure 3** summarizes our evidence that the crowd-sourced estimates of party policy positions can be used as substitutes for the expert estimates, which is our main concern in this article.

Our scaling model provides a theoretically well-grounded way to aggregate all the information in our expert or crowd data, relating the underlying position of the political text both to the “difficulty” of a particular sentence and to a reader’s propensity to identify the correct policy domain, and position within domain.<sup>29</sup> Because positions derived from the scaling model depend on parameters estimated using the full set of coders and codings, changes to the text corpus can affect the relative scaling. The simple mean of means method, however, is invariant to rescaling and always produces the same results, even for a single document.

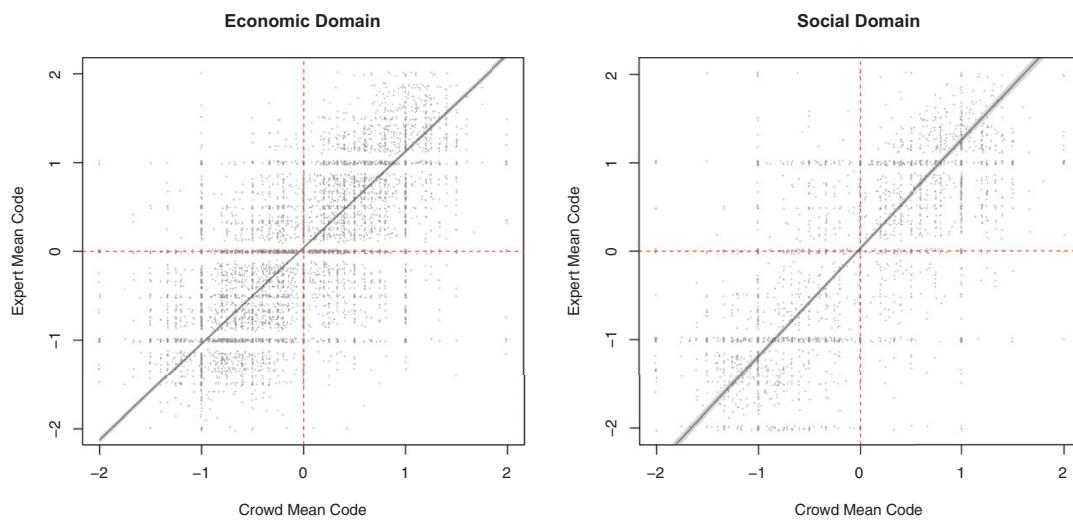
<sup>26</sup> Our final crowd-coded dataset was generated by deploying through a total of 26 CrowdFlower channels. The most common was Neodev (Neobux) (40%), followed by Mechanical Turk (18%), Bitcoinget (15%), Clixsense (13%), and Prodege (Swagbucks) (6%). Opening up multiple worker channels also avoided the restriction imposed by Mechanical Turk in 2013 to limit the labor pool to workers based in the United States and India. Full details along with the range of trust scores for coders from these platforms are presented in the Online Appendix (Section 4).

<sup>27</sup> Full point estimates are provided in the Online Appendix, Section 1.

<sup>28</sup> An alternative measure of correlation, Lin’s concordance correlation coefficient (Lin 1989, 2000), measures correspondence as well covariation, if our objective is to match the values on the identity line, although for many reasons here it is not. The economic and social measures for Lin’s coefficient are 0.95 and 0.84, respectively.

<sup>29</sup> We report more fully on diagnostic results for our coders on the basis of the auxiliary model quantity estimates in the Online Appendix (Section 1e).

**FIGURE 4. Expert and Crowd-sourced Estimates of Economic and Social Policy Codes of Individual Sentences, all Manifestos**



Note: Fitted line is the principal components or Deming regression line.

Comparing crowd-sourced estimates from the scaling model to those produced by a simple averaging of the mean of mean sentence scores, we find correlations of 0.96 for the economic and 0.97 for the social policy positions of the 18 manifestos. We present both methods as confirmation that our scaling method has not “manufactured” policy estimates. While this model does allow us to take proper account of reader and sentence fixed effects, it is also reassuring that a simple mean of means produced substantively similar estimates.

We have already seen that noisy expert judgments about sentences aggregate up to reliable and valid estimates for documents. Similarly, crowd-sourced document estimates reported in Figure 3 are derived from crowd-sourced sentence data that are full of noise. As we already argued, this is the essence of crowdsourcing. Figure 4 plots mean expert against mean crowd-sourced scores for each sentence. The scores are highly correlated, though crowd workers are substantially less likely to use extremes of the scales than experts. The first principal component and associated confidence intervals show a strong and significant statistical relationship between crowd sourced and expert assessments of individual manifesto sentences, with no evidence of systematic bias in the crowd-coded sentence scores.<sup>30</sup> Overall, despite the expected noise, our results show that crowd workers systematically tend to make the same judgments about individual sentences as experts.

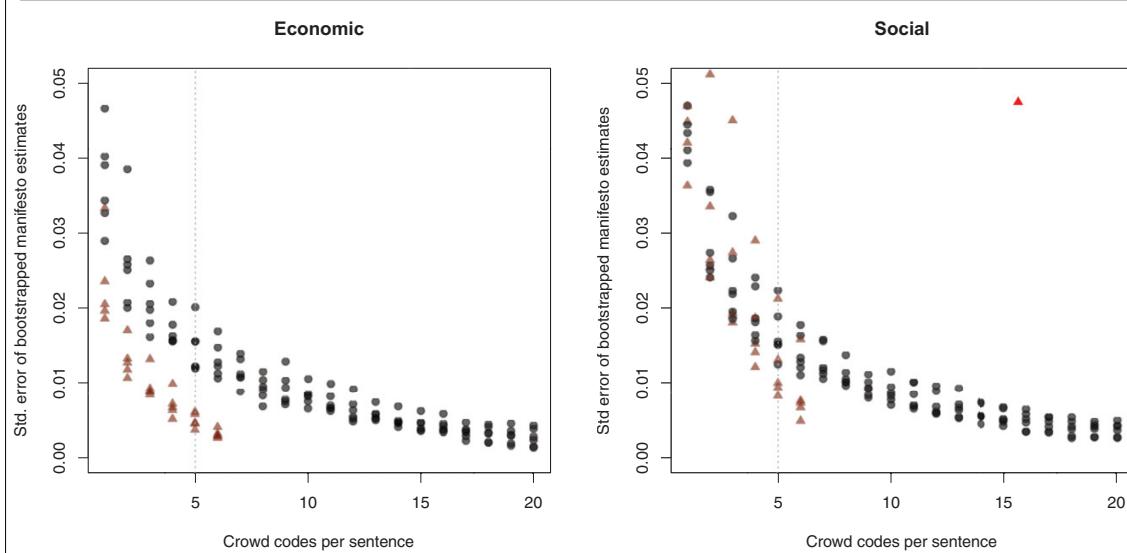
### Calibrating the number of crowd judgments per sentence

A key question for our method concerns *how many* noisier crowd-based judgments we need to generate reliable and valid estimates of fairly long documents such as party manifestos. To answer this, we turn to evidence from our oversampling of 1987 and 1997 manifestos. Recall we obtained a minimum of 20 crowd judgments for each sentence in each of these manifestos, allowing us to explore what our estimates of the position of each manifesto would have been, had we collected fewer judgments. Drawing random subsamples from our oversampled data, we can simulate the convergence of estimated document positions as a function of the number of crowd judgments per sentence. We did this by bootstrapping 100 sets of subsamples for each of the subsets of  $n = 1$  to  $n = 20$  workers, computing manifesto positions in each policy domain from aggregated sentence position means, and computing standard deviations of these manifesto positions across the 100 estimates. Figure 5 plots these for each manifesto as a function of the increasing number of crowd workers per sentence, where each point represents the empirical standard error of the estimates for a specific manifesto. For comparison, we plot the same quantities for the expert data in red.

The findings show a clear trend: uncertainty over the crowd-based estimates collapses as we increase the number of workers per sentence. Indeed, the only difference between experts and the crowd is that expert variance is smaller, as we would expect. Our findings vary somewhat with policy area, given the noisier character of social policy estimates, but adding additional

<sup>30</sup> Lack of bias is indicated by the fact that the fitted line crosses the origin.

**FIGURE 5. Standard Errors of Manifesto-level Policy Estimates as a Function of the Number of Workers, for the Oversampled 1987 and 1997 Manifestos**



*Note:* Each point is the bootstrapped standard deviation of the mean of means aggregate manifesto scores, computed from sentence-level random n subsamples from the codes.

crowd-sourced sentence judgments led to convergence with our expert panel of five to six coders at around 15 crowd coders. However, the steep decline in the uncertainty of our document estimates leveled out at around five crowd judgments per sentence, at which point the absolute level of error is already low for both policy domains. While increasing the number of unbiased crowd judgments will always give better estimates, we decided on cost-benefit grounds for the second stage of our deployment to continue coding in the crowd until we had obtained five crowd judgments per sentence. This may seem a surprisingly small number, but there are a number of important factors to bear in mind in this context. First, the manifestos comprise about 1000 sentences on average; our estimates of document positions aggregate codes for these. Second, sentences were randomly assigned to workers, so each sentence score can be seen as an independent estimate of the position of the manifesto on each dimension.<sup>31</sup> With five scores per sentence and about 1000 sentences per manifesto, we have about 5000 “little” estimates of the manifesto position, each a representative sample from the larger set of scores that would result from additional worker judgments about each sentence in each document. This sample is big enough to achieve a reasonable level of precision, given the large number of sentences per manifesto. While the *method* we use here could be used for much shorter documents, the *results* we infer here for the appropriate number of judgments per sentence

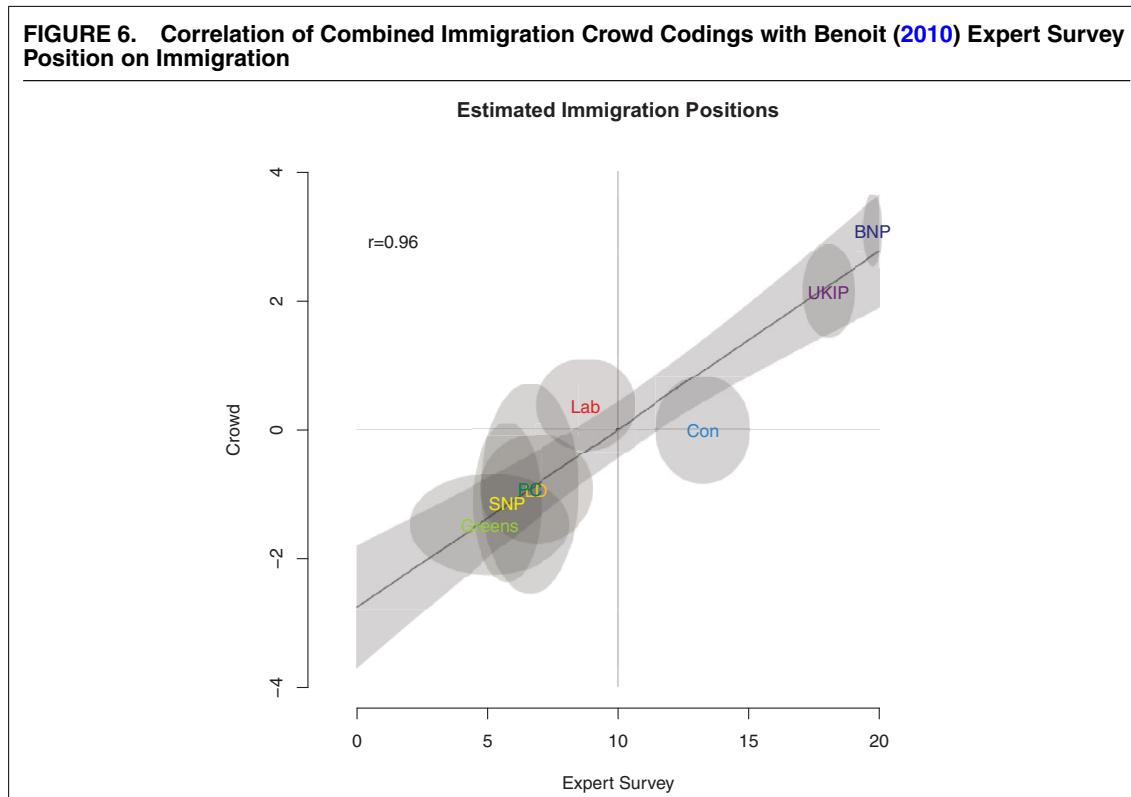
might well not apply, and would likely be higher. But, for large documents with many sentences, we find that the number of crowd judgments *per sentence* that we need is not high.

### CROWD-SOURCING DATA FOR SPECIFIC PROJECTS: IMMIGRATION POLICY

A key problem for scholars using “canonical” datasets, over and above the replication issues we discuss above, is that the data often do not measure what a modern researcher wants to measure. For example the widely used MP data, using a classification scheme designed in the 1980s, do not measure immigration policy, a core concern in the party politics of the 21st century (Ruedin 2013; Ruedin and Morales 2012). Crowd-sourcing data frees researchers from such “legacy” problems and allows them more flexibly to collect information on their precise quantities of interest. To demonstrate this, we designed a project tailored to measure British parties’ immigration policies during the 2010 election. We analyzed the manifestos of eight parties, including smaller parties with more extreme positions on immigration, such as the British National Party (BNP) and the UK Independence Party (UKIP). Workers were asked to label each sentence as referring to immigration policy or not. If a sentence did cover immigration, they were asked to rate it as pro- or anti-immigration, or neutral. We deployed a job with 7,070 manifesto sentences plus 136 “gold” questions and screeners devised specifically for this purpose. For this job, we used an adaptive sentence sampling strategy which set a minimum of

<sup>31</sup> Coding a sentence as referring to another dimension is a null estimate.

**FIGURE 6. Correlation of Combined Immigration Crowd Codings with Benoit (2010) Expert Survey Position on Immigration**



five crowd-sourced labels per sentence, unless the first three of these were unanimous in judging a sentence *not* to concern immigration policy. This is efficient when coding texts with only “sparse” references to the matter of interest; in this case most manifesto sentences (approximately 96%) were clearly not about immigration policy. Within just five hours, the job was completed, with 22,228 codings, for a total cost of \$360.<sup>32</sup>

We assess the external validity of our results using independent expert surveys by Benoit (2010) and the Chapel Hill Expert Survey (Bakker et al. 2015). Figure 6 compares the crowd-sourced estimates to those from expert surveys. The correlation with the Benoit (2010) estimates (shown) was 0.96, and 0.94 with independent expert survey estimates from the Chapel Hill survey.<sup>33</sup> To assess whether this data production exercise was as reproducible as we claim, we repeated the entire exercise with a second deployment two months after the first, with identical settings. This new job generated another 24,551 pieces of crowd-sourced data and was completed in just over three hours. The replication generated nearly identical estimates, detailed in Table 4, correlating at the same high

levels with external expert surveys, and correlating at 0.93 with party position estimates from the original crowd coding.<sup>34</sup> With just hours from deployment to dataset, and for very little cost, crowd sourcing enabled us to generate externally valid *and reproducible* data related to our precise research question.

### CROWD SOURCED TEXT ANALYSIS IN OTHER CONTEXTS AND LANGUAGES

As carefully designed official statements of a party’s policy stances, election manifestos tend to respond well to systematic text analysis. In addition, manifestos are written for popular consumption and tend to be easily understood by nontechnical readers. Much political information, however, can be found in texts generated from hearings, committee debates, or legislative speeches on issues that often refer to technical provisions, amendments, or other rules of procedure that might prove harder to analyze. Furthermore, a majority of the world’s political texts are not in English. Other widely studied political contexts, such as the European Union, are multilingual environments where researchers using automated methods designed for a single language must make hard choices. Schwarz et al. (forthcoming) applied unsupervised scaling methods to a multilingual debate in the Swiss parliament, for

<sup>32</sup> The job set 10 sentences per “task” and paid \$0.15 per task.

<sup>33</sup> CHES included two highly correlated measures, one aimed at “closed or open” immigration policy another aimed at policy toward asylum seekers and whether immigrants should be integrated into British society. Our measure averages the two. Full numerical results are given in the Online Appendix, Section 3.

<sup>34</sup> Full details are in the Online Appendix, Section 7.

**TABLE 4. Comparison Results for Replication of Immigration Policy Crowd Coding**

	Wave		
	Initial	Replication	Combined
Total Crowd Codings	24,674	24,551	49,225
Number of Coders	51	48	85
Total Sentences Coded as Immigration	280	264	283
Correlation with Benoit Expert Survey (2010)	0.96	0.94	0.96
Correlation with CHES 2010	0.94	0.91	0.94
Correlation of Results between Waves			0.93

instance, but had to ignore a substantial number of French and Italian speeches in order to focus on the majority German texts. In this section, we demonstrate that crowd-sourced text analysis, with appropriately translated instructions, offers the means to overcome these limitations by working in any language.

Our corpus comes from a debate in the European Parliament, a multilanguage setting where the EU officially translates every document into 24 languages. To test our method in a context very different from party manifestos, we chose a fairly technical debate concerning a Commission report proposing an extension to a regulation permitting state aid to uncompetitive coal mines. This debate concerned not only the specific proposal, involving a choice of letting the subsidies expire in 2011, permitting a limited continuation until 2014, or extending them until 2018 or even indefinitely.<sup>35</sup> It also served as debating platform for arguments supporting state aid to uncompetitive industries, versus the traditionally liberal preference for the free market over subsidies. Because a vote was taken at the end of the debate, we also have an objective measure of whether the speakers supported or objected to the continuation of state aid.

We downloaded all 36 speeches from this debate, originally delivered by speakers from 11 different countries in 10 different languages. Only one of these speakers, an MEP from the Netherlands, spoke in English, but all speeches were officially translated into each target language. After segmenting this debate into sentences, devising instructions and representative test sentences and translating these into each language, we deployed the same text analysis job in English, German, Spanish, Italian, Polish, and Greek, using crowd workers to read and label the same set of texts, but using the translation into their own language. *Figure 7* plots the score for each text against the eventual vote of the speaker. It shows that our crowd-sourced scores for each speech perfectly predict the voting behavior of each speaker, regardless of the language. In *Table 5*, we show correlations between our crowd-sourced estimates of the positions of the six different language versions of the same set of texts. The results are strik-

ing, with interlanguage correlations ranging between 0.92 and 0.96.<sup>36</sup> Our text measures from this technical debate produced reliable measures of the very specific dimension we sought to estimate, and the validity of these measures was demonstrated by their ability to predict the voting behavior of the speakers. Not only are these results straightforwardly reproducible, but this reproducibility is invariant to the language in which the speech was written. Crowd-sourced text analysis does not only work in English.

## CONCLUSIONS

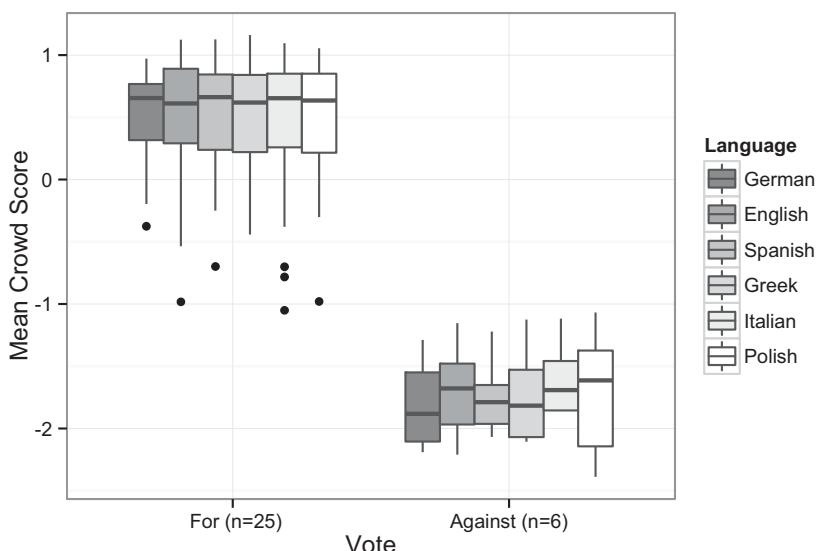
We have illustrated across a range of applications that crowd-sourced text analysis can produce valid political data of a quality indistinguishable from traditional expert methods. Unlike traditional methods, however, crowd-sourced data generation offers several advantages. Foremost among these is the possibility of meeting a replication standard far stronger than the current practice of facilitating reproducible *analysis*. By offering a published specification for feasibly replicating *the process of data generation*, the methods demonstrated here go much farther towards meeting a more stringent standard of *reproducibility* that is the hallmark of scientific inquiry. All of the data used in this article are of course available in a public archive for any reader to reanalyze at will. Crowd-sourcing our data allows us to do much more than this, however. Any reader can take our publicly available crowdsourcing code and deploy this code to *reproduce our data collection process and collect a completely new dataset*. This can be done many times over, by any researcher, anywhere in the world. This, to our minds, takes us significantly closer to a true scientific replication standard.

Another key advantage of crowd-sourced text analysis is that it can form part of an *agile* research process, precisely tailored to a specific research question rather than reflecting the grand compromise at the heart of the large canonical datasets so commonly deployed by political scientists. Because the crowd's resources can be tapped in a flexible fashion, text-based data on completely new questions of interest can be processed

<sup>35</sup> This was the debate from 23 November 2010, "State aid to facilitate the closure of uncompetitive coal mines." <http://bit.ly/EP-Coal-Aid-Debate>

<sup>36</sup> Lin's concordance coefficient has a similar range of values, from 0.90 to 0.95.

**FIGURE 7. Scored Speeches from a Debate over State Subsidies by Vote, from Separate Crowd-sourced Text Analysis in Six Languages**



Note: Aggregate scores are standardized for direct comparison.

**TABLE 5. Summary of Results from EP Debate Coding in Six Languages**

Language	Correlations of 35 Speaker Scores					
	English	German	Spanish	Italian	Greek	Polish
German	0.96	—	—	—	—	—
Spanish	0.94	0.95	—	—	—	—
Italian	0.92	0.94	0.92	—	—	—
Greek	0.95	0.97	0.95	0.92	—	—
Polish	0.96	0.95	0.94	0.94	0.93	—
Sentence N	414	455	418	349	454	437
Total Judgments	3,545	1,855	2,240	1,748	2,396	2,256
Cost	\$109.33	\$55.26	\$54.26	\$43.69	\$68.03	\$59.25
Elapsed Time (hrs)	1	3	3	7	2	1

only for the contexts, questions, and time periods required. Coupled with the rapid completion time of crowd-sourced tasks and their very low marginal cost, this opens the possibility of valid text processing to researchers with limited resources, especially graduate students. For those with more ambition or resources, its inherent *scalability* means that crowd-sourcing can tackle large projects as well. In our demonstrations, our method worked as well for hundreds of judgments as it did for hundreds of thousands.

Of course, retooling for any new technology involves climbing a learning curve. We spent considerable time pretesting instruction wordings, qualification tests, compensation schemes, gold questions, and a range of other detailed matters. Starting a new crowd-

sourcing project is by no means cost-free, though these costs are mainly denominated in learning time and effort spent by the researcher, rather than research dollars. Having paid the inevitable fixed start-up costs that apply to any rigorous new data collection project, whether or not this involves crowd-sourcing, the beauty of crowd-sourcing arises from two key features of the crowd. The pool of crowd workers is to all intents and purposes inexhaustible, giving crowd-sourcing projects a scalability and replicability unique among projects employing human workers. And the low *marginal* cost of adding more crowd workers to any given project puts ambitious high quality data generation projects in the realistic grasp of a wider range of researchers than ever before. We are still in the early days of crowd-sourced

data generation in the social sciences. Other scholars will doubtless find many ways to fortify the robustness and broaden the scope of the method. But, whatever these developments, we now have a new method for collecting political data that allows us to do things we could not do before.

## APPENDIX: METHODOLOGICAL DECISIONS ON SERVING POLITICAL TEXT TO WORKERS IN THE CROWD

*Text units: Natural sentences.* The MP specifies a “quasistence” as the fundamental text unit, defined as “an argument which is the verbal expression of one political idea or issue” (Volkens 2001). Recoding experiments by Däubler et al. (2012), however, show that using natural sentences makes no statistically significant difference to point estimates, but does eliminate significant sources of both unreliability and unnecessary work. Our dataset therefore consists of all natural sentences in the 18 UK party manifestos under investigation.<sup>37</sup>

*Text unit sequence: Random.* In “classical” expert text coding, experts process sentences in their natural sequence, starting at the beginning and ending at the end of a document. Most workers in the crowd, however, will never reach the end of a long policy document. Processing sentences in natural sequence, moreover, creates a situation in which one sentence coding may well affect priors for subsequent sentence codings, so that summary scores for particular documents are not aggregations of independent coder assessments.<sup>38</sup> An alternative is to randomly sample sentences from the text corpus for coding—with a fixed number of replacements per sentence across all coders—so that each coding is an independent estimate of the latent variable of interest. This has the big advantage in a crowdsourcing context of *scalability*. Jobs for individual coders can range from very small to very large; coders can pick up and put down coding tasks at will; every little piece of coding in the crowd contributes to the overall database of text codings. Accordingly our method for crowd-sourced text coding serves coders sentences randomly selected from the text corpus rather than in naturally occurring sequence. Our decision to do this was informed by coding experiments reported in the Online Appendix (Section 5), and confirmed by results reported above. Despite higher variance in individual sentence codings under random sequence coding, there is no systematic difference between point estimates of party policy positions depending on whether sentences were coded in natural or random sequence.

*Text authorship: Anonymous.* In classical expert coding, coders typically know the authorship of the document they are coding. Especially in the production of political data, coders likely bring nonzero priors to coding text units. Precisely the same sentence (“we must do all we can to make the

<sup>37</sup> Segmenting “natural” sentences, even in English, is never an exact science, but our rules matched those from Däubler et al. (2012), treating (for example) separate clauses of bullet pointed lists as separate sentences.

<sup>38</sup> Coded sentences do indeed tend to occur in “runs” of similar topics, and hence codes; however to ensure appropriate statistical aggregation it is preferable if the codings of those sentences are independent.

public sector more efficient”) may be coded in different ways if the coder knows this comes from a right- rather than a left-wing party. Codings are typically aggregated into document scores as if coders had zero priors, even though we do not know how much of the score given to some sentence is the coder’s judgment about the content of the sentence, and how much a judgment about its author. In coding experiments reported in the Online Appendix (Section 5), semiexpert coders coded the same manifesto sentences both knowing and not knowing the name of the author. We found slight systematic coding biases arising from knowing the identity of the document’s author. For example, we found coders tended to code precisely the same sentences from Conservative manifestos as more right wing, if they knew these sentences came from a Conservative manifesto. This informed our decision to withhold the name of the author of sentences deployed in crowd-sourcing text coding.

*Context units: +/- two sentences.* Classical content analysis has always involved coding an individual text unit in light of the text surrounding it. Often, it is this context that gives a sentence substantive meaning, for example because many sentences contain pronoun references to surrounding text. For these reasons, careful instructions for drawing on context have long formed part of coder instructions for content analysis (see Krippendorff 2013). For our coding scheme, on the basis of prerelease coding experiments, we situated each “target” sentence within a context of the two sentences on either side in the text. Coders were instructed to code target sentence not context, but to use context to resolve any ambiguity they might feel about the target sentence.

## SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit <http://dx.doi.org/10.1017/S0003055416000058>.

## REFERENCES

- Alonso, O., and R. Baeza-Yates. 2011. “Design and Implementation of Relevance Assessments Using Crowdsourcing.” In *Advances in Information Retrieval*, eds. P. Clough, C. Foley, C. Gurrin, G. Jones, W. Kraaij, H. Lee, and V. Mudo. Berlin: Springer.
- Alonso, O., and S. Mizzaro. 2009. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. Paper read at Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation.
- Ariely, D., W. T. Au, R. H. Bender, D. V. Budescu, C. B. Dietz, H. Gu, and G. Zauberman. 2000. “The effects of averaging subjective probability estimates between and within judges.” *Journal of Experimental Psychology: Applied* 6 (2): 130–47.
- Armstrong, J. S., ed. 2001. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. New York: Springer.
- Baker, Frank B., and Seock-Ho Kim. 2004. *Item Response Theory: Parameter Estimation Techniques*. Boca Raton: CRC Press.
- Bakker, Ryan, Catherine de Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen and Milada Vachudova. 2015. “Measuring Party Positions in Europe: The Chapel Hill Expert Survey Trend File, 1999–2010.” *Party Politics* 21 (1): 143–52.
- Benoit, Kenneth. 2005. “Policy Positions in Britain 2005: Results from an Expert Survey.” London School of Economics.

- Benoit, Kenneth. 2010. "Expert Survey of British Political Parties." Trinity College Dublin.
- Benoit, Kenneth, and Michael Laver. 2006. *Party Policy in Modern Democracies*. London: Routledge.
- Berinsky, A., G. Huber, and G. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20 (3): 351–68.
- Berinsky, A., M. Margolis, and M. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science*.
- Bohannon, J. 2011. "Social Science for Pennies." *Science* 334: 307.
- Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Eric Tannenbaum, Richard Fording, Derek Hearl, Hee Min Kim, Michael McDonald, and Silvia Mendes. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors and Governments 1945–1998*. Oxford: Oxford University Press.
- Budge, Ian, David Robertson, and Derek Hearl. 1987. *Ideology, Strategy and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*. Cambridge, UK: Cambridge University Press.
- Cao, J., S. Stokes, and S. Zhang. 2010. "A Bayesian Approach to Ranking and Rater Evaluation: An Application to Grant Reviews." *Journal of Educational and Behavioral Statistics* 35 (2): 194–214.
- Carpenter, B. 2008. "Multilevel Bayesian Models of Categorical Data Annotation." Unpublished manuscript.
- Chandler, Jesse, Pam Mueller, and Gabriel Paolacci. 2014. "Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers." *Behavior Research Methods* 46 (1): 112–30.
- Clemen, R., and R. Winkler. 1999. "Combining Probability Distributions From Experts in Risk Analysis." *Risk Analysis* 19 (2): 187–203.
- Conway, Drew. 2013. "Applications of Computational Methods in Political Science." Department of Politics, New York University.
- Däubler, Thomas, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2012. "Natural Sentences as Valid Units for Coded Political Text." *British Journal of Political Science* 42 (4): 937–51.
- Eickhoff, C., and A. de Vries. 2012. "Increasing Cheat Robustness of Crowdsourcing Tasks." *Information Retrieval* 15: 1–17.
- Fox, Jean-Paul. 2010. *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer.
- Galton, F. 1907. "Vox Populi." *Nature (London)* 75: 450–1.
- Goodman, Joseph, Cynthia Cryder, and Amar Cheema. 2013. "Data Collection in a Flat World: Strengths and Weaknesses of Mechanical Turk Samples." *Journal of Behavioral Decision Making* 26 (3): 213–24.
- Grimmer, Justin, and Brandon M Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97.
- Hambleton, Ronald K., Hariharan Swaminathan, and H. Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Thousand Oaks, CA: Sage.
- Hooghe, Liesbet, Ryan Bakker, Anna Brivevich, Catherine de Vries, Erica Edwards, Gary Marks, Jan Rovny, Marco Steenbergen, and Milada Vachudova. 2010. "Reliability and Validity of Measuring Party Positions: The Chapel Hill Expert Surveys of 2002 and 2006." *European Journal of Political Research.* 49 (5): 687–703.
- Horton, J., D. Rand, and R. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14: 399–425.
- Hsueh, P., P. Melville, and V. Sindhwani. 2009. "Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria." Paper read at Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing.
- Ipeirotis, Panagiotis G., Foster Provost, Victor S. Sheng, and Jing Wang. 2014. "Repeated Labeling using Multiple Noisy Labelers." *Data Mining and Knowledge Discovery* 28 (2): 402–41.
- Jones, Frank R., and Bryan D. Baumgartner. 2013. "Policy Agendas Project."
- Kapelner, A., and D. Chandler. 2010. "Preventing Satisficing in Online Surveys: A 'Kapcha' to Ensure Higher Quality Data." Paper read at The World's First Conference on the Future of Distributed Work (CrowdConf 2010).
- King, Gary. 1995. "Replication, Replication." *PS: Political Science & Politics* 28 (03): 444–52.
- Klingemann, Hans-Dieter, Richard I. Hofferbert, and Ian Budge. 1994. *Parties, Policies, and Democracy*. Boulder: Westview Press.
- Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge, and Michael McDonald. 2006. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990–2003*. Oxford: Oxford University Press.
- Krippendorff, Klaus. 2013. *Content Analysis: An Introduction to Its Methodology*. 3rd ed. Thousand Oaks, CA: Sage.
- Laver, M. 1998. "Party Policy in Britain 1997: Results from an Expert Survey." *Political Studies* 46 (2): 336–47.
- Laver, Michael, and Ian Budge. 1992. *Party Policy and Government Coalitions*. New York: St. Martin's Press.
- Laver, Michael, and W. Ben Hunt. 1992. *Policy and Party Competition*. New York: Routledge.
- Lawson, C., G. Lenz, A. Baker, and M. Myers. 2010. "Looking Like a Winner: Candidate Appearance and Electoral Success in New Democracies." *World Politics* 62 (4): 561–93.
- Lin, L. 1989. "A Concordance Correlation Coefficient to Evaluate Reproducibility." *Biometrics* 45: 255–68.
- Lin, L. 2000. "A Note on the Concordance Correlation Coefficient." *Biometrics* 56: 324–5.
- Lord, Frederic. 1980. *Applications of Item Response Theory to Practical Testing Problems*. New York: Routledge.
- Lyon, Aidan, and Eric Pacuit. 2013. "The Wisdom of Crowds: Methods of Human Judgement Aggregation." In *Handbook of Human Computation*, ed. P. Michelucci. New York: Springer.
- Mason, W., and S. Suri. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavior Research Methods* 44 (1): 1–23.
- Mikhaylov, Slava, Michael Laver, and Kenneth Benoit. 2012. "Coder Reliability and Misclassification in Comparative Manifesto Project Codings." *Political Analysis* 20 (1): 78–91.
- Nowak, S., and S. Rger. 2010. "How Reliable are Annotations via Crowdsourcing? A Study about Inter-Annotator Agreement for Multi-Label Image Annotation." Paper read at The 11th ACM International Conference on Multimedia Information Retrieval, 29–31 March 2010, Philadelphia.
- Paolacci, Gabriel, Jesse Chandler, and Panagiotis Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgement and Decision Making* 5: 411–9.
- Quoc Viet Hung, Nguyen, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. 2013. "An Evaluation of Aggregation Techniques in Crowdsourcing." In *Web Information Systems Engineering – WISE 2013*, eds. X. Lin, Y. Manolopoulos, D. Srivastava, and G. Huang. Berlin: Springer.
- Raykar, V. C., S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bo-goni, and L. Moy. 2010. "Learning from Crowds." *Journal of Machine Learning Research* 11: 1297–322.
- Ruedin, Didier. 2013. "Obtaining Party Positions on Immigration in Switzerland: Comparing Different Methods." *Swiss Political Science Review* 19 (1): 84–105.
- Ruedin, Didier, and Laura Morales. 2012. "Obtaining Party Positions on Immigration from Party Manifestos." Paper presented at the Elections, Public Opinion and Parties (EPOP) conference, Oxford, 7 Sept 2012.
- Schwarz, Daniel, Denise Traber, and Kenneth Benoit. Forthcoming. "Estimating Intra-Party Preferences: Comparing Speeches to Votes." *Political Science Research and Methods*.
- Sheng, V., F. Provost, and Panagiotis Ipeirotis. 2008. "Get Another Label? Improving Data Quality and Data Mining using Multiple, Noisy Labelers." Paper read at Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Snow, R., B. O'Connor, D. Jurafsky, and A. Ng. 2008. "Cheap and Fast—But is it Good?: Evaluating Non-expert Annotations for Natural Language Tasks." Paper read at Proceedings

- of the Conference on Empirical Methods in Natural Language Processing.
- Surowiecki, J. 2004. *The Wisdom of Crowds*. New York: W.W. Norton & Company, Inc.
- Turner, Brandon M., Mark Steyvers, Edgar C. Merkle, David V. Budescu, and Thomas S. Wallsten. 2014. "Forecast Aggregation via Recalibration." *Machine Learning* 95 (3): 261–89.
- Volkens, Andrea. 2001. "Manifesto Coding Instructions, 2nd revised ed." In *Discussion Paper (2001)*, ed. W. Berlin, p. 96, Andrea Volkens Berlin: Wissenschaftszentrum Berlin für Sozialforschung gGmbH (WZB).
- Welinder, P., S. Branson, S. Belongie, and P. Perona. 2010. "The Multidimensional Wisdom of Crowds." Paper read at Advances in Neural Information Processing Systems 23 (NIPS 2010).
- Welinder, P., and P. Perona. 2010. "Online CrowdSourcing: Rating Annotators and Obtaining Cost-effective Labels." Paper read at IEEE Conference on Computer Vision and Pattern Recognition Workshops (ACVHL).
- Whitehill, J., P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. 2009. "Whose Vote should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise." Paper read at Advances in Neural Information Processing Systems 22 (NIPS 2009).

## 2 *The term vocabulary and postings lists*

Recall the major steps in inverted index construction:

1. Collect the documents to be indexed.
2. Tokenize the text.
3. Do linguistic preprocessing of tokens.
4. Index the documents that each term occurs in.

In this chapter we first briefly mention how the basic unit of a document can be defined and how the character sequence that it comprises is determined (Section 2.1). We then examine in detail some of the substantive linguistic issues of tokenization and linguistic preprocessing, which determine the vocabulary of terms which a system uses (Section 2.2). Tokenization is the process of chopping character streams into tokens, while linguistic preprocessing then deals with building equivalence classes of tokens which are the set of terms that are indexed. Indexing itself is covered in Chapters 1 and 4. Then we return to the implementation of postings lists. In Section 2.3, we examine an extended postings list data structure that supports faster querying, while Section 2.4 covers building postings data structures suitable for handling phrase and proximity queries, of the sort that commonly appear in both extended Boolean models and on the web.

### 2.1 Document delineation and character sequence decoding

#### 2.1.1 Obtaining the character sequence in a document

Digital documents that are the input to an indexing process are typically bytes in a file or on a web server. The first step of processing is to convert this byte sequence into a linear sequence of characters. For the case of plain English text in ASCII encoding, this is trivial. But often things get much more

Online edition (c) 2009 Cambridge UP

complex. The sequence of characters may be encoded by one of various single byte or multibyte encoding schemes, such as Unicode UTF-8, or various national or vendor-specific standards. We need to determine the correct encoding. This can be regarded as a machine learning classification problem, as discussed in Chapter 13,<sup>1</sup> but is often handled by heuristic methods, user selection, or by using provided document metadata. Once the encoding is determined, we decode the byte sequence to a character sequence. We might save the choice of encoding because it gives some evidence about what language the document is written in.

The characters may have to be decoded out of some binary representation like Microsoft Word DOC files and/or a compressed format such as zip files. Again, we must determine the document format, and then an appropriate decoder has to be used. Even for plain text documents, additional decoding may need to be done. In XML documents (Section 10.1, page 197), character entities, such as &amp; ;, need to be decoded to give the correct character, namely & for &amp; ;. Finally, the textual part of the document may need to be extracted out of other material that will not be processed. This might be the desired handling for XML files, if the markup is going to be ignored; we would almost certainly want to do this with postscript or PDF files. We will not deal further with these issues in this book, and will assume henceforth that our documents are a list of characters. Commercial products usually need to support a broad range of document types and encodings, since users want things to just work with their data as is. Often, they just think of documents as text inside applications and are not even aware of how it is encoded on disk. This problem is usually solved by licensing a software library that handles decoding document formats and character encodings.

The idea that text is a linear sequence of characters is also called into question by some writing systems, such as Arabic, where text takes on some two dimensional and mixed order characteristics, as shown in Figures 2.1 and 2.2. But, despite some complicated writing system conventions, there is an underlying sequence of sounds being represented and hence an essentially linear structure remains, and this is what is represented in the digital representation of Arabic, as shown in Figure 2.1.

### 2.1.2 Choosing a document unit

#### DOCUMENT UNIT

The next phase is to determine what the *document unit* for indexing is. Thus far we have assumed that documents are fixed units for the purposes of indexing. For example, we take each file in a folder as a document. But there

---

1. A classifier is a function that takes objects of some sort and assigns them to one of a number of distinct classes (see Chapter 13). Usually classification is done by machine learning methods such as probabilistic models, but it can also be done by hand-written rules.

كِتَابٌ ← ا ب \*  
 un b ā t i k  
 /kitābun/ ‘a book’

► **Figure 2.1** An example of a vocalized Modern Standard Arabic word. The writing is from right to left and letters undergo complex mutations as they are combined. The representation of short vowels (here, /i/ and /u/) and the final /n/ (nunciation) departs from strict linearity by being represented as diacritics above and below letters. Nevertheless, the represented text is still clearly a linear ordering of characters representing sounds. Full vocalization, as here, normally appears only in the Koran and children’s books. Day-to-day text is unvocalized (short vowels are not represented but the letter for ā would still appear) or partially vocalized, with short vowels inserted in places where the writer perceives ambiguities. These choices add further complexities to indexing.

استقللت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

← → ← → ← START  
 ‘Algeria achieved its independence in 1962 after 132 years of French occupation.’

► **Figure 2.2** The conceptual linear order of characters is not necessarily the order that you see on the page. In languages that are written right-to-left, such as Hebrew and Arabic, it is quite common to also have left-to-right text interspersed, such as numbers and dollar amounts. With modern Unicode representation concepts, the order of characters in files matches the conceptual order, and the reversal of displayed characters is handled by the rendering system, but this may not be true for documents in older encodings.

are many cases in which you might want to do something different. A traditional Unix (mbox-format) email file stores a sequence of email messages (an email folder) in one file, but you might wish to regard each email message as a separate document. Many email messages now contain attached documents, and you might then want to regard the email message and each contained attachment as separate documents. If an email message has an attached zip file, you might want to decode the zip file and regard each file it contains as a separate document. Going in the opposite direction, various pieces of web software (such as `latex2html`) take things that you might regard as a single document (e.g., a Powerpoint file or a L<sup>A</sup>T<sub>E</sub>X document) and split them into separate HTML pages for each slide or subsection, stored as separate files. In these cases, you might want to combine multiple files into a single document.

INDEXING  
GRANULARITY

More generally, for very long documents, the issue of indexing *granularity* arises. For a collection of books, it would usually be a bad idea to index an

entire book as a document. A search for Chinese toys might bring up a book that mentions China in the first chapter and toys in the last chapter, but this does not make it relevant to the query. Instead, we may well wish to index each chapter or paragraph as a mini-document. Matches are then more likely to be relevant, and since the documents are smaller it will be much easier for the user to find the relevant passages in the document. But why stop there? We could treat individual sentences as mini-documents. It becomes clear that there is a precision/recall tradeoff here. If the units get too small, we are likely to miss important passages because terms were distributed over several mini-documents, while if units are too large we tend to get spurious matches and the relevant information is hard for the user to find.

The problems with large document units can be alleviated by use of explicit or implicit proximity search (Sections 2.4.2 and 7.2.2), and the trade-offs in resulting system performance that we are hinting at are discussed in Chapter 8. The issue of index granularity, and in particular a need to simultaneously index documents at multiple levels of granularity, appears prominently in XML retrieval, and is taken up again in Chapter 10. An IR system should be designed to offer choices of granularity. For this choice to be made well, the person who is deploying the system must have a good understanding of the document collection, the users, and their likely information needs and usage patterns. For now, we will henceforth assume that a suitable size document unit has been chosen, together with an appropriate way of dividing or aggregating files, if needed.

## 2.2 Determining the vocabulary of terms

### 2.2.1 Tokenization

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called *tokens*, perhaps at the same time throwing away certain characters, such as punctuation. Here is an example of tokenization:

Input: Friends, Romans, Countrymen, lend me your ears;  
Output: Friends Romans Countrymen lend me your ears

TOKEN	These tokens are often loosely referred to as terms or words, but it is sometimes important to make a type/token distinction. A <i>token</i> is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.
TYPE	A <i>type</i> is the class of all tokens containing the same character sequence.
TERM	A <i>term</i> is a (perhaps normalized) type that is included in the IR system's dictionary. The set of index terms could be entirely distinct from the tokens, for instance, they could be

Online edition (c) 2009 Cambridge UP

semantic identifiers in a taxonomy, but in practice in modern IR systems they are strongly related to the tokens in the document. However, rather than being exactly the tokens that appear in the document, they are usually derived from them by various normalization processes which are discussed in Section 2.2.3.<sup>2</sup> For example, if the document to be indexed is *to sleep perchance to dream*, then there are 5 tokens, but only 4 types (since there are 2 instances of *to*). However, if *to* is omitted from the index (as a stop word, see Section 2.2.2 (page 27)), then there will be only 3 terms: *sleep*, *perchance*, and *dream*.

The major question of the tokenization phase is what are the correct tokens to use? In this example, it looks fairly trivial: you chop on whitespace and throw away punctuation characters. This is a starting point, but even for English there are a number of tricky cases. For example, what do you do about the various uses of the apostrophe for possession and contractions?

Mr. O'Neill thinks that the boys' stories about Chile's capital aren't amusing.

For *O'Neill*, which of the following is the desired tokenization?

neill
oneill
o'neill
o' neill
o neill?

And for *aren't*, is it:

aren't
arent
are n't
aren t?

A simple strategy is to just split on all non-alphanumeric characters, but while 

o	neill
---	-------

 looks okay, 

aren	t
------	---

 looks intuitively bad. For all of them, the choices determine which Boolean queries will match. A query of *neill* AND *capital* will match in three cases but not the other two. In how many cases would a query of *o'neill* AND *capital* match? If no preprocessing of a query is done, then it would match in only one of the five cases. For either

---

2. That is, as defined here, tokens that are not indexed (stop words) are not terms, and if multiple tokens are collapsed together via normalization, they are indexed as one term, under the normalized form. However, we later relax this definition when discussing classification and clustering in Chapters 13–18, where there is no index. In these chapters, we drop the requirement of inclusion in the dictionary. A *term* means a normalized word.

Boolean or free text queries, you always want to do the exact same tokenization of document and query words, generally by processing queries with the same tokenizer. This guarantees that a sequence of characters in a text will always match the same sequence typed in a query.<sup>3</sup>

#### LANGUAGE IDENTIFICATION

These issues of tokenization are language-specific. It thus requires the language of the document to be known. *Language identification* based on classifiers that use short character subsequences as features is highly effective; most languages have distinctive signature patterns (see page 46 for references).

For most languages and particular domains within them there are unusual specific tokens that we wish to recognize as terms, such as the programming languages C++ and C#, aircraft names like B-52, or a T.V. show name such as M\*A\*S\*H – which is sufficiently integrated into popular culture that you find usages such as *M\*A\*S\*H-style hospitals*. Computer technology has introduced new types of character sequences that a tokenizer should probably tokenize as a single token, including email addresses (jblack@mail.yahoo.com), web URLs (<http://stuff.big.com/new/specials.html>), numeric IP addresses (142.32.48.231), package tracking numbers (1Z9999W99845399981), and more. One possible solution is to omit from indexing tokens such as monetary amounts, numbers, and URLs, since their presence greatly expands the size of the vocabulary. However, this comes at a large cost in restricting what people can search for. For instance, people might want to search in a bug database for the line number where an error occurs. Items such as the date of an email, which have a clear semantic type, are often indexed separately as document metadata (see Section 6.1, page 110).

#### HYPHENS

In English, *hyphenation* is used for various purposes ranging from splitting up vowels in words (*co-education*) to joining nouns as names (*Hewlett-Packard*) to a copyediting device to show word grouping (*the hold-him-back-and-drag-him-away maneuver*). It is easy to feel that the first example should be regarded as one token (and is indeed more commonly written as just *coeducation*), the last should be separated into words, and that the middle case is unclear. Handling hyphens automatically can thus be complex: it can either be done as a classification problem, or more commonly by some heuristic rules, such as allowing short hyphenated prefixes on words, but not longer hyphenated forms.

Conceptually, splitting on white space can also split what should be regarded as a single token. This occurs most commonly with names (*San Francisco, Los Angeles*) but also with borrowed foreign phrases (*au fait*) and com-

---

3. For the free text case, this is straightforward. The Boolean case is more complex: this tokenization may produce multiple terms from one query word. This can be handled by combining the terms with an AND or as a phrase query (see Section 2.4, page 39). It is harder for a system to handle the opposite case where the user entered as two terms something that was tokenized together in the document processing.

pounds that are sometimes written as a single word and sometimes space separated (such as *white space* vs. *whitespace*). Other cases with internal spaces that we might wish to regard as a single token include phone numbers ((800) 234-2333) and dates (Mar 11, 1983). Splitting tokens on spaces can cause bad retrieval results, for example, if a search for York University mainly returns documents containing *New York University*. The problems of hyphens and non-separating whitespace can even interact. Advertisements for air fares frequently contain items like *San Francisco-Los Angeles*, where simply doing whitespace splitting would give unfortunate results. In such cases, issues of tokenization interact with handling phrase queries (which we discuss in Section 2.4 (page 39)), particularly if we would like queries for all of *lowercase*, *lower-case* and *lower case* to return the same results. The last two can be handled by splitting on hyphens and using a phrase index. Getting the first case right would depend on knowing that it is sometimes written as two words and also indexing it in this way. One effective strategy in practice, which is used by some Boolean retrieval systems such as Westlaw and Lexis-Nexis (Example 1.1), is to encourage users to enter hyphens wherever they may be possible, and whenever there is a hyphenated form, the system will generalize the query to cover all three of the one word, hyphenated, and two word forms, so that a query for *over-eager* will search for *over-eager* OR “*over eager*” OR *overeager*. However, this strategy depends on user training, since if you query using either of the other two forms, you get no generalization.

Each new language presents some new issues. For instance, French has a variant use of the apostrophe for a reduced definite article ‘the’ before a word beginning with a vowel (e.g., *l’ensemble*) and has some uses of the hyphen with postposed clitic pronouns in imperatives and questions (e.g., *donne-moi* ‘give me’). Getting the first case correct will affect the correct indexing of a fair percentage of nouns and adjectives: you would want documents mentioning both *l’ensemble* and *un ensemble* to be indexed under *ensemble*. Other languages make the problem harder in new ways. German writes *compound nouns* without spaces (e.g., *Computerlinguistik* ‘computational linguistics’; *Lebensversicherungsgesellschaftsangestellter* ‘life insurance company employee’). Retrieval systems for German greatly benefit from the use of a *compound-splitter* module, which is usually implemented by seeing if a word can be subdivided into multiple words that appear in a vocabulary. This phenomenon reaches its limit case with major East Asian Languages (e.g., Chinese, Japanese, Korean, and Thai), where text is written without any spaces between words. An example is shown in Figure 2.3. One approach here is to perform *word segmentation* as prior linguistic processing. Methods of word segmentation vary from having a large vocabulary and taking the longest vocabulary match with some heuristics for unknown words to the use of machine learning sequence models, such as hidden Markov models or conditional random fields, trained over hand-segmented words (see the references

COMPOUNDS

COMPOUND-SPLITTER

WORD SEGMENTATION

莎拉波娃现在居住在美国东南部的佛罗里达。今年4月  
9日，莎拉波娃在美国第一大城市纽约度过了18岁生  
日。生日派对上，莎拉波娃露出了甜美的微笑。

► **Figure 2.3** The standard unsegmented form of Chinese text using the simplified characters of mainland China. There is no whitespace between words, not even between sentences – the apparent space after the Chinese period (。) is just a typographical illusion caused by placing the character on the left side of its square box. The first sentence is just words in Chinese characters with no spaces between them. The second and third sentences include Arabic numerals and punctuation breaking up the Chinese characters.

## 和尚

► **Figure 2.4** Ambiguities in Chinese word segmentation. The two characters can be treated as one word meaning ‘monk’ or as a sequence of two words meaning ‘and’ and ‘still’.

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will						

► **Figure 2.5** A stop list of 25 semantically non-selective words which are common in Reuters-RCV1.

in Section 2.5). Since there are multiple possible segmentations of character sequences (see Figure 2.4), all such methods make mistakes sometimes, and so you are never guaranteed a consistent unique tokenization. The other approach is to abandon word-based indexing and to do all indexing via just short subsequences of characters (character  $k$ -grams), regardless of whether particular sequences cross word boundaries or not. Three reasons why this approach is appealing are that an individual Chinese character is more like a syllable than a letter and usually has some semantic content, that most words are short (the commonest length is 2 characters), and that, given the lack of standardization of word breaking in the writing system, it is not always clear where word boundaries should be placed anyway. Even in English, some cases of where to put word boundaries are just orthographic conventions – think of *notwithstanding* vs. *not to mention* or *into* vs. *on to* – but people are educated to write the words with consistent use of spaces.

### 2.2.2 Dropping common terms: stop words

STOP WORDS  
COLLECTION FREQUENCY  
STOP LIST

Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called *stop words*. The general strategy for determining a stop list is to sort the terms by *collection frequency* (the total number of times each term appears in the document collection), and then to take the most frequent terms, often hand-filtered for their semantic content relative to the domain of the documents being indexed, as a *stop list*, the members of which are then discarded during indexing. An example of a stop list is shown in Figure 2.5. Using a stop list significantly reduces the number of postings that a system has to store; we will present some statistics on this in Chapter 5 (see Table 5.1, page 87). And a lot of the time not indexing stop words does little harm: keyword searches with terms like the and by don't seem very useful. However, this is not true for phrase searches. The phrase query "President of the United States", which contains two stop words, is more precise than President AND "United States". The meaning of flights to London is likely to be lost if the word to is stopped out. A search for Vannevar Bush's article *As we may think* will be difficult if the first three words are stopped out, and the system searches simply for documents containing the word think. Some special query types are disproportionately affected. Some song titles and well known pieces of verse consist entirely of words that are commonly on stop lists (*To be or not to be, Let It Be, I don't want to be, ...*).

The general trend in IR systems over time has been from standard use of quite large stop lists (200–300 terms) to very small stop lists (7–12 terms) to no stop list whatsoever. Web search engines generally do not use stop lists. Some of the design of modern IR systems has focused precisely on how we can exploit the statistics of language so as to be able to cope with common words in better ways. We will show in Section 5.3 (page 95) how good compression techniques greatly reduce the cost of storing the postings for common words. Section 6.2.1 (page 117) then discusses how standard term weighting leads to very common words having little impact on document rankings. Finally, Section 7.1.5 (page 140) shows how an IR system with impact-sorted indexes can terminate scanning a postings list early when weights get small, and hence common words do not cause a large additional processing cost for the average query, even though postings lists for stop words are very long. So for most modern IR systems, the additional cost of including stop words is not that big – neither in terms of index size nor in terms of query processing time.

Query term	Terms in documents that should be matched
Windows	Windows
windows	Windows, windows, window
window	window, windows

► **Figure 2.6** An example of how asymmetric expansion of query terms can usefully model users' expectations.

### 2.2.3 Normalization (equivalence classing of terms)

Having broken up our documents (and also our query) into tokens, the easy case is if tokens in the query just match tokens in the token list of the document. However, there are many cases when two character sequences are not quite the same but you would like a match to occur. For instance, if you search for *USA*, you might hope to also match documents containing *U.S.A.*

TOKEN  
NORMALIZATION  
EQUIVALENCE CLASSES

*Token normalization* is the process of canonicalizing tokens so that matches occur despite superficial differences in the character sequences of the tokens.<sup>4</sup> The most standard way to normalize is to implicitly create *equivalence classes*, which are normally named after one member of the set. For instance, if the tokens *anti-discriminatory* and *antidiscriminatory* are both mapped onto the term *antidiscriminatory*, in both the document text and queries, then searches for one term will retrieve documents that contain either.

The advantage of just using mapping rules that remove characters like hyphens is that the equivalence classing to be done is implicit, rather than being fully calculated in advance: the terms that happen to become identical as the result of these rules are the equivalence classes. It is only easy to write rules of this sort that remove characters. Since the equivalence classes are implicit, it is not obvious when you might want to add characters. For instance, it would be hard to know to turn *antidiscriminatory* into *anti-discriminatory*.

An alternative to creating equivalence classes is to maintain relations between unnormalized tokens. This method can be extended to hand-constructed lists of synonyms such as *car* and *automobile*, a topic we discuss further in Chapter 9. These term relationships can be achieved in two ways. The usual way is to index unnormalized tokens and to maintain a query expansion list of multiple vocabulary entries to consider for a certain query term. A query term is then effectively a disjunction of several postings lists. The alternative is to perform the expansion during index construction. When the document contains *automobile*, we index it under *car* as well (and, usually, also vice-versa). Use of either of these methods is considerably less efficient than equivalence classing, as there are more postings to store and merge. The first

4. It is also often referred to as *term normalization*, but we prefer to reserve the name *term* for the output of the normalization process.

method adds a query expansion dictionary and requires more processing at query time, while the second method requires more space for storing postings. Traditionally, expanding the space required for the postings lists was seen as more disadvantageous, but with modern storage costs, the increased flexibility that comes from distinct postings lists is appealing.

These approaches are more flexible than equivalence classes because the expansion lists can overlap while not being identical. This means there can be an asymmetry in expansion. An example of how such an asymmetry can be exploited is shown in Figure 2.6: if the user enters *windows*, we wish to allow matches with the capitalized *Windows* operating system, but this is not plausible if the user enters *window*, even though it is plausible for this query to also match lowercase *windows*.

The best amount of equivalence classing or query expansion to do is a fairly open question. Doing some definitely seems a good idea. But doing a lot can easily have unexpected consequences of broadening queries in unintended ways. For instance, equivalence-classing *U.S.A.* and *USA* to the latter by deleting periods from tokens might at first seem very reasonable, given the prevalent pattern of optional use of periods in acronyms. However, if I put in as my query term *C.A.T.*, I might be rather upset if it matches every appearance of the word *cat* in documents.<sup>5</sup>

Below we present some of the forms of normalization that are commonly employed and how they are implemented. In many cases they seem helpful, but they can also do harm. In fact, you can worry about many details of equivalence classing, but it often turns out that providing processing is done consistently to the query and to documents, the fine details may not have much aggregate effect on performance.

**Accents and diacritics.** Diacritics on characters in English have a fairly marginal status, and we might well want *cliché* and *cliche* to match, or *naïve* and *naïve*. This can be done by normalizing tokens to remove diacritics. In many other languages, diacritics are a regular part of the writing system and distinguish different sounds. Occasionally words are distinguished only by their accents. For instance, in Spanish, *peña* is ‘a cliff’, while *pena* is ‘sorrow’. Nevertheless, the important question is usually not prescriptive or linguistic but is a question of how users are likely to write queries for these words. In many cases, users will enter queries for words without diacritics, whether for reasons of speed, laziness, limited software, or habits born of the days when it was hard to use non-ASCII text on many computer systems. In these cases, it might be best to equate all words to a form without diacritics.

---

5. At the time we wrote this chapter (Aug. 2005), this was actually the case on Google: the top result for the query *C.A.T.* was a site about cats, the Cat Fanciers Web Site <http://www.fanciers.com/>.

## CASE-FOLDING

**Capitalization/case-folding.** A common strategy is to do *case-folding* by reducing all letters to lower case. Often this is a good idea: it will allow instances of *Automobile* at the beginning of a sentence to match with a query of *automobile*. It will also help on a web search engine when most of your users type in *ferrari* when they are interested in a *Ferrari* car. On the other hand, such case folding can equate words that might better be kept apart. Many proper nouns are derived from common nouns and so are distinguished only by case, including companies (*General Motors*, *The Associated Press*), government organizations (*the Fed* vs. *fed*) and person names (*Bush*, *Black*). We already mentioned an example of unintended query expansion with acronyms, which involved not only acronym normalization (*C.A.T.* → *CAT*) but also case-folding (*CAT* → *cat*).

For English, an alternative to making every token lowercase is to just make some tokens lowercase. The simplest heuristic is to convert to lowercase words at the beginning of a sentence and all words occurring in a title that is all uppercase or in which most or all words are capitalized. These words are usually ordinary words that have been capitalized. Mid-sentence capitalized words are left as capitalized (which is usually correct). This will mostly avoid case-folding in cases where distinctions should be kept apart. The same task can be done more accurately by a machine learning sequence model which uses more features to make the decision of when to case-fold. This is known as *truecasing*. However, trying to get capitalization right in this way probably doesn't help if your users usually use lowercase regardless of the correct case of words. Thus, lowercasing everything often remains the most practical solution.

## TRUECASING

**Other issues in English.** Other possible normalizations are quite idiosyncratic and particular to English. For instance, you might wish to equate *ne'er* and *never* or the British spelling *colour* and the American spelling *color*. Dates, times and similar items come in multiple formats, presenting additional challenges. You might wish to collapse together *3/12/91* and *Mar. 12, 1991*. However, correct processing here is complicated by the fact that in the U.S., *3/12/91* is *Mar. 12, 1991*, whereas in Europe it is *3 Dec 1991*.

**Other languages.** English has maintained a dominant position on the WWW; approximately 60% of web pages are in English (Gerrand 2007). But that still leaves 40% of the web, and the non-English portion might be expected to grow over time, since less than one third of Internet users and less than 10% of the world's population primarily speak English. And there are signs of change: Sifry (2007) reports that only about one third of blog posts are in English.

Other languages again present distinctive issues in equivalence classing.

ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるMOTTA I NAI キャンペーンの一環として、毎日新聞社とマガジンハウスは「私の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを800字以内の文章にまとめ、簡単な写真、イラスト、図などを添えて10月20日までにお送りください。大賞受賞者には、50万円相当の旅行券とエコ製品2点の副賞が贈られます。

► **Figure 2.7** Japanese makes use of multiple intermingled writing systems and, like Chinese, does not segment words. The text is mainly Chinese characters with the hiragana syllabary for inflectional endings and function words. The part in latin letters is actually a Japanese expression, but has been taken up as the name of an environmental campaign by 2004 Nobel Peace Prize winner Wangari Maathai. His name is written using the katakana syllabary in the middle of the first line. The first four characters of the final line express a monetary amount that we would want to match with ¥500,000 (500,000 Japanese yen).

The French word for *the* has distinctive forms based not only on the gender (masculine or feminine) and number of the following noun, but also depending on whether the following word begins with a vowel: *le, la, l', les*. We may well wish to equivalence class these various forms of *the*. German has a convention whereby vowels with an umlaut can be rendered instead as a two vowel digraph. We would want to treat *Schütze* and *Schuetze* as equivalent.

Japanese is a well-known difficult writing system, as illustrated in Figure 2.7. Modern Japanese is standardly an intermingling of multiple alphabets, principally Chinese characters, two syllabaries (hiragana and katakana) and western characters (Latin letters, Arabic numerals, and various symbols). While there are strong conventions and standardization through the education system over the choice of writing system, in many cases the same word can be written with multiple writing systems. For example, a word may be written in katakana for emphasis (somewhat like italics). Or a word may sometimes be written in hiragana and sometimes in Chinese characters. Successful retrieval thus requires complex equivalence classing across the writing systems. In particular, an end user might commonly present a query entirely in hiragana, because it is easier to type, just as Western end users commonly use all lowercase.

Document collections being indexed can include documents from many different languages. Or a single document can easily contain text from multiple languages. For instance, a French email might quote clauses from a contract document written in English. Most commonly, the language is detected and language-particular tokenization and normalization rules are applied at a predetermined granularity, such as whole documents or individual paragraphs, but this still will not correctly deal with cases where language changes occur for brief quotations. When document collections contain mul-

multiple languages, a single index may have to contain terms of several languages. One option is to run a language identification classifier on documents and then to tag terms in the vocabulary for their language. Or this tagging can simply be omitted, since it is relatively rare for the exact same character sequence to be a word in different languages.

When dealing with foreign or complex words, particularly foreign names, the spelling may be unclear or there may be variant transliteration standards giving different spellings (for example, *Chebyshev* and *Tchebycheff* or *Beijing* and *Peking*). One way of dealing with this is to use heuristics to equivalence class or expand terms with phonetic equivalents. The traditional and best known such algorithm is the Soundex algorithm, which we cover in Section 3.4 (page 63).

#### 2.2.4 Stemming and lemmatization

For grammatical reasons, documents are going to use different forms of a word, such as *organize*, *organizes*, and *organizing*. Additionally, there are families of derivationally related words with similar meanings, such as *democracy*, *democratic*, and *democratization*. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set.

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance:

am, are, is ⇒ be  
car, cars, car's, cars' ⇒ car

The result of this mapping of text will be something like:

the boy's cars are different colors ⇒  
the boy car be differ color

STEMMING	However, the two words differ in their flavor. <i>Stemming</i> usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. <i>Lemmatization</i> usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the <i>lemma</i> . If confronted with the token <i>saw</i> , stemming might return just <i>s</i> , whereas lemmatization would attempt to return either <i>see</i> or <i>saw</i> depending on whether the use of the token was as a verb or a noun. The two may also differ in that stemming most commonly collapses derivationally related words, whereas lemmatization commonly only collapses the different inflectional forms of a lemma.
LEMMATIZATION	
LEMMA	

Linguistic processing for stemming or lemmatization is often done by an additional plug-in component to the indexing process, and a number of such components exist, both commercial and open-source.

## PORTER STEMMER

The most common algorithm for stemming English, and one that has repeatedly been shown to be empirically very effective, is *Porter's algorithm* (Porter 1980). The entire algorithm is too long and intricate to present here, but we will indicate its general nature. Porter's algorithm consists of 5 phases of word reductions, applied sequentially. Within each phase there are various conventions to select rules, such as selecting the rule from each rule group that applies to the longest suffix. In the first phase, this convention is used with the following rule group:

(2.1)	Rule	Example
SSES	→ SS	caresses → caress
IES	→ I	ponies → poni
SS	→ SS	caress → caress
S	→	cats → cat

Many of the later rules use a concept of the *measure* of a word, which loosely checks the number of syllables to see whether a word is long enough that it is reasonable to regard the matching portion of a rule as a suffix rather than as part of the stem of a word. For example, the rule:

( $m > 1$ ) EMENT →

would map *replacement* to *replac*, but not *cement* to *c*. The official site for the Porter Stemmer is:

<http://www.tartarus.org/~martin/PorterStemmer/>

Other stemmers exist, including the older, one-pass Lovins stemmer (Lovins 1968), and newer entrants like the Paice/Husk stemmer (Paice 1990); see:

<http://www.cs.waikato.ac.nz/~eibe/stemmers/>

<http://www.comp.lancs.ac.uk/computing/research/stemming/>

Figure 2.8 presents an informal comparison of the different behaviors of these stemmers. Stemmers use language-specific rules, but they require less knowledge than a lemmatizer, which needs a complete vocabulary and morphological analysis to correctly lemmatize words. Particular domains may also require special stemming rules. However, the exact stemmed form does not matter, only the equivalence classes it forms.

## LEMMATIZER

Rather than using a stemmer, you can use a *lemmatizer*, a tool from Natural Language Processing which does full morphological analysis to accurately identify the lemma for each word. Doing full morphological analysis produces at most very modest benefits for retrieval. It is hard to say more,

**Sample text:** Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

**Lovins stemmer:** such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

**Porter stemmer:** such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

**Paice stemmer:** such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

► **Figure 2.8** A comparison of three stemming algorithms on a sample text.

because either form of normalization tends not to improve English information retrieval performance in aggregate – at least not by very much. While it helps a lot for some queries, it equally hurts performance a lot for others. Stemming increases recall while harming precision. As an example of what can go wrong, note that the Porter stemmer stems all of the following words:

*operate operating operates operation operative operatives operational*

to oper. However, since *operate* in its various forms is a common verb, we would expect to lose considerable precision on queries such as the following with Porter stemming:

operational AND research  
operating AND system  
operative AND dentistry

For a case like this, moving to using a lemmatizer would not completely fix the problem because particular inflectional forms are used in particular collocations: a sentence with the words *operate* and *system* is not a good match for the query *operating AND system*. Getting better value from term normalization depends more on pragmatic issues of word use than on formal issues of linguistic morphology.

The situation is different for languages with much more morphology (such as Spanish, German, and Finnish). Results in the European CLEF evaluations have repeatedly shown quite large gains from the use of stemmers (and compound splitting for languages like German); see the references in Section 2.5.

**Exercise 2.1**[*\**]

Are the following statements true or false?

- a. In a Boolean retrieval system, stemming never lowers precision.
- b. In a Boolean retrieval system, stemming never lowers recall.
- c. Stemming increases the size of the vocabulary.
- d. Stemming should be invoked at indexing time but not while processing a query.

**Exercise 2.2**[*\**]

Suggest what normalized form should be used for these words (including the word itself as a possibility):

- a. 'Cos
- b. Shi'ite
- c. cont'd
- d. Hawai'i
- e. O'Rourke

**Exercise 2.3**[*\**]

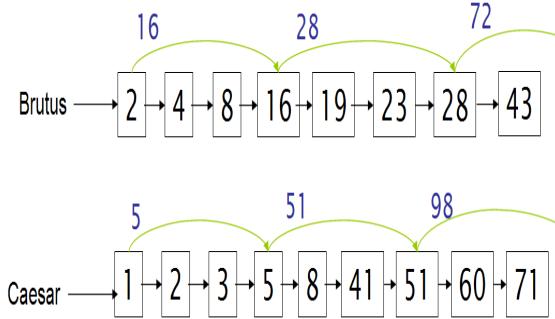
The following pairs of words are stemmed to the same form by the Porter stemmer. Which pairs would you argue shouldn't be conflated. Give your reasoning.

- a. abandon/abandonment
- b. absorbency/absorbent
- c. marketing/markets
- d. university/universe
- e. volume/volumes

**Exercise 2.4**[*\**]

For the Porter stemmer rule group shown in (2.1):

- a. What is the purpose of including an identity rule such as SS → SS?
- b. Applying just this rule group, what will the following words be stemmed to?  
*circus canaries boss*
- c. What rule should be added to correctly stem *pony*?
- d. The stemming for *ponies* and *pony* might seem strange. Does it have a deleterious effect on retrieval? Why or why not?



► **Figure 2.9** Postings lists with skip pointers. The postings intersection can use a skip pointer when the end point is still less than the item on the other list.

### 2.3 Faster postings list intersection via skip pointers

In the remainder of this chapter, we will discuss extensions to postings list data structures and ways to increase the efficiency of using postings lists. Recall the basic postings list intersection operation from Section 1.3 (page 10): we walk through the two postings lists simultaneously, in time linear in the total number of postings entries. If the list lengths are  $m$  and  $n$ , the intersection takes  $O(m + n)$  operations. Can we do better than this? That is, empirically, can we usually process postings list intersection in sublinear time? We can, if the index isn't changing too fast.

SKIP LIST

One way to do this is to use a *skip list* by augmenting postings lists with skip pointers (at indexing time), as shown in Figure 2.9. Skip pointers are effectively shortcuts that allow us to avoid processing parts of the postings list that will not figure in the search results. The two questions are then where to place skip pointers and how to do efficient merging using skip pointers.

Consider first efficient merging, with Figure 2.9 as an example. Suppose we've stepped through the lists in the figure until we have matched [8] on each list and moved it to the results list. We advance both pointers, giving us [16] on the upper list and [41] on the lower list. The smallest item is then the element [16] on the top list. Rather than simply advancing the upper pointer, we first check the skip list pointer and note that 28 is also less than 41. Hence we can follow the skip list pointer, and then we advance the upper pointer to [28]. We thus avoid stepping to [19] and [23] on the upper list. A number of variant versions of postings list intersection with skip pointers is possible depending on when exactly you check the skip pointer. One version is shown

```

INTERSECTWITHSKIP(p1, p2)
1  answer ← ⟨ ⟩
2  while p1 ≠ NIL and p2 ≠ NIL
3  do if docID(p1) = docID(p2)
4    then ADD(answer, docID(p1))
5    p1 ← next(p1)
6    p2 ← next(p2)
7  else if docID(p1) < docID(p2)
8    then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
9      then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10     do p1 ← skip(p1)
11    else p1 ← next(p1)
12    else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13      then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14      do p2 ← skip(p2)
15    else p2 ← next(p2)
16  return answer

```

► **Figure 2.10** Postings lists intersection with skip pointers.

in Figure 2.10. Skip pointers will only be available for the original postings lists. For an intermediate result in a complex query, the call *hasSkip(p)* will always return false. Finally, note that the presence of skip pointers only helps for AND queries, not for OR queries.

Where do we place skips? There is a tradeoff. More skips means shorter skip spans, and that we are more likely to skip. But it also means lots of comparisons to skip pointers, and lots of space storing skip pointers. Fewer skips means few pointer comparisons, but then long skip spans which means that there will be fewer opportunities to skip. A simple heuristic for placing skips, which has been found to work well in practice, is that for a postings list of length  $P$ , use  $\sqrt{P}$  evenly-spaced skip pointers. This heuristic can be improved upon; it ignores any details of the distribution of query terms.

Building effective skip pointers is easy if an index is relatively static; it is harder if a postings list keeps changing because of updates. A malicious deletion strategy can render skip lists ineffective.

Choosing the optimal encoding for an inverted index is an ever-changing game for the system builder, because it is strongly dependent on underlying computer technologies and their relative speeds and sizes. Traditionally, CPUs were slow, and so highly compressed techniques were not optimal. Now CPUs are fast and disk is slow, so reducing disk postings list size dominates. However, if you're running a search engine with everything in mem-

ory then the equation changes again. We discuss the impact of hardware parameters on index construction time in Section 4.1 (page 68) and the impact of index size on system speed in Chapter 5.


**Exercise 2.5**
[ $\star$ ]

Why are skip pointers not useful for queries of the form  $x \text{ OR } y$ ?

**Exercise 2.6**
[ $\star$ ]

We have a two-word query. For one term the postings list consists of the following 16 entries:

[4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180]

and for the other it is the one entry postings list:

[47].

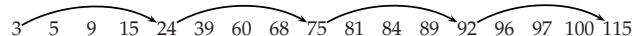
Work out how many comparisons would be done to intersect the two postings lists with the following two strategies. Briefly justify your answers:

a. Using standard postings lists

b. Using postings lists stored with skip pointers, with a skip length of  $\sqrt{P}$ , as suggested in Section 2.3.

**Exercise 2.7**
[ $\star$ ]

Consider a postings intersection between this postings list, with skip pointers:



and the following intermediate result postings list (which hence has no skip pointers):

3 5 89 95 97 99 100 101

Trace through the postings intersection algorithm in Figure 2.10 (page 37).

- How often is a skip pointer followed (i.e.,  $p_1$  is advanced to  $\text{skip}(p_1)$ )?
- How many postings comparisons will be made by this algorithm while intersecting the two lists?
- How many postings comparisons would be made if the postings lists are intersected without the use of skip pointers?

## 2.4 Positional postings and phrase queries

PHRASE QUERIES

Many complex or technical concepts and many organization and product names are multiword compounds or phrases. We would like to be able to pose a query such as Stanford University by treating it as a phrase so that a sentence in a document like *The inventor Stanford Ovshinsky never went to university.* is not a match. Most recent search engines support a double quotes syntax ("stanford university") for *phrase queries*, which has proven to be very easily understood and successfully used by users. As many as 10% of web queries are phrase queries, and many more are implicit phrase queries (such as person names), entered without use of double quotes. To be able to support such queries, it is no longer sufficient for postings lists to be simply lists of documents that contain individual terms. In this section we consider two approaches to supporting phrase queries and their combination. A search engine should not only support phrase queries, but implement them efficiently. A related but distinct concept is term proximity weighting, where a document is preferred to the extent that the query terms appear close to each other in the text. This technique is covered in Section 7.2.2 (page 144) in the context of ranked retrieval.

BIWORD INDEX

### 2.4.1 Biword indexes

One approach to handling phrases is to consider every pair of consecutive terms in a document as a phrase. For example, the text *Friends, Romans, Countrymen* would generate the *biwords*:

```
friends romans
romans countrymen
```

In this model, we treat each of these biwords as a vocabulary term. Being able to process two-word phrase queries is immediate. Longer phrases can be processed by breaking them down. The query *stanford university palo alto* can be broken into the Boolean query on biwords:

```
"stanford university" AND "university palo" AND "palo alto"
```

This query could be expected to work fairly well in practice, but there can and will be occasional false positives. Without examining the documents, we cannot verify that the documents matching the above Boolean query do actually contain the original 4 word phrase.

Among possible queries, nouns and noun phrases have a special status in describing the concepts people are interested in searching for. But related nouns can often be divided from each other by various function words, in phrases such as *the abolition of slavery* or *renegotiation of the constitution*. These needs can be incorporated into the biword indexing model in the following

way. First, we tokenize the text and perform part-of-speech-tagging.<sup>6</sup> We can then group terms into nouns, including proper nouns, (N) and function words, including articles and prepositions, (X), among other classes. Now deem any string of terms of the form NX\*N to be an extended biword. Each such extended biword is made a term in the vocabulary. For example:

renegotiation	of	the	constitution
N	X	X	N

To process a query using such an extended biword index, we need to also parse it into N's and X's, and then segment the query into extended biwords, which can be looked up in the index.

This algorithm does not always work in an intuitively optimal manner when parsing longer queries into Boolean queries. Using the above algorithm, the query

cost overruns on a power plant

is parsed into

“cost overruns” AND “overruns power” AND “power plant”

whereas it might seem a better query to omit the middle biword. Better results can be obtained by using more precise part-of-speech patterns that define which extended biwords should be indexed.

#### PHRASE INDEX

The concept of a biword index can be extended to longer sequences of words, and if the index includes variable length word sequences, it is generally referred to as a *phrase index*. Indeed, searches for a single term are not naturally handled in a biword index (you would need to scan the dictionary for all biwords containing the term), and so we also need to have an index of single-word terms. While there is always a chance of false positive matches, the chance of a false positive match on indexed phrases of length 3 or more becomes very small indeed. But on the other hand, storing longer phrases has the potential to greatly expand the vocabulary size. Maintaining exhaustive phrase indexes for phrases of length greater than two is a daunting prospect, and even use of an exhaustive biword dictionary greatly expands the size of the vocabulary. However, towards the end of this section we discuss the utility of the strategy of using a partial phrase index in a compound indexing scheme.

---

6. Part of speech taggers classify words as nouns, verbs, etc. – or, in practice, often as finer-grained classes like “plural proper noun”. Many fairly accurate (c. 96% per-tag accuracy) part-of-speech taggers now exist, usually trained by machine learning methods on hand-tagged text. See, for instance, Manning and Schütze (1999, ch. 10).

```

to, 993427:
  ⟨ 1, 6: ⟨7, 18, 33, 72, 86, 231⟩;
  2, 5: ⟨1, 17, 74, 222, 255⟩;
  4, 5: ⟨8, 16, 190, 429, 433⟩;
  5, 2: ⟨363, 367⟩;
  7, 3: ⟨13, 23, 191⟩; ...⟩

be, 178239:
  ⟨ 1, 2: ⟨17, 25⟩;
  4, 5: ⟨17, 191, 291, 430, 434⟩;
  5, 3: ⟨14, 19, 101⟩; ...⟩

```

► **Figure 2.11** Positional index example. The word *to* has a document frequency 993,477, and occurs 6 times in document 1 at positions 7, 18, 33, etc.

#### 2.4.2 Positional indexes

POSITIONAL INDEX

For the reasons given, a biword index is not the standard solution. Rather, a *positional index* is most commonly employed. Here, for each term in the vocabulary, we store postings of the form docID: ⟨position1, position2, ...⟩, as shown in Figure 2.11, where each position is a token index in the document. Each posting will also usually record the term frequency, for reasons discussed in Chapter 6.

To process a phrase query, you still need to access the inverted index entries for each distinct term. As before, you would start with the least frequent term and then work to further restrict the list of possible candidates. In the merge operation, the same general technique is used as before, but rather than simply checking that both terms are in a document, you also need to check that their positions of appearance in the document are compatible with the phrase query being evaluated. This requires working out offsets between the words.



**Example 2.1: Satisfying phrase queries.** Suppose the postings lists for *to* and *be* are as in Figure 2.11, and the query is “*to be or not to be*”. The postings lists to access are: *to*, *be*, or, *not*. We will examine intersecting the postings lists for *to* and *be*. We first look for documents that contain both terms. Then, we look for places in the lists where there is an occurrence of *be* with a token index one higher than a position of *to*, and then we look for another occurrence of each word with token index 4 higher than the first occurrence. In the above lists, the pattern of occurrences that is a possible match is:

```

to: ⟨..., 4:⟨..., 429, 433⟩; ...⟩
be: ⟨..., 4:⟨..., 430, 434⟩; ...⟩

```

```

POSITIONALINTERSECT( $p_1, p_2, k$ )
1   answer  $\leftarrow \langle \rangle$ 
2   while  $p_1 \neq \text{NIL}$  and  $p_2 \neq \text{NIL}$ 
3   do if  $\text{docID}(p_1) = \text{docID}(p_2)$ 
4     then  $l \leftarrow \langle \rangle$ 
5      $pp_1 \leftarrow \text{positions}(p_1)$ 
6      $pp_2 \leftarrow \text{positions}(p_2)$ 
7     while  $pp_1 \neq \text{NIL}$ 
8     do while  $pp_2 \neq \text{NIL}$ 
9       do if  $|\text{pos}(pp_1) - \text{pos}(pp_2)| \leq k$ 
10      then ADD( $l, \text{pos}(pp_2)$ )
11      else if  $\text{pos}(pp_2) > \text{pos}(pp_1)$ 
12        then break
13         $pp_2 \leftarrow \text{next}(pp_2)$ 
14        while  $l \neq \langle \rangle$  and  $|l[0] - \text{pos}(pp_1)| > k$ 
15        do DELETE( $l[0]$ )
16        for each  $ps \in l$ 
17        do ADD(answer,  $\langle \text{docID}(p_1), \text{pos}(pp_1), ps \rangle$ )
18         $pp_1 \leftarrow \text{next}(pp_1)$ 
19         $p_1 \leftarrow \text{next}(p_1)$ 
20         $p_2 \leftarrow \text{next}(p_2)$ 
21        else if  $\text{docID}(p_1) < \text{docID}(p_2)$ 
22          then  $p_1 \leftarrow \text{next}(p_1)$ 
23        else  $p_2 \leftarrow \text{next}(p_2)$ 
24   return answer

```

► **Figure 2.12** An algorithm for proximity intersection of postings lists  $p_1$  and  $p_2$ . The algorithm finds places where the two terms appear within  $k$  words of each other and returns a list of triples giving docID and the term position in  $p_1$  and  $p_2$ .

The same general method is applied for within  $k$  word proximity searches, of the sort we saw in Example 1.1 (page 15):

employment /3 place

Here,  $/k$  means “within  $k$  words of (on either side)”. Clearly, positional indexes can be used for such queries; biword indexes cannot. We show in Figure 2.12 an algorithm for satisfying within  $k$  word proximity searches; it is further discussed in Exercise 2.12.

**Positional index size.** Adopting a positional index expands required postings storage significantly, even if we compress position values/offsets as we

will discuss in Section 5.3 (page 95). Indeed, moving to a positional index also changes the asymptotic complexity of a postings intersection operation, because the number of items to check is now bounded not by the number of documents but by the total number of tokens in the document collection  $T$ . That is, the complexity of a Boolean query is  $\Theta(T)$  rather than  $\Theta(N)$ . However, most applications have little choice but to accept this, since most users now expect to have the functionality of phrase and proximity searches.

Let's examine the space implications of having a positional index. A posting now needs an entry for each occurrence of a term. The index size thus depends on the average document size. The average web page has less than 1000 terms, but documents like SEC stock filings, books, and even some epic poems easily reach 100,000 terms. Consider a term with frequency 1 in 1000 terms on average. The result is that large documents cause an increase of two orders of magnitude in the space required to store the postings list:

Document size	Expected postings	Expected entries in positional posting
1000	1	1
100,000	1	100

While the exact numbers depend on the type of documents and the language being indexed, some rough rules of thumb are to expect a positional index to be 2 to 4 times as large as a non-positional index, and to expect a compressed positional index to be about one third to one half the size of the raw text (after removal of markup, etc.) of the original uncompressed documents. Specific numbers for an example collection are given in Table 5.1 (page 87) and Table 5.6 (page 103).

### 2.4.3 Combination schemes

The strategies of biword indexes and positional indexes can be fruitfully combined. If users commonly query on particular phrases, such as Michael Jackson, it is quite inefficient to keep merging positional postings lists. A combination strategy uses a phrase index, or just a biword index, for certain queries and uses a positional index for other phrase queries. Good queries to include in the phrase index are ones known to be common based on recent querying behavior. But this is not the only criterion: the most expensive phrase queries to evaluate are ones where the individual words are common but the desired phrase is comparatively rare. Adding *Britney Spears* as a phrase index entry may only give a speedup factor to that query of about 3, since most documents that mention either word are valid results, whereas adding *The Who* as a phrase index entry may speed up that query by a factor of 1000. Hence, having the latter is more desirable, even if it is a relatively less common query.

## NEXT WORD INDEX

Williams et al. (2004) evaluate an even more sophisticated scheme which employs indexes of both these sorts and additionally a partial next word index as a halfway house between the first two strategies. For each term, a *next word index* records terms that follow it in a document. They conclude that such a strategy allows a typical mixture of web phrase queries to be completed in one quarter of the time taken by use of a positional index alone, while taking up 26% more space than use of a positional index alone.

**Exercise 2.8**[*\**]

Assume a biword index. Give an example of a document which will be returned for a query of New York University but is actually a false positive which should not be returned.

**Exercise 2.9**[*\**]

Shown below is a portion of a positional index in the format: term: doc1: ⟨position1, position2, ...⟩; doc2: ⟨position1, position2, ...⟩; etc.

```

angels: 2: ⟨36,174,252,651⟩; 4: ⟨12,22,102,432⟩; 7: ⟨17⟩;
fools: 2: ⟨1,17,74,222⟩; 4: ⟨8,78,108,458⟩; 7: ⟨3,13,23,193⟩;
fear: 2: ⟨87,704,722,901⟩; 4: ⟨13,43,113,433⟩; 7: ⟨18,328,528⟩;
in: 2: ⟨3,37,76,444,851⟩; 4: ⟨10,20,110,470,500⟩; 7: ⟨5,15,25,195⟩;
rush: 2: ⟨2,66,194,321,702⟩; 4: ⟨9,69,149,429,569⟩; 7: ⟨4,14,404⟩;
to: 2: ⟨47,86,234,999⟩; 4: ⟨14,24,774,944⟩; 7: ⟨199,319,599,709⟩;
tread: 2: ⟨57,94,333⟩; 4: ⟨15,35,155⟩; 7: ⟨20,320⟩;
where: 2: ⟨67,124,393,1001⟩; 4: ⟨11,41,101,421,431⟩; 7: ⟨16,36,736⟩;
```

Which document(s) if any match each of the following queries, where each expression within quotes is a phrase query?

- a. “fools rush in”
- b. “fools rush in” AND “angels fear to tread”

**Exercise 2.10**[*\**]

Consider the following fragment of a positional index with the format:

```

word: document: ⟨position, position, ...⟩; document: ⟨position, ...⟩
...
```

```

Gates: 1: ⟨3⟩; 2: ⟨6⟩; 3: ⟨2,17⟩; 4: ⟨1⟩;
IBM: 4: ⟨3⟩; 7: ⟨14⟩;
Microsoft: 1: ⟨1⟩; 2: ⟨1,21⟩; 3: ⟨3⟩; 5: ⟨16,22,51⟩;
```

The *k* operator, word1 /*k* word2 finds occurrences of word1 within *k* words of word2 (on either side), where *k* is a positive integer argument. Thus *k* = 1 demands that word1 be adjacent to word2.

- a. Describe the set of documents that satisfy the query Gates /2 Microsoft.
- b. Describe each set of values for *k* for which the query Gates /*k* Microsoft returns a different set of documents as the answer.

**Exercise 2.11**

[\*\*]

Consider the general procedure for merging two positional postings lists for a given document, to determine the document positions where a document satisfies a  $/k$  clause (in general there can be multiple positions at which each term occurs in a single document). We begin with a pointer to the position of occurrence of each term and move each pointer along the list of occurrences in the document, checking as we do so whether we have a hit for  $/k$ . Each move of either pointer counts as a step. Let  $L$  denote the total number of occurrences of the two terms in the document. What is the big-O complexity of the merge procedure, if we wish to have postings including positions in the result?

**Exercise 2.12**

[\*\*]

Consider the adaptation of the basic algorithm for intersection of two postings lists (Figure 1.6, page 11) to the one in Figure 2.12 (page 42), which handles proximity queries. A naive algorithm for this operation could be  $O(PL_{\max}^2)$ , where  $P$  is the sum of the lengths of the postings lists (i.e., the sum of document frequencies) and  $L_{\max}$  is the maximum length of a document (in tokens).

- Go through this algorithm carefully and explain how it works.
- What is the complexity of this algorithm? Justify your answer carefully.
- For certain queries and data distributions, would another algorithm be more efficient? What complexity does it have?

**Exercise 2.13**

[\*\*]

Suppose we wish to use a postings intersection procedure to determine simply the list of documents that satisfy a  $/k$  clause, rather than returning the list of positions, as in Figure 2.12 (page 42). For simplicity, assume  $k \geq 2$ . Let  $L$  denote the total number of occurrences of the two terms in the document collection (i.e., the sum of their collection frequencies). Which of the following is true? Justify your answer.

- The merge can be accomplished in a number of steps linear in  $L$  and independent of  $k$ , and we can ensure that each pointer moves only to the right.
- The merge can be accomplished in a number of steps linear in  $L$  and independent of  $k$ , but a pointer may be forced to move non-monotonically (i.e., to sometimes back up)
- The merge can require  $kL$  steps in some cases.

**Exercise 2.14**

[\*\*]

How could an IR system combine use of a positional index and use of stop words? What is the potential problem, and how could it be handled?

## 2.5 References and further reading

EAST ASIAN  
LANGUAGES

Exhaustive discussion of the character-level processing of East Asian languages can be found in Lunde (1998). Character bigram indexes are perhaps the most standard approach to indexing Chinese, although some systems use word segmentation. Due to differences in the language and writing system, word segmentation is most usual for Japanese (Luk and Kwok 2002, Kishida

et al. 2005). The structure of a character  $k$ -gram index over unsegmented text differs from that in Section 3.2.2 (page 54): there the  $k$ -gram dictionary points to postings lists of entries in the regular dictionary, whereas here it points directly to document postings lists. For further discussion of Chinese word segmentation, see Sproat et al. (1996), Sproat and Emerson (2003), Tseng et al. (2005), and Gao et al. (2005).

Lita et al. (2003) present a method for truecasing. Natural language processing work on computational morphology is presented in (Sproat 1992, Beesley and Karttunen 2003).

Language identification was perhaps first explored in cryptography; for example, Konheim (1981) presents a character-level  $k$ -gram language identification algorithm. While other methods such as looking for particular distinctive function words and letter combinations have been used, with the advent of widespread digital text, many people have explored the character  $n$ -gram technique, and found it to be highly successful (Beesley 1998, Dunning 1994, Cavnar and Trenkle 1994). Written language identification is regarded as a fairly easy problem, while spoken language identification remains more difficult; see Hughes et al. (2006) for a recent survey.

Experiments on and discussion of the positive and negative impact of stemming in English can be found in the following works: Salton (1989), Harman (1991), Krovetz (1995), Hull (1996). Hollink et al. (2004) provide detailed results for the effectiveness of language-specific methods on 8 European languages. In terms of percent change in mean average precision (see page 159) over a baseline system, diacritic removal gains up to 23% (being especially helpful for Finnish, French, and Swedish). Stemming helped markedly for Finnish (30% improvement) and Spanish (10% improvement), but for most languages, including English, the gain from stemming was in the range 0–5%, and results from a lemmatizer were poorer still. Compound splitting gained 25% for Swedish and 15% for German, but only 4% for Dutch. Rather than language-particular methods, indexing character  $k$ -grams (as we suggested for Chinese) could often give as good or better results: using within-word character 4-grams rather than words gave gains of 37% in Finnish, 27% in Swedish, and 20% in German, while even being slightly positive for other languages, such as Dutch, Spanish, and English. Tomlinson (2003) presents broadly similar results. Bar-Ilan and Gutman (2005) suggest that, at the time of their study (2003), the major commercial web search engines suffered from lacking decent language-particular processing; for example, a query on [www.google.fr](http://www.google.fr) for l'électricité did not separate off the article 'l' but only matched pages with precisely this string of article+noun.

#### SKIP LIST

The classic presentation of skip pointers for IR can be found in Moffat and Zobel (1996). Extended techniques are discussed in Boldi and Vigna (2005). The main paper in the algorithms literature is Pugh (1990), which uses multilevel skip pointers to give expected  $O(\log P)$  list access (the same expected

efficiency as using a tree data structure) with less implementational complexity. In practice, the effectiveness of using skip pointers depends on various system parameters. [Moffat and Zobel \(1996\)](#) report conjunctive queries running about five times faster with the use of skip pointers, but [Bahle et al. \(2002, p. 217\)](#) report that, with modern CPUs, using skip lists instead slows down search because it expands the size of the postings list (i.e., disk I/O dominates performance). In contrast, [Strohman and Croft \(2007\)](#) again show good performance gains from skipping, in a system architecture designed to optimize for the large memory spaces and multiple cores of recent CPUs.

[Johnson et al. \(2006\)](#) report that 11.7% of all queries in two 2002 web query logs contained phrase queries, though [Kammenhuber et al. \(2006\)](#) report only 3% phrase queries for a different data set. [Silverstein et al. \(1999\)](#) note that many queries without explicit phrase operators are actually implicit phrase searches.

Online edition (c) 2009 Cambridge UP

# Snowball: A language for stemming algorithms

## Links

[Snowball main page](#)

[Porter stemmer page](#)

M.F. Porter  
October 2001

## Summary

Algorithmic stemmers continue to have great utility in IR, despite the promise of out-performance by dictionary-based stemmers. Nevertheless, there are few algorithmic descriptions of stemmers, and even when they exist they are liable to misinterpretation. Here we look at the ideas underlying stemming, and on this website define a language, Snowball, in which stemmers can be exactly defined, and from which fast stemmer programs in ANSI C or Java can be generated. A range of stemmers is presented in parallel algorithmic and Snowball form, including the original Porter stemmer for English.

## 1 Introduction

There are two main reasons for creating Snowball. One is the lack of readily available stemming algorithms for languages other than English. The other is the consciousness of a certain failure on my part in promoting exact implementations of the stemming algorithm described in (Porter 1980), which has come to be called the Porter stemming algorithm. The first point needs some qualification: a great deal of work has been done on stemmers in a wide range of natural languages, both

in their development and evaluation, (a complete bibliography cannot be attempted here). But it is rare to see a stemmer laid out in an unambiguous algorithmic form from which encodings in C, Java, Perl etc might easily be made. When exact descriptions are attempted, it is often with approaches to stemming that are relatively simple, for example the Latin stemmer of Schinke (Shinke 1996), or the Slovene stemmer of Popovic (Popovic 1990). A more complex, and therefore more characteristic stemmer is the Kraaij-Pohlmann stemmer for Dutch (Kraaij 1994), which is presented as open source code in ANSI C. To extract an algorithmic description of their stemmer from the source code proves to be quite hard.

The disparity between the Porter stemmer definition and many of its purported implementations is much wider than is generally realised in the IR community. Three problems seem to compound: one is a misunderstanding of the meaning of the original algorithm, another is bugs in the encodings, and a third is the almost irresistible urge of programmers to add improvements. For example, a Perl script advertised on the Web as an implementation of the Porter algorithm was tested in October 2001, and it was found that 14 percent of words were stemmed incorrectly when given a large sample vocabulary. Most words of English have very simple endings, so this means that it was effectively getting everything wrong. At certain points on the Web are demonstrations of the Porter stemmer. You type some English into a box and the stemmed words are displayed. These are frequently faulty. (A good test is to type in *agreement*. It should stem to *agreement* — the same word. If it stems to *agreeem* there is an error.) Researchers frequently pick up faulty versions of the stemmer and report that they have applied ‘Porter stemming’, with the result that their experiments are not quite repeatable. Researchers who work on stemming will sometimes give incorrect examples of the behaviour of the Porter stemmer in their published works.

To address all these problems I have tried to develop a rigorous system for defining stemming algorithms. A language, Snowball, has been invented, in which the rules of stemming algorithms can be expressed in a natural way. Snowball is quite small, and can be learned by an experienced programmer in an hour or so. On this website a number of foreign language stemmers is presented (a) in Snowball, and (b) in a less formal English-language description. (b) can be thought of as the program comments for (a). A Snowball compiler translates each Snowball definition into (c) an equivalent program in ANSI C or Java. Finally (d) standard vocabularies of words and their stemmed equivalents are provided for each stemmer. The combination of (a), (b), (c) and (d) can be used to pin down the definition of a stemmer exactly, and it is hoped that Snowball itself will be a useful resource in creating stemmers in the future.

## 2 Some ideas underlying stemming

Work in stemming has produced a number of different approaches, albeit tied together by a number of common assumptions. It is worthwhile looking at some of them to see exactly where Snowball fits into the whole picture.

A point tacitly assumed in almost all of the stemming literature is that stemmers are based upon the written, and not the spoken, form of the language. This is also the assumption here. Historically, grammarians often regarded the written language as the real language and the spoken as a mere derivative form. Almost in reaction, many modern linguists have taken a precisely opposite view (Farmer, 1965 pp 2-3).

A more balanced position is that the two languages are distinct though connected, and require separate treatment. One can in fact imagine parallel stemming algorithms for the spoken language, or rather for the phoneme sequence into which the spoken language is transformed. Stress and intonation could be used as clues for an indexing process in the same way that punctuation and capitalisation are used as clues in the written language. But currently stemmers work on the written language for the good reason that there is so much of it available in machine readable form from which to build our IR systems. Inevitably therefore the stemmers get caught up in accidental details of orthography. In English, removing the *ing* from *rotting* should be followed by undoubling the *tt*, whereas in *rolling* we do not undouble the *ll*. In French, removing the *er* from *ennuyer* should be followed by changing the *y* to *i*, so that the resulting word conflates with *ennui*, and so on.

The idea of stemming is to improve IR performance generally by bringing under one heading variant forms of a word which share a common meaning. Harman (1991) was first to present compelling evidence that it may not do so, when her experiments discovered no significant improvement with the use of stemming. Similarly Lennon (1981) discovered no appreciable difference between different stemmers running on a constant collection. Later work has modified this position however. Krovetz (1995) found significant, although sometimes small, improvements across a range of test collections. What he did discover is that the degree of improvement varies considerably between different collections. These tests were however done on collections in English, and the reasonable assumption of IR researchers has always been that for languages that are more highly inflected than English (and nearly all are), greater improvements will be observed when stemming is applied. My own view is that stemming helps regularise the vocabulary of an IR system, and this leads to advantages that are not easily quantifiable through standard IR experiments. For example, it helps in presenting lists of terms associated with the query back to the IR user in a relevance feedback cycle, which is one of the underlying ideas of the probabilistic model. More will be said on the use of a stemmed vocabulary in section 5.

Stemming is not a concept applicable to all languages. It is not, for example, applicable in Chinese. But to languages of the Indo-European ([\\*](#)) group (and most of the stemmers on this site are for Indo-European languages), a common pattern of word structure does emerge. Assuming words are written left to right, the stem, or root of a word is on the left, and zero or more suffixes may be added on the right. If the root is modified by this process it will normally be at its right hand end. And also prefixes may be added on the left. So *unhappiness* has a prefix *un*, a suffix *ness*, and the *y* of *happy* has become *i* with the addition of the suffix. Usually, prefixes alter meaning radically, so they are best left in place (German and Dutch *ge* is an exception here). But suffixes can, in certain circumstances, be removed. So for example *happy* and *happiness* have closely related meanings, and we may wish to stem both forms to *happy*, or *happi*. Infixes can occur, although rarely: *ge* in German and Dutch, and *zu* in German.

One can make some distinction between *root* and *stem*. Lovins (1968) sees the root as the stem minus any prefixes. But here we will think of the stem as the residue of the stemming process, and the root as the inner word from which the stemmed word derives, so we think of root to some extent in an etymological way. It must be admitted that when you start thinking hard about these concepts *root*, *stem*, *suffix*, *prefix* ... they turn out to be very difficult indeed to define. Nor do definitions, even if we arrive at them, help us much. After all, suffix stripping is a practical aid in IR, not an exercise in linguistics or etymology. This is especially true of the central

concept of *root*. We think of the etymological root of a word as something we can discover with certainty from a dictionary, forgetting that etymology itself is a subject with its own doubts and controversies (Jesperson 1922, Chapter XVI). Indeed, Jesperson goes so far as to say that

‘It is of course impossible to say how great a proportion of the etymologies given in dictionaries should strictly be classed under each of the following heads: (1) certain, (2) probable, (3) possible, (4) improbable, (5) impossible — but I am afraid the first two classes would be the least numerous.’

Here we will simply assume a common sense understanding of the basic idea of stem and suffix, and hope that this proves sufficient for designing and discussing stemming algorithms.

We can separate suffixes out into three basic classes, which will be called *d*-, *i*- and *a*-suffixes.

An *a*-suffix, or *attached* suffix, is a particle word attached to another word. (In the stemming literature they sometimes get referred to as ‘enclitics’.) In Italian, for example, personal pronouns attach to certain verb forms:

mandargli = mandare + *gli* = to send + to him  
 mandarglielo = mandare + *gli* + *lo* = to send + it + to him

*a*-suffixes appear in Italian and Spanish, and also in Portuguese, although in Portuguese they are separated by hyphen from the preceding word, which makes them easy to eliminate.

An *i*-suffix, or *inflectional* suffix, forms part of the basic grammar of a language, and is applicable to all words of a certain grammatical type, with perhaps a small number of exceptions. In English for example, the past of a verb is formed by adding *ed*. Certain modifications may be required in the stem:

fit + *ed* → fitted (double *t*)  
 love + *ed* → loved (drop the final *e* of love)

but otherwise the rule applies in a regular way to all verbs in contemporary English, with about 150 (Palmer, 1965) exceptional forms,

bear beat become begin bend ....  
 bore beat became began bent

A *d*-suffix, or *derivational* suffix, enables a new word, often with a different grammatical category, or with a different sense, to be built from another word. Whether a *d*-suffix can be attached is discovered not from the rules of grammar, but by referring to a dictionary. So in English, *ness* can be added to certain adjectives to form corresponding nouns (*littleness, kindness, foolishness ...*) but not to all adjectives (not for example, to *big, cruel, wise ...*) *d*-suffixes can be used to change meaning, often in rather exotic ways. So in Italian **astro** means a sham form of something else:

medico + **astro** = medicastro = quack doctor  
 poeta + **astro** = poetastro = poetaster

Generally *i*-suffixes follow *d*-suffixes. *i*-suffixes can precede *d*-suffixes, for example *lovingly*, *devotedness*, but such cases are exceptional. To be a little more precise, *d*-suffixes can sometimes be added to participles. *devoted*, used adjectivally, is a participle derived from the verb *devote*, and *ly* can be added to turn the adjective into an adverb, or *ness* to turn it into a noun. The same feature occurs in other Indo-European languages.

Sometimes it is hard to say whether a suffix is a *d*-suffix or *i*-suffix, the comparative and superlative endings *er*, *est* of English for example.

A *d*-suffix can serve more than one function. In English, for example, *ly* standardly turns an adjective into an adverb (*greatly*), but it can also turn a noun into an adjective (*kingly*). In French, *ement* also standardly turns an adjective into an adverb (*grandement*), but it can also turn a verb into a noun (*rapprochement*). (Referring to the French stemmer, this double use is ultimately why *ement* is tested for being in the *RV* rather than the *R2* region of the word being stemmed.)

It is quite common for an *i*-suffix to serve more than one function. In English, *s* can either be (1) a verb ending attached to third person singular forms (*runs*, *sings*), (2) a noun ending indicating the plural (*dogs*, *cats*) or (3) a noun ending indicating the possessive (*boy's*, *girls'*). By an orthographic convention now several hundred years old, the possessive is written with an apostrophe, but nowadays this is frequently omitted in familiar phrases (*a girls school*). (Usage (3) is relatively rare compared with (1) and (2): there are only nine uses of 's in this document.)

Since the normal order of suffixes is *d*, *i* and *a*, we can expect them to be removed from the right in the order *a*, *i* and *d*. Usually we want to remove all *a*- and *i*-suffixes, and some of the *d*-suffixes.

If the stemming process reduces two words to the same stem, they are said to be *conflated*.

### 3 Stemming errors, and the use of dictionaries

One way of thinking of the relation between terms and documents in an IR system is to see the documents as being about concepts, and the terms as words that describe the concepts. Then, of course, one word can cover many concepts, so *pound* can mean a unit of currency, a weight, an enclosure, or a beating. *Pound* is a homonym. And one concept can be described by many words, as with *money*, *capital*, *cash*, *currency*. These words are synonyms. There is a many-many mapping therefore between the set of terms and the set of concepts. Stemming is a process that transforms this mapping to advantage, on the whole reducing the number of synonyms, but occasionally creating new homonyms. It is worth remembering that what are called stemming errors are usually just the introduction of new homonyms into vocabularies that already contain very large numbers of homonyms.

Words which have no place in this term-concept mapping are those which describe no concepts. The particle words of grammar, *the*, *of*, *and* ..., known in IR as *stopwords*, fall into this category. Stopwords can be useful for retrieval but only in searching for phrases, '*to be or not to be*', '*do as you would be done by*' etc. This suggests that stemming stopwords is not useful. More will be said on stopwords in

## section 7.

In the literature, a distinction is often made between under-stemming, which is the error of taking off too small a suffix, and over-stemming, which is the error of taking off too much. In French, for example, *croûtons* is the plural of *croûton*, ‘a crust’, so to remove *ons* would be over-stemming, while *croulons* is a verb form of *crouler*, ‘to totter’, so to remove *s* would be under-stemming. We would like to introduce a further distinction between mis-stemming and over-stemming. Mis-stemming is taking off what looks like an ending, but is really part of the stem. Over-stemming is taking off a true ending which results in the conflation of words of different meanings.

So for example *ly* can be removed from *cheaply*, but not from *reply*, because in *reply ly* is not a suffix. If it was removed, *reply* would conflate with *rep*, (the commonly used short form of *representative*). Here we have a case of mis-stemming.

To illustrate over-stemming, look at these four words,

### **verb adjective**

First pair: prove provable

Second pair: probe probable

Morphologically, the two pairs are exactly parallel (in the written, if not the spoken language). They also have a common etymology. All four words derive from the Latin *probare*, ‘to prove or to test’, and the idea of testing connects the meanings of the words. But the meanings are not parallel. *provable* means ‘able to be proved’; *probable* does not mean ‘able to be probed’. Most people would judge conflation of the first pair as correct, and of the second pair, incorrect. In other words, to remove *able* from *probable* is a case of over-stemming.

We can try to avoid mis-stemming and over-stemming by using a dictionary. The dictionary can tell us that *reply* does not derive from *rep*, and that the meanings of *probe* and *probable* are well separated in modern English. It is important to realise however that a dictionary does not give a complete solution here, but can be a tool to improve the conflation process.

In Krovetz’s dictionary experiments (Krovetz 1995), he noted that in looking up a past participle like *suitied*, one is led either to *suit* or to *suite* as plausible infinitive forms. *suite* can be rejected, however, because the dictionary tells us that although it is a word of English it is not a verb form. Cases like this (and Krovetz found about 60) had to be treated as exceptions. But the form *routed* could either derive from the verb *rout* or the verb *route*:

At Waterloo Napoleon’s forces were routed  
The cars were routed off the motorway

Such cases in English are extremely rare, but they are commoner in more highly inflected languages. In French for example, *affiliez* can either be the verb *affiler*, to sharpen, with imperfect ending *iez*, or the verb *affilier*, to affiliate, with present indicative ending *ez*:

vous affiliez = vous affil-iez = you sharpened  
vous affiliez = vous affili-ez = you affiliate

If the second is intended, removal of *iez* is mis-stemming.

With over-stemming we must rely upon the dictionary to separate meanings. There are different ways of doing this, but all involve some degree of reliance upon the lexicographers. Krovetz's methods are no doubt best, because the most objective: he uses several measures, but they are based on the idea of measuring the similarity in meaning of two words by the degree of overlap among the words used to define them, and this is at a good remove from a lexicographer's subjective judgement about semantic similarity.

There is an interesting difference between mis-stemming and over-stemming to do with language history. The morphology of a language changes less rapidly than the meanings of the words in it. When extended to include a few archaic endings, such as *ick* as an alternative to *ic*, a stemmer for contemporary English can be applied to the English of 300 years ago. Mis-stemmings will be roughly the same, but the pattern of over-stemming will be different because of the changing meaning of words in the language. For example, *relativity* in the 19th century merely meant 'the condition of being relative to'. With that meaning, it is acceptable to conflate it with *relative*. But with the 20th century meaning brought to it by Einstein, stemming to *relativ* is over-stemming. Here we see the word with the suffix changing its meaning, but it can happen the other way round. *transpire* has come to mean 'happen', and its old meaning of 'exhalation' or 'breathing out' is now effectively lost. (That is the bitter reality, although dictionaries still try to persuade us otherwise). But *transpiration* still carries the earlier meaning. So what was formerly an acceptable stemming may be judged now as an over-stemming, not because the word being stemmed has changed its meaning, but because some cognate word has changed its meaning.

In these examples we are presenting words as if they had single meanings, but the true picture is more complicated. Krovetz uses a model of word meanings which is extremely helpful here. He makes a distinction between *homonyms* and *polysemes*. The meaning of homonyms are quite unrelated. For example, *ground* in the sense of 'earth', and 'ground' as the past participle of 'grind' are homonyms. Etymologically homonyms have different stories, and they usually have separate entries in a dictionary. But each homonym form can have a range of polysemic forms, corresponding to different shades of meaning. So *ground* can mean the earth's surface, or the bottom of the sea, or soil, or any base, and so the basis of an argument, and so on. Over time new polysemes appear and old ones die. At any moment, the use of a word will be common in some polysemic forms and rare in others. If a suffix is attached to a word the new word will get a different set of polysemes. For example, *grounds* = *ground* + *s* acquires the sense of 'dregs' and 'estate lands', loses the sense of 'earth', and shares the sense of 'basis'. Consider the conflation of *mobility* with *mobile*. *mobile* has acquired two new polysemes not shared with *mobility*. One is the 'mobile art object', common in the nursery. This arrived in the 1960s, and is still in use. The other is the 'mobile phone' which is now very dominant, although it may decline in the future when it has been replaced by some new gadget with a different name. We might draw a graph of the degree of separation of the meanings of *mobility* and *mobile* against time, which would depend upon the number of polysemes and the intensity of their use. What seemed like a valid conflation of the two words in 1940 may seem to be invalid today.

In general therefore one can say that judgements about whether words are over-stemmed change with time as the meanings of words in the language change.

The use of a dictionary should reduce errors of mis-stemming and errors of over-stemming. And, for English at least, the mis-stemming errors should reduce well, even if there are problems with over-stemming errors. Of course, it depends on the quality of the dictionary. A dictionary will need to be very comprehensive, fully up-to-date, and with good word definitions to achieve the best results.

Historically, stemmers have often been thought of as either dictionary-based or algorithmic. The presentation of studies of stemming in the literature has perhaps helped to create this division. In the Lovins' stemmer the algorithmic description is central. In accounts of dictionary-based stemmers the emphasis tends to be on dictionary content and structure, and IR effectiveness. Savoy's French stemmer (Savoy, 1993) is a good example of this. But the two approaches are not really distinct. An algorithmic stemmer can include long exception lists that are effectively mini-dictionaries, and a dictionary-based stemmer usually needs a process for removing at least *i*-suffixes to make the look-up in the dictionary possible. In fact in a language in which proper names are inflected (Latin, Finnish, Russian ...), a dictionary-based stemmer will need to remove *i*-suffixes independently of dictionary look-up, because the proper names will not of course be in the dictionary.

The stemmers available on the Snowball website are all purely algorithmic. They can be extended to include built-in exception lists, they could be used in combination with a full dictionary, but they are still presented here in their simplest possible form. Being purely algorithmic, they are, or ought to be, inferior to the performance of well-constructed dictionary-based stemmers. But they are still very useful, for the following reasons:

- 1) Algorithmic stemmers are (or can be made) very lean and very fast. The stemmers presented here generate code that will process about a million words in six seconds on a conventional 500MHz PC. Nowadays we can generate very large IR systems with quite modest resources, and tools that assist in this have value.
- 2) Despite the errors they can be seen to make, algorithmic stemmers still give good practical results. As Krovetz (1995) says in surprise of the algorithmic stemmer, 'Why does it do so well?' (page 89).
- 3) Dictionary-based stemmers require dictionary maintenance, to keep up with an ever-changing language, and this is actually quite a problem. It is not just that a dictionary created to assist stemming today will probably require major updating in a few years time, but that a dictionary in use for this purpose today may already be several years out of date.

We can hazard an answer to Krovetz's question, as to why algorithmic stemmers perform as well as they do, when they reveal so many cases of under-, over- and mis-stemming. Under-stemming is a fault, but by itself it will not degrade the performance of an IR system. Because of under-stemming words may fail to conflate that ought to have conflated, but you are, in a sense, no worse off than you were before. Mis-stemming is more serious, but again mis-stemming does not really matter unless it leads to false conflations, and that frequently does not happen. For example, removing the *ate* ending in English, can result in useful conflations (*luxury, luxuriate; affection, affectionate*), but very often produces stems that are not English words (*enerv-ate, accommod-ate, deliber-ate* etc). In the literature, these are normally classed as stemming errors — overstemming — although in our nomenclature they are examples of mis-stemming. However these

residual stems, *enerv*, *accommode*, *deliber* ... do not conflate with other word forms, and so behave in an IR system in the same way as if they still retained their *ate* ending. No false conflations arise, and so there is no over-stemming here.

To summarise, one can say that just as a word can be over-stemmed but not mis-stemmed (*relativity* → *relative*), so it can be mis-stemmed but not over-stemmed (*enervate* → *enerv*). And, of course, even over-stemming does not matter, if the over-stemmed word falsely conflates with other words that exist in the language, but are not encountered in the IR system which is being used.

Of the three types of error, over-stemming is the most important, and using a dictionary does not eliminate all over-stemmings, but does reduce their incidence.

## 4 Stemming as part of an indexing process

Stemming is part of a composite process of extracting words from text and turning them into index terms in an IR system. Because stemming is somewhat complex and specialised, it is usually studied in isolation. Even so, it cannot really be separated from other aspect of the indexing process:

- 1) What is a word? For indexing purposes, a word in a European language is a sequence of letters bounded by non-letters. But in English, an internal apostrophe does not split a word, although it is not classed as a letter. The treatment of these word boundary characters affects the stemmer. For example, the Kraaij Pohlmann stemmer for Dutch (Kraaij, 1994, 1995) removes hyphen and treats apostrophe as part of the alphabet (so 's, 'tje and 'je are three of their endings). The Dutch stemmer presented here assumes hyphen and apostrophe have already been removed from the word to be stemmed.
- 2) What is a letter? Clearly letters define words, but different languages use different letters, much confusion coming from the varied use of accented Roman letters.

English speakers, perhaps influenced by the ASCII character set, typically regard their alphabet of *a* to *z* as the norm, and other forms (for example, Danish *å* and *ø*, or German *ß*) as somewhat abnormal. But this is an insular point of view. In Italian, for example, the letters *j*, *k*, *w*, *x* and *y* are not part of the alphabet, and are only seen in foreign words. We also tend to regard other alphabets as only used for isolated languages, and that is not strictly true. Cyrillic is used for a range of languages other than Russian, among which additional letters and accented forms abound.

In English, a broad definition of letter would be anything that could be accepted as a pronounceable element of a word. This would include accented Roman letters (*naïve*, *Fauré*), and certain ligature forms (*encyclopaedia*). It would exclude letters of foreign alphabets, such as Greek and Cyrillic. The *a* to *z* alphabet is one of those where letters come in two styles, upper and lower case, which historically correspond (very roughly) to the shapes you get if you use a chisel or a pen. Across all languages, the exact relation of upper to lower case is not so easy to define. In Italian, for example, an accented lower case letter is sometimes represented in upper case by an unaccented letter followed by an apostrophe. (I have seen this convention used in modern Italian news stories in machine readable form.)

In fact the Porter stemmer (which is for English) assumes the word being stemmed is unaccented and in lower case. More exactly, *a, e, i, o, u*, and sometimes *y*, are treated as vowels, and any other character gets treated as a consonant. Each stemmer presented here assumes some degree of normalisation before it receives the word, which is roughly (a) put all letters into lower case, and (b) remove accents from letter-accent combinations that do not form part of the alphabet of the language. Each stemmer declares the letter-accent combinations for its language, and this can be used as a guide for the normalisation, but even so, we can see from the discussion above that (a) and (b) are not trivial operations, and need to be done with care.

(Incidentally, because the stemmers work on lower case words, turning letters to upper case is sometimes used internally for flagging purposes.)

3) Identifying stopwords. Invariant stopwords are more easily found before stemming is applied, but inflecting stopwords (for example, German *kein, keine, keinem, keinen* ...) may be easier to find after — because there are fewer forms. There is a case for building stopword identification into the stemming process. See section 7.

4) Conflating irregular forms. More will be said on this in section 6.

## 5 The use of stemmed words

The idea of how stemmed words might be employed in an IR system has evolved slightly over the years. The Lovins stemmer (Lovins 1968) was developed not for indexing document texts, but the subject terms attached to them. With queries stemmed in the same way, the user needed no special knowledge of the form of the subject terms. Rijksen (1979, Chapter 2) assumes document text analysis: stopwords are removed, the remaining words are stemmed, and the resulting set of stemmed word constitute the IR index (and this style of use is widespread today). More flexibility however is obtained by indexing *all* words in a text in an unstemmed form, and keeping a separate two-column relation which connects the words to their stemmed equivalents. The relation can be denoted by  $R(s, w)$ , which means that  $s$  is the stemmed form of word  $w$ . From the relation we can get, for any word  $w$ , its unique stemmed form,  $stem(w)$ , and for any stem  $s$ , the set of words,  $words(s)$ , that stem to  $s$ .

The user should not have to see the stemmed form of a word. If a list of stems is to be presented back for query expansion, in place of a stem,  $s$ , the user should be shown a single representative from the set  $words(s)$ , the one of highest frequency perhaps. The user should also be able to choose for the whole query, or at a lower level for each word in a query, whether or not it should be stemmed. In the absence of such choices, the system can make its own decisions. Perhaps single word queries would not undergo stemming; long queries would; stopwords would be removed except in phrases. In query expansion, the system would work with stemmed forms, ignoring stopwords.

Query expansion with stemming results in a much cleaner vocabulary list than without, and this is a main strength of using a stemming process.

A question arises: if the user never sees the stemmed form, does its appearance matter? The answer must be no, although the Porter stemmer tries to make the unstemmed forms guessable from the stemmed forms. For example, from *appropriate* you can guess *appropriate*. At least, trying to achieve this effect acts as a useful control. Similarly with the other stemmers presented here, an attempt has been made to keep the appearance of the stemmed forms as familiar as possible.

## 6 Irregular grammatical forms

All languages contain irregularities, but to what extent should they be accommodated in a stemming algorithm? An English stemmer, for example, can convert regular plurals to singular form without difficulty (*boys, girls, hands* ...). Should it do the same with irregular plurals (*men, children, feet*, ...)? Here we have irregular cases with *i*-suffixes, but there are irregularities with *d*-suffixes, which Lovins calls ‘spelling exceptions’. *absorb/absorption* and *conceive/conception* are examples of this. Etymologically, the explanation of the first is that the Latin root, *sorberē*, is an irregular verb, and of the second that the word *conceive* comes to us from the French rather than straight from the Latin. It is interesting that, even with no knowledge of the etymology, we do recognise the connection between the words.

Lovins tries to solve spelling exceptions by formulating general respelling rules (turn *rpt* into *rb* for example), but it might be easier to have simply a list of exceptional stems.

The Porter stemmer does not handle irregularities at all, but from the author’s own experience, this has never been an area of complaint. Complaints in fact are always about false conflations, for example *new* and *news*.

Possibly Lovins was right in wanting to resolve *d*-suffix irregularities, and not being concerned about *i*-suffix irregularities. *i*-suffix irregularities in English go with short, old words, that are either in very common use (*man/men, woman/women, see/saw* ...) or are used only rarely (*ox/oxen, louse/lice, forsake/forsook* ...). The latter class can be ignored, and the former has its own problems which are not always solved by stemming. For example *man* is a verb, and *saw* can mean a cutting instrument, or, as a verb, can mean to use such an instrument. Conflation of these forms frequently leads to an error like mis-stemming therefore.

An algorithmic stemmer really needs holes where the irregular forms can be plugged in as necessary. This is more serviceable than attempting to embed special lists of these irregular forms into software.

## 7 Stopwords

We have suggested that stemming stopwords is not useful. There is a grammatical connection between *being* and *be*, but conflation of the two forms has little use in IR because they have no shared meaning that would entitle us to think of them as synonyms. *being* and *be* have a morphological connection as well, but that is not true of *am* and *was*, although they have a grammatical connection. Generally

speaking, inflectional stopwords exhibit many irregularities, which means that stemming is not only not useful, but not possible, unless one builds into the stemmer tables of exceptions.

Switching from English to French, consider *être*, the equivalent form of *be*. It has about 40 different forms, including,

*suis es sommes serez étaient fus furent sois été*

(and *suis* incidentally is a homonym, as part of the verb *suivre*.) Passing all forms through a rule-based stemmer creates something of a mess. An alternative approach is to recognise this group of words, and other groups, and take special action. The recognition could take place inside the stemmer, or be done before the stemmer is called. One special action would be to stem (perhaps one should say ‘map’) all the forms to a standard form, ETRE, to indicate that they are parts of the verb *être*. Deciding what to do with the term ETRE, and it would probably be to discard it, would be done outside the stemming process. Another special action would be to recognize a whole class of stopwords and simply discard them.

The strategy adopted will depend upon the underlying IR model, so what one needs is the flexibility to create modified forms of a standard stemmer. Usually we present Snowball stemmers in their unadorned form. Thereafter, the addition of stopword tables is quite easy.

## 8 Rare forms

Stemmers do not need to handle linguistic forms that turn up only very rarely, but in practice it is hard to design a stemmer with all rare forms eliminated without there appearing to be some gaps in the thinking. For this reason one should not worry too much about their occasional presence. For example, in contemporary Portuguese, use of the second person plural form of verbs has almost completely disappeared. Even so, endings for those forms are included in the Portuguese stemmer. They appear in all the grammar books, and will in any case be found in older texts. The habit of putting in rare forms to ‘complete the picture’ is well established, and usually passes unnoticed. An example is the list of English stopwords in van Rijsbergen (1979). This includes *yourselves*, by analogy with *himself, herself* etc., although *yourselves* is actually quite a rare word in English.

## References

Farber DJ, Griswold RE and Polonsky IP (1964) SNOBOL, a string manipulation language. *Journal of the Association for Computing Machinery*, **11**: 21-30.

Griswold RE, Poage JF and Polonsky IP (1968) *The SNOBOL4 programming language*. Prentice-Hall, New Jersey.

Harman D (1991) How effective is suffixing? *Journal of the American Society for Information Science*, **42**: 7-15.

Jesperson O (1921) *Language, its nature, origin and development*. George Allen & Unwin, London.

Kraaij W and Pohlmann R. (1994) Porter's stemming algorithm for Dutch. In Noordman LGM and de Vroomen WAM, eds. *Informatiewetenschap 1994: Wetenschappelijke bijdragen aan de derde STINFON Conferentie*, Tilburg, 1994. pp. 167-180.

Kraaij W and Pohlmann R (1995) Evaluation of a Dutch stemming algorithm. Rowley J, ed. *The New Review of Document and Text Management*, volume 1, Taylor Graham, London, 1995. pp. 25-43,

Krovetz B (1995) *Word sense disambiguation for large text databases*. PhD Thesis. Department of Computer Science, University of Massachusetts Amherst.

Lennon M, Pierce DS, Tarry BD and Willett P (1981) An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science*, **3**: 177-183.

Lovins JB (1968) Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, **11**: 22-31.

Palmer FR (1965) *A linguistic study of the English verb*. Longmans, London.

Popovic M and Willett P (1990) Processing of documents and queries in a Slovene language free text retrieval system. *Literary and Linguistic Computing*, **5**: 182-190.

Porter MF (1980) An algorithm for suffix stripping. *Program*, **14**: 130-137.

Rijsbergen CJ (1979) *Information retrieval*. Second edition. Butterworths, London.

Savoy J (1993) Stemming of French words based on grammatical categories. *Journal of the American Society for Information Science*, **44**: 1-9.

Schinke R, Greengrass M, Robertson AM and Willett P (1996) A stemming algorithm for Latin text databases. *Journal of Documentation*, **52**: 172-187.



## Text Analysis in R

Kasper Welbers<sup>a</sup>, Wouter Van Atteveldt<sup>b</sup>, and Kenneth Benoit <sup>c</sup>

<sup>a</sup>Institute for Media Studies, University of Leuven, Leuven, Belgium; <sup>b</sup>Department of Communication Science, VU University Amsterdam, Amsterdam, The Netherlands; <sup>c</sup>Department of Methodology, London School of Economics and Political Science, London, UK

### ABSTRACT

Computational text analysis has become an exciting research field with many applications in communication research. It can be a difficult method to apply, however, because it requires knowledge of various techniques, and the software required to perform most of these techniques is not readily available in common statistical software packages. In this teacher's corner, we address these barriers by providing an overview of general steps and operations in a computational text analysis project, and demonstrate how each step can be performed using the R statistical software. As a popular open-source platform, R has an extensive user community that develops and maintains a wide range of text analysis packages. We show that these packages make it easy to perform advanced text analytics.

With the increasing importance of computational text analysis in communication research (Boumans & Trilling, 2016; Grimmer & Stewart, 2013), many researchers face the challenge of learning how to use advanced software that enables this type of analysis. Currently, one of the most popular environments for computational methods and the emerging field of “data science”<sup>1</sup> is the R statistical software (R Core Team, 2017). However, for researchers that are not well-versed in programming, learning how to use R can be a challenge, and performing text analysis in particular can seem daunting. In this teacher's corner, we show that performing text analysis in R is not as hard as some might fear. We provide a step-by-step introduction into the use of common techniques, with the aim of helping researchers get acquainted with computational text analysis in general, as well as getting a start at performing advanced text analysis studies in R.

R is a free, open-source, cross-platform programming environment. In contrast to most programming languages, R was specifically designed for statistical analysis, which makes it highly suitable for data science applications. Although the learning curve for programming with R can be steep, especially for people without prior programming experience, the tools now available for carrying out text analysis in R make it easy to perform powerful, cutting-edge text analytics using only a few simple commands. One of the keys to R’s explosive growth (Fox & Leanage, 2016; TIOBE, 2017) has been its densely populated collection of extension software libraries, known in R terminology as *packages*, supplied and maintained by R’s extensive user community. Each package extends the functionality of the base R language and core packages, and in addition to functions and data must include documentation and examples, often in the form of vignettes demonstrating the use of the package. The best-known package repository, the Comprehensive R Archive Network (CRAN), currently has over 10,000 packages that are published, and which have gone through an extensive

**CONTACT** Kasper Welbers  [kasperwelbers@gmail.com](mailto:kasperwelbers@gmail.com)  Institute for Media Studies, University of Leuven, Sint-Andriesstraat 2 – box 15530, Antwerp 2000, Belgium.

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/hcms](http://www.tandfonline.com/hcms).

<sup>1</sup>The term “data science” is a popular buzzword related to “data-driven research” and “big data” (Provost & Fawcett, 2013).

© 2017 Taylor & Francis Group, LLC

screening for procedural conformity and cross-platform compatibility before being accepted by the archive.<sup>2</sup> R thus features a wide range of inter-compatible packages, maintained and continuously updated by scholars, practitioners, and projects such as RStudio and rOpenSci. Furthermore, these packages may be installed easily and safely from within the R environment using a single command. R thus provides a solid bridge for developers and users of new analysis tools to meet, making it a very suitable programming environment for scientific collaboration.

Text analysis in particular has become well established in R. There is a vast collection of dedicated text processing and text analysis packages, from low-level string operations (Gagolewski, 2017) to advanced text modeling techniques such as fitting Latent Dirichlet Allocation models (Blei, Ng, & Jordan, 2003; Roberts et al., 2014) — nearly 50 packages in total at our last count. Furthermore, there is an increasing effort among developers to cooperate and coordinate, such as the rOpenSci special interest group.<sup>3</sup> One of the main advantages of performing text analysis in R is that it is often possible, and relatively easy, to switch between different packages or to combine them. Recent efforts among the R text analysis developers' community are designed to promote this interoperability to maximize flexibility and choice among users.<sup>4</sup> As a result, learning the basics for text analysis in R provides access to a wide range of advanced text analysis features.

### Structure of this Teacher's Corner

This teacher's corner covers the most common steps for performing text analysis in R, from data preparation to analysis, and provides easy to replicate example code to perform each step. The example code is also digitally available in our online appendix, which is updated over time.<sup>5</sup> We focus primarily on bag-of-words text analysis approaches, meaning that only the frequencies of words per text are used and word positions are ignored. Although this drastically simplifies text content, research and many real-world applications show that word frequencies alone contain sufficient information for many types of analysis (Grimmer & Stewart, 2013).

Table 1 presents an overview of the text analysis operations that we address, categorized in three sections. In the *data preparation* section we discuss five steps to prepare texts for analysis. The first step, *importing text*, covers the functions for reading texts from various types of file formats (e.g., txt, csv, pdf) into a *raw text* corpus in R. The steps *string operations* and *preprocessing* cover techniques for manipulating raw texts and processing them into *tokens* (i.e., units of text, such as words or word stems). The tokens are then used for creating the *document-term matrix* (DTM), which is a common format for representing a bag-of-words type corpus, that is used by many R text analysis packages. Other non-bag-of-words formats, such as the tokenlist, are briefly touched upon in the *advanced topics* section. Finally, it is a common step to *filter and weight* the terms in the DTM. These steps are generally performed in the presented sequential order (see Figure 1 for conceptual illustration). As we will show, there are R packages that provide convenient functions that manage multiple data preparation steps in a single line of code. Still, we first discuss and demonstrate each step separately to provide a basic understanding of the purpose of each step, the choices that can be made and the pitfalls to watch out for.

The *analysis* section discusses four text analysis methods that have become popular in communication research (Boumans & Trilling, 2016) and that can be performed with a DTM as input. Rather than being competing approaches, these methods have different advantages and disadvantages, so choosing the best method for a study depends largely on the research question, and

<sup>2</sup>Other programming environments have similar archives, such as pip for python. However, CRAN excels in how it is strictly maintained, with elaborate checks that packages need to pass before they will be accepted.

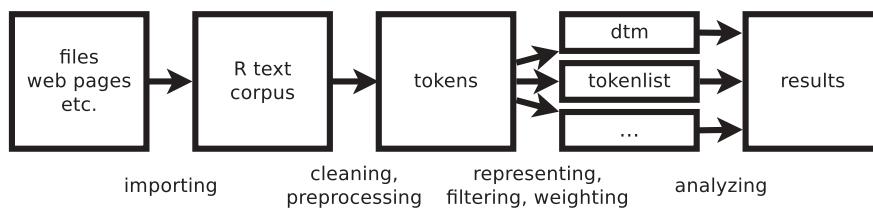
<sup>3</sup>The London School of Economics and Political Science recently hosted a workshop (<http://textworkshop17.ropensci.org/>), forming the beginnings of an rOpenSci special interest group for text analysis.

<sup>4</sup>For example, the tif (Text Interchange Formats) package (rOpenSci Text Workshop, 2017) describes and validates standards for common text data formats.

<sup>5</sup>[https://github.com/kasperwelbers/text\\_analysis\\_in\\_R](https://github.com/kasperwelbers/text_analysis_in_R).

**Table 1.** An overview of text analysis operations, with the R packages used in this Teacher's Corner.

Operation	example	R packages
		alternatives
<b>Data preparation</b>		
importing text	<i>readtext</i>	<i>jsonlite, XML, antiword, readxl, pdftools</i>
string operations	<i>stringi</i>	<i>stringr</i>
preprocessing	<i>quanteda</i>	<i>stringi, tokenizers, snowballC, tm, etc.</i>
document-term matrix (DTM)	<i>quanteda</i>	<i>tm, tidytext, Matrix</i>
filtering and weighting	<i>quanteda</i>	<i>tm, tidytext, Matrix</i>
<b>Analysis</b>		
dictionary	<i>quanteda</i>	<i>tm, tidytext, koRpus, corpustools</i>
supervised machine learning	<i>quanteda</i>	<i>RTextTools, kerasR, austin</i>
unsupervised machine learning	<i>topicmodels</i>	<i>quanteda, stm, austin, text2vec</i>
text statistics	<i>quanteda</i>	<i>koRpus, corpustools, textreuse</i>
<b>Advanced topics</b>		
advanced NLP	<i>spacyr</i>	<i>coreNLP, cleanNLP, koRpus</i>
word positions and syntax	<i>corpustools</i>	<i>quanteda, tidytext, koRpus</i>

**Figure 1.** Order of text analysis operations for data preparation and analysis.

sometimes different methods can be used complementarily (Grimmer & Stewart, 2013). Accordingly, our recommendation is to become familiar with each type of method. To demonstrate the general idea of each type of method, we provide code for typical analysis examples. Furthermore, it is important to note that different types of analysis can also have different implications for how the data should be prepared. For each type of analysis we therefore address general considerations for data preparation.

Finally, the additional *advanced topics* section discusses alternatives for data preparation and analysis that require external software modules or that go beyond the bag-of-words assumption, using word positions and syntactic relations. The purpose of this section is to provide a glimpse of alternatives that are possible in R, but might be more difficult to use.

Within each category we distinguish several groups of operations, and for each operation we demonstrate how they can be implemented in R. To provide parsimonious and easy to replicate examples, we have chosen a specific selection of packages that are easy to use and broadly applicable. However, there are many alternative packages in R that can perform the same or similar operations. Due to the open-source nature of R, different people from often different disciplines have worked on similar problems, creating some duplication in functionality across different packages. This also offers a range of choice, however, providing alternatives to suit a user's needs and tastes. Depending on the research project, as well as personal preference, other packages might be better suited to different readers. While a fully comprehensive review and comparison of text analysis packages for R is beyond our scope here—especially given that existing and new packages are constantly being developed—we have tried to cover, or at least mention, a variety of alternative packages for each text analysis operation.<sup>6</sup> In general, these

<sup>6</sup>For a list that includes more packages, and that is also maintained over time, a good source is the CRAN Task View for Natural Language Processing (Wild, 2017). CRAN Task Views are expert curated and maintained lists of R packages on the Comprehensive R Archive Network, and are available for various major methodological topics.

packages often use the same standards for data formats, and thus are easy to substitute or combine with the other packages discussed in this teacher's corner.

## Data preparation

Data preparation is the starting point for any data analysis. Not only is computational text analysis no different in this regard, but also frequently presents special challenges for data preparation that can be daunting for novice and advanced practitioners alike. Furthermore, preparing texts for analysis requires making choices that can affect the accuracy, validity, and findings of a text analysis study as much as the techniques used for the analysis (Crone, Lessmann, & Stahlbock, 2006; Günther & Quandt, 2016; Leopold & Kindermann, 2002). Here we distinguish five general steps: importing text, string operations, preprocessing, creating a document-term matrix (DTM), and filtering and weighting the DTM.

### Importing text

Getting text into R is the first step in any R-based text analytic project. Textual data can be stored in a wide variety of file formats. R natively supports reading regular flat text files such as CSV and TXT, but additional packages are required for processing formatted text files such as JSON (Ooms, 2014), HTML, and XML (Lang & the CRAN Team, 2017), and for reading complex file formats such as Word (Ooms, 2017a), Excel (Wickham & Bryan, 2017) and PDF (Ooms, 2017b). Working with these different packages and their different interfaces and output can be challenging, especially if different file formats are used together in the same project. A convenient solution for this problem is the *readtext* package, that wraps various import packages together to offer a single catch-all function for importing many types of data in a uniform format. The following lines of code illustrate how to read a CSV file with the *readtext* function, by providing the path to the file as the main argument (the path can also be an URL, as used in our example; online appendix with copyable code available from [https://github.com/kasperwelbers/text\\_analysis\\_in\\_R](https://github.com/kasperwelbers/text_analysis_in_R)).

```
install.packages("readtext")
library(readtext)

## url to Inaugural Address demo data that is provided by the readtext package
filepath <- "http://bit.ly/2uhqjJE?.csv"

rt <- readtext(filepath, text_field = "texts")
rt
```

```
readtext object consisting of 5 documents and 3 docvars.
# data.frame [5 x 5]
  doc_id          text      Year President FirstName
  <chr>          <chr>    <int>   <chr>    <chr>
1 2uhqjJE?.csv.1 "\"Fellow-Cit\"..." 1789 Washington George
2 2uhqjJE?.csv.2 "\"Fellow cit\"..." 1793 Washington George
3 2uhqjJE?.csv.3 "\"When it wa\"..." 1797 Adams      John
4 2uhqjJE?.csv.4 "\"Friends an\"..." 1801 Jefferson Thomas
5 2uhqjJE?.csv.5 "\"Proceeding\"..." 1805 Jefferson Thomas
```

The same function can be used for importing all formats mentioned above, and the path can also reference a (zip) folder to read all files within. In most cases, the only thing that has to be specified is

the name of the field that contains the texts. Not only can multiple files be references using simple, “glob”-style pattern matches, such as `~/myfiles/* .txt`, but also the same command will recurse through sub-directories to locate these files. Each file is automatically imported according to its format, making it very easy to import and work with data from different input file types.

Another important consideration is that texts can be represented with different character encodings. Digital text requires binary code to be mapped to semantically meaningful characters, but many different such mappings exist, with widely different methods of encoding “extended” characters, including letters with diacritical marks, special symbols, and emoji. In order to be able to map all known characters to a single scheme, the Unicode standard was proposed, although it also requires a digital encoding format (such as the UTF-8 format, but also UTF-16 or UTF-32). Our recommendation is simple: in R, ensure that all texts are encoded as UTF-8, either by reading in UTF-8 texts, or converting them from a known encoding upon import. If the encoding is unknown, `readtext`’s `encoding` function can be used to guess the encoding. `readtext` can convert most known encodings (such as ISO-8859-2 for Central and Eastern European languages, or Windows-1250 for Cyrillic—although there are hundreds of others) into the common UTF-8 standard. R also offers additional low-level tools for converting character encodings, such as a bundled version of the GNU `libiconv` library, or conversion though the `stringi` package.

### **String operations**

One of the core requirements of a framework for computational text analysis is the ability to manipulate digital texts. Digital text is represented as a sequence of characters, called a string. In R, strings are represented as objects called “character” types, which are vectors of strings. The group of string operations refers to the low-level operations for working with textual data. The most common string operations are joining, splitting, and extracting parts of strings (collectively referred to as *parsing*) and the use of *regular expressions* to find or replace patterns.

Although R has numerous built-in functions for working with character objects, we recommend using the `stringi` package (Gagolewski, 2017) instead. Most importantly, because `stringi` uses the International Components for Unicode (ICU) library for proper Unicode support, such as implementing Unicode character categories (such as punctuation or spacing) and Unicode-defined rules for case conversion that work correctly in all languages. An alternative is the `stringr` package, which uses `stringi` as a backend, but has a simpler syntax that many end users will find sufficient for their needs.

It is often unnecessary to perform manual, low-level string operations, because the most important applications of string operations for text analysis are built into common text analysis packages. Nevertheless, access to low-level string operations provides a great deal of versatility, which can be crucial when standardized solutions are not an option for a specific use case. The following example shows how to perform some basic cleaning with `stringi` functions: removing boilerplate content in the form of markup tags, stripping extraneous whitespace, and converting to lower case.

```
library(stringi)
x <- c("The first string", ' The <font size="6">second string</font>')
x <- stri_replace_all(x, "", regex = "<.*?>") ## remove html tags
x <- stri_trim(x) ## strip surrounding whitespace
x <- stri_trans_tolower(x) ## transform to lower case
x
[1] "the first string" "the second string"
```

As with most functions in R, *stringi* operations are *vectorized*, meaning they apply to each element of a vector. Manipulation of vectors of strings is the recommended approach in R, since looping over each element and processing it separately in R is very inefficient.

### **Preprocessing**

For most computational text analysis methods, full texts must be *tokenized* into smaller, more specific text features, such as words or word combinations. Also, the computational performance and accuracy of many text analysis techniques can be improved by normalizing features, or by removing “stopwords”: words designated in advance to be of no interest, and which are therefore discarded prior to analysis. Taken together, these preparatory steps are commonly referred to as “preprocessing”. Here we first discuss several of the most common preprocessing techniques, and show how to perform each technique with the *quanteda* package.

In practice, all of these preprocessing techniques can be applied in one function when creating a document-term matrix, as we will demonstrate in the *DTM* section. Here, we show each step separately to illustrate what each technique does.

#### **Tokenization**

Tokenization is the process of splitting a text into tokens. This is crucial for computational text analysis, because full texts are too specific to perform any meaningful computations with. Most often tokens are words, because these are the most common semantically meaningful components of texts.

For many languages, splitting texts by words can mostly be done with low-level string processing due to clear indicators of word boundaries, such as white spaces, dots and commas. A good tokenizer, however, must also be able to handle certain exceptions, such as the period in the title “Dr.”, which can be confused for a sentence boundary. Furthermore, tokenization is more difficult for languages where words are not clearly separated by white spaces, such as Chinese and Japanese. To deal with these cases, some tokenizers include dictionaries of patterns for splitting texts. In R, the *stringi* package is often used for sentence and word disambiguation, for which it leverages dictionaries from the ICU library. There is also a dedicated package for text tokenization, called *tokenizers* (Mullen, 2016b).

The following code uses *quanteda*’s (Benoit et al., 2017) *tokens* function to split a single sentence into words. The *tokens* function returns a list whose elements each contain the tokens of the input texts as a character vector.

```
install.packages("quanteda")
library(quanteda)

text <- "An example of preprocessing techniques"
toks <- tokens(text) ## tokenize into unigrams
toks

tokens from 1 document.
text1 :
[1] "An"   "example" "of"    "preprocessing" "techniques"
```

#### **Normalization: Lowercasing and stemming**

The process of normalization broadly refers to the transformation of words into a more uniform form. This can be important if for a certain analysis a computer has to recognize when two words have (roughly) the same meaning, even if they are written slightly differently. Another advantage is that it reduces the size of the vocabulary (i.e., the full range of features used in the analysis). A simple

but important normalization techniques is to make all text lower case. If we do not perform this transformation, then a computer will not recognize that two words are identical if one of them was capitalized because it occurred at the start of a sentence.

Another argument for normalization is that a base word might have different morphological variations, such as the suffixes from conjugating a verb, or making a noun plural. For purposes of analysis, we might wish to consider these variations as equivalent because of their close semantic relation, and because reducing the feature space is generally desirable when multiple features are in fact closely related. A technique for achieving this is *stemming*, which is essentially a rule-based algorithm that converts inflected forms of words into their base forms (stems). A more advanced technique is *lemmatization*, which uses a dictionary to replace words with their morphological root form. However, lemmatization in R requires external software modules (see the *advanced preprocessing* section for instructions) and for weakly inflected languages such as modern English, stemming is often sufficient. In R, the SnowballC (Bouchet-Valat, 2014; Porter, 2001) package is used in many text analysis packages (such as *quanteda* and *tm*) to implement stemming, and currently supports 15 different languages. Lowercasing and stemming of character, tokens, or feature vectors can be performed in *quanteda* with the `*_tolower` and `*_wordstem` functions, such as `char_tolower` to convert character objects to lower case, or `tokens_wordstem` to stem tokens.

```
toks <- tokens_tolower(toks)
toks <- tokens_wordstem(toks)
toks
[1] "an"    "exampl"   "of"     "preprocess" "techniqu"
```

In this example we see that the difference between “an” and “An” is eliminated due to lowercasing. The words “example” and “techniques” are reduced to “exampl” and “techniqu”, such that any distinction between singular and plural forms is removed.

### **Removing stopwords**

Common words such as “the” in the English language are rarely informative about the content of a text. Filtering these words out has the benefit of reducing the size of the data, reducing computational load, and in some cases also improving accuracy. To remove these words beforehand, they are matched to predefined lists of “stop words” and deleted. Several text analysis packages provide stopword lists for various languages, that can be used to manually filter out stopwords. In *quanteda*, the `stopwords` function returns a character vector of stopwords for a given language. A total of 17 languages are currently supported.

```
sw <- stopwords("english") ## get character vector of stopwords
head(sw) ## show head (first 6) stopwords
[1] "i"    "me"   "my"   "myself"  "we"   "our"
tokens_remove(toks, sw)
text1 :
[1] "exampl"   "preprocess" "techniqu"
```

Care should be taken to perform some preprocessing steps in the correct order, for instance removing stopwords prior to stemming, otherwise “during” will be stemmed into “dure” and not matched to a stopword “during”. Case conversion may also create sequencing issues, although the default stopword matching used by *quanteda* is case-insensitive.

Conveniently, the preprocessing steps discussed above can all be performed with a single function that will automatically apply the correct order of operations. We will demonstrate this in the next section.

### **Document-term matrix**

The document term matrix (DTM) is one of the most common formats for representing a text corpus (i.e. a collection of texts) in a bag-of-words format. A DTM is a matrix in which rows are documents, columns are terms, and cells indicate how often each term occurred in each document. The advantage of this representation is that it allows the data to be analyzed with vector and matrix algebra, effectively moving from text to numbers. Furthermore, with the use of special matrix formats for sparse matrices, text data in a DTM format is very memory efficient and can be analyzed with highly optimized operations.

Two of the most established text analysis packages in R that provide dedicated DTM classes are *tm* and *quanteda*. Of the two, the venerable *tm* package is the more commonly used, with a user base of almost 10 years (Meyer, Hornik, & Feinerer, 2008) and several other R packages using its DTM classes (*DocumentTermMatrix* and *TermDocumentMatrix*) as inputs for their analytic functions. The *quanteda* package is a more recently developed package, built by a team supported by an ERC grant to provide state-of-the-art, high performance text analysis. Its sparse DTM class, known as a *dFM* or *document-feature matrix*, is based on the powerful *Matrix* package (Bates & Maechler, 2015) as a backend, but includes functions to convert to nearly every other sparse document-term matrix used in other R packages (including the *tm* formats). The performance and flexibility of *quanteda*'s *dFM* format lends us to recommend it over the *tm* equivalent.

Another notable alternative is the *tidytext* package (Silge & Robinson, 2016). This is a text analysis package that is part of the *Tidyverse*<sup>7</sup>—a collection of R packages with a common philosophy and format. Central to the Tidyverse philosophy is that all data is arranged as a table, where (1) “each variable forms a column”, (2) “each observation forms a row”, and (3) “each type of observational unit forms a table” (Wickham et al., 2014, p. 4). As such, *tidytext* does not strictly use a document term *matrix*, but instead represents the same data in a long format, where each (non-zero) value of the DTM is a row with the columns document, term, and count (note that this is essentially a triplet format for sparse matrices, with the columns specifying the row, column and value). This format can be less memory efficient and make matrix algebra less easily applicable, but has the advantage of being able to add more variables (e.g., a sentiment score) and enables the use of the entire Tidyverse arsenal. Thus, for users that prefer the tidy data philosophy, *tidytext* can be a good alternative package to *quanteda* or *tm*, although these packages can also be used together quite nicely depending on the particular operations desired.

Consistent with the other examples in this teacher’s corner, we demonstrate the creation of DTMs using the *quanteda* package. Its *dFM* function provides a single line solution for creating a DTM from raw text, that also integrates the preprocessing techniques discussed above. These may also be built up through a sequence of lower-level functions, but many users find it convenient to go straight from a text or corpus to a DTM using this single function.

---

<sup>7</sup><http://www.tidyverse.org/>.

```

text <- c(d1 = "An example of preprocessing techniques",
          d2 = "An additional example",
          d3 = "A third example")
dtm <- dfm(text,
            tolower = TRUE, stem = TRUE,      ## set lowercasing and stemming to TRUE
            remove = stopwords("english"))   ## provide the stopwords for deletion
dtm

Document-feature matrix of: 3 documents, 5 features (53.3% sparse).
3 x 5 sparse Matrix of class "dfmSparse"
   features
docs      exempl preprocess techniqu addit  third
d1           1         1        1     0     0
d2           1         0        0     1     0
d3           1         0        0     0     1

```

The DTM can also be created from a *quanteda* corpus object, which stores text and associated meta-data, including document-level variables. When a corpus is tokenized or converted into a DTM, these document-level variables are saved in the object, which can be very useful later when the documents in the DTM need to be used as covariates in supervised machine learning. The stored document variables also make it possible to aggregate *quanteda* objects by groups, which is extremely useful when texts are stored in small units—like Tweets—but need to be aggregated in a DTM by grouping variables such as users, dates, or combinations of these.

Because *quanteda* is compatible with the *readtext* package, creating a corpus from texts on disk takes only a single additional step. In the following example we create a DTM from the *readtext* data as imported above.

```

fulltext <- corpus(rt)                                ## create quanteda corpus
dtm <- dfm(fulltext, tolower = TRUE, stem = TRUE,    ## create dtm with preprocessing
           remove_punct = TRUE, remove = stopwords("english"))
dtm

Document-feature matrix of: 5 documents, 1,405 features (67.9% sparse).

```

### **Filtering and weighting**

Not all terms are equally informative for text analysis. One way to deal with this is to remove these terms from the DTM. We have already discussed the use of stopword lists to remove very common terms, but there are likely still other common words and this will be different between corpora. Furthermore, it can be useful to remove very rare terms for tasks such as category prediction (Yang & Pedersen, 1997) or topic modeling (Griffiths & Steyvers, 2004). This is especially useful for improving efficiency, because it can greatly reduce the size of the vocabulary (i.e., the number of unique terms), but it can also improve accuracy. A simple but effective method is to filter on document frequencies (the number of documents in which a term occurs), using a threshold for minimum and maximum number (or proportion) of documents (Griffiths & Steyvers, 2004; Yang & Pedersen, 1997).

Instead of removing less informative terms, an alternative approach is assign them variable weights. Many text analysis techniques perform better if terms are weighted to take an estimated information value into account, rather than directly using their occurrence frequency. Given a sufficiently large corpus, we can use information about the distribution of terms in the corpus to estimate this information value. A popular weighting scheme that does so is term frequency-inverse document frequency (*tf-idf*), which down-weights that occur in many documents in the corpus.

Using a document frequency threshold and weighting can easily be performed on a DTM. *quanteda* includes the functions `docfreq`, `tf`, and `tfidf`, for obtaining document frequency, term frequency, and *tf-idf* respectively. Each function has numerous options for implementing the SMART weighting scheme Manning et al. (2008). As a high-level wrapper to these, *quanteda* also provides the `dfm_weight` function. In the example below, the word “senat[e]” has a higher weight than the less informative term “among”, which both occur once in the first document.

```
doc_freq <- docfreq(dtm)      ## document frequency per term (column)
dtm <- dtm[, doc_freq >= 2]    ## select terms with doc_freq >= 2
dtm <- dfm_weight(dtm, "tfidf") ## weight the features using tf-idf
head(dtm)

Document-feature matrix of: 5 documents, 524 features (46.6% sparse).
(showing first 5 documents and first 6 features)
             features
docs      fellow-citizen   senat     hous   repres   :
2uhqjJE?.csv.1 0.2218487 0.39794 0.79588 0.4436975 0.2218487 0.09691001
2uhqjJE?.csv.2 0.0000000 0.00000 0.00000 0.0000000 0.2218487 0.00000000
2uhqjJE?.csv.3 0.6655462 0.39794 1.19382 0.6655462 0.0000000 0.38764005
2uhqjJE?.csv.4 0.4436975 0.00000 0.00000 0.2218487 0.2218487 0.09691001
2uhqjJE?.csv.5 0.0000000 0.00000 0.00000 0.0000000 0.0000000 0.67837009
```

## Analysis

For an overview of text analysis approaches we build on the classification proposed by Boumans and Trilling (2016) in which three approaches are distinguished: *counting and dictionary* methods, *supervised machine learning*, and *unsupervised machine learning*. They position these approaches, in this order, on a dimension from most deductive to most inductive. Deductive, in this scenario, refers to the use of an *a priori* defined coding scheme. In other words, the researchers know beforehand what they are looking for, and only seek to automate this analysis. The relation to the concept of deductive reasoning is that the researcher assumes that certain rules, or premises, are true (e.g., a list of words that indicates positive sentiment) and thus can be applied to draw conclusions about texts. Inductive, in contrast, means here that instead of using an *a priori* coding scheme, the computer algorithm itself somehow extracts meaningful codes from texts. For example, by looking for patterns in the co-occurrence of words and finding latent factors (e.g., topics, frames, authors) that explain these patterns—at least mathematically. In terms of inductive reasoning, it can be said that the algorithm creates broad generalizations based on specific observations.

In addition to these three categories, we also consider a *statistics* category, encompassing all techniques for describing a text or corpus in numbers. Like unsupervised learning, these techniques are inductive in the sense that no *a priori* coding scheme is used, but they do not use machine learning.

For the example code for each type of analysis, we use the Inaugural Addresses of US presidents ( $N = 58$ ) that is included in the *quanteda* package.

```
dtm <- dfm(data_corpus_ inaugural, stem = TRUE, remove = stopwords("english"),
            remove_punct = TRUE)
dtm

Document-feature matrix of: 58 documents, 5,405 features (89.2% sparse).
```

## Counting and dictionary

The dictionary approach broadly refers to the use of patterns—from simple keywords to complex Boolean queries and regular expressions—to count how often certain concepts occur in texts. This is

a deductive approach, because the dictionary defines *a priori* what codes are measured and how, and this is not affected by the data.<sup>8</sup> Using dictionaries is a computationally simple but powerful approach. It has been used to study subjects such as media attention for political actors (Schuck, Xezonakis, Elenbaas, Banducci, & De Vreese, 2011; Vliegenthart, Boomgaarden, & Van Spanje, 2012) and framing in corporate news (Schultz, Kleinnijenhuis, Oegema, Utz, & Van Atteveldt, 2012). Dictionaries are also a popular approach for measuring sentiment (De Smedt & Daelemans, 2012; Mostafa, 2013; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011) as well as other dimensions of subjective language (Tausczik & Pennebaker, 2010). By combining this type of analysis with information from advanced NLP techniques for identifying syntactic clauses, it also becomes possible to perform more fine-grained analyses, such as sentiment expressions attributed to specific actors (Van Atteveldt, 2008), or actions and affections from one actor directed to another (Van Atteveldt, Sheafer, Shenhav, & Fogel-Dror, 2017).

The following example shows how to apply a dictionary to a quanteda DTM. The first step is to create a dictionary object (here called myDict), using the dictionary function. For simplicity our example uses a very simple dictionary, but it is also possible to import large, pre-made dictionaries, including files in other text analysis dictionary formats such as LIWC, Wordstat, and Lexicoder. Dictionaries can also be written and imported from YAML files, and can include patterns of fixed matches, regular expressions, or the simpler “glob” pattern match (using just \* and ? for wildcard characters) common in many dictionary formats. With the dfm\_lookup function, the dictionary object can then be applied on a DTM to create a new DTM in which columns represent the dictionary codes.

```
myDict <- dictionary(list(terror = c("terror*"),
                           economy = c("job*", "business*", "econom*")))
dict_dtm <- dfm_lookup(dtm, myDict, nomatch = "_unmatched")
tail(dict_dtm)
```

Document-feature matrix of: 58 documents, 3 features (37.4% sparse).  
(showing last 6 documents and last 3 features)

docs	features		
	terror	economy	_unmatched
1997-Clinton	2	3	1125
2001-Bush	0	2	782
2005-Bush	0	1	1040
2009-Obama	1	7	1165
2013-Obama	0	6	1030
2017-Trump	1	5	709

### Supervised machine learning

The supervised machine learning approach refers to all classes of techniques in which an algorithm learns patterns from an annotated set of training data. The intuitive idea is that these algorithms can learn how to code texts if we give them enough examples of how they should be coded. A straightforward example is sentiment analysis, using a set of texts that are manually coded as *positive*, *neutral*, or *negative*, based on which the algorithm can learn which features (words or word combinations) are more likely to occur in positive or negative texts. Given an unseen text (from which the algorithm was not trained), the sentiment of the text can then be estimated based on

---

<sup>8</sup>Notably, there are techniques for automatically expanding a dictionary based on the semantic space of a text corpus (see, e.g., Watanabe, 2017). This can be said to add an inductive layer to the approach, because the coding rules (i.e., the dictionary) are to some extent learned from the data.

the presence of these features. The deductive part is that the researchers provide the training data, which contains good examples representing the categories that the researchers are attempting to predict or measure. However, the researchers do not provide explicit rules for how to look for these codes. The inductive part is that the supervised machine learning algorithm learns these rules from the training data. To paraphrase a classic syllogism: if the training data is a list of people that are either mortal or immortal, then the algorithm will learn that all men are extremely likely to be mortal, and thus would estimate that Socrates is mortal as well.

To demonstrate this, we train a model to predict whether an Inaugural Address was given before World War II—which we expect because prominent issues shift over time, and after wars in particular. Some dedicated packages for supervised machine learning are *RTextTools* (Jurka, Collingwood, Boydston, Grossman, & Van Atteveldt, 2014) and *kerasR* (Arnold, 2017b). For this example, however, we use a classifier that is included in *quanteda*. Before we start, we set a custom seed for R’s random number generator so that the results of the random parts of the code are always the same. To prepare the data, we add the document (meta) variable *is\_prewar* to the DTM that indicates which documents predate 1945. This is the variable that our model will try to predict. We then split the DTM into training (*train\_dtm*) and test (*test\_dtm*) data, using a random sample of 40 documents for training and the remaining 18 documents for testing. The training data is used to train a multinomial Naive Bayes classifier (Manning et al., 2008, Ch. 13) which we assign to *nb\_model*. To test how well this model predicts whether an Inaugural Address predates the war, we predict the code for the test data, and make a table in which the rows show the prediction and the columns show the actual value of the *is\_prewar* variable.

```
set.seed(2)
## create a document variable indicating pre or post war
docvars(dtm, "is_prewar") <- docvars(dtm, "Year") < 1945

## sample 40 documents for the training set and use remaining (18) for testing
train_dtm <- dfm_sample(dtm, size = 40)
test_dtm <- dtm[setdiff(docnames(dtm), docnames(train_dtm)), ]

## fit a Naive Bayes multinomial model and use it to predict the test data
nb_model <- textmodel_NB(train_dtm, y = docvars(train_dtm, "is_prewar"))
pred_nb <- predict(nb_model, newdata = test_dtm)

## compare prediction (rows) and actual is_prewar value (columns) in a table
table(prediction = pred_nb$nb.predicted, is_prewar = docvars(test_dtm, "is_prewar"))

  is_prewar
prediction FALSE TRUE
  FALSE     8    0
  TRUE      0   10
```

The results show that the predictions are perfect. Of the eight times that FALSE (i.e. Inaugural Address does not predate the war) was predicted, and the 10 times that TRUE was predicted, this was actually the case.

### **Unsupervised machine learning**

In unsupervised machine learning approaches, no coding rules are specified and no annotated training data is provided. Instead, an algorithm comes up with a model by identifying certain patterns in text. The only influence of the researcher is the specification of certain parameters, such as the number of categories into which documents are classified. Popular examples are topic

modeling for automatically classifying documents based on an underlying topical structure (Blei et al., 2003; Roberts et al., 2014) and the “Wordfish” parametric factor model (Proksch & Slapin, 2009) for scaling documents on a single underlying dimension, such as left-right ideology.

Grimmer and Stewart (2013) argue that supervised and unsupervised machine learning are not competitor methods, but fulfill different purposes and can very well be used to complement each other. Supervised methods are the most suitable approach if documents need to be placed in predetermined categories, because it is unlikely that an unsupervised method will yield a categorization that reflects these categories and how the researcher interprets them. The advantage of the somewhat unpredictable nature of unsupervised methods is that it can come up with categories that the researchers had not considered. (Conversely, this may also present challenges for post-hoc interpretation when results are unclear.)

To demonstrate the essence of unsupervised learning, the example below shows how to fit a topic model in R using the *topicmodels* package (Grun & Hornik, 2011). To focus more specifically on topics within the inaugural addresses, and to increase the number of texts to model, we first split the texts by paragraph and create a new DTM. From this DTM we remove terms with a document frequency of five and lower to reduce the size of the vocabulary (less important for current example) and use *quanteda*'s convert function to convert the DTM to the format used by *topicmodels*. We then train a vanilla LDA topic model (Blei et al., 2003) with five topics—using a fixed seed to make the results reproducible, since LDA is non-deterministic.

```
install.packages("topicmodels")
library(topicmodels)

texts = corpus_reshape(data_corpus_inaugural, to = "paragraphs")

par_dtm <- dfm(texts, stem = TRUE,                      ## create a document-term matrix
                remove_punct = TRUE, remove = stopwords("english"))
par_dtm <- dfm_trim(par_dtm, min_count = 5)      ## remove rare terms
par_dtm <- convert(par_dtm, to = "topicmodels") ## convert to topicmodels format

set.seed(1)
lda_model <- topicmodels::LDA(par_dtm, method = "Gibbs", k = 5)
terms(lda_model, 5)
```

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
[1,]	"govern"	"nation"	"great"	"us"	"shall"
[2,]	"state"	"can"	"war"	"world"	"citizen"
[3,]	"power"	"must"	"secur"	"new"	"peopl"
[4,]	"constitut"	"peopl"	"countri"	"american"	"duti"
[5,]	"law"	"everi"	"unit"	"america"	"countri"

The results show the first five terms of the five topics. Although this is a basic example, the idea of “topics” being found bottom-up from the data can be seen in the semantic coherence of terms within the same topic. In particular, topic one seems to revolve around governance, with the terms “govern[ance]”, “power”, “state”, “constitut[ion]”, and “law”.

### Statistics

Various statistics can be used to describe, explore and analyze a text corpus. An example of a popular technique is to rank the information value of words inside a corpus and then visualize the most informative words as a word cloud to get a quick indication of what a corpus is about. Text statistics (e.g., average word and sentence length, word and syllable counts) are also commonly used as an operationalization of concepts such as readability (Flesch, 1948) or lexical diversity (McCarthy & Jarvis, 2010). A wide range of such measures is available in R with the *koRpus* package (Michalke,

2017). Furthermore, there are many useful applications of calculating term and document similarities (which are often based on the inner product of a DTM or transposed DTM), such as analyzing semantic relations between words or concepts and measuring content homogeneity. Both techniques are supported in *quanteda*, *corpustools* (Welbers & Van Atteveldt, 2016), or dedicated packages such as *textreuse* (Mullen, 2016a) for text overlap.

A particularly useful technique is to compare the term frequencies of two corpora, or between two subsets of the same corpus. For instance, to see which words are more likely to occur in documents about a certain topic. In addition to providing a way to quickly explore how this topic is discussed in the corpus, this can provide input for developing better queries. In the following example we show how to perform this technique in *quanteda*.

```
## create DTM that contains Trump and Obama speeches
corpus_pres = corpus_subset(data_corpus_ inaugural,
                           President %in% c("Obama", "Trump"))
dtm_pres = dfm(corpus_pres, groups = "President",
               remove = stopwords("english"), remove_punct = TRUE)

## compare target (in this case Trump) to rest of DTM (in this case only Obama).
keyness = textstat_keyness(dtm_pres, target = "Trump")
textplot_keyness(keyness)

## results in Figure 2.
```

Here, the signed  $\chi^2$  measure of association indicates that “america”, “american”, and “first” were used with far greater frequency by Trump than Obama, while “us”, “can”, “freedom”, “peace”, and “liberty” were among the words much more likely to be used by Obama than by Trump.

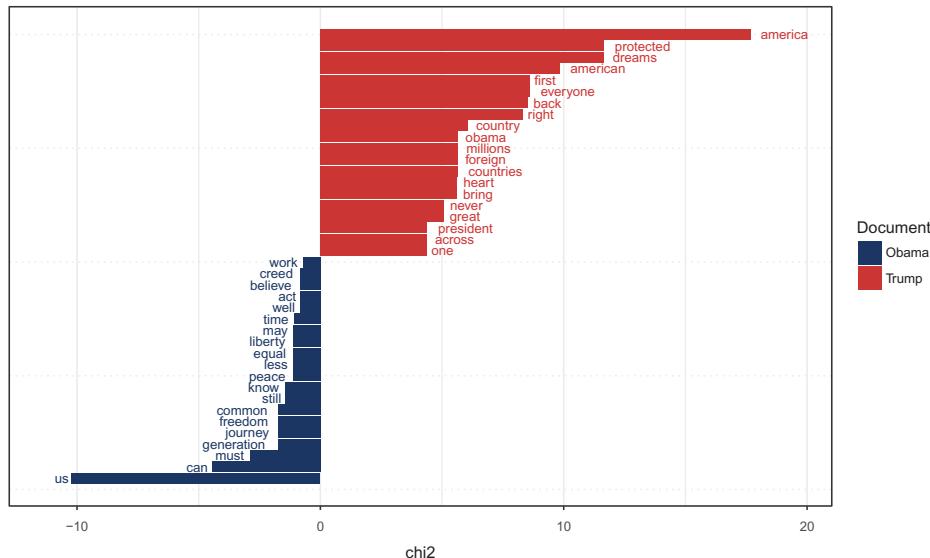
## Advanced topics

The data preparation and bag-of-words analysis techniques discussed above are the basis for the majority of the text analysis approaches that are currently used in communication research. For certain types of analyses, however, techniques might be required that rely on external software modules, are more computationally demanding, or that are more complicated to use. In this section we briefly elaborate on some of these advanced approaches that are worth taking note of.

### Advanced NLP

In addition to the preprocessing techniques discussed in the data preparation section, there are powerful preprocessing techniques that rely on more advanced natural language processing (NLP). At present, these techniques are not available in native R, but rely on external software modules that often have to be installed outside of R. Many of the advanced NLP techniques are also much more computationally demanding, thus taking more time to perform. Another complication is that these techniques are language specific, and often only available for English and a few other big languages.

Several R packages provide interfaces for external NLP modules, so that once these modules have been installed, they can easily be used from within R. The *coreNLP* (Arnold & Tilton, 2016) package provides bindings for Stanford CoreNLP java library (Manning et al., 2014), which is a full NLP parser for English, that also supports (albeit with limitations) Arabic, Chinese, French, German, and Spanish. The *spacyr* (Benoit & Matsuo, 2017) package provides an interface for the spaCy module for Python, which is comparable to CoreNLP but is faster, and supports English and (again with some limitations) German and French. A third package, *cleanNLP* (Arnold, 2017a), conveniently wraps



**Figure 2.** Keyness plot comparing relative word frequencies for Trump and Obama.

both CoreNLP and spaCy, and also includes a minimal back-end that does not rely on external dependencies. This way it can be used as a swiss army knife, choosing the approach that best suits the occasion and for which the back-end is available, but with standardized output and methods.

Advanced NLP parsers generally perform all techniques in one go. In the following example we use the *spacyr* package to parse a sentence, that will be used to illustrate four advanced NLP techniques: lemmatization, part-of-speech (POS) tagging, named entity recognition (NER) and dependency parsing.

```
## first install spaCy. See instructions on spacyr GitHub page:
## https://github.com/kbenoit/spacyr
install.packages("spacyr")
library(spacyr)

spacy_initialize()
d = spacy_parse("Bob Smith gave Alice his login information.", dependency = TRUE)
d[,-c(1,2)]
```

token_id	token	lemma	pos	head_token_id	dep_rel	entity
1	Bob	bob	PROPN	2	compound	PERSON_B
2	Smith	smith	PROPN	3	nsubj	PERSON_I
3	gave	give	VERB	3	ROOT	
4	Alice	alice	PROPN	3	dative	PERSON_B
5	his	-PRON-	ADJ	7	poss	
6	login	login	NOUN	7	compound	
7	information	information	NOUN	3	dobj	
8	.	.	PUNCT	3	punct	

### Lemmatization

Lemmatization fulfills a similar purpose as stemming, but instead of cutting off the ends of terms to normalize them, a dictionary is used to replace terms with their lemma. The main advantage of this

approach is that it can more accurately normalize different verb forms—such as “gave” and “give” in the example—which is particularly important for heavily inflected languages such as Dutch or German.

### **Part-of-speech tagging**

POS tags are morpho-syntactic categories for words, such as nouns, verbs, articles and adjectives. In the example we see three proper nouns (PROPN), a verb (VERB) an adjective (ADJ), two nouns (NOUN), and punctuation (PUNCT). This information can be used to focus an analysis on certain types of grammar categories, for example, using nouns and proper names to measure similar events in news items (Welbers, Van Atteveldt, Kleinnijenhuis, & Ruigrok, 2016), or using adjectives to focus on subjective language (De Smedt & Daelemans, 2012). Similarly, it is a good approach for filtering out certain types of words, such as articles or pronouns.

### **Named entity recognition**

Named entity recognition is a technique for identifying whether a word or sequence of words represents an entity and what type of entity, such as a person or organization. Both “Bob Smith” and “Alice” are recognized as persons. Ideally, named entity recognition is paired with co-reference resolution. This is a technique for grouping different references to the same entity, such as anaphora (e.g., he, she, the president). In the example sentence, the word “his” refers to “Bob Smith”. Co-reference resolution is currently only supported by Stanford CoreNLP, but discussion on the spaCy GitHub page suggests that this feature is on the agenda.

### **Dependency parsing**

Dependency parsing provides the syntactic relations between tokens, which can be used to analyze texts at the level of syntactic clauses (Van Atteveldt, 2008). In the *spacyr* output this information is given in the *head\_token\_i* and *dep\_rel* columns, where the former indicates to what token a token is related and the latter indicates the type of relation. For example, we see that “Bob” is related to “Smith” (*head\_token\_i* 2) as a compound, thus recognizing “Bob Smith” as a single entity. Also, since “Smith” is the nominal subject (nsubj) of the verb “gave”, and Alice is the dative case (dative) we know that “Bob Smith” is the one who gives to “Alice”. This type of information can for instance be used to analyze who is attacking whom in news coverage about the Gaza war (Van Atteveldt et al., 2017).

### **Word positions and syntax**

As discussed above, the bag-of-words representation of texts is memory-efficient and convenient for various types of analyses, and this often outweighs the disadvantage of losing information by dropping the word positions. For some analyses, however, the order of words and syntactical properties can be highly beneficial if not crucial. In this section we address some text representations and analysis techniques where word positions are maintained.

A simple but potentially powerful solution is to use higher order *n*-grams. That is, instead of tokenizing texts into single words ( $n = 1$ ; *unigrams*), sequences of two words ( $n = 2$ ; *bigrams*), three words ( $n = 3$ ; *trigrams*) or more are used.<sup>9</sup> The use of higher order n-grams is often optional in tokenization functions. *quanteda* makes it possible to form n-grams when tokenizing, or to form ngrams from tokens already formed. Other options include the formation of “skip-grams”, or n-grams from words with variable windows of adjacency. Such non-adjacent collocations form the

---

<sup>9</sup>The term *n*-grams can be used more broadly to refer to sequences, and is also often used for sequences of individual characters. In this teacher’s corner we strictly use *n*-grams to refer to sequences of words.

basis for counting weighted proximity vectors, used in vector-space network-based models built on deep learning techniques (Mikolov, Chen, Corrado, & Dean, 2013; Selivanov, 2016). Below, we illustrate how to form both tri-grams and skipgrams of size three using a vector of both 0 and 1 skips.

```
text <- "an example of preprocessing techniques"
tokens(text, ngrams = 3, skip = 0:1)

tokens from 1 document. text1 :
[1] "an_example_of"                  "an_example_preprocessing"
[3] "an_of_preprocessing"            "an_of_techniques"
[5] "example_of_preprocessing"       "example_of_techniques"
[7] "example_preprocessing_techniques" "of_preprocessing_techniques"
```

The advantage of this approach is that these n-grams can be used in the same way as unigrams: we can make a DTM with n-grams, and perform all the types of analyses discussed above. Functions for creating a DTM in R from raw text therefore often allow the use of n-grams other than unigrams. For some analysis this can improve performance. Consider, for instance, the importance of negations and amplifiers in sentiment analysis, such as “not good” and “very bad” (Aue & Gamon, 2005). An important disadvantage, however, is that using n-grams is more computationally demanding, since there are many more unique sequences of words than individual words. This also means that more data is required to get a good estimate of the distribution of n-grams.

Another approach is to preserve the word positions after tokenization. This has three main advantages. First, the order and distance of tokens can be taken into account in analyses, enabling analyses such as the co-occurrence of words within a word window. Second, the data can be transformed into both a fulltext corpus (by pasting together the tokens) and a DTM (by dropping the token positions). This also enables the results of some text analysis techniques to be visualized in the text, such as coloring words based on a word scale model (Slapin & Proksch, 2008), or to produce browsers for topic models (Gardner et al., 2010). Third, each token can be annotated with token specific information, such as obtained from advanced NLP techniques. This enables, for instance, the use of dependency parsing to perform an analysis at the level of syntactic clauses (Van Atteveldt et al., 2017). The main disadvantage of preserving positions is that it is very memory inefficient, especially if all tokens are kept and annotations are added.

A common way to represent tokens with positions maintained is a data frame in which rows represent tokens, ordered by their position, and columns represent different variables pertaining to the token, such as the literal text, its lemma form and its POS tag. An example of this type of representation was shown above in the advanced NLP section, in the *spacyr* token output. Several R packages provide dedicated classes for tokens in this format. One is the *koRpus* (Michalke, 2017) package, which specializes in various types of text statistics, in particular lexical diversity and readability. Another is *corpustools* (Welbers & Van Atteveldt, 2016), which focuses on managing and querying annotated tokens, and on reconstructing texts to visualize quantitative text analysis results in the original text content for qualitative investigation. A third option is *tidytext* (Silge & Robinson, 2016), which does not focus on this format of annotated tokens, but provides a framework for working with tokenized text in data frames.

For a brief demonstration of utilizing word positions, we perform a dictionary search with the *corpustools* package (Welbers & Van Atteveldt, 2016), that supports searching for words within a given word distance. The results are then viewed in key word in context (KWIC) listings. In the example, we look for the queries “freedom” and “america” within a distance of five words, using the State of the Union speeches from George W. Bush and Barack Obama.

```

install.packages("corpustools")
library(corpustools)

tc <- create_tcorpus(sotu_texts, doc_column = "id")
hits <- tc$search_features('"freedom americ*~5')
kwic <- tc$kwic(hits, ntokens = 3)
head(kwic$kwic, 3)

[1] ...making progress toward <freedom> will find <America> is their friend...
[2] ...friends, and <freedom> in Iraq will make <America> safer for generations...
[3] ...men who despise <freedom>, despise <America>, and aim...

```

## Conclusion

R is a powerful platform for computational text analysis, that can be a valuable tool for communication research. First, its well developed packages provide easy access to cutting edge text analysis techniques. As shown here, not only are most common text analysis techniques implemented, but in most cases, multiple packages offer users choice when selecting tools to implement them. Many of these packages have been developed by and for scholars, and provide established procedures for data preparation and analysis. Second, R's open source nature and excellent system for handling packages make it a convenient platform for bridging the gap between research and tool development, which is paramount to establishing a strong computational methods paradigm in communication research. New algorithms do not have to be confined to abstract and complex explanations in journal articles aimed at methodology experts, or made available through arcane code that many interested parties would not know what to do with. As an R package, algorithms can be made readily available in a standardized and familiar format.

For new users, however, choosing from the wide range of text analysis packages in R can also be daunting. With various alternatives for most techniques, it can be difficult to determine which packages are worth investing the effort to learn. The primary goal of this teacher's corner, therefore, has been to provide a starting point for scholars looking for ways to incorporate computational text analysis in their research. Our selection of packages is based on our experience as both users and developers of text analysis packages in R, and should cover the most common use cases. In particular, we advise users to become familiar with at least one established and well-maintained package that handles data preparation and management, such as *quanteda*, *tidytext* (Silge & Robinson, 2016) or *tm* (Feinerer & Hornik, 2017). From here, it is often a small step to convert data to formats that are compatible with most of the available text analysis packages.

It should be emphasized that the selection of packages presented in this teacher's corner is not exhaustive, and does not represent which packages are the most suitable for the associated functionalities. Often, the best package for the job depends largely on specific features and problem specific priorities such as speed, memory efficiency and accuracy. Furthermore, when it comes to establishing a productive workflow, the importance of personal preference and experience should not be underestimated. A good example is the workflow of the *tidytext* package (Silge & Robinson, 2016), which could be preferred by people that are familiar with the tidyverse philosophy (Wickham et al., 2014). Accordingly, the packages recommended in this teacher's corner provide a good starting point, and for many users could be all they need, but there are many other great package out there. For a more complete list of packages, a good starting point is the CRAN Task View for Natural Language Processing (Wild, 2017).

Marshall McLuhan (1964) famously stated that “we shape our tools, and thereafter our tools shape us”, and in science the same can be said for how tools shape our findings. We thus argue that the establishment of a strong computational methods paradigm in communication research goes hand-in-hand with embracing open-source tool development as an inherent part of scientific practice. As such, we conclude with a call for researchers to cite R packages similar to how one



would cite other scientific work.<sup>10</sup> This gives due credit to developers, and thereby provides a just incentive for developers to publish and maintain new code, including proper testing and documentation to facilitate the correct use of code by others. Citing packages is also paramount for the transparency of research, which is especially important when using new computational techniques, where results might vary depending on implementation choices and where the absence of bugs is often not guaranteed. Just as our theories are shaped through collaboration, transparency and peer feedback, so should we shape our tools.

## Declaration of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

## ORCID

Kenneth Benoit <http://orcid.org/0000-0002-0797-564X>

## References

- Arnold, T. (2017a). *cleanNLP: A tidy data model for natural language processing* [Computer software manual] (R package version 1.9.0). Retrieved from <https://CRAN.R-project.org/package=cleanNLP>
- Arnold, T. (2017b). *kerasR: R interface to the keras deep learning library* [Computer software manual] (R package version 0.6.1). Retrieved from <https://CRAN.R-project.org/package=kerasR>
- Arnold, T., & Tilton, L. (2016). *coreNLP: Wrappers around Stanford CoreNLP tools* [Computer software manual] (R package version 0.4-2). Retrieved from <https://CRAN.R-project.org/package=coreNLP>
- Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*. Retrieved from [http://research.microsoft.com/pubs/65430/new\\_domain\\_sentiment.pdf](http://research.microsoft.com/pubs/65430/new_domain_sentiment.pdf)
- Bates, D., & Maechler, M. (2015). *Matrix: Sparse and dense matrix classes and methods* [Computer software manual] (R package version 1.2-3). Retrieved from <https://CRAN.R-project.org/package=Matrix>
- Benoit, K., & Matsuo, A. (2017). *spacyr: R Wrapper to the spaCY NLP Library* [Computer software manual] (R package version 0.9.0). Retrieved from <https://CRAN.R-project.org/package=spacyr>
- Benoit, K., Watanabe, K., Nulty, P., Obeng, A., Wang, H., Lauderdale, B., & Lowe, W. (2017). *quanteda: Quantitative analysis of textual data* [Computer software manual] (R package version 0.99). Retrieved from <http://quanteda.io>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.
- Bouchet-Valat, M. (2014). *SnowballC: Snowball Stemmers based on the C Libstemmer UTF-8 Library* [Computer software manual] (R package version 0.5.1). Retrieved from <https://CRAN.R-project.org/package=SnowballC>
- Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, 4(1), 8–23. doi:[10.1080/21670811.2015.1096598](https://doi.org/10.1080/21670811.2015.1096598)
- Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, 173(3), 781–800. doi:[10.1016/j.ejor.2005.07.023](https://doi.org/10.1016/j.ejor.2005.07.023)
- De Smedt, T., & Daelemans, W. (2012). “vreselijk mooi!” (terribly beautiful): A subjectivity lexicon for dutch adjectives. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, May 2012, 3568–3572.
- Feinerer, I., & Hornik, K. (2017). *tm: Text mining package* [Computer software manual] (R package version 0.7-1). Retrieved from <https://CRAN.R-project.org/package=tm>
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233. doi:[10.1037/h0057532](https://doi.org/10.1037/h0057532)
- Fox, J., & Leanage, A. (2016). R and the journal of statistical software. *Journal of Statistical Software*, 73(2), 1–13.
- Gagolewski, M. (2017). *R package stringi: Character string processing facilities* [Computer software manual]. Retrieved from <http://www.gagolewski.com/software/stringi/>

<sup>10</sup>To view how to cite a package, the *citation* function can be used—e.g., `citation("quanteda")` for citing quanteda, or `citation()` for citing the R project. This either provides the citation details provided by the package developer or auto-generated details.

- Gardner, M. J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., & Seppi, K. (2010). The topic browser: An interactive tool for browsing topic models. In *Nips workshop on Challenges of Data Visualization*. Retrieved from <http://cseweb.ucsd.edu/~lvdmaaten/workshops/nips2010/papers/gardner.pdf>
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. In *Proceedings of the National Academy of Sciences*, 5228–5235. doi:[10.1073/pnas.0307752101](https://doi.org/10.1073/pnas.0307752101)
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297. doi:[10.1093/pan/mps028](https://doi.org/10.1093/pan/mps028)
- Grun, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40 (13), 1–30. doi:[10.18637/jss.v040.i13](https://doi.org/10.18637/jss.v040.i13)
- Günther, E., & Quandt, T. (2016). Word counts and topic models: Automated text analysis methods for digital journalism research. *Digital Journalism*, 4(1), 75–88. doi:[10.1080/21670811.2015.1093270](https://doi.org/10.1080/21670811.2015.1093270)
- Jurka, T. P., Collingwood, L., Boydston, A. E., Grossman, E., & Van Atteveldt, W. (2014). *RTextTools: Automatic text classification via supervised learning* [Computer software manual] (R package version 1.4.2). Retrieved from <https://CRAN.R-project.org/package=RTextTools>
- Lang, D. T., & the CRAN Team. (2017). *XML: Tools for parsing and generating XML within R and S-plus* [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=XML>
- Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines. How to represent texts in input space? *Machine Learning*, 46(1), 423–444. doi:[10.1023/a:1012491419635](https://doi.org/10.1023/a:1012491419635)
- Manning, C. D., Manning, C. D., Raghavan, P., Raghavan, P., Schütze, H., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press. doi:[10.1017/cbo9780511809071](https://doi.org/10.1017/cbo9780511809071)
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). doi: [10.3115/v1/p14-5010](https://doi.org/10.3115/v1/p14-5010)
- McCarthy, P. M., & Jarvis, S. (2010). Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392. doi:[10.3758/brm.42.2.381](https://doi.org/10.3758/brm.42.2.381)
- McLuhan, M. (1964). *Understanding Media: The Extensions of Man*. New York: Penguin Press.
- Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in r. *Journal of Statistical Software*, 25(5), 1–54. doi:[10.18637/jss.v025.i05](https://doi.org/10.18637/jss.v025.i05)
- Michalke, M. (2017). *koRpus: An R package for text analysis* [Computer software manual] (Version 0.10-2). Retrieved from <https://reaktanz.de/?c=hacking&s=koRpus>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of International Conference of Learning Representations. arXiv preprint arXiv:1301.3781*, Scottsdale, Arizona, May 2013.
- Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expertat Systems with Applications*, 40(10), 4241–4251. doi:[10.1016/j.eswa.2013.01.019](https://doi.org/10.1016/j.eswa.2013.01.019)
- Mullen, L. (2016a). *textruese: Detect text reuse and document similarity* [Computer software manual] (R package version 0.1.4). Retrieved from <https://CRAN.R-project.org/package=textruese>.
- Mullen, L. (2016b). *tokenizers: A consistent interface to tokenize natural language text* [Computer software manual] (R package version 0.1.4). Retrieved from <https://CRAN.R-project.org/package=tokenizers>
- Ooms, J. (2014). *The jsonlite package: A practical and consistent mapping between json data and r objects* [Computer software manual]. Retrieved from <https://arxiv.org/abs/1403.2805>
- Ooms, J. (2017a). *antiword: Extract text from microsoft word documents* [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=antiword>
- Ooms, J. (2017b). *pdftools: Text extraction, rendering and converting of pdf documents* [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=pdftools>
- Porter, M. F. (2001). *Snowball: A language for stemming algorithms*. Retrieved from <http://snowball.tartarus.org/texts/introduction.html>
- Proksch, S.-O., & Slapin, J. B. (2009). How to avoid pitfalls in statistical analysis of political texts: The case of germany. *German Politics*, 18(3), 323–344. doi:[10.1080/09644000903055799](https://doi.org/10.1080/09644000903055799)
- Provost, F., & Fawcett, T. (2013). Data science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), 51–59. doi:[10.1089/big.2013.1508](https://doi.org/10.1089/big.2013.1508)
- R Core Team. (2017). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064–1082. doi:[10.1111/ajps.12103](https://doi.org/10.1111/ajps.12103)
- rOpenSci Text Workshop. (2017). *tif: Text interchange format* [Computer software manual]. Retrieved from <https://github.com/ropensci/tif>
- Schuck, A. R., Xezonakis, G., Elenbaas, M., Banducci, S. A., & De Vreese, C. H. (2011). Party contestation and Europe on the news agenda: The 2009 European Parliamentary Elections. *Electoral Studies*, 30(1), 41–52. doi:[10.1016/j.electstud.2010.09.021](https://doi.org/10.1016/j.electstud.2010.09.021)



- Schultz, F., Kleinnijenhuis, J., Oegema, D., Utz, S., & Van Atteveldt, W. (2012). Strategic framing in the BP crisis: A semantic network analysis of associative frames. *Public Relations Review*, 38(1), 97–107. doi:10.1016/j.pubrev.2011.08.003
- Selivanov, D. (2016). *text2vec: Modern text mining framework for R* [Computer software manual] (R package version 0.4.0). Retrieved from <https://CRAN.R-project.org/package=text2vec>
- Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, 1, 3. doi:10.21105/joss.00037
- Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3), 705–722. doi:10.1111/j.1540-5907.2008.00338.x
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267–307. doi:10.1162/coli\_a\_00049
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54. doi:10.1177/0261927X09351676
- TIOBE. (2017). *The R programming language*. Retrieved from <https://www.tiobe.com/tiobe-index/r/>
- Van Atteveldt, W. (2008). Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content (Dissertation). Charleston, SC: BookSurge.
- Van Atteveldt, W., Sheafet, T., Shenhav, S. R., & Fogel-Dror, Y. (2017). Clause analysis: Using syntactic information to automatically extract source, subject, and predicate from texts with an application to the 2008–2009 Gaza War. *Political Analysis*, 25(2), 207–222. doi:10.1017/pan.2016.12
- Vliegenthart, R., Boomgaarden, H. G., & Van Spanje, J. (2012). Anti-immigrant party support and media visibility: A cross-party, over-time perspective. *Journal of Elections, Public Opinion & Parties*, 22(3), 315–358. doi:10.1080/17457289.2012.693933
- Watanabe, K. (2017). The spread of the Kremlin's narratives by a western news agency during the Ukraine crisis. *The Journal of International Communication*, 23(1), 138–158. doi:10.1080/13216597.2017.1287750
- Welbers, K., & Van Atteveldt, W. (2016). *corpustools: Tools for managing, querying and analyzing tokenized text* [Computer software manual] (R package version 0.201). Retrieved from <http://github.com/kasperwelbers/corpustools>
- Welbers, K., Van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, N. (2016). A Gatekeeper among Gatekeepers: News Agency Influence in Print and Online Newspapers in the Netherlands. *Journalism Studies*, 1–19 (online first). doi:10.1080/1461670x.2016.1190663
- Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10), 1–23. doi:10.18637/jss.v059.i10
- Wickham, H., & Bryan, J. (2017). *readxl: Read excel files* [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=readxl>
- Wild, F. (2017). Cran task view: Natural language processing. CRAN. Version: 2017-01-17. Retrieved from <https://CRAN.R-project.org/view=NaturalLanguageProcessing>
- Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)* (pp. 412–420), Nashville, TN, July 1997.

# Dictionnaires



---

# The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods

Journal of Language and Social Psychology

29(1) 24–54

© 2010 SAGE Publications

DOI: 10.1177/0261927X09351676

<http://jls.sagepub.com>



**Yla R. Tausczik<sup>1</sup> and James W. Pennebaker<sup>1</sup>**

## Abstract

We are in the midst of a technological revolution whereby, for the first time, researchers can link daily word use to a broad array of real-world behaviors. This article reviews several computerized text analysis methods and describes how Linguistic Inquiry and Word Count (LIWC) was created and validated. LIWC is a transparent text analysis program that counts words in psychologically meaningful categories. Empirical results using LIWC demonstrate its ability to detect meaning in a wide variety of experimental settings, including to show attentional focus, emotionality, social relationships, thinking styles, and individual differences.

## Keywords

computerized text analysis, LIWC, relationships, dominance, deception, attention, pronouns

James J. Bradac (1986, 1999) celebrated the many ways that scientists could simultaneously study both language and human communication. He understood the value of highly controlled laboratory studies and, at the same time, the importance of exploring the ways people naturally talk in the real world. Of particular importance to him, however, was that language research replicates its theories and findings across a wide array of methods and samples. This article draws heavily from Bradac's approach to research by applying a new array of computer-based text analysis tools to the study of everyday language.

---

<sup>1</sup>University of Texas at Austin, Austin, TX, USA

## Corresponding Author:

James W. Pennebaker, Department of Psychology, University of Texas at Austin, Austin, TX 78712, USA  
Email: [Pennebaker@mail.utexas.edu](mailto:Pennebaker@mail.utexas.edu)

The words we use in daily life reflect who we are and the social relationships we are in. This is neither a new nor surprising insight. Language is the most common and reliable way for people to translate their internal thoughts and emotions into a form that others can understand. Words and language, then, are the very stuff of psychology and communication. They are the medium by which cognitive, personality, clinical, and social psychologists attempt to understand human beings.

The simultaneous development of high-speed personal computers, the Internet, and elegant new statistical strategies have helped usher in a new age of the psychological study of language. By drawing on massive amounts of text, researchers can begin to link everyday language use with behavioral and self-reported measures of personality, social behavior, and cognitive styles. Beginning in the early 1990s, we stumbled on the remarkable potential of computerized text analysis through the development of our own computer program—Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007). We are now witnessing new generations of text analysis coming from computer sciences and computational linguistics.

This article is divided into three sections. The first is a brief history of text analysis in psychology. The second focuses on our own efforts to develop LIWC along with some of the basic psychometrics of words. The third explores the links between word usage and basic social and personality processes.

## Computerized Text Analysis: A Brief History

The roots of modern text analysis go back to the earliest days of psychology. Freud (1901) wrote about slips of the tongue whereby a person's hidden intentions would reveal themselves in apparent linguistic mistakes. Rorschach and others (e.g., Holtzman, 1950; Rorschach, 1921) developed projective tests to detect people's thoughts, intentions, and motives from the way they described ambiguous inkblots. McClelland and a generation of thematic apperception test (TAT) researchers (e.g., McClelland, 1979; Winter, 1998) found that the stories people told in response to drawings of people could provide important clues to their needs for affiliation, power, and achievement. In all cases, trained raters read the transcripts of people's descriptions and tagged words or phrases that represented the dimensions the investigators were studying.

More general and less stimulus-bound approaches began to evolve in the 1950s. Gottschalk and his colleagues (e.g., Gottschalk & Gleser, 1969; Gottschalk, Gleser, Daniels, & Block, 1958) developed a content-analysis method by which to track Freudian themes in text samples. The original Gottschalk method required patients to talk in a stream of consciousness way into a tape recorder for 5 minutes. The language samples were transcribed and broken down into grammatical phrases. Judges, then, evaluated each phrase to determine the degree it might reflect one or more themes related to anxiety (e.g., death, castration), hostility toward self or others, and various interpersonal and psychological topics. The Gottschalk method later was used in the psychiatric diagnoses of cognitive impairments, alcohol abuse, brain damage, and mental disorders. Attempts to translate the original Gottschalk–Gleser scoring scheme

to a computer program have proven difficult with modest correlations to the judge-based “gold standard” (e.g., Gottschalk & Bechtel, 1993).

The first general purpose computerized text analysis program in psychology was developed by Philip Stone and his colleagues (Rosenberg & Tucker, 1978; Stone, Dunphy, Smith, & Ogilvie, 1966). Using a mainframe computer, the authors built a complex program that adapted McClelland’s need-based coding schemes to any open-ended text. The program, called General Inquirer, relied on a series of author-developed algorithms. The General Inquirer and other programs like it (e.g., Hart’s, 1984, DICTON program; Martindale, 1990) have proven valuable in distinguishing mental disorders, assessing personality dimensions, and evaluating speeches. One limitation of these approaches is that they have relied on the manipulation and weighting of language variables that were not visible to the user.

The first truly transparent text analysis method was pioneered by Walter Weintraub (1981, 1989). Weintraub, a physician by training, became fascinated by the everyday words people used—words such as pronouns and articles. Over the span of a decade, he hand-counted people’s words in texts such as political speeches and medical interviews. He noticed that first-person singular pronouns (e.g., I, me, my) were reliably linked to people’s levels of depression. Although his methods were straightforward and his findings consistently related to important outcome measures, his work was largely ignored. His observation that the simple words of everyday speech reflected psychological state nevertheless was prescient. (See also the work of Mergenthaler, 1996, who developed a computer program TAS/C that taps abstraction and emotion in psychotherapy sessions.)

### **The Development of LIWC and the Psychometrics of Words**

In the 1980s, we discovered that when people were asked to write about emotional upheavals in their lives they subsequently evidenced improvements in physical health (e.g., Pennebaker & Beall, 1986). The first group of writing studies generated hundreds of writing samples that revealed deeply moving human stories. Intuitively, the ways the stories were written should have been related to whether people’s health improved or not. In an attempt to link the stories with health outcomes, judges were asked to read the emotional essays and to rate them along multiple dimensions. Some of the categories included the degree to which the stories were organized, coherent, personal, emotional, vivid, optimistic, and evidenced insight.

Relying on judges’ ratings yielded three important findings: (a) even with in-depth training, judges do not agree with each other in rating most dimensions when evaluating a broad range of deeply personal stories; (b) rating essays by multiple judges is extremely slow and expensive; and (c) judges tend to get depressed when reading depressing stories.

To find a more efficient evaluation method, we turned to the promise of computerized text analysis programs to assess the essays. At the time, no simple text analysis program existed. Consequently, Martha Francis and the second author began the task

of developing one. Our goal was to create a program that simply looked for and counted words in psychology-relevant categories across multiple text files. The result has been an ever-changing computer program named Linguistic Inquiry and Word Count, or LIWC (pronounced “Luke”).

### *The Logic and Development of LIWC*

The LIWC program has two central features—the processing component and the dictionaries. The processing feature is the program itself, which opens a series of text files—which can be essays, poems, blogs, novels, and so on—and then goes through each file word by word. Each word in a given text file is compared with the dictionary file.

For example, if LIWC were analyzing the first line of the novel *Paul Clifford* by Edward Bulwer-Lytton (1842):

It was a dark and stormy night

the program would first look at the word “it” and then see if “it” was in the dictionary.

It is and is coded as a function word, a pronoun, and, more specifically, an impersonal pronoun. All three of these LIWC categories would then be incremented. Next, the word “was” would be checked and would be found to be associated with the categories of verbs, auxiliary verbs, and past tense verbs.

After going through all the words in the novel, LIWC would calculate the percentage of each LIWC category. So, for example, we might discover that 2.34% of all the words in a given book were impersonal pronouns and 3.33% were auxiliary verbs. The LIWC output, then, lists all LIWC categories and the rates that each category was used in the given text.

The dictionaries are the heart of the LIWC program. A dictionary refers to the collection of words that define a particular category. When LIWC was first created, the goal was fairly modest. We simply wanted the computer to calculate the percentage of positive and negative emotion words within a text. To do this, we needed to specify exactly which words to look for. Based on our judges’ ratings, we also wanted to include measures of thinking styles—for example, signs of self-reflection, and causal thinking. Over several weeks, the number of categories we were interested in expanded from the original 2 to more than 80.

Across the 80 categories, several language dimensions are straightforward. For example, the category of articles is made up of three words: “a,” “an,” and “the.” Other dimensions are more subjective. For example, the emotion word categories required human judges to evaluate which words were suited for which categories. For all subjective categories, an initial selection of word candidates was gleaned from dictionaries, thesauruses, questionnaires, and lists made by research assistants. Groups of three judges then independently rated whether each word candidate was appropriate to the overall word category.

All category word lists were updated by the following set of rules: (a) a word remained in the category list if two out of three judges agreed it should be included; (b) a word was deleted from the category list if at least two of the three judges agreed it should be excluded; and (c) a word was added to the category list if two out of three judges agreed it should be included. This entire process was then repeated a final time by a separate group of three judges. The final percentages of judges' agreement for this second rating phase ranged from 93% to 100% agreement.

The initial LIWC judging took place between 1992 and 1994. A significant LIWC revision was undertaken in 1997 and again in 2007 to streamline the original program and dictionaries. Text files from several dozen studies, totaling more than 100 million words were analyzed. Some low base rate word categories were deleted and others were added. For details of the process and specific findings, see Pennebaker, Chung, Ireland, Gonzales, and Booth (2007).

### *The Psychometrics of Word Usage*

Unlike the typical development of a new measurement instrument, verifying the validity and reliability of word usage is trickier. Consider how psychologists typically develop and test a new measurement instrument. For questionnaires, for example, after specific questions have been generated and initially tested, the investigator computes reliability statistics to be sure that all items are correlated with the sum of the remaining items. Generally, a factor analysis of the items is run to see if the items reflect more than one dimension. Next, the investigator computes the test-retest reliability of the questionnaire. And, finally, there are a series of validation tests to see if the questionnaire correlates with or predicts real-world behaviors that it is supposed to measure.

Word categories are unlike questionnaire items. Words are rarely normally distributed, they generally have low base rates, and standard measures of reliability are not always appropriate. Consider, for example, the category of articles—"a," "an," and "the." All three words serve the same function, which is to signal the upcoming use of a concrete noun. From a classically trained psychometric perspective, for us to consider "articles" to be a coherent, internally consistent category, use of the three words should be highly correlated with each other—with Cronbach's  $\alpha$  of at least .60 or .70, it is hoped. Tragically, words do not adhere to traditional psychometric laws that we see in questionnaires. For example, our lab frequently relies on a random assortment of about 2,800 text files that includes a wide range of text genres, including blogs, experimental essays, poetry, books, science articles, and natural speech transcripts to examine the psychometrics of words. Within this text corpus, articles represent 5.43% of all words used (where "a" = 1.96, "an" = 0.19, "the" = 3.27). The intercorrelation among these words is low but highly significant ("a" with "an" = .13, "a" with "the" = .09, "an" with "the" = .09), resulting in Cronbach's  $\alpha$  of .14 (for a summary of all reliability statistics, see Pennebaker et al., 2007).

Note that assessing the psychometrics of word use is even more complicated than what the above statistics suggest. To get reliability data for a questionnaire, we typically give people the same test of often-redundant questionnaire items on two occasions.

In theory, the questionnaire has exactly the same meaning on the two administrations. Asking people to, say, describe themselves on two occasions will generally evoke different types of responses. For example, within the open-ended response itself, people generally don't repeat themselves (meaning one rarely gets good split-half reliability). Second, if people tell an experimenter who they are today, they will likely change their stories next time either because they have changed a bit or they want the experimenter to have a fuller sense of who they were from the previous time. Furthermore, saying the same thing as they did to the person on the first occasion would be redundant and, perhaps, a bit rude. In short, the psychometrics of word use pose a new set of problems that questionnaires avoid.

### ***Content Versus Style Words***

When LIWC was first developed, the goal was to devise an efficient system that could tap both psychological processes and the content of what people were writing or talking about. Within a few years, it became clear that there are two very broad categories of words that have different psychometric and psychological properties. *Content words* are generally nouns, regular verbs, and many adjectives and adverbs. They convey the content of a communication. To go back to the phrase "It was a dark and stormy night" the content words are: "dark," "stormy," and "night." Intertwined through these content words are *style words*, often referred to as function words. Style or function words are made up of pronouns, prepositions, articles, conjunctions, auxiliary verbs, and a few other esoteric categories. In the phrase these words are "it," "was," "a," and "and."

Although we tend to have almost 100,000 English words in our vocabulary, only about 500 (or 0.05%) are style words. Nevertheless, style words make up about 55% of all the words we speak, hear, and read. Furthermore, content and style words tend to be processed in the brain very differently (Miller, 1995).

From a psychological perspective, style words reflect how people are communicating, whereas content words convey what they are saying. It is not surprising, then, that style words are much more closely linked to measures of people's social and psychological worlds. Indeed, the ability to use style words requires basic social skills. Consider the sentence, "I will meet you here later." Although grammatically correct, the sentence has no real meaning unless the reader knows who "I" and "you" refer to. Where is "here" and what is meant by "later"? These are all referents that are shared by two people in a particular conversation taking place at a particular time. To say this implies that the speaker knows that the listener shares the same knowledge of these style words (cf. Chung & Pennebaker, 2007).

*Caveats concerning computer text analysis.* Psychologists are always looking for measures that reveal the secret, hidden, or distorted "real" self. Freud's popularity was partly attributable to his assertion that subconscious thoughts, emotions, and experiences drove our behavior. People continue to be enthralled with his methods of dream analysis, slips of the tongue, and other psychoanalytic claims. This trend continues with a new generation of measures and theories that rely on a host of implicit measures such as the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998),

priming strategies, and various imaging techniques such as functional MRI that all hold out the promise of discovering the “real” person. Many people consider the analysis of language—especially function or style words—to do the same. And, indeed, they sometimes can reveal social psychological processes that people are not able to easily conceal.

Despite the appeal of computerized language measures, they are still quite crude. Programs such as LIWC ignore context, irony, sarcasm, and idioms. The word “mad,” for example, is currently coded as an anger word. When people say things such as “I’m mad about him,” or “He’s as mad as a hatter” the meaning and intent of their utterances will be miscoded. LIWC, like any computerized text analysis program, is a probabilistic system.

The study of word use as a reflection of psychological state is in its earliest stages. As described below, studies are providing evidence that function words can detect emotional and biological states, status, honesty, and a host of individual differences. Nevertheless, the imprecise measurement of word meaning and psychological states themselves should give pause to anyone who relies too heavily on accurately detecting people’s true selves through their use of words.

## **The Social and Psychological Meaning of Words**

The words we use in daily life reflect what we are paying attention to, what we are thinking about, what we are trying to avoid, how we are feeling, and how we are organizing and analyzing our worlds. The 80 language categories in LIWC have been linked in hundreds of studies to interesting psychological processes. In this section, we give a brief discussion of psychological processes and a small set of related language categories. The section concludes with a comprehensive summary of findings about the correlates of word categories from a large group of studies.

### ***Attentional Focus: Pronouns and Verb Tense***

Tracking people’s attention reveals information about their priorities, intentions, and thoughts. Infants, for example, focus on objects that display novelty, complexity, and motion (Berlyne, 1960), which shows the extent to which they are focused on learning. Our attention can oscillate from our external worlds to our internal feelings or sensations (e.g., Pennebaker, 1982). If we are playing a game of tennis, we might bruise our arm and not notice because our full attention is on the game itself. Alternatively, if the injury is significant, the pain may be so attention grabbing that we no longer are aware of the game at all.

Tracking language use such as tracking people’s gaze can tell us where they are attending. At the most superficial level, content word categories explicitly reveal where individuals are focusing. Those thinking about death, sex, money, or friends will refer to them in their writing or conversation. Function words, such as personal pronouns, also reflect attentional allocation. People who are experiencing physical or emotional pain tend to have their attention drawn to themselves and subsequently use more first-person singular pronouns (e.g., Rude, Gortner, & Pennebaker, 2004). When

people sit in front of a mirror and complete a questionnaire, they use more words such as “I” and “me” than when the mirror is not present (Davis & Brock, 1975). As we might expect, positive ads focus on the political candidate producing the ad and negative ads focus on their opponent; use of pronouns quickly reveals these differences (Gunsch, Brownlow, Haynes, & Mabe, 2000). Gunsch and colleagues show that more self-references (e.g., “I,” “we”) were present in positive political ads compared with mixed and negative political ads, whereas more other-references (e.g., “he,” “she,” “they”) were present in negative ads compared with positive and mixed ads.

Attention can reveal not just who someone is attending to but how they are processing the situation. Students who wrote about their experiences with teasing varied in the pronouns they used depending on whether they were teasing others or were being teased by others (Kowalski, 2000). Participants used more first-person singular and fewer third-person pronouns (e.g., “he,” “she”) when describing an event when they were being teased compared with when they described an event where they were teasing someone else. In both cases, the focus is on the person who was teased—the victim of the event. There was a significant interaction with sex and use of third-person pronouns; male participants used more third-person pronouns when describing an event in which they were being teased than female participants. Compared with women, men may focus more on the perpetrator of the event when they are the victim, although it remains unclear why this is the case.

Whereas personal pronouns provide information about the subject of attention, analyses of the tense of common verbs can tell us about the temporal focus of attention. In the same study of political ads, the authors found that positive ads used more present and future tense verbs, and negative ads used more past tense verbs (Gunsch et al., 2000). From the tense of the verbs and the personal pronouns used, we can infer that negative ads focus on past actions of the opponent, and positive ads focus on the present and future acts of the candidate.

Studying attention also gives us a deeper understanding of how people are processing a situation or event. Participants were asked to either recall an event that they had discussed with someone else, or an undisclosed event; there were significant differences in the verb tense used in the two conditions (Pasupathi, 2007). Participants used greater past tense in discussing a disclosed event and greater present tense in discussing an undisclosed event. Verb tense differences could indicate increased psychological distance and a higher degree of resolution for disclosed events compared with undisclosed events.

Pronouns and verb tense are useful linguistic elements that can help identify focus, which, in turn, can show priorities, intentions, and processing. Some care should be taken in evaluating how pronouns and verbs are used. An exception to the pronoun-attention rule concerns first-person plural pronouns—“we,” “us,” and “our.” Sometimes “we” can signal a sense of group identity, such as when couples are asked to evaluate their marriages to an interviewer, the more the participants use “we,” the better their marriage (Simmons, Gordon, & Chambliss, 2005). “We” can also be used as the Royal We, such as when the advisor announces to his or her graduate students that “we need to analyze that data.” The use of “we” in this case actually means “you students” rather than “you students and I” (see also use of the Royal We by political figures, such as Rudolph Giuliani in Pennebaker & Lay, 2002).

### ***Emotionality: Positive and Negative Emotions***

The degree to which people express emotion, how they express emotion, and the valence of that emotion can tell us how people are experiencing the world. People react in radically different ways to traumatic or important events; how people react may say a lot about how they cope with the event and the extent to which the event plays a role in the future. At the heart of reacting and coping with events is people's emotional response.

Research suggests that LIWC accurately identifies emotion in language use. For example, positive emotion words (e.g., love, nice, sweet) are used in writing about a positive event, and more negative emotion words (e.g., hurt, ugly, nasty) are used in writing about a negative event (Kahn, Tobin, Massey, & Anderson, 2007). LIWC ratings of positive and negative emotion words correspond with human ratings of the writing excerpts (Alpers et al., 2005).

Use of emotion words has also been used as a measure of the degree of immersion. Holmes et al. (2007) found that among women trying to cope with intimate partner violence, using more positive and negative emotion words to describe the violence led to increased feelings of physical pain over the four writing sessions. The authors conclude that higher use of emotion words showed more immersion in the traumatic event, which led to increased experience of physical pain.

Language emotionality extends beyond the simple expression of more or less emotion; use of emotion words relate to other key language elements. In an examination of the random assortment of around 2,800 texts described earlier, emotion words were negatively correlated with articles ( $r = -.33$ ), prepositions ( $r = -.38$ ), and relativity words ( $r = -.40$ ). These language features as we discuss later, may be important in cognitive complexity and thinking styles. Emotion words were positively correlated with pronoun use ( $r = .29$ ), auxiliary verb use ( $r = .29$ ) and negation use ( $r = .32$ ). All correlations are highly significant,  $p < .001$ . The nature of these correlations suggests a deeper importance of the expression of emotion and thinking styles, and social awareness.

### ***Social Relationships***

Language at its most basic function is to communicate. Words provide information about social processes—who has more status, whether a group is working well together, if someone is being deceptive, and the quality of a close relationship. Word choice provides information about person perception (Semin & Fiedler, 1988). Certain language clues give away relationships. Pronouns reveal how an individual is referencing those in the interaction and outside of it. Word count explains who is dominating the conversation and how engaged they are in the conversation. Assents and positive emotion words measure levels of agreement. Other language cues are specific to the interaction; here we offer a few situations that have been studied.

### *Status, Dominance, and Social Hierarchy*

Higher-status individuals speak more often and freely make statements that involve others. Lower-status language is more self-focused and tentative. In a study of groups of three crew members, a captain, a first lieutenant, and a second lieutenant engaging in several flight simulations, the use of greater first-person plural correlated with higher rank (Sexton & Helmreich, 2000). The authors found the opposite pattern for question marks: Higher-ranked crew members asked fewer questions compared with lower-ranked crew members. Across five studies in which status was either experimentally manipulated, determined by partner ratings, or based on existing titles, increased use of first-person plural was a good predictor of higher status, and in four of the studies increased use of first-person singular was a good predictor of lower status (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2009). Leshed, Hancock, Cosley, McLeod, and Gay (2007) reported that members of small groups are rated as being more involved and task focused by their teammates if they use more words; supporting the assertion that total word count may also indicate status.

### *Social Coordination and Group Processes*

More communication, more unity, and positive feedback may promote better group performance. Word count can act as a proxy for amount of communication; in some circumstances, more first-person plural may show group cohesion; and assents and question marks show how individuals are responding to each other. In the study of flight crews simulating easy and difficult flights, increased group word count, increased use of first-person plural, and increased use of question marks in early simulations predicted better team performance (Sexton & Helmreich, 2000). However, groups of 4 to 6 participants working on a joint task that used less first-person plural rated their group as having more group cohesion, although first-person plural was unrelated to group performance (Gonzales, Hancock, & Pennebaker, in press). The type of first-person plural pronouns may be important, if “we” is being used to promote interdependence as in “we can do this;” it may increase group cohesion if, on the other hand, it is being used to indirectly assign tasks as it may lead to resentment. Increased use of assents (e.g., agree, OK, yes) could signal increased group consensus and agreement; however, the timing of assents is important. Later in a group task, assents may signal consensus, early assents may indicate blind agreement by unmotivated group members (Leshed, Hancock, Cosley, McLeod, & Gay, 2007).

### *Honesty and Deception*

Deceptive statements compared with truthful ones are moderately descriptive, distanced from self, and more negative. Newman, Pennebaker, Berry, and Richards (2003) investigated lying behavior in five experiments; in each experiment, lying was operationalized differently. Across the studies when participants were lying they used more negative emotion, more motion words (e.g., arrive, car, go), fewer exclusion

words, and less first-person singular. More motion words and fewer third-person pronouns were also significant predictors of deception by prisoners instructed to lie or tell the truth about videos they had watched (Bond & Lee, 2005). Hancock, Curry, Goorha, and Woodworth (2008) expanded these findings to study lying within pairs of participants over instant messenger. They found a similar pattern of language use when a participant was lying. They also found that the people being deceived, the partners of the participants lying, also changed their language. When one participant was lying both used a higher total word count, less first-person singular, and more sense words. Motion, exclusion, and sense words all indicate the degree to which an individual elaborated on the description of the scenario. Deceptive statements are balanced in descriptiveness because enough description is required to convince the other person of an untruthful statement but too much information might reveal inaccuracies. Using different linguistic measures, researchers found that non-naïve individuals assigned to be deceptive compared with non-naïve individuals assigned to be truthful or naïve individuals who were truthful used some language features that showed less diversity and complexity (Zhou, Burgoon, Nunamaker, & Twitchell, 2004). Exclusive words are also a marker of complexity. Complexity may be reduced in deceptive speech because of the cognitive load required to maintain a story that is contrary to experience, and the effort taken to try to convince someone else that something false is true.

### ***Close Relationships***

Pronoun use is very important in showing the quality of a close relationship, because it shows how individuals are referring to each other. Surprisingly, first-person plural ("we") has not been found to be related to higher relationship quality, instead use of second person ("you") is more important in predicting lower-quality relationships. Simmons, Chambliss, and Gordon (2008) found that use of second-person pronouns was negatively related to relationship quality. They found in a study of relatives of participants suffering from either obsessive-compulsive disorder or panic attacks with agoraphobia that there were differences in the use of pronouns and that these differences signaled the extent to which they had a poor relationship with the patient. Relatives who used more second person in a taped interview with the patient scored higher on measures of criticism and having an overinvolved emotional reaction to the patient's condition. In this study, use of second person showed hostility and willingness to confront the patient. In a study of archived instant message conversation between heterosexual romantic partners shows a marginal trend that increased use of second person by the male participant predicted lower ratings of relationship satisfaction (Slatcher, Vazire, & Pennebaker, 2008). Researchers have hypothesized that increased use of first-person plural in conversations between romantic partners should lead to increased ratings of relationship satisfaction and stability. In fact in the study of instant message transcripts of romantic partners shows that increased use of first-person singular by the women leads to higher ratings of satisfaction for both individuals, use of first-person plural is unrelated to the satisfaction. Higher positive emotion words for men lead to increased relationship satisfaction as well.

These are only a few possible interactions and related language categories. Patterns of language use are a rich tool in studying interactions, because so much of the interplay between individuals is carried out through language. However, language use depends on the situational context. For example, in a cooperative coordination context, higher total word count may signal better communication and agreement, whereas in a negotiation context it may signal a breakdown in agreement.

### *Thinking Styles: Conjunctions, Nouns, Verbs, and Cognitive Mechanisms*

Language can track what information people are selecting from their environment by monitoring attentional focus. By the same token, natural language use provides important clues as to how people process that information and interpret it to make sense of their environment. Thinking can vary in depth and complexity; this is reflected in the words people use to connect thoughts. Language changes when people are actively reevaluating a past event. It can also differ depending on the extent to which an event has already been evaluated.

Depth of thinking can vary between people and situations; certain words can reveal these differences. Cognitive complexity can be thought of as a richness of two components of reasoning: the extent to which someone differentiates between multiple competing solutions and the extent to which someone integrates among solutions (Tetlock, 1981). These two processes are captured by two LIWC categories—exclusion words and conjunctions. Exclusive words (e.g., but, without, exclude) are helpful in making distinctions. Indeed, people use exclusion words when they are attempting to make a distinction between what is in a category and what is not in a category. Exclusive words are used at higher rates among people telling the truth (Newman et al., 2003) and by Gore compared with Kerry and Edwards (Pennebaker, Slatcher, & Chung, 2005). Conjunctions (e.g., and, also, although) join multiple thoughts together and are important for creating a coherent narrative (Graesser, McNamara, Louwerse, & Cai, 2004).

Prepositions (e.g., to, with, above), cognitive mechanisms (e.g., cause, know, ought), and words greater than six letters are all also indicative of more complex language. Prepositions, for example, signal that the speaker is providing more complex and, often, concrete information about a topic. “The keys are *in* the box *by* the lamp *under* the painting.” Within published journal articles, authors use more prepositions in the discussion than the introduction or abstract. Discussions are often the most complex part of an article because results must be integrated and differentiated from past findings (Hartley, Pennebaker, & Fox, 2003).

The use of causal words (e.g., because, effect, hence) and insight words (e.g., think, know, consider), two subcategories of cognitive mechanisms, in describing a past event can suggest the active process of reappraisal. In a reanalysis of six expressive writing studies, Pennebaker, Mayne, and Francis (1997) found that increasing use of causal and insight words led to greater health improvements. This finding suggests that changing from not processing to actively processing an event in combination of emotional writing leads to better outcomes. In these experiments, increasing use of

casual and insight words may be analogous to making reconstrual statements. In other work, use of reconstrual in combination with discussion of a traumatic events has shown to have the best health outcomes (Kross & Ayduk, 2008). Participants in describing a painful relationship breakup used more cognitive mechanisms, particularly causal words, in describing the breakup and postbreakup compared with the prebreakup (Boals & Klein, 2005). The authors argue that causal words are used in the most traumatic parts, the breakup and postbreakup, because they are being used to create causal explanations to organize the participant's thoughts.

The language that people use to discuss an event can reveal something about the extent to which a story may have been established or is still being formed. When people are uncertain or insecure about their topic, they use tentative language (e.g., maybe, perhaps, guess) and more filler words (e.g., blah, I mean, you know). Participants who recounted an event that they had already disclosed to someone else used fewer words from the tentative category than participants who recounted an undisclosed event (Pasupathi, 2007). Possibly, higher use of tentative words suggests that a participant has not yet processed an event and formed it into a story. Similarly, Beaudreau, Storandt, and Strube (2006) found that in recounting a personal story younger participants used more filler words compared with older participants. However, there was no difference in filler words when the two groups described a story based on a picture. In this experiment, use of filler words may suggest the degree to which the story was well formed, presumably older participants had more perspective on the personal life events and may have recounted them many more times than the younger participants.

### *Individual Differences*

The self-focus, cognitive complexity, social references, and emotional tone inherent in language use can help identify individual differences. These linguistic characteristics differ with age, sex, personality, and mental health. Language use, like any behavioral manifestation, can reflect individual differences. These language features can be used to make predictions about individuals and also may underlie causal processes that create some individual differences.

As people age, they become less self-focused, refer more to the moment, and do not decline in verbal complexity. Pennebaker and Stone (2003) examined the writing of participants of varying ages in emotional writing studies. In a second experiment, the authors examined the text of published authors from the span of their writing career. Across these two studies, first-person singular decreased with time, whereas insight words, future tense verbs, and exclusive words increased. The authors observe these patterns of language use both in studies of different individuals at different points in their lives, and of authors over the course of their life. From the results, they reason that there are shifts in self-focus as people age and, counter to expectations, attention to time is more present and future oriented, and verbal complexity may increase or at least stay the same as people age, evidenced by insight words and exclusive words.

Sex differences in language use show that women use more social words and references to others, and men use more complex language. A meta-analysis of the texts

from many studies shows that the largest language differences between males and females are in the complexity of the language used and the degree of social references (Newman, Groom, Handelman, & Pennebaker, 2008). Males had higher use of large words, articles, and prepositions. Females had higher use of social words, and pronouns, including first-person singular and third-person pronouns. There were also large effect sizes for use of swear words, feeling words, and present tense verbs. The fact that there are predictable differences in language used between sexes makes it possible to predict the sex of the user without knowledge of the true sex. An open research question remains what it means if a participant uses sex atypical language.

Studies measuring personality in participants through writing samples (Pennebaker & King, 1999) and spoken dialogue (Mehl, Gosling, & Pennebaker, 2006) have shown that some LIWC categories correspond with big-five personality traits. For example, Mehl and colleagues found that for both males and females higher word count and fewer large words predicted extraversion. Pennebaker and King showed that other LIWC categories showing complexity of language (such as articles, exclusive words, causal words, and negations) were less frequent in the writing of people who scored high on extraversion. Social and emotional language also differed with respect to extraversion; people who scored high on extraversion used more social words, more positive emotion, and less negative emotion. The findings from these two studies partially support traditional personality models. Models of extraversion would predict that extraverts engage in more social interaction, and have a more positive response to that engagement. Also, these models would predict that people high in extroversion would be less inhibited in their language production, possibly leading to less complex language.

Depressed and suicidal individuals are more self-focused, express more negative emotion and sometime use more death-related words. Studies on depression and suicide show that language features can be markers of mental health. Depressed patients are more likely to use more first-person singular and more negative emotion words than participants who have never been depressed in emotional writings (Rude et al., 2004). Suicidal poets in their published works compared with matched nonsuicidal poets use more first-person singular and more death-related words (Stirman & Pennebaker, 2001). This individual difference may show an attentional difference, that is, more self-focus in response to emotional pain, or it may indicate a thinking pattern that is a predilection for experiencing depression (see also work by Wolf, Sedway, Bulik, & Kordy, 2007, dealing with the language of anorexia).

## Conclusion

The function and emotion words people use provide important psychological cues to their thought processes, emotional states, intentions, and motivations. We have summarized some of the LIWC dimensions that reflect language correlates of attentional focus, emotional state, social relationships, thinking styles, and individual differences. This review is, by definition, brief and selective. Word use is highly contextual and many of the findings may not hold with different groups of people or across a wide range of settings. More of the research results have come from labs in the United

States working with college-aged students, often in highly contrived settings. Very little work has explored the differences between spoken and written language.

As can be seen in the appendix, an increasing number of studies are beginning to link daily word use to broader social and psychological processes. What is most striking has been the relatively fast growth of the language-behavior research endeavor.

The connections between language and social psychology are changing at an accelerating rate. When journals such as the *Journal of Language and Social Psychology* were founded, most research was based on written text or transcriptions of spoken text, all of which were hand-typed, hand-scored, and stored in a filing cabinet for later analyses. Researchers interested in language and social processes have historically been trained in laboratory methods whereby participants were run, one at a time, in highly controlled settings to best capture the links between language use, cognitive processing, and communication dynamics.

Innovations in word analysis—as exemplified by Google and Yahoo—are challenging the social psychological methodologies most of us have grown up with. In the amount of time it takes to run a single participant in a social psychology language study, we can now download thousands of personal writings, interaction transcripts, or other forms of text that can be analyzed in seconds. The Internet world provides a far more diverse population from which to draw as well as access to a wide range of languages.

The availability of natural language use and our computational resources are transforming language analysis and modern social science. LIWC represents only a transitional text analysis program in the shift from traditional language analysis to a new era of language analysis. Newer text analysis will be able to analyze more complex language structure while retaining LIWC's transparency. Studies have begun to look at *n*-grams, groups of two or more words together in the same way we have used LIWC to look at frequencies of single words (Oberlander & Gill, 2006). Text analysis methods should also increase in flexibility, allowing the researcher to examine language categories specific to his or her research program. New techniques to automatically extract conceptually related words should be expanded to incorporate related patterns of language style with related content words. From research using LIWC, it has become clear that language style information is critical to understanding a person's state of mind.

Research using these new text analysis methods will also be expanded to capture cultural differences mirrored in language use. Language style conveys subtle information about social relations. The relevant social information can vary greatly between language and cultures (cf. Maass, Karasawa, Politi, & Suga, 2006). Indeed, some of the most striking cultural differences in language—such as markers of politeness, formality, and social closeness—are inherent in function words rather than content words (Boroditsky, Schmidt, & Phillips, 2003).

We are standing on the threshold of a new era of language analysis. One can easily imagine how Jim Bradac would have celebrated the possibilities of tracking natural language across hundreds of millions of people and an unknown number of contexts. The expanding galaxy of computer-based text analysis methods have the potential to add to our current ways of thinking about language and, in Bradac's (1999) words, "burn ever brighter and illuminate the universe increasingly from their different places" (p. 11).

## Appendix

**Summary Table Linking LIWC Word Categories to Published Research Studies**

Category	Examples	Words in Category	Psychological Correlates	Published Articles
Linguistic processes				
Word count			Talkativeness, verbal fluency	2, 9, 18, 19, 20, 24, 32, 35, 36, 39, 40, 48, 53, 54, 57, 60, 66, 70, 72, 73, 74, 86, 89, 103, 115
Words/sentence			Verbal fluency, cognitive complexity	3, 7, 39, 43
Dictionary words	(Percentage of all words captured by the program)		Informal, nontechnical language	19, 42, 43, 65, 66, 85, 89
Words >6 letters	(Percentage of all words longer than 6 letters)		Education, social class	3, 19, 20, 27, 35, 36, 42, 43, 73, 74, 79, 89, 90, 93, 103, 115
Total function words		464	Informal, personal Personal, social Honest, depressed, low status, personal, emotional, informal	1, 19, 36, 43, 55, 89, 90, 119 58, 79 1, 3, 4, 5, 11, 13, 18, 27, 35, 36, 46, 55, 56, 64, 65, 66, 68, 69, 72, 73, 74, 78, 80, 81, 87, 89, 90, 92, 93, 94, 100, 101, 105, 108, 109, 112, 113, 115
Total pronouns	I, them, itself	116		
Personal pronouns	I, them, her	70		
First-person singular	I, me, mine	12		
First-person plural	We, us, our	12	Detached, high status, socially connected to group (sometimes)	1, 4, 13, 18, 35, 46, 55, 64, 65, 74, 78, 81, 87, 90, 93, 94, 97, 100, 103, 104, 105, 106, 113
Second person	You, your; thou	20	Social, elevated status	1, 18, 27, 41, 55, 90, 100, 105, 106
Third-person singular	She, her, him	17	Social interests, social support	1, 3, 14, 36, 39, 55, 64, 66, 80, 87, 88, 90, 95
Third-person plural	They, their, they'd	10	Social interests, out-group awareness (sometimes)	1, 3, 14, 39, 55, 64, 80, 87, 88, 95

(continued)

## Appendix (continued)

Category	Examples	Words in Category	Psychological Correlates	Published Articles
Indefinite pronouns Articles	It, it's, those A, an, the	46 3	Use of concrete nouns, interest in objects and things	19, 36, 43, 74, 79, 80, 89, 92, 115
Common verbs Auxiliary verbs Past tense	Walk, went, see Am, will, have Went, ran, had	383 144 145	Informal, passive voice Focus on the past	1, 13, 37, 62, 68, 73, 79, 87, 89, 91, 93, 115
Present tense	Is, does, hear	169	Living in the here and now	13, 36, 37, 42, 62, 68, 73, 87, 89, 90, 93, 115
Future tense Adverbs Prepositions	Will, gonna Very, really, quickly To, with, above	48 69 60	Future and goal oriented Education, concern with precision	13, 26, 37, 41, 62, 64, 76, 90, 93, 114 58 43, 79, 89, 92, 115
Conjunctions Negations Quantifiers Numbers Swear words	And, but, whereas No, not, never Few, many, much Second, thousand Damn, piss, fuck	28 57 89 34 53	Inhibition Informal, aggression,	24, 39, 40, 48, 79, 89, 90, 114, 115 19, 79 58, 73, 74, 81, 98
Psychological processes Social processes	Mate, talk, they, child	455	Social concerns, social support	1, 18, 23, 27, 32, 35, 41, 55, 78, 79, 85, 88, 89, 90, 93, 95, 97, 115, 116
Family Friends Humans Affective Processes	Daughter, husband Buddy, friend, neighbor Adult, baby, boy Happy, cried, abandon	64 37 61 915	Emotionality	18, 95 18, 95 1, 1 12, 27, 28, 32, 33, 34, 40, 44, 50, 54, 57, 58, 60, 62, 69, 77, 85, 86, 119

(continued)

## Appendix (continued)

Category	Examples	Words in Category	Psychological Correlates	Published Articles
Positive emotion	Love, nice, sweet	406		2, 3, 4, 5, 6, 8, 10, 12, 15, 17, 21, 22, 23, 25, 28, 30, 31, 33, 36, 37, 38, 41, 45, 46, 47, 48, 49, 50, 51, 53, 54, 55, 57, 59, 60, 61, 62, 64, 66, 67, 68, 69, 70, 71, 73, 74, 75, 76, 77, 81, 82, 85, 89, 91, 93, 94, 96, 99, 107, 108, 109, 110, 113, 115, 117, 118
Negative emotion	Hurt, ugly, nasty	499		2, 3, 4, 6, 10, 12, 13, 16, 17, 20, 21, 22, 25, 28, 29, 30, 31, 33, 35, 37, 40, 44, 45, 46, 47, 48, 50, 51, 52, 53, 55, 57, 59, 61, 62, 63, 64, 66, 67, 70, 71, 72, 73, 74, 76, 79, 80, 81, 82, 84, 85, 89, 91, 92, 93, 94, 96, 99, 102, 107, 113, 115, 117, 119, 121
Anxiety	Worried, nervous	91		6, 28, 50, 66, 68, 77, 84, 85, 92
	Hate, kill, annoyed	184		6, 28, 33, 50, 58, 66, 72, 74, 92
	Crying, grief, sad	101		6, 28, 33, 38, 50, 63, 66, 77, 84, 90
Cognitive processes	Cause, know, ought	730		2, 3, 5, 8, 13, 18, 21, 23, 31, 32, 34, 46, 47, 49, 55, 58, 61, 68, 69, 71, 75, 83, 84, 85, 86, 89, 92, 93, 102, 104, 119, 120
Insight	Think, know, consider	195		1, 4, 18, 19, 25, 35, 37, 45, 53, 59, 68, 73, 76, 89, 90, 91, 92, 93, 97, 99, 111, 113, 115, 118, 119, 121
Causation	Because, effect, hence	108		10, 13, 16, 20, 35, 37, 39, 45, 53, 72, 76, 89, 90, 91, 93, 97, 99, 115, 121, 122
Discrepancy	Should, would, could	76		10, 16, 18, 19, 49, 63, 74, 89, 115
Tentative	Maybe, perhaps, guess	155		18, 19, 24, 37, 38, 49, 73, 87, 89, 98, 115
Certainty	Always, never	83	Social/verbal skills, emotional stability	38
Inhibition	Block, constrain, stop	111		1, 16, 18, 19, 49, 90, 111
Inclusive	And, with, include	18		41, 60, 73, 74, 89, 115
Exclusive	But, without, exclude	17	Cognitive complexity, honesty	24, 49, 73, 80, 89, 92, 93, 115

(continued)

## Appendix (continued)

Category	Examples	Words in Category	Psychological Correlates	Published Articles
Perceptual processes	Observing, heard, feeling	273		14, 37, 120
See	View, saw, seen	72		36
Hear	Listen, hearing	51		13, 41
Feel	Feels, touch	75		13, 88
Biological processes	Eat, blood, pain	567		36
Body	Cheek, hands, spit	180		34, 36, 37, 49, 116
Health	Clinic, flu, pill	236		
Sexual	Horny, love, incest	96		36, 94, 96, 112
Ingestion	Dish, eat, pizza	111		68, 94
Relativity	Area, bend, go	638		49, 110
Motion	Arrive, car, go	168		14, 37, 80
Space	Down, in, thin	220		14, 120
Time	End, until, season	239		1, 13, 41, 64, 93, 119, 120
Personal concerns				
Work	Job, majors, xerox	327		36
Achievement	Earn, hero, win	186		36, 60, 103
Leisure	Cook, chat, movie	229		
Home	Apartment, kitchen, family	93		79
Money	Audit, cash, owe	173		
Religion	Altar, church, mosque	159		41, 94
Death	Bury, coffin, kill	62		1, 2, 4, 35, 64, 68, 91, 94
Spoken categories				
Assent	Agree, OK, yes	30	Agreement, passivity	48, 60, 81
Nonfluencies	Er, hm, umm	8		74
Fillers	Blah, I mean, yaknow	9	Informal, Unprepared speech	9, 74

## Appendix (continued)

### References Cited in the Table

1. Alexander-Emery, S., Cohen, L. M., & Prensky, E. H. (2005). Linguistic analysis of college aged smokers and never smokers. *Journal of Psychopathology and Behavioral Assessment*, 27, 11-16.
2. Alvarez-Conrad, J., Zoellner, L. A., & Foa, E. B. (2001). Linguistic predictors of trauma pathology and physical health. *Applied Cognitive Psychology*, 15, 159-170.
3. Arguello, J., Butler, B. S., Joyce, E., Kraut, R., Ling, K. S., Rosé, C., et al. (2006). Talk to me: Foundations for successful individual-group interactions in online communities. In *Proceedings of the CHI'06 conference on human factors in computing systems* (pp. 959-968). New York: Association for Computing Machinery Press.
4. Baddeley, J. L., & Singer, J. A. (2008). Telling losses: Functions and personality correlates of bereavement narratives. *Journal of Research in Personality*, 42, 421-438.
5. Baikie, K. A., Wilhelm, K., Johnson, B., Boskovic, M., Wedgwood, L., Finch, A., et al. (2006). Expressive writing for high-risk drug dependent patients in a primary care clinic: A pilot study. *Harm Reduction Journal*, 3, 34-42.
6. Bantum, E. O., & Owen, J. E. (2009). Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological Assessment*, 21, 79-88.
7. Barnes, D. H. (2007). Letters from a suicide. *Death Studies*, 31, 671-678.
8. Batten, S. V., Follette, V. M., Rasmussen Hall, M. L., & Palm, K. M. (2002). Physical and psychological effects of written disclosure among sexual abuse survivors. *Behavior Therapy*, 33, 107-122.
9. Beaudreau, S. A., Storandt, M., & Strube, M. J. (2006). A comparison of narratives told by younger and older adults. *Experimental Aging Research*, 32, 105-117.
10. Beevers, C. G., & Scott, W. D. (2001). Ignorance may be bliss, but thought suppression promotes superficial cognitive processing. *Journal of Research in Personality*, 35, 546-553.
11. Block-Lerner, J., Adair, C., Plumb, J. C., Rhatigan, D. L., & Orsillo, S. M. (2007). The case for mindfulness-based approaches in the cultivation of empathy: Does nonjudgmental, present-moment awareness increase capacity for perspective-taking and empathic concern? *Journal of Marital & Family Therapy*, 33, 501-516.
12. Blonder, L. X., Heilman, K. M., Ketterson, T., Rosenbek, J., Raymer, A., Crosson, B., et al. (2005). Affective facial and lexical expression in a prosodic versus aphasic stroke patients. *Journal of the International Neuropsychological Society*, 11, 677-685.
13. Boals, A., & Klein, K. (2005). Word use in emotional narratives about failed romantic relationships and subsequent mental health. *Journal of Language and Social Psychology*, 24, 252-268.
14. Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, 19, 313-329.

---

(continued)

**Appendix (continued)**

15. Bono, J. E., & Ilies, R. (2006). Charisma, positive emotions and mood contagion. *The Leadership Quarterly, 17*, 317-334.
16. Brett, J. M., Olekalns, M., Friedman, R., Goates, N., Anderson, C., & Lisco, C. C. (2007). Sticks and stones: Language, face, and online dispute resolution. *Academy of Management Journal, 50*, 85-99.
17. Broderick, J. E., Junghaenel, D. U., & Schwartz, J. E. (2005). Written emotional expression produces health benefits in fibromyalgia patients. *Psychosomatic Medicine, 67*, 326-334.
18. Burke, P. A., & Dollinger, S. J. (2005). A picture's worth a thousand words: Language use in autophotographic essay. *Personality and Social Psychology Bulletin, 31*, 536-548.
19. Centerbar, D. B., Schnall, S., Clore, G. L., & Garvin, E. D. (2008). Affective incoherence: When affective concepts and embodied reactions clash. *Journal of Personality and Social Psychology, 94*, 560-578.
20. Chung, C. K., & Pennebaker, J. W. (2008). Variations in the spacing of expressive writing sessions. *British Journal of Health Psychology, 13*, 15-21.
21. Cohen, A. S., Minor, K. S., Baillie, L. E., & Dahir, A. M. (2008). Clarifying the linguistic signature: Measuring personality from natural speech. *Journal of Personality Assessment, 90*, 559-563.
22. Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science, 15*, 687-693.
23. Carter, A. L., & Petrie, K. J. (2008). Expressive writing in context: The effects of a confessional setting and delivery of instructions on participant experience and language in writing. *British Journal of Health Psychology, 13*, 27-30.
24. Creswell, J. D., Lam, S., Stanton, A. L., Taylor, S. E., Bower, J. E., & Sherman, D. K. (2007). Does self-affirmation, cognitive processing, or discovery of meaning explain cancer-related health benefits of expressive writing? *Personality and Social Psychology Bulletin, 33*, 238-250.
25. DiNardo, A. C., Schober, M. F., & Stuart, J. (2005). Chair and couch discourse: A study of visual copresence in psychoanalysis. *Discourse Processes, 40*, 209-238.
26. Dino, A., Reysen, S., & Branscombe, N. R. (2009). Online interactions between group members who differ in status. *Journal of Language and Social Psychology, 28*, 85-94.
27. Djikic, M., Oatley, K., & Peterson, J. B. (2006). The bitter-sweet labor of emoting: The linguistic comparison of writers and physicists. *Creativity Research Journal, 18*, 191-197.
28. D'Souza, P., Lumley, M., Kraft, C., & Dooley, J. (2008). Relaxation training and written emotional disclosure for tension or migraine headaches: A randomized, controlled trial. *Annals of Behavioral Medicine, 36*, 21-32.
29. Eid, J., Johnsen, B., Helge, R. N., & Saus, E. R. (2005). Trauma narratives and emotional processing. *Scandinavian Journal of Psychology, 46*, 503-510.
30. Epstein, E. M., Sloan, D. M., & Marx, B. P. (2005). Getting to the heart of the matter: Written disclosure, gender, and heart rate. *Psychosomatic Medicine, 67*, 413-419.

---

*(continued)*

## Appendix (continued)

---

31. Friedman, S. R., Rapport, L. J., Lumley, M., Tzelepis, A., VanVoorhis, A., Stettner, L., et al. (2003). Aspects of social and emotional competence in adult attention-deficit/hyperactivity disorder. *Neuropsychology, 17*, 50-58.
32. Gill, A. J., French, R. M., Gergle, D., & Oberlander, J. (2008). The language of emotion in short blog texts. In *Proceedings of the CSCW'08 computer supported cooperative work* (pp. 299-302). New York: Association for Computing Machinery Press.
33. Gillis, M. E., Lumley, M. A., Mosley-Williams, A., Leisen, J. C. C., & Roehrs, T. (2006). The health effects of at-home written emotional disclosure in fibromyalgia: A randomized trial. *Annals of Behavioral Medicine, 32*, 135-146.
34. Gortner, E. M., & Pennebaker, J. W. (2003). The archival anatomy of a disaster: Media coverage and community-wide health effects of the Texas A&M bonfire tragedy. *Journal of Social and Clinical Psychology, 22*, 580-603.
35. Groom, C. J., & Pennebaker, J. W. (2005). The language of love: Sex, sexual orientation, and language use in online personal advertisements. *Sex Roles, 52*, 447-461.
36. Guastella, A. J., & Dadds, M. R. (2006). Cognitive-behavioral models of emotional writing: A validation study. *Cognitive Therapy and Research, 30*, 397-414.
37. Hamilton-West, K. E. (2007). Effects of written emotional disclosure on health outcomes in patients with ankylosing spondylitis. *Psychology & Health, 22*, 637-657.
38. Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 45*, 1-23.
39. Hancock, J. T., Landrigan, C., & Silver, C. (2007). Expressing emotion in text-based communication. In *Proceedings of the CHI'07 conference on human factors in computing systems* (pp. 929-932). New York: Association for Computing Machinery Press.
40. Handelman, L. D., & Lester, D. (2007). The content of suicide notes from attempters and completers. *Crisis, 28*, 102-104.
41. Hartley, J. (2003). Improving the clarity of journal abstracts in psychology: The case for structure. *Science Communication, 24*, 366-379.
42. Hartley, J., Pennebaker, J. W., & Fox, C. (2003). Abstracts, introductions and discussions: How far do they differ in style? *Scientometrics, 57*, 389-398.
43. Heberlein, A. S., Adolphs, R., Pennebaker, J. W., & Tranel, D. (2003). Effects of damage to right-hemisphere brain structures on spontaneous emotional and social judgments. *Political Psychology, 24*, 705-726.
44. Hemenover, S. H. (2003). The good, the bad, and the healthy: Impacts of emotional disclosure of trauma on resilient self-concept and psychological distress. *Personality and Social Psychology Bulletin, 29*, 1236-1244.
45. Hoyt, T., & Pasupathi, M. (2008). Blogging about trauma: Linguistic measures of apparent recovery [Electronic version]. *Journal of Applied Psychology, 93*, 4.
46. Jones, S. M., & Wirtz, J. G. (2006). How does the comforting process work? An empirical test of an appraisal-based model of comforting. *Human Communication Research, 32*, 217-243.

---

(continued)

**Appendix (continued)**

47. Joyce, E., & Kraut, R. E. (2006). Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication, 11*, 723-747.
48. Junghaenel, D. U., Smyth, J. M., & Santner, L. (2008). Linguistic dimensions of psychopathology: A quantitative analysis. *Journal of Social and Clinical Psychology, 27*, 36-55.
49. Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *American Journal of Psychology, 120*, 263-286.
50. Kiesler, S., Lee, S., & Kramer, A. D. I. (2006). Relationship effects in psychological explanations of nonhuman behavior. *Anthrozoos, 19*, 335-352.
51. King, E. B., Shapiro, J. R., Hebl, M. R., Singletary, S. L., & Turner, S. (2006). The stigma of obesity in customer service: A mechanism for remediation and bottom-line consequences of interpersonal discrimination. *Journal of Applied Psychology, 91*, 579-593.
52. Klein, K., & Boals, A. (2001). Expressive writing can increase working memory capacity. *Journal of Experimental Psychology: General, 130*, 520-533.
53. Knight, J. L., & Hebl, M. R. (2005). Affirmative reaction: The influence of type of justification on nonbeneficiary attitudes toward affirmative action plans in higher education. *Journal of Social Issues, 61*, 547-568.
54. Kramer, A. D. I., Oh, L. M., & Fussell, S. R. (2006). Using linguistic features to measure presence in computer-mediated communication. In *Proceedings of the CHI'06 conference on human factors in computing systems* (pp. 913-916). New York: Association for Computing Machinery Press.
55. Kross, E., & Ayduk, O. (2008). Facilitating adaptive emotional analysis: Distinguishing distanced-analysis of depressive experiences from immersed-analysis and distraction. *Personality and Social Psychology Bulletin, 34*, 924-938.
56. Lambie, J. A., & Baker, K. L. (2003). Article details Intentional avoidance and social understanding in repressors and nonrepressors: Two functions for emotion experience? *Consciousness and Emotion, 4*, 17-42.
57. Lee, C. H., Kim, K., Seo, Y. S., & Chung, C. K. (2007). The relations between personality and language use. *Journal of General Psychology, 134*, 405-413.
58. Lepore, S. J. (1997). Expressive writing moderates the relation between intrusive thoughts and depressive symptoms. *Journal of Personality and Social Psychology, 73*, 1030-1037.
59. Leshed, G., Hancock, J. T., Cosley, D., McLeod, P. L., & Gay, G. (2007). Feedback for guiding reflection on teamwork practices. In *Proceedings of the GROUP'07 conference on supporting group work* (pp. 217-220). New York: Association for Computing Machinery Press.
60. Lieberman, M. A. (2008). Effects of disease and leader type on moderators in online support groups. *Computers in Human Behavior, 24*, 2446-2455.
61. Liehr, P., Takahashi, R., Nishimura, C., Frazier, L., Kuwajima, I., & Pennebaker, J. W. (2002). Expressing health experience through embodied language. *Journal of Nursing Scholarship, 34*, 27-32.

---

*(continued)*

## Appendix (continued)

---

62. Liess, A., Simon, W., Yutsis, M., Owen, J. E., Piemme, K. A., Golant, M., et al. (2008). Detecting emotional expression in face-to-face and online breast cancer support groups. *Journal of Consulting and Clinical Psychology*, 76, 517-523.
  63. Lightman, E. J., McCarthy, P. M., Dufty, D. F., & McNamara, D. S. (2007). Using computational text analysis tools to compare the lyrics of suicidal and non-suicidal songwriters. In D. S. McNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
  64. Lillard, A., Nishida, T., Massaro, D., Vaish, A., Ma, L., & McRoberts, G. (2007). Signs of pretense across age and scenario. *Infancy*, 11, 1-30.
  65. Lockenhoff, C. E., Costa, P. T., Jr., & Lane, R. D. (2008). Age differences in descriptions of emotional experiences in oneself and others. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 63, 92-99.
  66. Luterek, J. A., Orsillo, S. M., & Marx, B. P. (2005). An experimental examination of emotional experience, expression, and disclosure in women reporting a history of childhood sexual abuse. *Journal of Traumatic Stress*, 18, 237-244.
  67. Lyons, E. J., Mehl, M. R., & Pennebaker, J. W. (2006). Pro-anorexics and recovering anorexics differ in their linguistic Internet self-presentation. *Journal of Psychosomatic Research*, 60, 253-256.
  68. Mackenzie, C. S., Wiprzycka, U. J., Hasher, L., & Goldstein, D. (2007). Does expressive writing reduce stress and improve health for family caregivers of older adults? *The Gerontologist*, 47, 296-306.
  69. Manne, S. (2002). Language use and post-traumatic stress symptomatology in parents of pediatric cancer survivors 1. *Journal of Applied Social Psychology*, 32, 608-629.
  70. McCullough, M. E., Root, L. M., & Cohen, A. D. (2006). Writing about the benefits of an interpersonal transgression facilitates forgiveness. *Journal of Consulting and Clinical Psychology*, 74, 887-897.
  71. Mehl, M. R. (2006). The lay assessment of subclinical depression in daily life. *Psychological Assessment*, 18, 340-345.
  72. Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90, 862-877.
  73. Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, 84, 857-870.
  74. van Middendorp, H., & Geenen, R. (2008). Poor cognitive-emotional processing may impede the outcome of emotional disclosure interventions. *British Journal of Health Psychology*, 13, 49-52.
  75. van Middendorp, H., Sorbi, M. J., van Doornen, L. J. P., Bijlsma, J. W. J., & Geenen, R. (2007). Feasibility and induced cognitive-emotional change of an emotional disclosure intervention adapted for home application. *Patient Education and Counseling*, 66, 177-187.
- 

(continued)

**Appendix (continued)**

76. Morgan, N. P., Graves, K. D., Poggi, E. A., & Cheson, B. D. (2008). Implementing an expressive writing study in a cancer clinic. *The Oncologist, 13*, 196-204.
77. Neff, K. D., Kirkpatrick, K. L., & Rude, S. S. (2007). Self-compassion and adaptive functioning. *Journal of Research in Personality, 41*, 139-154.
78. Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes, 45*, 211-236.
79. Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665-675.
80. Oliver, E. J., Markland, D., Hardy, J., & Petherick, C. M. (2008). The effects of autonomy-supportive versus controlling environments on self-talk. *Motivation and Emotion, 32*, 200-212.
81. Orsillo, S. M., Batten, S. V., Plumb, J. C., Luterek, J. A., & Roessner, B. M. (2004). An experimental study of emotional responding in women with posttraumatic stress disorder related to interpersonal violence. *Journal of Traumatic Stress, 17*, 241-248.
82. Owen, J. E., Giese-Davis, J., Cordova, M., Kronenwetter, C., Golant, M., & Spiegel, D. (2006). Self-report and linguistic indicators of emotional expression in narratives as predictors of adjustment to cancer. *Journal of Behavioral Medicine, 29*, 335-345.
83. Owen, J. E., Klapow, J. C., Roth, D. L., Shuster, J. L., Bellis, J., Meredith, R., et al. (2005). Randomized pilot of a self-guided Internet coping group for women with early-stage breast cancer. *Annals of Behavioral Medicine, 30*, 54-64.
84. Owen, J. E., Klapow, J. C., Roth, D. L., & Tucker, D. C. (2004). Use of the internet for information and support: disclosure among persons with breast and prostate cancer. *Journal of Behavioral Medicine, 27*, 491-505.
85. Owen, J. E., Yarbrough, E. J., Vaga, A., & Tucker, D. C. (2003). Investigation of the effects of gender and preparation on quality of communication in Internet support groups. *Computers in Human Behavior, 19*, 259-275.
86. Pasupathi, M. (2007). Telling and the remembered self: Linguistic differences in memories for previously disclosed and previously undisclosed events. *Memory, 15*, 258-270.
87. Pennebaker, J. W., Groom, C. J., Loew, D., & Dabbs, J. M. (2004). Testosterone as a social inhibitor: two case studies of the effect of testosterone treatment on language. *Journal of Abnormal Psychology, 113*, 172-175.
88. Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*, 1296-1312.
89. Pennebaker, J. W., & Lay, T. C. (2002). Language use and personality during crises: Analyses of mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality, 36*, 271-282.
90. Pennebaker, J. W., Mayne, T. J., & Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology, 72*, 863-871.

(continued)

## Appendix (continued)

91. Pennebaker, J. W., Slatcher, R. B., & Chung, C. K. (2005). Linguistic markers of psychological state through media interviews: John Kerry and John Edwards in 2004, Al Gore in 2000. *Analyses of Social Issues and Public Policy*, 5, 197-204.
92. Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85, 291-301.
93. Pennebaker, J. W., & Stone, L. D. (2004). What was she trying to say? A linguistic analysis of Katie's diaries. In D. Lester (Ed.), *Katie's diary: Unlocking the mystery of a suicide* (pp. 55-80). New York: Brunner-Routledge.
94. Pressman, S. D., & Cohen, S. (2007). Use of social words in autobiographies and longevity. *Psychosomatic Medicine*, 69, 262-269.
95. Rellini, A. H., & Meston, C. M. (2007). Sexual desire and linguistic analysis: A comparison of sexually-abused and non-abused women. *Archives of Sexual Behavior*, 36, 67-77.
96. Rew, L. (2007). A linguistic investigation of mediators between religious commitment and health behaviors in older adolescents. *Issues in Comprehensive Pediatric Nursing*, 30, 71-86.
97. Robertson, K., & Murachver, T. (2006). Intimate partner violence: Linguistic features and accommodation behavior of perpetrators and victims. *Journal of Language and Social Psychology*, 25, 406-422.
98. Rogers, L. J., Wilson, K. G., Gohm, C. L., & Merwin, R. M. (2007). Revisiting written disclosure: The effects of warm versus cold experimenters. *Journal of Social and Clinical Psychology*, 26, 556-574.
99. Rohrbaugh, M. J., Mehl, M. R., Shoham, V., Reilly, E. S., & Ewy, G. A. (2008). Prognostic significance of spouse we talk in couples coping with heart failure. *Journal of Consulting and Clinical Psychology*, 76, 781-789.
100. Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18, 1121-1133.
101. Schwartz, L., & Drotar, D. (2004). Linguistic analysis of written narratives of caregivers of children and adolescents with chronic illness: Cognitive and emotional processes and physical and psychological health outcomes. *Journal of Clinical Psychology in Medical Settings*, 11, 291-301.
102. Sexton, J. B., & Helmreich, R. L. (2000). Analyzing cockpit communications: The links between language, performance, and workload. *Human Performance in Extreme Environments*, 5, 63-68.
103. Sharp, W. G., & Hargrove, D. S. (2004). Emotional expression and modality: An analysis of affective arousal and linguistic output in a computer versus paper paradigm. *Computers in Human Behavior*, 20, 461-475.
104. Simmons, R. A., Chambliss, D. L., & Gordon, P. C. (2008). How do hostile and emotionally overinvolved relatives view relationships? What relatives' pronoun use tells us. *Family Process*, 47, 405-419.
105. Simmons, R. A., Gordon, P. C., & Chambliss, D. L. (2005). Pronouns in marital interaction. *Psychological Science*, 16, 932-936.

(continued)

**Appendix (continued)**

- 
106. Slatcher, R. B., & Pennebaker, J. W. (2006). How do I love thee? Let me count the words: The social effects of expressive writing. *Psychological Science, 17*, 660-664.
  107. Slatcher, R. B., Vazire, S., & Pennebaker, J. W. (2008). Am "I" more important than "we"? Couples' word use in instant messages. *Personal Relationships, 15*, 407-424.
  108. Sloan, D. M. (2005). It's all about me: Self-focused attention and depressed mood. *Cognitive Therapy and Research, 29*, 279-288.
  109. Soliday, E., Garofalo, J. P., & Rogers, D. (2004). Expressive writing intervention for adolescents' somatic symptoms and mood. *Journal of Clinical Child and Adolescent Psychology, 33*, 792-801.
  110. Stephenson, G. M., Laszlo, J., Ehmann, B., Lefever, R. M. H., & Lefever, R. (1997). Diaries of significant events: Socio-linguistic correlates of therapeutic outcomes in patients with addiction problems. *Journal of Community and Applied Social Psychology, 7*, 389-411.
  111. Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine, 63*, 517-522.
  112. Stone, L. D., & Pennebaker, J. W. (2002). Trauma in real time: Talking and avoiding online conversations about the death of Princess Diana. *Basic and Applied Social Psychology, 24*, 173-183.
  113. Swaab, R. I., Phillips, K. W., Diermeier, D., & Husted Medvec, V. (2008). The pros and cons of dyadic side conversations in small groups: The impact of group norms and task type. *Small Group Research, 39*, 372-390.
  114. Taylor, P. J., & Thomas, S. (2008). Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research, 1*, 263-281.
  115. Tsai, J. L., Simeonova, D. I., & Watanabe, J. T. (2004). Somatic and social: Chinese Americans talk about emotion. *Personality and Social Psychology Bulletin, 30*, 1226-1238.
  116. Tull, M. T., Medaglia, E., & Roemer, L. (2005). An investigation of the construct validity of the 20-Item Toronto Alexithymia Scale through the use of a verbalization task. *Journal of Psychosomatic Research, 59*, 77-84.
  117. VandeCreek, L., Janus, M. D., Pennebaker, J. W., & Binau, B. (2002). Praying about difficult experiences as self-disclosure to God. *International Journal for the Psychology of Religion, 12*, 29-39.
  118. Vedhara, K., Morris, R. M., Booth, R., Horgan, M., Lawrence, M., & Birchall, N. (2007). Changes in mood predict disease activity and quality of life in patients with psoriasis following emotional disclosure. *Journal of Psychosomatic Research, 62*, 611-619.
  119. Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. *Law and Human Behavior, 31*, 499-518.
  120. Warner, L. J., Lumley, M. A., Casey, R. J., Pierantoni, W., Salazar, R., Zoratti, E. M., et al. (2006). Health effects of written emotional disclosure in adolescents with asthma: A randomized, controlled trial. *Journal of Pediatric Psychology, 31*, 557-568.
  121. Watkins, E. (2004). Adaptive and maladaptive ruminative self-focus during emotional processing. *Behaviour Research and Therapy, 42*, 1037-1052.
-

### Authors' Note

The original version of this article was presented as part of the James J. Bradac Memorial Lecture at the University of California at Santa Barbara in 2008.

### Declaration of Conflicting Interests

The text analysis program, LIWC, is a commercial product co-owned by Pennebaker. Proceeds from his share of the profits are all donated to the University of Texas at Austin. The authors declared no other conflicts of interests with respect to authorship and/or publication of this article.

### Funding

The authors disclosed receipt of the following financial support for the research and/or authorship of this article:

Army Research Institute (W91WAW-07-C002), DOD-CIFA (H9C104-07-C0019), and Sandia National Laboratories (26-3963-70).

### References

- Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., et al. (2005). Evaluation of computerized text analysis in an Internet breast cancer support group. *Computers in Human Behavior*, 21, 361-376.
- Beaudreau, S. A., Storandt, M., & Strube, M. J. (2006). A comparison of narratives told by younger and older adults. *Experimental Aging Research*, 32, 105-117.
- Berlyne, D. E. (1960). *Conflict, arousal, and curiosity*. New York: McGraw-Hill.
- Boals, A., & Klein, K. (2005). Word use in emotional narratives about failed romantic relationships and subsequent mental health. *Journal of Language and Social Psychology*, 24, 252-268.
- Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, 19, 313-329.
- Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 61-79). Cambridge: MIT Press.
- Bradac, J. J. (1986). Threats to generalization in the use of elicited, purloined, and contrived messages in human communication research. *Communication Quarterly*, 34, 55-65.
- Bradac, J. J. (1999). Language1 . . . n and Social Interaction1 . . . n: Nature abhors uniformity. *Research on Language and Social Interaction*, 32, 11-20.
- Bulwer-Lytton, E. (1842). *Paul Clifford*. Leipzig, Germany: B. Tauchnitz.
- Chung, C. K., & Pennebaker, J. W. (2007). The psychological function of function words. In K. Fiedler (Ed.), *Social communication: Frontiers of social psychology* (pp. 343-359). New York: Psychology Press.
- Davis, D., & Brock, T. C. (1975). Use of first person pronouns as a function of increased objective self-awareness and performance feedback. *Journal of Experimental Social Psychology*, 11, 381-388.

- Freud, S. (1901). *Psychopathology of everyday life*. New York: Basic Books.
- Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (in press). Language indicators of social dynamics in small groups. *Communication Research*.
- Gottschalk, L. A., & Bechtel, R. (1993). *Computerized content analysis of natural language or verbal texts*. Palo Alto, CA: Mind Garden.
- Gottschalk, L. A., & Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior*. Berkeley: University of California Press.
- Gottschalk, L. A., Gleser, G. C., Daniels, R., & Block, S. (1958). The speech patterns of schizophrenic patients: a method of assessing relative degree of personal disorganization and social alienation. *Journal of Nervous and Mental Disease*, 127, 153-166.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36, 193-202.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74, 1464-1480.
- Gunsch, M. A., Brownlow, S., Haynes, S. E., & Mabe, Z. (2000). Differential linguistic content of various forms of political advertising. *Journal of Broadcasting & Electronic Media*, 44, 27-42.
- Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45, 1-23.
- Hart, R. P. (1984). *Verbal style and the presidency: A computer-based analysis*. New York: Academic Press.
- Hartley, J., Pennebaker, J. W., & Fox, C. (2003). Abstracts, introductions and discussions: How far do they differ in style? *Scientometrics*, 57, 389-398.
- Holmes, D., Alpers, G. W., Ismailji, T., Classen, C., Wales, T., Cheasty, V., et al. (2007). Cognitive and emotional processing in narratives of women abused by intimate partners. *Violence Against Women*, 13, 1192-1205.
- Holtzman, W. H. (1950). Validation studies of the Rorschach test: Shyness and gregariousness in the normal superior adult. *Journal of Clinical Psychology*, 6, 343-347.
- Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2009). *The language of social hierarchies*. Manuscript submitted for publication.
- Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *American Journal of Psychology*, 120, 263-286.
- Kowalski, R. M. (2000). "I was Only Kidding!" Victims' and perpetrators' perceptions of teasing. *Personality and Social Psychology Bulletin*, 26, 231-241.
- Kross, E., & Ayduk, O. (2008). Facilitating adaptive emotional analysis: Distinguishing distanced-analysis of depressive experiences from immersed-analysis and distraction. *Personality and Social Psychology Bulletin*, 34, 924-938.
- Leshed, G., Hancock, J. T., Cosley, D., McLeod, P. L., & Gay, G. (2007). Feedback for guiding reflection on teamwork practices. In *Proceedings of the 2007 international ACM conference on supporting group work* (pp. 217-220). New York: Association for Computing Machinery Press.

- Maass, A., Karasawa, M., Politi, F., & Suga, S. (2006). Do verbs and adjectives play different roles in different cultures? A cross-linguistic analysis of person representation. *Journal of Personality and Social Psychology, 90*, 734-750.
- Martindale, C. (1990). *The clockwork muse: The predictability of artistic change*. New York: Basic Books.
- McClelland, D. C. (1979). Inhibited power motivation and high blood pressure in men. *Journal of Abnormal Psychology, 88*, 182-190.
- Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology, 90*, 862-877.
- Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology, 64*, 1306-1315.
- Miller, G. (1995). *The science of words*. New York: Scientific American Library.
- Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes, 45*, 211-236.
- Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665-675.
- Oberlander, J., & Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes, 42*, 239-270.
- Pasupathi, M. (2007). Telling and the remembered self: Linguistic differences in memories for previously disclosed and previously undisclosed events. *Memory, 15*, 258-270.
- Pennebaker, J. W. (1982). *The psychology of physical symptoms*. New York: Springer-Verlag.
- Pennebaker, J. W., & Beall, K. S. (1986). Confronting a traumatic event: Toward an understanding of inhibition and disease. *Journal of Abnormal Psychology, 95*, 274-281.
- Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC [Computer software]. Austin, TX: LIWC.net.
- Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007* [LIWC manual]. Austin, TX: LIWC.net.
- Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*, 1296-1312.
- Pennebaker, J. W., & Lay, T. C. (2002). Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality, 36*, 271-282.
- Pennebaker, J. W., Mayne, T. J., & Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology, 72*, 863-871.
- Pennebaker, J. W., Slatcher, R. B., & Chung, C. K. (2005). Linguistic markers of psychological state through media interviews: John Kerry and John Edwards in 2004, Al Gore in 2000. *Analyses of Social Issues and Public Policy, 5*, 197-204.
- Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology, 85*, 291-301.
- Rorschach, H. (1921). *Psychodiagnostik*. Leipzig, Germany: Ernst Bircher Verlag.
- Rosenberg, S. D., & Tucker, G. J. (1978). Verbal behavior and schizophrenia: The semantic dimension. *Archives of General Psychiatry, 36*, 1331-1337.

- Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18*, 1121-1133.
- Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology, 54*, 558-568.
- Sexton, J. B., & Helmreich, R. L. (2000). Analyzing cockpit communications: The links between language, performance, and workload. *Human Performance in Extreme Environments, 5*, 63-68.
- Simmons, R. A., Chambliss, D. L., & Gordon, P. C. (2008). How do hostile and emotionally overinvolved relatives view relationships? What relatives' pronoun use tells us. *Family Process, 47*, 405-419.
- Simmons, R. A., Gordon, P. C., & Chambliss, D. L. (2005). Pronouns in marital interaction. *Psychological Science, 16*, 932-936.
- Slatcher, R. B., Vazire, S., & Pennebaker, J. W. (2008). Am "I" more important than "we"? Couples' word use in instant messages. *Personal Relationships, 15*, 407-424.
- Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine, 63*, 517-522.
- Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. Cambridge: MIT Press.
- Tetlock, P. E. (1981). Pre- to post-election shifts in presidential rhetoric: Impression management or cognitive adjustment. *Journal of Personality and Social Psychology, 41*, 207-212.
- Weintraub, W. (1981). *Verbal behavior: Adaptation and psychopathology*. New York: Springer.
- Weintraub, W. (1989). *Verbal behavior in everyday life*. New York: Springer.
- Winter, D. G. (1998). A motivational analysis of the Clinton first term and the 1996 presidential campaign. *The Leadership Quarterly, 9*, 367-376.
- Wolf, M., Sedway, J., Bulik, C. M., & Kordy, H. (2007). Linguistic analyses of natural written language: Unobtrusive assessment of cognitive style in eating disorders. *International Journal of Eating Disorders, 40*, 711-717.
- Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation, 13*, 81-106.

## Bios

**Yla R. Tausczik** is a doctoral student in the Department of Psychology at the University of Texas at Austin. She received her BA at the University of California at Berkeley in 2005. Her research interests include using language to understand group dynamics and natural language use in the workplace.

**James W. Pennebaker** (PhD, University of Texas, Austin) is a professor and chair of the Department of Psychology at the University of Texas at Austin. He is the author of multiple books, including *Opening Up: The Healing Power of Expressing Emotions* (1997). He has recently published in *Science*, *Psychological Science*, and *Journal of Personality and Social Psychology*.

## Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents

Peter Sheridan Dodds · Christopher M. Danforth

Published online: 17 July 2009  
© The Author(s) 2009. This article is published with open access at Springerlink.com

**Abstract** The importance of quantifying the nature and intensity of emotional states at the level of populations is evident: we would like to know how, when, and why individuals feel as they do if we wish, for example, to better construct public policy, build more successful organizations, and, from a scientific perspective, more fully understand economic and social phenomena. Here, by incorporating direct human assessment of words, we quantify happiness levels on a continuous scale for a diverse set of large-scale texts: song titles and lyrics, weblogs, and State of the Union addresses. Our method is transparent, improvable, capable of rapidly processing Web-scale texts, and moves beyond approaches based on coarse categorization. Among a number of observations, we find that the happiness of song lyrics trends downward from the 1960s to the mid 1990s while remaining stable within genres, and that the happiness of blogs has steadily increased from 2005 to 2009, exhibiting a striking rise and fall with blogger age and distance from the Earth's equator.

**Keywords** Happiness · Hedonometer · Measurement · Emotion · Written expression · Remote sensing · Blogs · Song lyrics · State of the Union addresses

### 1 Introduction

The desire for well-being and the avoidance of suffering arguably underlies all behavior (Argyle 2001; Layard 2005; Gilbert 2006; Snyder and Lopez 2009). Indeed, across a wide range of cultures, people regularly rank happiness as what they want most in life (Argyle 2001; Layard 2005; Lyubomirsky 2007) and numerous countries have attempted to introduce indices of well-being, such as Bhutan's National Happiness Index. Such a focus is not new: Plato held that achieving eudaimonia (flourishing) was an individual's true goal

---

P. S. Dodds (✉) · C. M. Danforth (✉)

Department of Mathematics and Statistics, Vermont Advanced Computing Center, Complex Systems Center, University of Vermont, Burlington, VT, USA  
e-mail: peter.dodds@uvm.edu

C. M. Danforth

e-mail: chris.danforth@uvm.edu

(Jones 1970), Bentham's hedonistic calculus and John Stuart Mill's refinements (Russell 1961) sought to codify collective happiness maximization as the determinant of all moral action, and in the United States Declaration of Independence, Jefferson famously asserted the three unalienable rights of 'life, liberty, and the pursuit of happiness.'

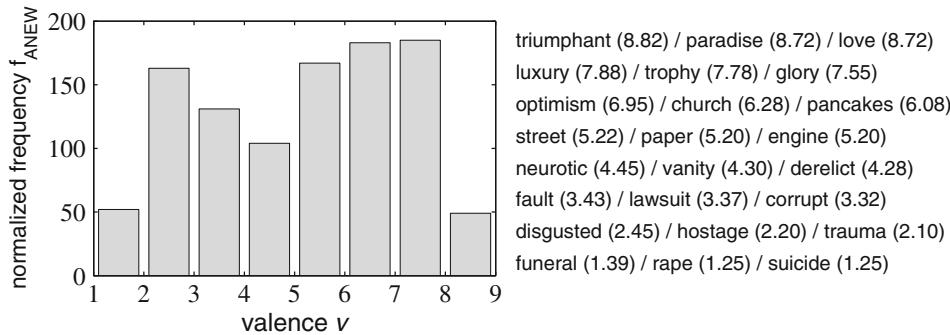
In recognizing the importance of quantifying well-being, we have seen substantial interest and progress in measuring how individuals feel in a wide range of contexts, particularly in the fields of psychology (Osgood et al. 1957; Csikszentmihalyi et al. 1977; Csikszentmihalyi 1990; Gilbert 2006) and behavioral economics (Kahneman et al. 2004; Layard 2005). Most methods, such as experience sampling (Conner Christensen et al. 2003) and day reconstruction (Kahneman et al. 2004), are based on self-reported assessments of happiness levels and are consequently invasive to some degree; dependent on memory and self-perception, which degrades reliability (Killworth and Bernard 1976); likely to induce misreporting (Martinelli and Parker 2009); and limited to small sample sizes due to costs.

Complementing these techniques, we would ideally also have some form of transparent, non-reactive, population-level hedonometer (Edgeworth 1881) which would remotely sense and quantify emotional levels, either post hoc or in real time (Mishne and de Rijke 2005). Our method for achieving this goal based on large-scale texts is to use human evaluations of the emotional content of individual words within a given text to generate an overall score for that text. Our method could be seen as a form of data mining (Witten and Frank 2005; Tan et al. 2005), but since it involves human assessment and not just statistical or machine learning techniques, could be more appropriately classed as 'sociotechnical data mining.' In what follows, we explain the evaluations we use, how we combine these evaluations in analysing written expression, and address various issues concerning our measure.

For human evaluations of the 'happiness' level of individual words, we draw directly on the Affective Norms for English Words (ANEW) study (Bradley and Lang 1999). For this study, participants graded their reactions to a set of 1034 words with respect to three standard semantic differentials (Osgood et al. 1957) of good-bad (psychological valence), active-passive (arousal), and strong-weak (dominance) on a 1–9 point scale with half integer increments. The specific words tested had been previously identified as bearing meaningful emotional content (Mehrabian and Russell 1974; Bellezza et al. 1986). Here, we focus specifically on ratings of psychological valence. [We note that other scales are possible, for example ones that do not presume a single dimension of good-bad, but rather independent scales for good and bad (Diener and Emmons 1984).]

Of great utility to our present work was the study's explanation of the psychological valence scale to participants as a 'happy-unhappy scale.' Participants were further told that "At one extreme of this scale, you are happy, pleased, satisfied, contented, hopeful. ...The other end of the scale is when you feel completely unhappy, annoyed, unsatisfied, melancholic, despaired, or bored" (Bradley and Lang 1999). We can thus reasonably take the average psychological valence scores for the ANEW study words as measures of average happiness experienced by a reader. For consistency with the literature, we will use the term valence for the remainder of the paper.

The measured average valence of the ANEW study words is well distributed across the entire 1–9 scale, as shown by the bar graph in Fig. 1. This suggests we may be able to fashion a measurement instrument based on the ANEW words that has sufficient sensitivity to be of use in evaluating and discriminating texts. Figure 1 also provides some example words employed in the ANEW study along with their average valence scores.

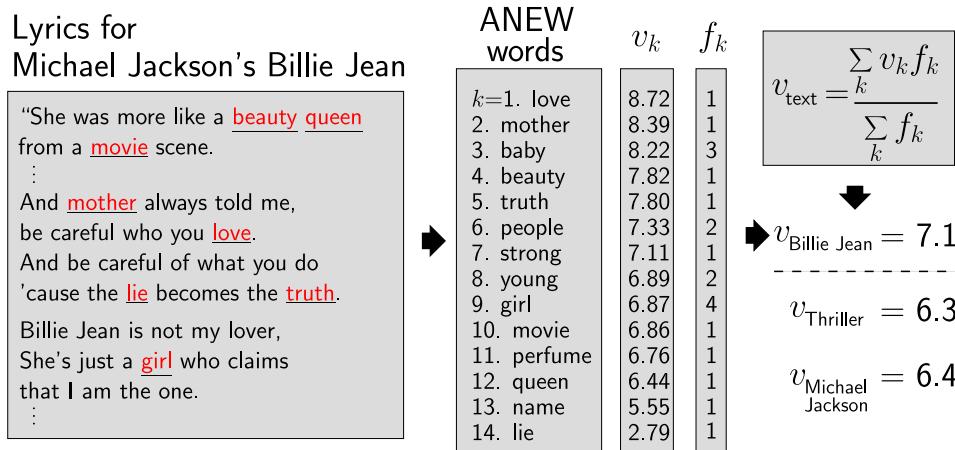


**Fig. 1** Psychological valence (happiness) distribution for words in the Affective Norms for English Words (ANEW) study (Bradley and Lang 1999) along with representative words. Word valence was scored by study participants on a scale ranging from 1 (lowest valence) to 9 (highest valence) with resolution 0.5

To estimate the overall valence score for a text, which we denote by  $v_{\text{text}}$ , we (a) determine the frequency  $f_i$  that the  $i$ th word from the ANEW study word list appears in the text; and (b) compute a weighted average of the valence of the ANEW study words as

$$v_{\text{text}} = \frac{\sum_{i=1}^n v_i f_i}{\sum_{i=1}^n f_i} \quad (1)$$

where  $v_i$  is the ANEW study's recorded average valence for word  $i$ . As a simple example, take the pangram “The quick brown fox jumps over the lazy dog.” The three underlined words appear in the ANEW study word list with average valences 6.64, 4.38, and 7.57, respectively. We would therefore assign an overall valence score for the sentence of  $v_{\text{text}} = \frac{1}{3} \times (1 \times 6.64 + 1 \times 4.38 + 1 \times 7.57) \approx 6.20$ . We hasten to add that our method is only reasonable for large-scale texts where we can demonstrate robustness, and we discuss this in detail below. In Fig. 2, we also outline our measurement schematically, using the example of Michael Jackson’s lyrics. To give a sense of range, for the texts we analyse



**Fig. 2** A schematic example of our method for measuring the average psychological valence of a text, in this case the lyrics of Michael Jackson’s Billie Jean. Average valences for the song Billie Jean, the album Thriller, and all of Jackson’s lyrics are given at right

here, we find that average valence typically falls between 4.5 and 7.5 (our results for lyrics below will give concrete examples for these limiting values).

Our general focus is thus on quantifying how writings are received rather than on what an author may have intended to convey emotionally. Nevertheless, as we discuss below, we attempt to understand the latter with our investigations of blogs.

In using the ANEW data set, we also take the viewpoint that direct human assessment remains, in many complex contexts, superior to artificial intelligence methods. Describing the content of an image, for example, remains an extremely difficult computational problem, yet is trivial for people (von Ahn 2006).

Since our method does not account for the meaning of words in combination, it is suitable only for large-scale texts. We argue that the results from even sophisticated natural language parsing algorithms (Riloff and Wiebe 2003) cannot be entirely trusted for small-scale texts, as individual expression is simply too variable (Lee 2004) and must therefore be viewed over long time scales (or equivalently via large-scale texts). Problematically, the desired scalability is a barrier for such parsing algorithms which run slowly and still suffer from considerable inaccuracy. With our method based on the ANEW data set, we are able to collect and rapidly analyze very large corpuses, giving strength to any statistical assessment. Indeed, with advances in cloud computing, we see no practical limit to the size of meaningful corpuses we can analyse.

A key aspect of our method is that it allows us to quantify happiness on a continuum. By comparison, previous analyses have focused on differences in frequency of words belonging to coarse, broad categories (Cohn et al. 2004), such as ‘negative emotion’, ‘no emotion’, and ‘positive emotion’. For example, using a category-based approach and covering a smaller scale in time and population size than we do here, studies of blogs over a single day have found that content and style vary with age and gender, suggesting automated identification of author demographics is feasible (Schler et al. 2006). However, comparisons between data sets using broad categorical variables are not robust, even if the categories can be ordered. Consider two texts that have the same balance of positive and negative emotion words. Without a value of valence for individual words, we are unable to distinguish further between these texts, which may easily be distinct in emotional content. By using the ANEW data set, we are able to numerically quantify emotional content in a principled way that can be refined with future studies of human responses to words.

## 2 Description of Large-Scale Texts Studied

We use our method to study four main corpuses: song lyrics, song titles, blog sentences written in the first person and containing the word “feel”, and State of the Union addresses. Before exploring valence patterns in depth for these data sets, we first provide some summary statistics relevant to our particular interests, and we also detail our sources.

Table 1 records the total number of words and ANEW words in each data set, along with the number of individual authors. The relative proportions of ANEW words within the four corpuses range from 3.5% (State of the Union) to 9.2% (song titles). These percentages are not insubstantial due to Zipf’s law (Zipf 1949) and the high prevalence of articles, prepositions, etc., in language. Approximately 175 words account for half of all words in the British National Corpus, for example, with the five distinctly neutral words ‘the’, ‘of’, ‘and’, ‘a’, and ‘in’ comprising over 15%.

Table 2, shows the five most frequent ANEW words for each data set, presenting a kind of essence for each corpus. The top five words in song lyrics and titles (which we obtained

**Table 1** Total number of words in each corpus along with the number and percentage of words found in the ANEW database

Counts	Song lyrics	Song titles	Blogs	SOTU
All words	58,610,849	60,867,223	157,853,709	1,796,763
ANEW words	3,477,575 (5.9%)	5,612,708 (9.2%)	8,697,633 (5.5%)	61,926 (3.5%)
Individuals	~20,000	~632,000	~2,400,000	43

Individuals refers to the number of distinct artists, blogs, and presidents

*SOTU* State of the Union addresses

**Table 2** Top five most frequently occurring ANEW words in each corpus with frequency expressed as a percentage of all ANEW words

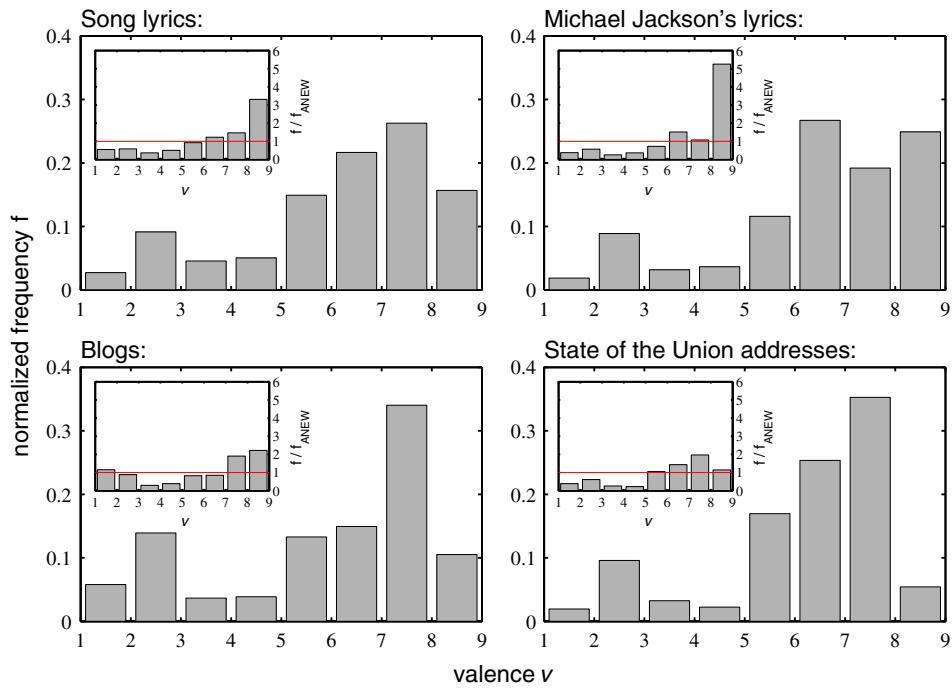
Rank	Song lyrics	Song titles	Blogs	SOTU
1	Love (7.37%)	Love (7.39%)	Good (4.89%)	People (5.49%)
2	Time (4.18%)	Time (4.19%)	Time (4.72%)	Time (4.09%)
3	Baby (2.75%)	Baby (2.75%)	People (3.94%)	Present (3.45%)
4	Life (2.59%)	Life (2.60%)	Love (3.31%)	World (3.10%)
5	Heart (2.14%)	Heart (2.15%)	Life (3.13%)	War (2.98%)

*SOTU* State of the Union addresses

from different databases, see below) are very similar in prevalence, with ‘love’ unsurprisingly being the dominant word. Blogs evince a more social aspect with ‘people’ and ‘life’ in the top five, while the nature of State of the Union addresses is reflected in the disproportionate appearance of ‘world’ and ‘war.’

Figure 3 shows the normalized abundances of ANEW study words appearing in our various corpuses, as a function of their average valence. We include the example of Michael Jackson’s lyrics for reference. The insets for each plot show the same distributions but now normalized by the underlying frequency distribution of ANEW words (Fig. 1). These insets reveal that song lyrics are weighted towards high valence words, and the mode bin is 8–9. Blogs, by contrast, have more low valence words resulting in a bimodal distribution, though the mode bin is again 8–9. State of the Union addresses favor high valence words in the 7–8 bin and show less negativity than blogs.

We obtained our four data sets as follows. We downloaded lyrics to 232,574 songs composed by 20,025 artists between 1960 and 2007 from the website <http://www.hotlyrics.net> and tagged them with their release year and genre using the Compact Disc Data Base available online at <http://www.freedb.org>. We separately obtained from <http://www.freedb.org> a larger database of song titles and genre classifications. Starting August, 2005, first person sentences using the word *feel* (or a conjugated form) were extracted from blogs and made available through the website <http://www.wefelfine.org>, via a public API (Harris and Kamvar 2009). Demographic data was furnished by the site when available. These sentences appeared in over 2.3 million unique blogs during a 47 month span starting in August 2005. In total, we retrieved 9,113,772 sentences which appeared during the period August 26, 2005 to June 30, 2009, inclusive. For each day, we removed repeat sentences of six words or more to eliminate substantive copied material. We obtained State of the Union messages from the American Presidency Project at <http://www.presidency.ucsb.edu>. Finally, we accessed the British National Corpus at <http://www.natcorp.ox.ac.uk>.

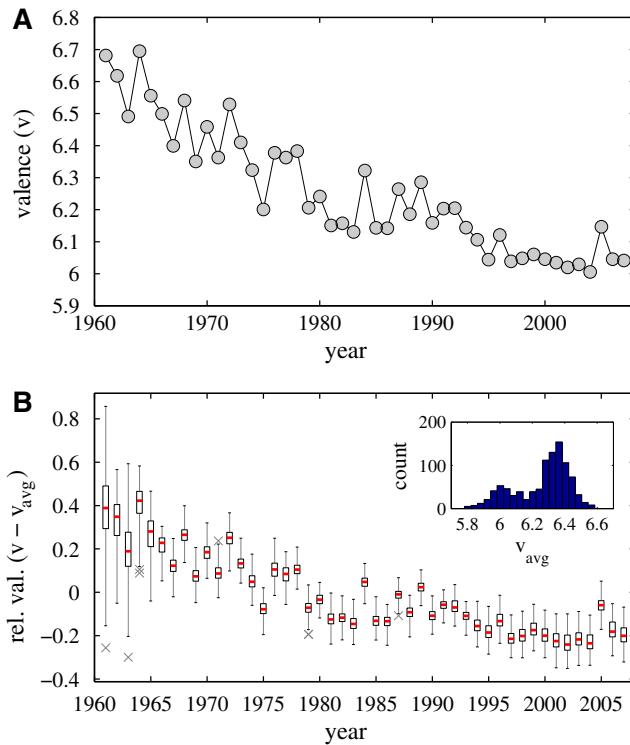


**Fig. 3** Normalized frequency distribution of the ANEW study words binned by valence for the main corpora (excluding song titles) we study here, along with the more specific example of Michael Jackson's lyrics. Insets show ratios of normalized frequencies for corpora to that of the ANEW study word set

### 3 Results

We analyse song lyrics first, in part to demonstrate the robustness of our approach. In Fig. 4a, we show how the average valence of lyrics declines from the years 1961 to 2007. The decline is strongest up until around 1985 and appears to level off after 1995. Since our estimate is based on a partial sample of all words, we need a way of checking its stability. In Fig. 4b, we repeat our analysis using 100 random subsets of the ANEW word list with 750 words, removing the overall average valence from each time series to facilitate comparison of the relative change of valence. The downward trend remains for each measurement while the overall average valence shifts (as shown by the inset). For example, as we have noted, love is the most frequent word in song lyrics, and with its high valence, its inclusion or exclusion from the measurement has the most significant impact on the overall average valence. Thus, we are confident that our estimates of relative as opposed to absolute valence are reasonable.

We more finely examine the reason for this decline in valence in Fig. 5 where we compare individual word prevalence changes in lyrics before and after 1980 using what we term a ‘Valence Shift Word Graph.’ For these graphs, we rank words by their descending absolute contribution to the change in average valence between the two eras,  $\delta$ . Word  $i$ ’s contribution depends on its change in relative frequency, and its valence relative to the pre-1980 era average. In general, in comparing some text  $b$  with respect to a given text  $a$ , we define the valence difference as



**Fig. 4** **a** Valence time series for song lyrics, showing a clear downward trend over the 47 year period starting in 1961. Valence is measured by averaging over the valences of individual words from the ANEW study (Bradley and Lang 1999) found in songs released in each year. **b** Box and whisker plot of relative valence time series for song lyrics for 1000 random sets of 750 ANEW words. The overall mean valence  $v_{avg}$  is removed from each time series for comparison; the *inset* histogram shows the distribution of overall means. Excluding the most frequent words such as ‘love’ (see Table 2) shifts the time series vertically but the downward trend remains apparent in all cases. *Boxes* indicate first and third quartiles and the median; *whiskers* indicate extent of data or  $1.5 \times$  interquartile range; and *outliers* are marked by a gray  $\times$

$$\delta(b, a) = v_b - v_a \quad (2)$$

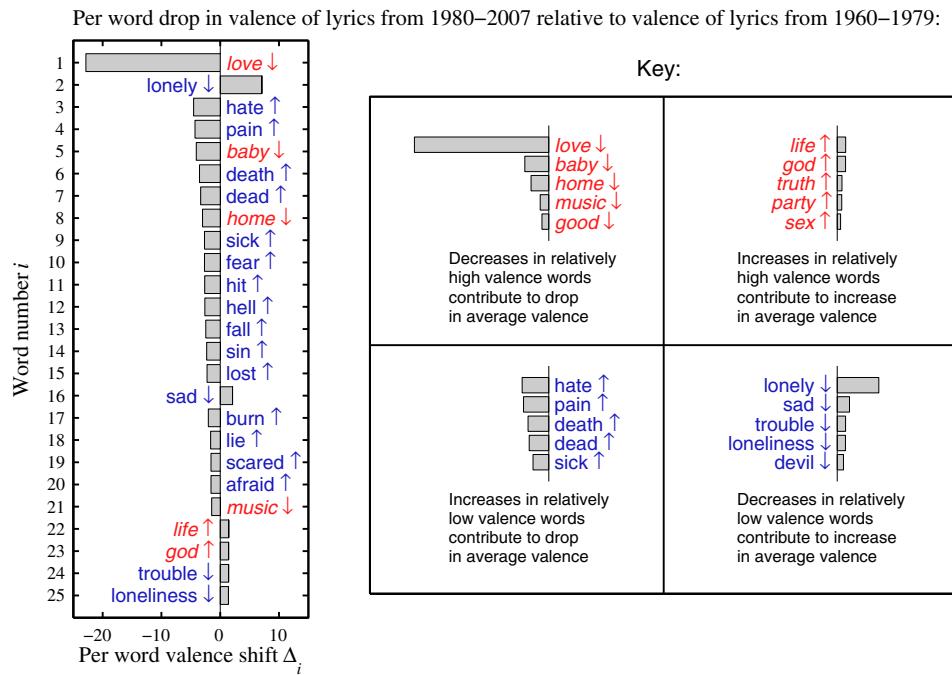
and the percentage contribution to this difference by word  $i$  as

$$\Delta_i(b, a) = 100 \times \frac{(p_{i,b} - p_{i,a})(v_i - v_a)}{\delta(b, a)} \quad (3)$$

where  $p_{i,a}$  and  $p_{i,b}$  are the fractional abundances of word  $i$  in texts  $a$  and  $b$ . As required, summing  $\Delta_i(b,a)$  over all  $i$  gives +100% or -100% depending on whether  $\delta(b,a)$  is positive or negative.

Four basic possibilities arise for each word’s contribution, as indicated by the key in Fig. 5. A word may have higher or lower valence than the average of text  $a$ , and it may also increase or decrease in relative abundance. Further, the contribution of word  $i$  to  $\Delta_i(b,a)$  will be 0 if either the relative prevalences are the same, or the average valence of word  $i$  matches the average of text  $a$ . Note that  $\Delta_i(b,a)$  is not symmetric in  $b$  and  $a$  and is meant only to be used to describe one text ( $b$ ) with respect to another ( $a$ ).

Ranking words according to the above definition of  $\Delta_i$  gives us Fig. 5. We see that the decrease in average valence for lyrics after 1980 is due to a loss of positive words such as

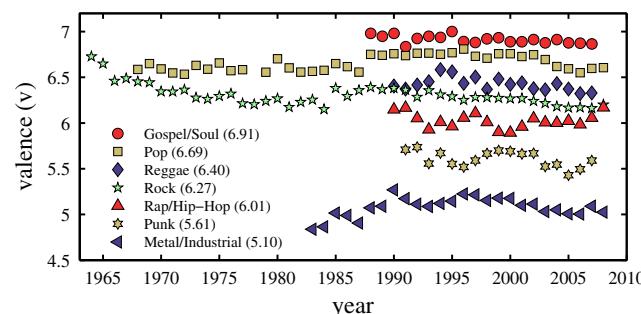


**Fig. 5** Valence Shift Word Graph: Words ranked by their absolute contributions to the drop in average valence of song lyrics from January 1, 1980 onwards relative to song lyrics from before January 1, 1980. The contribution of word  $i$  is defined in Eq. 3 and explained in the surrounding text

‘love’, ‘baby’, and ‘home’ (italicized and in red) and a gain in negative words such as ‘hate’, ‘pain’, and ‘death’ (normal font and in blue). These drops are countered by the trends of less ‘lonely’ and ‘sad’, and more ‘life’ and ‘god’. The former dominates the latter and the average valence decreases from approximately 6.4 to 6.1. Even though the contribution of ‘love’ is clearly the largest, the overall drop is due to changes in many word frequencies. And while we are unable to assess words for which we do not have valence, we can make qualitative observations. For example, the word ‘not’, a generally negative word, accounts for 0.22% of all words prior to 1980 and 0.28% of all words after 1980, in keeping with the overall drop in valence.

To help further unravel this decline in song lyric valence, we show the valence time series for some important music genres in Fig. 6. For this plot, we move to examining song titles for which we have a more complete data set involving genres. We observe that the valence of individual genres is relatively stable over time, with only rock showing a minor decrease. The ordering of genres by measured average valence is sensible: gospel and soul are at the top while several subgenres of rock including metal and punk, and related variants which emerged through the 1970s exhibit much lower valences. Rap and hip-hop, two other notable genres that appear halfway through the time series, are lower in valence than the main genres of rock and pop, but not to the same degree as metal and punk. Thus, the decline in overall valence does not occur within particular genres, but rather in the evolutionary appearance of new genres that accessed more negative emotional niches. Finally, we show the top ten and bottom ten artists ranked according to valence in Table 3, given a certain minimum sampling of each artist’s lyrics.

**Fig. 6** Valence time series for song titles broken down by representative genres. For each genre, we have omitted years in which less than 1000 ANEW words appear



**Table 3** Average valence scores for the top and bottom 10 artists for which we have the lyrics to at least 50 songs and at least 1000 samples of (nonunique) words from the ANEW study word list

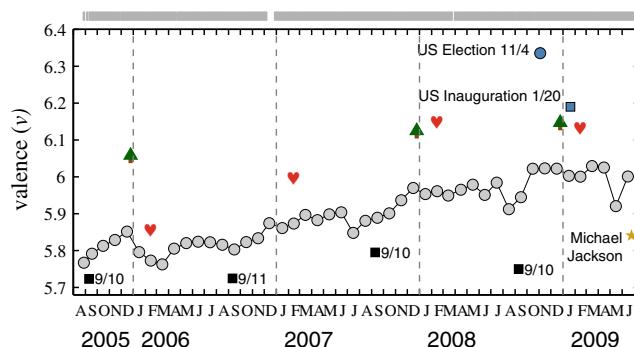
Rank	Top artists	Valence	Bottom artists	Valence
1	All 4 One	7.15	Slayer	4.80
2	Luther Vandross	7.12	Misfits	4.88
3	S Club 7	7.05	Staind	4.93
4	K Ci & JoJo	7.04	Slipknot	4.98
5	Perry Como	7.04	Darkthrone	4.98
6	Diana Ross & the Supremes	7.03	Death	5.02
7	Buddy Holly	7.02	Black Label Society	5.05
8	Faith Evans	7.01	Pig	5.08
9	The Beach Boys	7.01	Voivod	5.14
10	Jon B	6.98	Fear Factory	5.15

While of considerable intrinsic interest, song lyrics of popular music provide us with a limited reflection of society’s emotional state, and we move now to exploring more directly the valence of human expression. The proliferation of personal online writing such as blogs gives us the opportunity to measure emotional levels in real time. At the end of 2008, the blog tracking website <http://www.technorati.com> reported it had indexed 133 million blog records. Blogger demographics are broad with an even split between genders and high racial diversity with some skew towards the young and educated (Lenhart and Fox 2006).

We have examined nearly 10 million blog sentences retrieved via the website <http://www.wefefine.org>, as we have described in detail above. In focusing on this subset of sentences, we are attempting to use our valence measures not only to estimate perceived valence but also the revealed emotional states of blog authors. We are thus able to present results from what might be considered a very basic remote-sensing hedonometer.

In Fig. 7, we plot average monthly valence as a function of time for blogs. We first see that over the time period examined, our subset of blog sentences gradually increase in valence, rising from an average of around 5.75 to over 6.0. Within individual years, there is generally an increase in valence over the last part of the year. In 2008, after a midyear dip, perhaps due to the economic recession, valence notably peaks in the last part of the year and appears to correlate with the US presidential election.

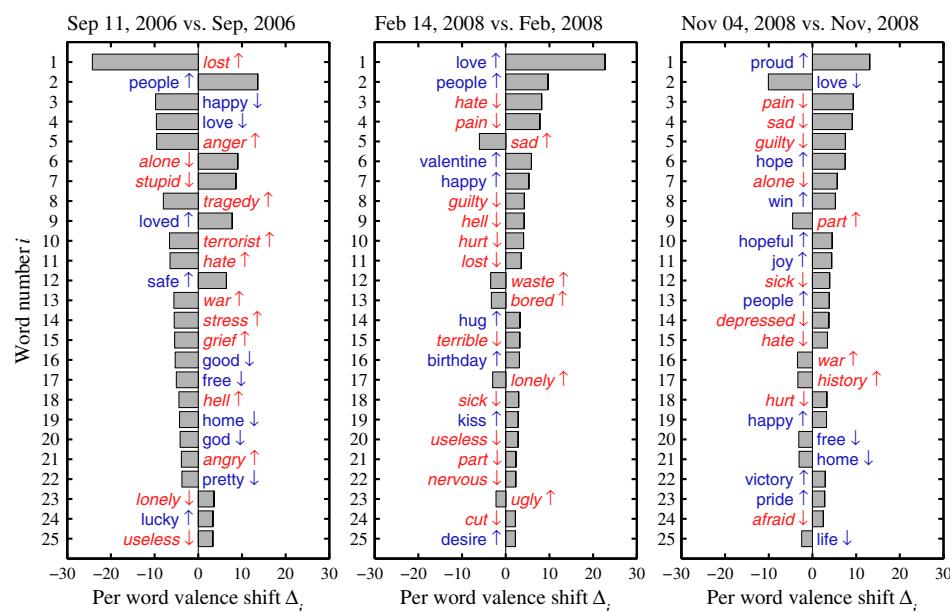
We highlight a number of specific dates which most sharply depart from their month’s average: Christmas Day; Valentine’s Day; September 11, 2006, the fifth anniversary of the World Trade Center and Pentagon attacks in the United States; September 10 in other



**Fig. 7** Time series of average monthly valence for blog sentences starting with “I feel...” show a gradual upward trend over 3 years and 8 months. Notable individual days that differ strongly in valence from that of the surrounding month are indicated, including Christmas Day (*trees*), Valentine’s Day (*hearts*), 9/11 or 9/10 (*squares*), the US Presidential election and inauguration (*circles* and *square*), and Michael Jackson’s death (*star*). The gray bar at the top of the graph indicates the days for which we have data with white gaps corresponding to missing data (we have no estimate of valence for Christmas Day, 2006, hence its absence)

years; the US Presidential Election, November 4, 2008; the US Presidential Inauguration, January 20, 2009; and the day of Michael Jackson's death, June 25, 2009 (June 26 and 27 were also equally low).

In Fig. 8, we show three Valence Shift Word Graphs corresponding to September 11, 2006; Valentine's Day, 2008; and US Presidential Election Day, November 4, 2008. The first panel in Fig. 8 shows that the negative words most strongly driving down the average valence of the fifth anniversary of the 9/11 attacks are ‘lost’, ‘anger’, ‘hate’, and ‘tragedy’



**Fig. 8** Valence Shift Word Graphs for three example dates which are markedly different from the general valence trend shown in Fig. 7. The content of each date's blogs is compared with that of the surrounding month. See Fig. 5 for an explanation of these graphs

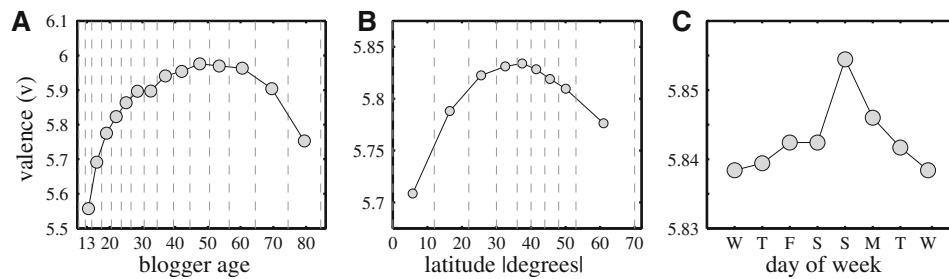
(‘terrorist’ ranks 10th in valence shift). The impact of these words is augmented by a decrease in frequency of ‘love’ and ‘happy’, overwhelming the appearance of more ‘people’ and less ‘stupid’ and ‘alone.’ In other years, September 10 rather than September 11 appears to be more clearly negative in tone, perhaps indicating an anticipatory aspect.

Christmas Day and Valentine’s Day are largely explained by the increase in frequency of the words Christmas and Valentine, both part of the ANEW word list. But other words contribute strongly. For Christmas Day, there is more ‘family’ and less ‘pain’, with an increase in ‘guilty’ going against the trend. As shown for Valentine’s day in 2008 in the second panel of Fig. 8, ‘love’ and ‘people’ are more prevalent, ‘hate’ and ‘pain’ less so, countervailed by more ‘sad,’ ‘lonely,’ and ‘bored.’

The strongest word driving the spike in valence for the 2008 US Election, the happiest individual day in the entire dataset, is ‘proud’ (third panel of Fig. 8). Valence increases also due to a mixture of more positive words such as ‘hope’ and ‘win’ as well as a decrease in the appearances of ‘pain,’ ‘sad,’ and ‘guilty.’

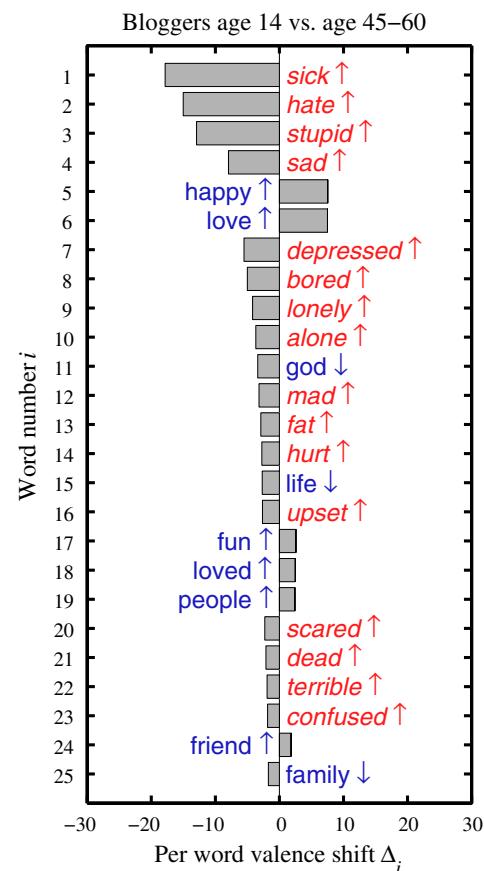
For some blogs, we also have self-reported demographic and contextual information allowing us to make some deeper observations. Figure 9a shows that the average valence of blog sentences follows a pronounced single maximum, convex curve as a function of age. Thirteen and fourteen year-olds produce the lowest average valence sentences (5.58 and 5.55 respectively). As age increases, valence rises until leveling off near 6.0 for ages 45–60, and then begins to trend downwards. Figure 10 compares 14 year-olds to those of age 45–60, and we see the former disproportionately using low valence words ‘sick’, ‘hate’, ‘stupid’, ‘sad’, ‘depressed’, ‘bored’, ‘lonely’, ‘mad’, and ‘fat.’ The increase is most marked throughout the teenage years with 20 year-olds (5.83) closer in average valence to 45–60 year-olds than to 14 year-olds. At the other end of the age spectrum, individuals in the 75 to 84 age range produce sentences with valence similar to those of 17 year olds.

Our age dependent estimates of valence comport with and extend previous observations of blogs that suggested an increase in valence over the age range 10–30 (Schler et al. 2006). Our results are however at odds with those of studies based on self-reports which largely find little or no change in valence over life times (Easterlin 2001, 2003). These latter results have been considered surprising as a rise and fall in valence—precisely what we find here—would be expected due to changes in income (rising) and health (eventually declining) (Easterlin 2001). Our results do not preclude that self-perception of happiness



**Fig. 9** **a** Average valence as a function of blogger’s self-reported age. We use the age a blogger will turn in the year of his or her post and an approximately logarithmically growing bin size such that all data points are based on at least 3000 ANEW words. Bin boundaries are indicated by the *dashed vertical lines*. **b** Average valence as a function of blogger’s absolute latitude. Bins are indicated by *vertical gray lines*. The first two bins have 8,390 and 26,071 ANEW words and the remainder all have approximately  $10^5$  or more. **c** Valence averaged over days of the week for blogs showing a subtle seven-day cycle peaking on Sunday with a trough in the middle of the week. Each day’s average is based on at least  $10^6$  ANEW words

**Fig. 10** Valence Shift Word Graph comparing bloggers who list themselves as turning 14 in year of post to those turning 45–60. The average valences are 5.55 and 5.98 respectively



may indeed be stable, but since our results are based on measured behavior, they strongly suggest individuals do present differently throughout their lifespan. And while we have no data regarding income, because income typically rises with age, our results are sympathetic to recent work that finds happiness increases with income (Stevenson and Wolfers 2008), going against the well known Easterlin Paradox, popularized as the notion that ‘money does not buy happiness’ (Easterlin 1974).

Figure 9b shows that the average valence of blog sentences gently rolls over as a function of absolute latitude (i.e., combining both the Northern and Southern Hemispheres). Average valence ranges from 5.71 (for 0–11.5°) up to 5.83 (for 29.5–44.5°) and then back to 5.78 (for 52.5–69.5°). Seasonal Affective Disorder (Rosenthal et al. 1984) may be the factor behind the small drop for higher latitudes, though a different mechanism would need to be invoked to account for lower valence near the equator. One possible explanation could be that the relatively higher population of the mid-latitudes leads to stronger social structures (Layard 2005). We find some support for the social argument for individuals near the equator (absolute latitude  $\leq 11.5$ ), who we observe more frequently use the words ‘sad’, ‘bored’, ‘lonely’, ‘stupid’ and ‘guilty’ and avoid using ‘good’ and ‘people.’ On the other hand, the valence drop at higher latitudes (between 52.5 and 69.5° absolute latitude) is reflected in the frequency changes of a mixture of social, psychological, and some conditions-related words: ‘sick’, ‘guilty’, ‘cold’, ‘depressed’, and ‘headache’ all increase, ‘love’ and ‘life’ decrease, offset by less ‘hurt’ and ‘pain’ and more ‘bed’ and ‘sleep.’

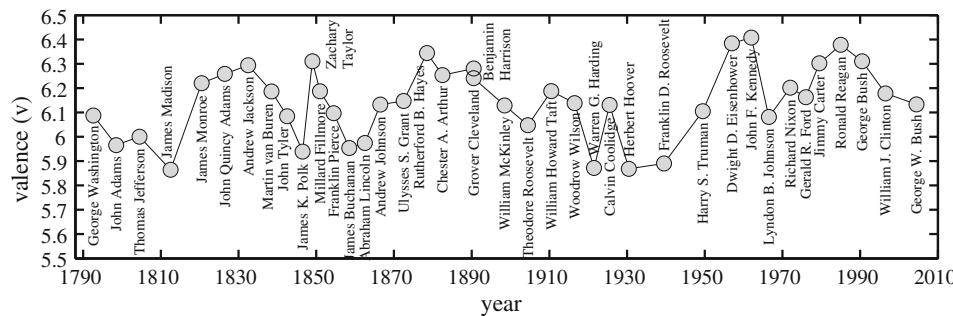
At a much more subtle level, a weekly cycle in valence is visible in blog sentences (Fig. 9c). A relatively sharp peak in valence occurs on Sunday, after which valence steadily drops daily to its lowest point on Wednesday before climbing back up. Monday, contrary to commonly held perceptions but consistent with previous studies (Stone et al. 1984), exhibits the highest average valence after Sunday, perhaps indicating a lag effect.

We also observe some variation among countries. Of the four countries with at least 1% representation, the United States has the highest average valence (5.83) followed by Canada (5.78), the United Kingdom (5.77), and Australia (5.74).

In terms of gender, males exhibit essentially the same average valence as females (5.89 vs. 5.91). Females however show a larger variance than males (4.75 vs. 4.44) in agreement with past research (Snyder and Lopez 2009). We further find females disproportionately use the most impactful high and low valence words separating the two genders: ‘love’, ‘baby’, ‘loved’ and ‘happy’ on the positive end, and ‘hurt’, ‘hate’, ‘sad’, and ‘alone’ on the negative end. In fact, of the top 15 words contributing to  $\delta(\text{female}, \text{male})$ , the only one used more frequently by males is the rather perfunctory word ‘good.’

We turn to our last data set, State of the Union (SOTU) addresses for the United States. These addresses, which include both speeches and written reports, grant us a starting point for assessing the emotional temperature of the United States over its 220 year history, as may or may not have been intended by the authors. Figure 11 shows a valence time series for SOTU addresses binned by President. In comparison to our lyrics and blog data sets, SOTU addresses comprise far fewer words and the observations we make are consequently more speculative. Nevertheless, we do find some resonance between the valence level of SOTU addresses and major historical events.

The presidents with the highest average valence scores are Kennedy (6.41), Eisenhower (6.38), and Reagan (6.38), all of whose speeches are tightly clustered around their means. Eisenhower and Kennedy reach a high point after a period of relatively low valence starting with the First World War through and beyond which Wilson’s speeches steeply drop from an initial 6.58 in 1913 to 5.88 in 1920. The mean valence of Coolidge’s addresses provide the single exception during this time. Coolidge’s successor Hoover’s low average is largely due to his speech in 1930, the first one given after the stock market crash of October 29, 1929—Black Tuesday—which marked the beginning of the Great Depression; his speeches are burdened with ‘depression’, ‘debt’, ‘crisis’, and ‘failure.’ While Franklin Roosevelt’s overall average valence is low, the first eight of his four term stay in office range from 6.06 to 6.34. His last four speeches, coming during the Second World War (1942–1945), are sharply lower in valence, ranging from 5.48 to 5.60; ‘war’ naturally dominates these later



**Fig. 11** Average valence of State of the Union addresses, binned by President and plotted against the average of the years the President was in office

speeches and along with ‘fight’ and ‘destroy’, overwhelm the positives of ‘peace’ and ‘victory.’

The large-scale pattern of the 19th Century shows two periods of relatively high valence, 1820–1840 and 1880–1890. The years before and during the American Civil War form a local minimum in valence corresponding to Buchanan and Lincoln.

The recent era shows a drop from Eisenhower and Kennedy’s level to that of Johnson (6.08), the latter’s first SOTU speech coming just seven weeks after the assassination of Kennedy, and the remainder through the heightening Vietnam War. Valence rises through the 1970s to reach the high of Reagan in the 1980s, from which it has since declined.

#### 4 Concluding Remarks

Undoubtedly, the online recording of social interactions and personal experiences will continue to grow, providing ever richer data sets and the consequent opportunity and need for a wide range of scientific investigations. A natural extension of our work here would be to examine the dynamics of emotions in online interactive contexts, particularly in the realm of contagion (Hatfield et al. 1993, Fowler and Christakis 2008). If emotional contagion is observable, we would then be in a position to characterize its nature on the spectrum from analogous to an infectious disease (Murray 2002) to the more complex threshold-based contagion (Granovetter 1978; Dodds and Watts 2004). Our technique could also be useful in testing predictive theories of social interactions such as Heise’s affect control theory (Heise 1979) and Burke’s identity control theory (Stets and Tsushima 2001).

While we have been able to make and support a range of observations with our method for measuring the emotional content of large-scale texts, our approach can be improved in a number of ways. A first step would be to perform experiments and surveys to gather emotional content estimates for a more extensive set of individual words. The instrumental lens can also be made more sophisticated by coupling word assessments with detailed demographics of participants. Other approaches not necessarily based on semantic differentials in the manner of the ANEW study could also be naturally explored. Game-based experiments could also be used to assess the emotional content of common word groups and phrases (von Ahn 2006), allowing us to better characterize the micro-macro connection between the atoms of words and sentences, and differences in interpretations among various age groups and cultures.

**Acknowledgements** The authors are grateful to Jonathan Harris and Sep Kamvar, the creators of <http://www.wefelfine.org>; for helpful discussions with John Tucker, Lilian Lee, Andrew G. Reece, Josh Bongard, Mary Lou Zeeman, and Elizabeth Pinel; and for the suggestions of three anonymous reviewers.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

#### References

- Argyle, M. (2001). *The psychology of happiness*, 2nd edn. New York: Routledge.  
Bellezza, F. S., Greenwald, A. G., & Banaji, M. R. (1986). Words high and low in pleasantness as rated by male and female college students. *Behavior Research Methods, Instruments & Computers*, 18, 299–203.

- Bradley, M., & Lang, P. (1999). *Affective norms for english words (anew): Stimuli, instruction manual and affective ratings*. Technical report c-1, Gainesville, FL: University of Florida.
- Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, 15, 687–693.
- Conner Christensen, T., Feldman Barrett, L., Bliss-Moreau, E., Lebo, K., & Kaschub, C. (2003). A practical guide to experience-sampling procedures. *Journal of Happiness Studies*, 4, 53–78.
- Csikszentmihalyi, M. (1990). *Flow*. New York: Harper & Row.
- Csikszentmihalyi, M., Larson, R., & Prescott, S. (1977). The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, 6, 281–294.
- Diener, E., & Emmons, R. A. (1984). The independence of positive and negative affect. *Journal of Personality and Social Psychology*, 47, 1105–1117.
- Dodds, P. S., & Watts, D. J. (2004). Universal behavior in a generalized model of contagion. *Physical Review Letters*, 92. Article #218701.
- Easterlin, R. (1974). Does economic growth improve the human lot? Some empirical evidence. In P. A. David & M. W. Reder (Eds.), *Nations and households in economic growth: Essays in honour of Moses Abramowitz* (pp. 89–125). New York: Academic Press.
- Easterlin, R. A. (2001). Income and happiness: Towards a unified theory. *The Economic Journal*, 111, 465–484.
- Easterlin, R. A. (2003). Explaining happiness. *Proceedings of the National Academy of Sciences of the United States of America*, 100, 11176–11183.
- Edgeworth, F. Y. (1881). *Mathematical physics: An essay into the application of mathematics to moral sciences*. London, UK: Kegan Paul.
- Fowler, J. H. & Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *British Medical Journal*, 337. Article #2338.
- Gilbert, D. (2006). *Stumbling on happiness*. New York: Knopf.
- Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83(6), 1420–1443.
- Harris, J., & Kamvar, S. (2009). *We feel fine: An almanac of human emotion*. New York, NY: Scribner.
- Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). *Emotional contagion. Studies in emotion and social interaction*. Cambridge, UK: Cambridge University Press.
- Heise, D. R. (1979). *Understanding events: Affect and the construction of social action*. New York: Cambridge University Press.
- Jones, W. T. (1970). *The classical mind*. New York: Harcourt, Brace, Jovanovich.
- Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science*, 306(5702), 1776–1780.
- Killworth, P. D., & Bernard, H. R. (1976). Informant accuracy in social network data. *Human Organization*, 35, 269–286.
- Layard, R. (2005). *Happiness*. London: The Penguin Press.
- Lee, L. (2004). “I’m sorry Dave, I’m afraid I can’t do that”: Linguistics, statistics, and natural language processing circa 2001. In On the Fundamentals of Computer Science: Challenges C, Opportunities CS, Telecommunications Board NRC (Eds.), *Computer science: Reflections on the field, reflections from the field* (pp. 111–118). Washington, DC: The National Academies Press.
- Lenhart, A., & Fox, S. (2006). *Bloggers: A portrait of the Internet’s new storytellers*. Technical report. Pew Internet & American Life Project.
- Lyubomirsky, S. (2007). *The how of happiness*. New York: The Penguin Press.
- Martinelli, C., & Parker, S. W. (2009). Deception and misreporting in a social program. *Journal of the European Economic Association*, 7, 886–908.
- Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. Cambridge, MA: MIT Press.
- Mishne, G., & de Rijke, M. (2005). Capturing global mood levels using blog posts. AAAI 2006 Spring symposium on computational approaches to analysing weblogs.
- Murray, J. D. (2002). *Mathematical biology*, 3rd edn. New York: Springer.
- Osgood, C., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois.
- Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. Conference on empirical methods in natural language processing (EMNLP-03), ACL SIGDAT (pp. 105–112).
- Rosenthal, N. E., Sack, D. A., Gillin, J. C., Lewy, A. J., Goodwin, F. K., Davenport, Y., et al. (1984). Seasonal affective disorder: A description of the syndrome and preliminary findings with light therapy. *Archives of General Psychiatry*, 41(1), 72–80.

- Russell, B. (1961). *A history of western philosophy*. London: Allen & Unwin.
- Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). Effects of age and gender on blogging. In *Computational approaches to analyzing weblogs: Papers from the 2006 AAAI spring symposium* (pp. 199–205). Menlo Park, CA: AAAI Press.
- Snyder, C. R., & Lopez, S. J. (2009). *Positive psychology*, 2nd edn. New York, NY: Oxford University Press.
- Stets, J. E., & Tsushima, T. M. (2001). Negative emotion and coping responses within identity control theory. *Social Psychology Quarterly*, 64, 283–295.
- Stevenson, B., & Wolfers, J. (2008). Economic growth and subjective well-being: Reassessing the Easterlin Paradox, Brookings papers on economic activity.
- Stone, A. A., Hedges, S., Neale, J. M., & Satin, M. S. (1984). Prospective and cross-sectional mood reports offer no evidence of a “Blue Monday” phenomenon. *Journal of Personality and Social Psychology*, 49, 129–134.
- Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston, MA: Addison Wesley.
- von Ahn, L. (2006). Games with a purpose. *IEEE Computer Magazine*, 39(6), 96–98.
- Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*, 2nd edn. San Francisco, CA: Morgan Kaufmann.
- Zipf, G. K. (1949). *Human behaviour and the principle of least-effort*. Cambridge, MA: Addison-Wesley.

## Affective News: The Automated Coding of Sentiment in Political Texts

LORI YOUNG and STUART SOROKA

*An increasing number of studies in political communication focus on the “sentiment” or “tone” of news content, political speeches, or advertisements. This growing interest in measuring sentiment coincides with a dramatic increase in the volume of digitized information. Computer automation has a great deal of potential in this new media environment. The objective here is to outline and validate a new automated measurement instrument for sentiment analysis in political texts. Our instrument uses a dictionary-based approach consisting of a simple word count of the frequency of keywords in a text from a predefined dictionary. The design of the freely available Lexicoder Sentiment Dictionary (LSD) is discussed in detail here. The dictionary is tested against a body of human-coded news content, and the resulting codes are also compared to results from nine existing content-analytic dictionaries. Analyses suggest that the LSD produces results that are more systematically related to human coding than are results based on the other available dictionaries. The LSD is thus a useful starting point for a revived discussion about dictionary construction and validation in sentiment analysis for political communication.*

**Keywords** content analysis, media tone, methodology

Political discourse cannot be reduced to mere factual information—the tone of a text may be as influential as its substantive content. Indeed, numerous studies have focused on the tone or sentiment of news content, political speeches, and advertisements.<sup>1</sup> Moreover, a substantial and growing body of research suggests that affect<sup>2</sup> is a central component of individual decision making and political judgment generally, as well as the processing of media information in particular.<sup>3</sup> Negative affect appears to be particularly prominent in the human psyche, and in politics as well.<sup>4</sup> The reliable and valid analysis of sentiment is, in short, a critical component of a burgeoning field of research in political communication, and political science more broadly.

The growing interest in, and importance of, measuring sentiment coincides with a dramatic increase in the volume of digitized information. Computer automation has a great deal of potential in this new media environment. Automation is very efficient and

Lori Young is a doctoral candidate at the Annenberg School for Communication, University of Pennsylvania. Stuart Soroka is Associate Professor and William Dawson Scholar in the Department of Political Science, McGill University.

The authors are grateful to Mark Daku, who programmed Lexicoder; to Marc André Bodet and Blake Andrew for their work using the LSD in its early stages; to Christopher Wlezien and Robert Erikson, for providing us some U.S. polling data with which to further test the dictionary; and to the editor and anonymous reviewers, whose comments were critical to this final version of the article.

Address correspondence to Stuart Soroka, Department of Political Science, McGill University, 855 Sherbrooke St. West, Montreal, QC, H3A 2T7 Canada. E-mail: stuart.soroka@mcgill.ca

becoming easier to implement as new software is developed and lexical resources become more widely available. Our objective here, then, is to outline and validate a new automated measurement instrument for sentiment analysis in political texts. Our instrument uses a dictionary-based approach consisting of a simple word count of the frequency of keywords in a text from a predefined dictionary. There are a number of machine-readable sentiment lexicons currently available for automation. However, each has been compiled for specific types of research in various disciplines using diverse methodologies. Consequently, they vary widely with respect to sentiment categories, coding schemes, and scope of coverage. Indeed, there are to our knowledge no comparative studies of existing lexicons used in sentiment analysis in political communication. Moreover, we find that proprietary restrictions occasionally impinge on the assessment and use of such resources, raising concerns about replicability and development in the field. Many lexicons are also temporally or corporally specific. Our goal is to develop a sentiment dictionary that is more broadly applicable, across a wide range of research foci in political communication.

We do so by combining and standardizing three of the largest and most widely used lexical resources to create a comprehensive valence dictionary of positive and negative words. The scope and performance of the resulting dictionary is then compared to six other commonly used sentiment lexicons, as well as to the three from which it was composed. We automate the analysis of positive and negative tone in *New York Times* coverage across four topics: economy, environment, crime, and international affairs. We compare results not just across sentiment dictionaries, but more importantly with results from trained human coders. Results suggest that the dictionary developed here, the Lexicoder Sentiment Dictionary (LSD), performs somewhat better than others. This, combined with the fact that the LSD is freely available and easily adaptable, makes the dictionary proposed here a valuable step forward in automated content analysis of political communication.

The following sections outline the design and reliability of the dictionary in some detail. A final section then provides one example of how the dictionary can be used, by extending previous work on the relationship between campaign-period vote intentions and the tone of media content. We extend previous work on the 2006 Canadian election campaign, comparing directly the previous manually coded results with those using the LSD; results demonstrate the strength of the dictionary and also speak to the role of media in election campaigns. First, however, the next sections review the relevant literature concerning media effects, describe in some detail the state of research in automated content-analytic techniques as they pertain to current work in political communication, and consider some of the advantages and limitations of a dictionary-based approach.

## Media Affect

The development of a dictionary for automated content analysis below stems first and foremost from our interest in media effects, and more generally the role of media in representative democracy. It is axiomatic that, across all modern representative democracies, mass media play a central role in everyday politics. Media both reflect and inform public opinion; many of us are dependent on mass media for much of the information we require in order to be effective democratic citizens. Scholars are thus perennially concerned with the content of mass media, as well as the potential consequences that content may have on political judgment and behavior. There are vast bodies of work detailing the many ways in which media can affect public preferences on political issues. Much of this work has been interested in large-scale content analysis of media content.

Most relevant to our work here is research focused on capturing the tone of media content. This body of research is wide and varied—it includes, for instance, work on the tone of news coverage of presidents and political parties (e.g., Eshbaugh-Soha, 2010; Farnsworth & Lichter, 2010; Ottati, Steenbergen, & Riggle, 1992; Soroka, Bodet, Young, & Andrew, 2009), research dealing with the tone of economic news coverage (e.g., Gentzkow & Shapiro, 2010; Lowry, 2008; Nadeau, Niemi, Fan, & Amato, 1999; Soroka, 2006), and work reflecting other diverse interests, such as Cho et al.'s (2003) research on the "emotionality" of television and print media coverage of the 9/11 terrorism attacks.

The affective content of news is also related to the body of literature focused on symbolic language and issue frames. This work views modern politics largely as a struggle over language, a battle to define terms and frame the debate (see, e.g., Edelman, 1985; Hart, 2000b). It suggests that symbolic language and framing influences the way people think about particular issues (see, e.g., Iyengar, 1996; Quattrone & Tversky, 1988). Shifting frames can change the affective composition of the media, creating narratives that construct and shape perceptions of social and political reality (e.g., Johnson-Cartee, 2005; Shenhav, 2006). Driven by dramatic, novel, and negative information, affective narratives can inform judgments, sometimes quite independent of real-world events (e.g., McComas & Shanahan, 1999). And while the automated identification of frames is not our focus here, we view sentiment as one central component in the empirical study of issue frames.

In short, a wide range of work suggests that the "tone" or "sentiment" of text matters to our understanding of both the content and effects of mass media in modern representative democracy. It matters for our understanding of media content, political behavior, and policy-making. Given the widespread interest in, and demonstrated importance of, tone in media content, coupled with the increasing use of computer automation, we believe that it is important to consider the extent to which tone can be captured reliably and validly using automated systems.

### **Computer Automation**

Computer automation has become a mainstay of empirical research in the study of political communication. Since the 1950s scholars have been developing computer-assisted methods to analyze textual information in new and interesting ways. As mass media have expanded, so too has the volume of political text, and this text is increasingly readily available electronically. Increasing volumes of digital information have been met with increasingly sophisticated content-analytic methodologies. Many of these are computer automated. Automation can facilitate the analysis of enormous bodies of data in meaningful ways, where labor-intensive manual content analyses often fall short, either because of time and budgetary constraints or because of difficulties obtaining intercoder reliability.

Broadly speaking, automated content analysis can be undertaken as a statistical or a nonstatistical endeavor. Machine-learning techniques using statistical classification exploded with computational advances in the 1990s. This method does not rely on predefined dictionaries; rather, data are generated from the text itself using statistical classifiers. Supervised machine learning (SML) involves identifying various features (prevalent words or word patterns) in a set of "reference" texts with a known *a priori* class or category. Reference texts are manually classified or selected to be the best possible representation of the category to be coded. Much rests, then, on the quality and representativeness of the reference texts, which are used to "train" classifiers to recognize or predict the class of unknown texts according to the presence of linguistic features "learned" from the reference texts.<sup>5</sup>

Unsupervised machine learning (UML) does not rely on reference texts or predefined categories. Based on latent semantic analysis, the method uses matrix algebra to measure word associations (i.e., word clusters, local co-occurrence, pairwise patterns) within and between texts to infer unknown categories, much like factor analysis (Hogenraad, McKenzie, & Péladeau, 2003; Landauer & Dumais, 1997). It is relatively easy to implement and does not require extensive pre-coding or a priori dictionaries. SML (as well as the dictionary-based approach discussed below) relies on predefined classification schemes that map onto the text, giving meaning to the words; in contrast, UML is used to discover new or unknown categories inductively, using the correlation of words to give meaning to a text.<sup>6</sup>

The nonstatistical dictionary-based approach (also referred to as frequency or categorical analysis) is, in terms of implementation, much simpler: It involves counting the frequency of definitive keywords in a text. A key feature of this approach is use of a machine-readable dictionary of a priori categories. Herein lies the challenge: A good dictionary, particularly for something like sentiment, is very difficult to develop. Nevertheless, numerous content-analytic dictionaries have been developed for automated analysis featuring a range of topic and sentiment lexicons.

With a well-defined and comprehensive dictionary, a basic word count can provide a powerful and reliable analysis of the topical and affective composition of a text. Existing applications range from analysis of children's writing to discern their affect toward police (Bolasco & Ratta-Rinaldi, 2004) to predicting the variance of firm account earnings and stock returns by counting negative words near the keyword "earn" in economic journals (Tetlock, Saar-Tsechansky, & Macskassy, 2007). The approach has been used to analyze political communications since the 1960s. Stone and colleagues (Stone, Bales, Namenwirth, & Ogilvie, 1962; Stone, Dumphy, & Ogilvie, 1966) first used the General Inquirer (GI) dictionary to compare the tone of political speeches by various candidates. Hart (1984, 2000a) has used a similar approach since the 1980s to analyze president rhetoric and campaign style. Hart's DICTION program has become a mainstay in content-analytic work on political rhetoric and discourse and has been used in over 50 studies to analyze political speech, differentiate news by genre, study religious ideology, analyze corporate publications, and so forth.<sup>7</sup>

### ***Advantages and Challenges in Automated Analysis***

There are some clear advantages to using computer automation for text analysis, including efficiency, scope, and reliability. Dictionary-based results have the additional advantage of parsimony. Constructing a dictionary is quite an investment, to be sure. Once constructed, however, computation is very easy to implement. Certain dictionaries may be better suited for some texts than others, of course, but in each case we know exactly what is being applied, and all cases are thus directly comparable. They are also perfectly reliable, in the sense that they produce exactly the same results at the article level, whether one analyzes 10 articles or 10 years of news.

This increase in reliability does not necessarily reflect greater validity, of course. Automation is typically capable of lexical and syntactic analysis and certain types of semantic or discourse analysis (see further discussion below). But some types of textual analysis are much less well suited for automation, dictionary-based or otherwise. Automation counts but does not rate entries; it identifies but does not interpret semantic patterns; it quantifies concepts but not symbols. To be clear: We readily acknowledge that

there are many questions of interest in content analysis that are still beyond the capability of computers.

That said, automated analysis clearly does have value in particular contexts. In one sense, it is simply a different level of analysis. Borrowing Hart's (2001) analogy, manual coding may be likened to the perspective of a beat cop in a specific neighborhood, rich in context and detail-oriented, while computer automation offers a bird's eye view, like a helicopter pilot circling the city to monitor overall crime patterns. The methods are complementary—each perspective generates useful information the other cannot see.

We should also not regard computer automation as somehow less refined than manual forms of textual analysis. Automation may be especially well-suited for certain questions that would be difficult for humans to code. For instance, human coders may be limited in their ability to identify certain types of latent (rather than manifest) content. And as we will see below, many dictionaries are constructed to capture complicated latent cognitive and affective concepts, such as the degree of primordial versus conceptual thought (Martindale, 1975, 1990), or "certainty" operationalized by uses of the verb "to be" (Hart, 1984). Indeed, the psycholinguistic underpinnings of many content-analytic dictionaries point to a level of sophistication in computer automation not always noted by its critics.

That said, there are least two main challenges to automated techniques, each of which deserves mention here. First, most automated systems process words regardless of order or context using the so-called "bag-of-words" approach; that is, they assume "semantic independence." The strategy is "based on the assumption that the words people use convey psychological information over and above their literal meaning and independent of their semantic context" (Pennebaker, Mehl, & Niederhoffer, 2003, p. 550). Obviously, this assumption does not always hold. For instance, many scholars have noted that tone is far more dependent on the relation between words than topic (Murphy, Bowler, Burgess, & Johnson, 2006; Pang, Lee, & Vaithyanathan, 2002; Thomas, Pang, & Lee, 2006; Wilson, Wiebe, & Hoffman, 2005). For instance, multiple speakers and/or topics can make the attribution of tone difficult—the tone of coverage about a political actor or topic is not necessarily reflected in the overall tone of an article.<sup>8</sup> Depending on the goals of the research at hand, however, it may not be necessary to fully disentangle the semantic relationship between actors, topics, and opinions. For texts with multiple topics and speakers, unattributed tone at the document level simply reflects the overall tone of a document. (We discuss this in more detail below.)

A tougher challenge to this assumption is the tendency for the linguistic markers of tone to be context specific. Consider how the meaning of the word "happy" changes when it follows the word "not" or the difficulty determining the meaning of homographs such as "right," "lie," or "well." One strategy to mitigate the impact of contextual language is the preprocessing of text in order to standardize words and phrases, disambiguate homographs, and account for basic negation patterns. In our analysis we apply extensive preprocessing, the development of which is described below. In this way we are able, at least modestly, to move beyond a simple bag of words.

The second general concern with automation is the assumption of additivity—that is, every instance of every word contributes isomorphically to the output. In natural language, of course, certain words may carry more weight than others. "Evil" may matter more than "bad," for instance. The point is well taken; however, it is not clear that this is an issue of content analysis so much as one of psychology. Technically speaking, weights are easy to apply. There are numerous examples of automated dictionaries that use weights or apply modifiers to try to overcome the assumption of additivity (e.g., Subasic & Huettner, 2001). What seems to be lacking is a good theory as to how the weights should be applied based

on the varied effects of particular words. This does not negate the fact that the potentially differential weighting of words can present real problems for automated analyses. On the contrary, it is an important reminder that automation is simply a lexical scan of the frequency of words used. And there is, of course, no substitute for careful data interpretation.

### Sentiment Lexicons

Since the 1960s, scholars have been developing psycholinguistic lexicons coded for basic affective and cognitive dimensions or tagged for valence to categorize the positive and negative connotations they carry. There are now numerous machine-readable dictionaries available for research, but they vary widely with respect to categories and scope of coverage. They include, for example, the following: from political science, the GI (Stone et al., 1966); from communication, DICTION (Hart, 2000a); from psychology, Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, & Booth, 2001), the Regressive Imagery Dictionary (RID) (Martindale, 1975, 1990), and TAS/C (Mergenthaler, 1996, 2008); from behavioral science, Affective Norms for English Words (ANEW) (Bradley & Lang, 1999); from literature, Whissell's Dictionary of Affect in Language (DAL) (Whissell, 1989); from linguistics, WordNet-Affect (WNA) (Strapparava & Valitutti, 2004); and from computational linguistics, Turney and Littman's (2003) pointwise mutual (PMI) information wordlist, as well as the ubiquitous Roget's Thesaurus (hereafter Roget's) (Roget, 1911).

These dictionaries have been compiled for a variety of research projects across disciplines. Consequently, there has been a range of methodological approaches to dictionary construction. In some cases dictionaries are compiled from previously generated word lists (e.g., GI)<sup>9</sup>; in others codes are manually attributed by expert coders or panels of judges (e.g., LIWC); in others words are tagged using computer automation based on patterns in natural language (e.g., PMI) or the linguistic properties of words (e.g., WNA); and others still are derived from experimental methods (e.g., ANEW, DAL) or iterative processes combining a number of different approaches (e.g., DICTION, TAS/C). Each method measures something slightly different. Generally speaking, expert coding seeks to capture the definitive meaning of words, automation captures contextual or common usage, and experimental methods capture something closer to perceptions of words.

Construction of a valence lexicon is particularly challenging because the semantic category of valence itself appears to be structurally fuzzy (Andreevskaia & Bergler, 2006; Subasic & Huettner, 2001). The ambiguity of the category poses a challenge to researchers who rely on discrete categories (positive versus negative) for frequency analysis, and efforts to resolve ambiguity tend to further limit the scope of coverage. Two main approaches are adopted to address the ambiguity of valence in dictionary construction: Researchers either disambiguate the dictionary or attempt to score entries according to their centrality to a given category.

For instance, the GI is the oldest and most expansive dictionary of its kind. By consequence, its large valence categories of positive and negative words tend to be overly general and lack discriminative capacity. Few researchers use the dictionary without encountering the need to generate expanded and/or disambiguated versions (e.g., Hogenraad et al., 2003; Kennedy & Inkpen, 2006; Pennebaker et al., 2003; Scharl, Pollach, & Bauer, 2003).<sup>10</sup>

Rather than removing ambiguity, some seek to preserve it by adopting “fuzzy logic” and “continuous valence” categories. Andreevskaia and Bergler (2006) maintain that discrepancies in the dictionaries and inter-annotator disagreements are “not really errors but

a reflection of the natural ambiguity of the words that are located on the periphery of the sentiment category"; disagreement reflects the "structural property of the semantic category" (p. 4). Other work reflects a similar belief; researchers adopting this approach have thus calculated the degree of centrality to a category by (manually or statistically) weighing a word according to all possible meanings, magnitude or intensity, or lexical relations with other words (Andreevskaia & Bergler, 2006; Subasic & Huettner, 2001).<sup>11</sup> Resulting dictionaries may be more precise, but they are often quite limited in their scope. Moreover, they tend to be computationally sophisticated, and only a fraction of such lexicons are available for research, making replication and improvement difficult.

Indeed, notwithstanding concerns about ambiguity, most dictionaries are generated for a particular purpose or genre of text, and as a consequence tend to be temporally and corporally specific. For instance, TAS/C was created to measure emotional tone and abstraction in psychotherapy sessions; DAL was developed to analyze the affective content of poetry and literature; RID was designed to distinguish between primordial and conceptual thinking; and DICTION was developed primarily to understand the rhetoric of speechmakers. Closest to our purposes are GI and LIWC, both of which were developed to analyze various affect categories in political texts.

Indeed, the dictionaries listed above show stunningly little overlap, and where they do overlap codes are often discrepant. (There are only two words that appear in all nine of the dictionaries analyzed here.) To be clear: Despite their varying uses, each dictionary relies on similar underlying constructs relating to various sentiment categories, yet the universe of words and the coding schemes vary greatly. Scholars have yet to construct a universal sentiment lexicon that can be exported across diverse corpora; despite many advances, a definitive lexicon does not exist (Athanaselis et al., 2005; Grefenstette, Qu, Evans, & Shanahan, 2004). The challenge remains, then, to expand the scope of a sentiment dictionary without compromising its accuracy. Below we outline a method to merge, standardize, and disambiguate three of the largest and most widely used lexical resources to create a comprehensive valence dictionary, which we hope meets some of the challenges presented above and which proves broadly applicable for scholarship on the tone of political communication.

### The Lexicoder Sentiment Dictionary (LSD)

In a first effort to produce a comprehensive dictionary coded for valence, aimed primarily at news content but potentially useful elsewhere as well, we merge and standardize three widely used and publicly available affective lexical resources from political science, linguistics, and psychology. The resulting Lexicoder Sentiment Dictionary (LSD) is a broad lexicon scored for positive and negative tone and tailored primarily to political texts.

LSD is comprised of words from Roget's Thesaurus, the GI, and the RID.<sup>12</sup> Roget's is the only truly comprehensive word list scored for sentiment. Our goal was to attribute the valence code that a word takes in most contexts. From Roget's, then, we include words from all categories that we identified as positive or negative. This includes, for example, positive categories such as "benevolence," "vindication," "respect," "cheerfulness," and "intelligence" and negative categories such as "insolence," "malevolence," "painfulness," "disappointment," and "neglect" ( $n = 47,596$ ).<sup>13</sup> From the GI, we include two broad valence categories labeled "Positiv" and "Negativ" ( $n = 4,295$ ). From the RID, we include the positive categories "positive affect" and "glory" and the negative categories "chaos," "aggression," "diffusion," "anxiety," and "sadness" ( $n = 1,056$ ).

We first sought to attribute a single code to every word per dictionary.<sup>14</sup> Words found in more positive than negative categories were coded as positive, and vice versa. Ambiguous and neutral words were not included—that is, those found in an equal number of positive and negative categories and those independently coded as neutral or ambiguous by any of the dictionaries. Each word was then classified as positive or negative in the LSD if (a) the word appeared in all three dictionaries and was consistently coded as positive or negative, or (b) the word appeared in just two of the dictionaries but was consistently coded as positive or negative. Additional analysis was required for the remaining words—those mentioned in just one dictionary or those for which there were coding discrepancies across the three dictionaries. Each of these words was considered for inclusion manually. Ambiguously coded words were included if the discrepancy was easy to resolve. Otherwise, contextual analysis (described below) was employed to make final decisions about the remaining words.

Few automated methods attempt to disambiguate homographs,<sup>15</sup> and standard bag-of-words approaches gloss over context entirely. We employed a number of strategies to address these issues, at least minimally. First and foremost, we made liberal use of WordStat's<sup>16</sup> keyword-in-context (KWIC) feature. KWIC functionality is available in many software packages, allowing content analysts to examine different uses of the same word in a corpus. LSD entries were analyzed in context using some 10,000 newspaper articles on a wide range of topics (selected randomly from a database of front-page news stories in major Canadian dailies over a 20-month period), enabling us to identify dictionary entries with multiple word senses and tricky contextual usage. Ambiguous entries were confirmed, dropped, or disambiguated in one of two ways. In some cases ambiguous words were replaced with contextual phrases that capture a particular use or sense of a word. For example, the homograph “lie” is only negative in certain contexts. To capture negative senses only, the dictionary entry “lie” is replaced with several phrases, including “a lie” and “lie to.” In other cases disambiguation occurs in the preprocessing phase, to which we turn in a moment.

Contextual analysis was also employed to analyze several problematic word categories. A number of direction words indicating an “increase” or “decrease” initially found their way into the dictionary, even though they do not have a clear tone. For instance, the verb “augment” is listed in many positive categories, though the tone clearly depends on what is being augmented. Likewise, “decline” is often listed as negative, even though it is not so if something negative is declining (e.g., unemployment). A list of economic terms was also analyzed, given the prevalence of economic terms to which tone is attributed. For instance, this process removed words such as “profit” and “credit.” Additionally, common stop-words were removed and alternative spellings added. Contextual analysis also facilitated the addition of numerous dictionary entries—many particular to political news reporting—that were not present in any of the core lexicons ( $n = 1,021$ ). Finally, all dictionary entries were truncated to capture inflected variations, provided that the inflected forms retained the same tone. In cases where the tone of inflected words differed from the original entry, truncation was not applied; instead, inflected variations were individually added to the dictionary with appropriate codes for tone. The final dictionary comprised 4,567 positive and negative words.<sup>17</sup>

Finally, contextual information was employed to generate several preprocessing modules, which format the text prior to content analysis to facilitate word sense disambiguation and contextual analysis of affective language.<sup>18</sup> The first preprocessing module standardizes and/or removes punctuation. The second removes capitalized words (other than the first word of a sentence). The logic here is to remove proper nouns, which should by

definition not have tone. The third module processes basic negation phrases. Standardizing negation allows for a variety of negated phrases to be captured with a limited number of dictionary entries. The preprocessor first replaces negation words such as “no,” “never,” “neither,” “hardly,” “less,” and so forth with “not.” Second, various verbs and pronouns following negation words are removed—“not going to,” “not feeling,” “hardly any,” “nor were they,” “no one was,” “without much,” and so forth are replaced with “not.” Standardized in this manner, the dictionary entry “not good” captures a range of negative phrases including “not at all good,” “hardly a good idea,” “no one good,” “nor was it good,” “without goodness,” and many more. Finally, the main body of the preprocessor removes “false hits” for dictionary entries, where false hits are topical, multi-toned, or non-tonal instances of a dictionary entry. For example, multi-toned phrases such as “good grief,” “losing hope” and “tears of joy” are processed to remove the toned words “good,” “hope,” and “tears”; non-tonal phrases such as “an awful lot,” “crude oil,” and “child care” are processed to remove the toned words “awful,” “crude,” and “care.” The preprocessing of over 1,500 words and phrases facilitates basic word sense disambiguation and the contextualization of many commonly used sentiment words and phrases.<sup>19</sup> Some of the most nuanced entries in the dictionary rely on a combination of contextual phrases, truncation, negation, and preprocessing.

### Testing: Data and Results

Does the LSD work? More precisely, does the LSD produce codes that are (a) consistent with human coding and (b) more consistent with human coding than other available sentiment dictionaries?

LSD is just the dictionary itself, of course—with some minor reformatting, it can be used by any number of available software packages. We implement it, and all other dictionaries used here, in Lexicoder, which is a freely available java-based, multiplatform software that implements frequency analysis for any number of user-written categorical dictionaries.<sup>20</sup> As the name suggests, it was developed alongside the LSD.

Our aim in this section, then, is to compare the reliability and validity of the LSD to several commonly used lexicons. More specifically, we compare results using the LSD with results using the following<sup>21</sup>:

- LIWC, from which we include the positive category “positive emotion” and the negative categories “negative emotion,” “anxiety,” “anger,” and “sadness” ( $n = 1,502$ )
- WNA, from which we use a subset of affective words generated from WordNet synsets labeled “positive” or “negative” ( $n = 1,640$ )
- TAS/C’s “emotional tone” word list, which is labeled on the dimension pleasure-displeasure ( $n = 4,058$ )
- Mean scores from DAL, which labels words along a scale of pleasantness ( $n = 8,743$ )
- Mean scores from ANEW, which labels words along a scale of pleasure ( $n = 1,034$ )
- Point-wise mutual information scores from the PMI, which are based on the proximity of entries to positive and negative seed words in a text ( $n = 1,719$ )

WNA and TAS/C did not require recoding to generate a valence dictionary. In the case of LIWC, we simply collapsed the four negative categories. DAL, ANEW, and PMI posed a challenge, since they are measured on a continuous scale. Here, we had little indication of where the cut-points between positive, neutral, and negative might be. Thus, we divided

each dictionary into terciles based on the scale capturing sentiment, putting the top tercile into the positive category, the bottom tercile into the negative category, and omitting the middle third (which upon inspection was indeed comprised mostly of ambiguous or neutral words).

We also include in our comparison the GI, RID, and Roget's—the three dictionaries from which the LSD was derived. Part of the intention in creating our own dictionary was to resolve some of the ambiguity and imprecision in these large dictionaries. And given the method by which we combined the three dictionaries (described above), we do expect the LSD to produce somewhat different—and ultimately improved—results.

Note that not all, indeed few, of these dictionaries provide categories clearly aimed at capturing valence or positive-negative sentiment in text. We approach each dictionary as researchers interested in capturing positive and negative tone. Thus, it is important to note that we are not always using them exactly as intended. For instance, several are constructed with multiple dimensions (that were not conducive to valence codes), which have simply been omitted. And obviously our coding of the continuous measures is somewhat crude. Nevertheless, we regard our tests as a good indication of the extent to which one can achieve a valid measure of tone using a range of currently available content-analytic dictionaries.

The measure of tone outlined below captures, simply, the degree of positive or negative coverage in news stories. Recall that the unit of analysis in the automated system is un-contextualized words, aggregated in this case to the document level. Since this process does not distinguish among words (other than accounting for their polarity), the measure necessarily reflects a combination of objective content about the various issues and events being reported on and subjective opinions or attitudes about the content itself. Positive coverage may result from attention to objectively positive events or policy successes; it might equally result from avid support for, or praise of, government policies. Likewise, negative coverage may result from attention to objectively tragic events or the failure of a policy; it might equally reflect criticism of government policies. Thus, tone should be considered a composite measure of (a) the relative negativity of the actual events or issues being covered and (b) the opinions and attitudes of newsmakers about those events and issues.<sup>22</sup> In terms of media and public opinion research, we expect that in many cases both components of tone matter; this general measure of the tone of coverage thus reflects both. In the event that one must distinguish between the two, researchers should clearly proceed with caution (and, more to the point, some additional data processing).

Simply comparing results across dictionaries is not enough, of course—we need to compare them with some other externally valid analysis, specifically human-coded content-analytic data. We do so here using a body of data coded by three trained human coders. The data include 900 articles from the *New York Times*. Four hundred fifty were randomly drawn from an existing database on all economics stories published in the *Times* from 1988–2008. The other 450 were randomly drawn from a previously topic-coded database of all front-page stories in the *Times* from 2007–2009. In this case, we drew 150 randomly from within each of three topic categories: environment, foreign affairs, and crime. The selection of these topics provides, we believe, a good basis for an initial test of the LSD. We have chosen a range of domestic and international policy issues; we also intentionally use issues that have seen a particularly large amount of attention in the political communication literature.

Coders were directed to read each article and then assign to each article a tone of positive, negative, or neutral. The tone is intended to reflect the overall sentiment of the article—not the tone for a particular individual or paragraph or the coders' own feelings

about the news content. Directing the coders in this general way is critical to our endeavor, since we need to have coders produce coding that is consistent with what we then ask of our dictionary.

Assigning identical codes is clearly important when we are assigning topics, or frames, but may not be as feasible where tone is concerned. As noted above, for some computational linguists, small differences across human coders are regarded as capturing real variation, or ambiguity, given the natural and structural ambiguity in categories of sentiment (Andreevskaia & Bergler, 2006; Subasic & Huettner, 2001). Following this approach, codes from the three human coders are arranged here into a 5-point scale: negative, where all three coders selected negative; mildly negative, where two coders selected negative; neutral, where two or more coders selected neutral; mildly positive, where two coders selected positive; and positive, where all three coders selected positive.<sup>23</sup>

The resulting distribution of stories across tone categories and topics is shown in Table 1. Overall, more than half of the stories were coded into one of the negative categories; roughly 25% of stories were coded as positive, and roughly 20% were coded as neutral. The distribution changes as we move from issue to issue, of course. Environmental coverage was comparatively positive in our database; crime and foreign affairs were somewhat more negative. This is apparent not just in the distribution of human codes, but in the automated tone measure shown in the final column of Table 1. “Net tone,” our core measure of automated tone, is the proportion of positive words minus the proportion of negative words in an article, that is: (# positive words/all words) – (# negative words/all words).<sup>24</sup> So a score of –2.4 for crime means that, on average, in crime stories there is a 2.4-percentage-point gap between the number of negative words and the number of positive words.

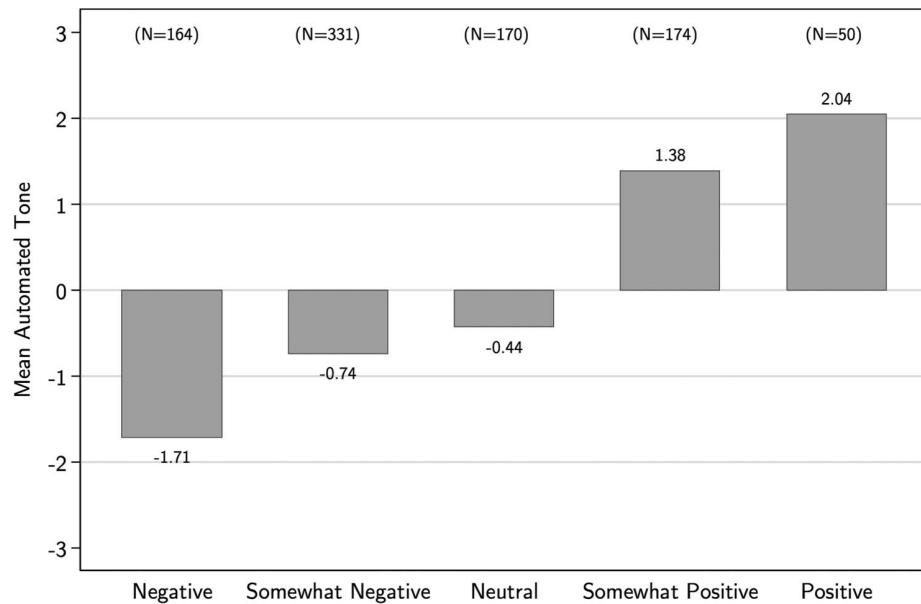
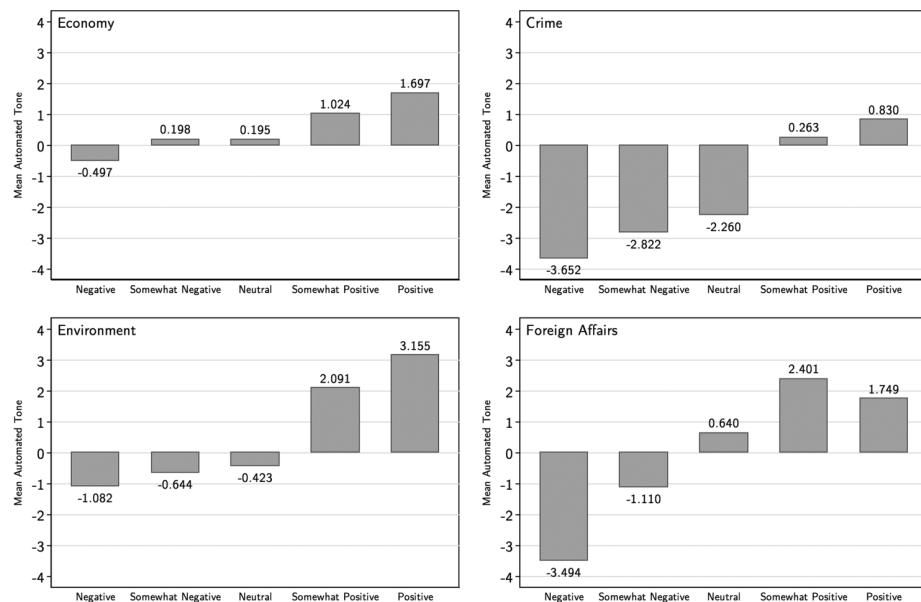
How do LSD net tone scores differ across manual tone categories? This is the central test of the success of the LSD dictionary; it appears in Figure 1. The figure shows the average net tone score for each of the five codes resulting from the manual coding. Results are as we would hope—in the aggregate, more negative stories receive, in short, more negative scores. The difference in mean net tone is statistically significant, even using the more stringent two-tailed test, across all categories except one. The LSD does not distinguish especially well here between somewhat negative and neutral. There is a difference in the right direction, to be sure, but it narrowly misses statistical significance.

The issue of whether the dictionary works equally well across all four topics is the focus of Figure 2. Results are somewhat noisier, in part due to much-reduced sample sizes. And the varying tone of coverage across issues is in evidence here. The range of tone in economic articles is rather narrower than the others, and crime and foreign affairs coverage lean more toward the negative than do economic and environmental coverage (at least during the time period investigated here). Even so, all topics show the correct basic trends. The difference between neutral and somewhat negative is not as wide as we would like for economic stories or the environment, and there is a drop in tone from somewhat positive to positive in foreign affairs articles that we would not expect. But overall, we regard these results as promising.

Table 2 and Figure 3 provide some basic information comparing results using the LSD with those using other dictionaries. Table 2 makes readily apparent the fact that these dictionaries do indeed capture different things. The table shows bivariate correlations between the net tone measures that result from using one dictionary versus another. The first column is the most important for our purposes, as it shows the bivariate correlations between the LSD and all others. As expected, the correlations are relatively high between the LSD and the three dictionaries on which it is based (GI, Roget’s, and RID). Roget’s is slightly

**Table 1**  
Data set descriptives

	Manual tone					n	Mean tone
	Negative (%)	Somewhat negative (%)	Neutral (%)	Somewhat positive (%)	Positive (%)		
All	18.45	37.23	19.12	19.57	5.62	900	-0.284
Economy	19.37	35.36	18.02	21.4	5.86	450	0.318
Crime	18.67	40.67	28	10	2.67	150	-2.414
Environment	11.03	31.03	16.55	31.72	9.66	150	0.589
Foreign	22.67	45.33	16	12	4	150	-0.836

**Figure 1.** Automated versus manual tone.**Figure 2.** Comparing results across topics.

lower, and this is likely due to its breadth—many of the more arcane entries were simply not applicable for our purposes. The high correlation with LIWC makes good sense. Like the LSD, LIWC was constructed using a methodology based on definitive codes. It is one of the few to contain large positive and negative valence categories; it is also one of the

**Table 2**  
Pairwise correlations, automated dictionaries

	LSD	GI	ROG	RID	ANEW	DAL	LIWC	PMI	TAS/C
GI	0.672								
ROG	0.471	0.469							
RID	0.669	0.480	0.350						
ANEW	0.500	0.464	0.236	0.367					
DAL	0.519	0.481	0.285	0.385	0.482				
LIWC	0.753	0.598	0.428	0.663	0.488	0.490			
PMI	0.228	0.172	0.093	0.128	0.115	0.201	0.159		
TAS/C	0.663	0.601	0.455	0.513	0.438	0.432	0.635	0.178	
WNA	0.230	0.220	0.102	0.068	0.076	0.155	0.224	0.176	0.178

*Note.*  $N = 900$ . All correlations are significant at  $p < .001$ .

only dictionaries making liberal use of truncation. (In our own analyses, we have found that truncation has a huge impact on performance, due to the large boost in coverage.)

Overall, bivariate correlations between results from many of these dictionaries are not especially high. This is not particularly surprising, given that the dictionaries were built for very different purposes, even as they sought to capture similar concepts. Whether they serve our purpose best is the focus of Figure 3.

Figure 3 shows directly comparable tests of external validity—replications of Figure 2, but using each of nine other content-analysis dictionaries. The scales on the y-axes vary widely, since the various dictionaries have different numbers of words in them. That the RID produces only negative values is a function of that dictionary being heavily weighted toward negative words; the opposite is true for TAS/C, which leans toward the positive. The raw values are not of primary interest here, however. What matters is whether the dictionaries produce net tone codes that systematically increase alongside results from manual coding.

In some cases, they certainly do. There is a clear and nearly monotonic increase in net tone for the GI, LIWC, and TAS/C dictionaries, and the other dictionaries perform sufficiently as well, though with somewhat rougher results. This is true in spite of the fact that the dictionaries are in some cases only barely correlated with each other and contain vastly different word lists. This highlights two facts. First, valence, as captured by manual coders, likely depends on a variety of factors, only some of which are adequately captured by any one of these dictionaries. So, two dictionaries with rather different word lists can produce aggregate code statistics that match, at least in part, human-coded results. Second, automated valence codes can be relatively successful in the aggregate even as they are noisy at the individual level. The relative success of each of these dictionaries is in part due to a large sample of articles; were we to look at any one single article, one dictionary might produce quite a different estimate of tone than another. Overall, however, the language in each of these dictionaries captures something to do with valence, and as a consequence we can find relatively sensible results in the aggregate.

How can we better judge the relative success of each of these dictionaries? Table 3 presents what we regard as the critical test. The table presents basic ANOVA results in which the variance of each net tone measure is analyzed as a function of the five-category manual results. A first version assumes a linear effect—results are essentially the  $R^2$  values

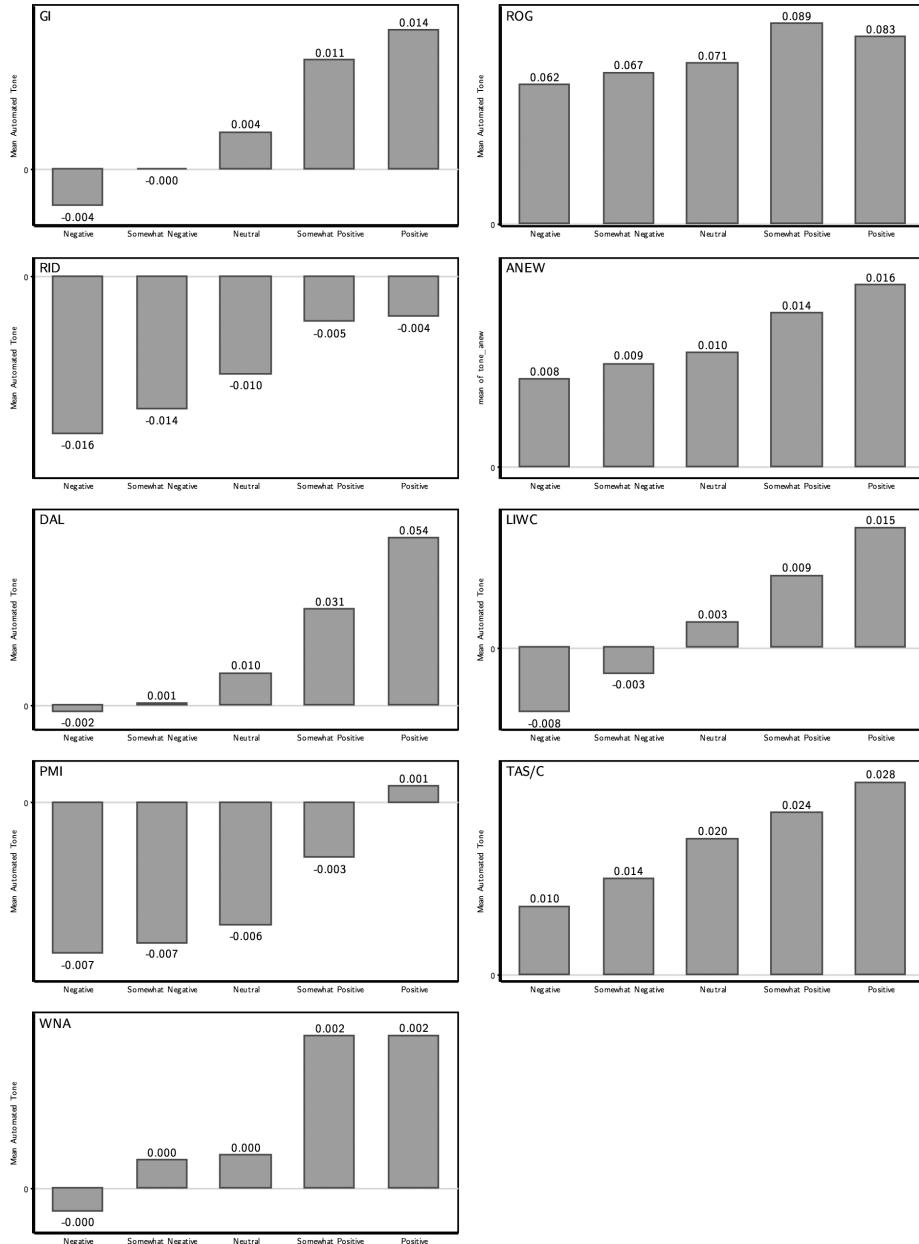


Figure 3. Comparing results across dictionaries.

from an OLS model regressing net tone on the five-category scale. A second version relaxes the assumption of linearity, and results in that case are essentially the  $R^2$  values from a model regressing net tone on a set of four dummy variables (and one residual category) from the manual coding. Results change very little from one column to the next. In each case, the LSD matches human codes better than the other dictionaries; more precisely,

**Table 3**  
Automated dictionaries and manual coding

	Percentage variance explained	
	Linear	Nonlinear
LSD	14.3	15.6
GI	7.0	7.2
ROG	5.3	6.2
RID	10.2	10.6
ANEW	2.3	2.5
DAL	6.4	7.3
LIWC	12.7	12.7
PMI	3.6	4.2
TAS/C	7.7	7.7
WNA	1.3	1.4

*Note.* Cells contain the percentage of variance in the various automated measures that is accounted for by the manual 5-point coding.

human codes account for a greater proportion of the variance in net tone as estimated using the LSD than in net tone estimated by any other means.

This is not to say that there isn't room for improvement. Results in Table 3 suggest that 15.6%<sup>25</sup> of the variance in net tone as determined using the LSD can be accounted for by human codes. There is a good deal of variance that is not clearly accounted for, and there surely are a good number of errors at the level of individual articles. This is actually quite a difficult thing to gauge, since the interval-level net tone measure does not have obvious cut-points at which we can easily distinguish between negative, neutral, and positive. That said, a preliminary test is illustrative. If we allow zero to be the neutral point for the LSD net tone score,<sup>26</sup> and then allow all net tone values that are significantly different from zero (based on our sample) to fall into either the positive or negative categories, we can compare the resulting three-way categorization of articles, as determined by the LSD, with a similar distribution using human codes. Doing so suggests that of the 224 stories categorized as positive by at least two of the three human coders, LSD results assign 74% to the positive category and just 12% to the negative category. Of the 495 articles that are categorized as negative by at least two coders, LSD results assign 53% to the negative category and 32% to the positive category.<sup>27</sup> Thus, it seems that the LSD performs better in the attribution of positive tone than in the attribution of negative tone. Only further testing can reveal exactly why this is the case.

### **Media Tone and Vote Intentions**

What are the uses of the LSD? There are, we believe, many. The LSD can be used to capture the tone as well as, more generally, the use of affective language in a wide range of contexts. We explore one such context here: news about political parties and candidates in election campaigns.

Past work suggests that the tone of news content is strongly related to variations in vote intentions during election campaigns. This could be because media drive vote intentions;

it could also be that media simply reflect the tone and content of public debate at the time. Most likely, media do a little of both. In any case, the relationship between media and vote intentions tends to be rather strong.

Recent work using manually coded data points to a strong association between vote intentions and lagged media content in the 2004 and 2006 Canadian federal elections (Soroka et al., 2009). We extend this work here, comparing results from human-coded data for the 2006 election with results based on automated data using the LSD. Doing so allows us to examine not just the convergent validity of the LSD (above), but the predictive validity as well.

Models in the 2009 article predict vote shares for both the Conservative and Liberal parties in the 2006 Canadian federal election campaign, using a combination of 4-, 5-, and 6-day lags of media content (lagged in this way to allow for predictions 3 days ahead), alongside a 4-day lag of vote intentions and a set of dummy variables to capture house effects.<sup>28</sup> The original data were based on manually coded newspaper content, drawn directly from hard copies of newspapers during the campaign. To compare these results with automated data, we created a new database of campaign-related stories drawn from full-text indices in Nexis for the five English-language newspapers used in the original article.<sup>29</sup> The samples will not be identical, of course. To facilitate comparison, we matched articles by date, newspaper, and title, capturing 1,590—roughly half—of the original human-coded stories to be preprocessed and coded with the LSD. We thus rely on that subsample here to replicate the original model, and then compare results to those using the LSD.<sup>30</sup>

All 1,590 stories were preprocessed as above and coded for tone using the LSD. In order to attribute tone to one party/leader or the other, we look for the co-occurrence of party/leader names and positive or negative keywords in the same sentence. One consequence of this proximity-based search is that dictionary terms that occur in sentences that mention both Liberals and Conservatives are attributed to both parties. Using “net tone” for the automated measure ensures that these common words cancel each other out, which is desirable, since we do not know to whom they should be attributed. Consequently, the measure below captures the relative tone of coverage toward each actor over the campaign period.<sup>31</sup>

Our automated measure of leader or party tone is calculated as follows: # positive words – # negative words, using only those words that co-occur in sentences that mention the party or leader’s name.<sup>32</sup> The measure takes on positive values when a party/leader mention co-occurs with more positive than negative words, and negative values when a party/leader mention co-occurs with more negative than positive words. Note that more party/leader mentions in a given article increase the number of words analyzed, and thus the potential value of this measure<sup>33</sup>; in these data, the article-level measure ranges from about –4 to 6.<sup>34</sup>

Results are shown in Table 4. The table includes coefficients for the media tone variables; the  $R^2$  and adjusted  $R^2$  values for model fit, and to assess predictive accuracy, the mean average error (MAE) of the estimate, which captures the average gap between the prediction and the actual vote intentions.<sup>35</sup> Control variables, including lagged vote intentions, and dummy variables capturing house effects are included in the appendix.

Model 1 is the baseline model, and includes no media variables; Model 2 includes the original manually coded media variables for each leader and party tone; and Model 3 includes the same set of media variables, though produced using the LSD. For all media variables, the table shows the summed coefficients (and standard errors) for the 4-, 5-, and 6-day lags, leaders and parties combined.

**Table 4**  
Media content and vote intentions, 2006 Canadian election

	Conservatives			Liberals		
	1	2	3	1	2	3
<b>Media tone coefficients</b>						
$\sum \text{CPC}_{t-(4,5,6)}$	21.080** (7.624)	-.020 (.921)		-8.858 (5.540)	-1.161 (.971)	
$\sum \text{LPC}_{t-(4,5,6)}$	-8.485 (9.670)	-1.177* (.636)		28.188** (6.508)	2.047** (-1.161)	
<b>Variance explained</b>						
$(R^2)$	.725	.868	.840	.745	.944	.886
<b>Accuracy</b>						
(MAE)	1.548 (1.097)	1.107 (.709)	1.087 (.954)	1.654 (1.166)	.742 (.587)	1.085 (.812)

*Note.*  $N = 47$  for all models. Media tone coefficients cells contain OLS coefficients with standard errors in parentheses; MAE cells contain mean average errors with standard deviations in parentheses.

\* $p < .10$ ; \*\* $p < .05$ .

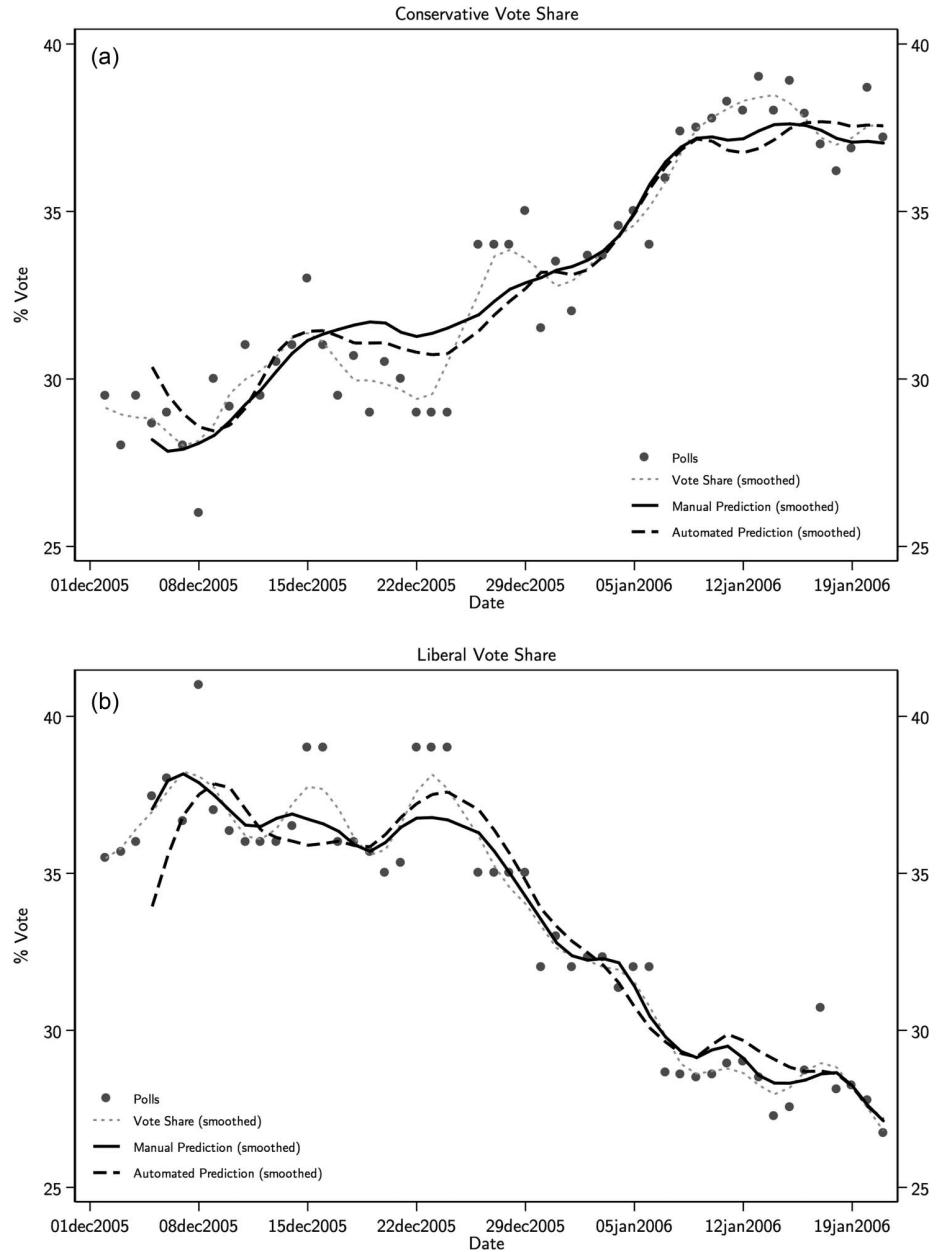
Models 2 and 3 show the expected relationship between lagged media content and current vote intentions.<sup>36</sup> The variance in the automated measure is very different from that in the manual measure, so we cannot easily compare the magnitude of coefficients. We can compare the significance of coefficients and the predictive capacity of the models, however. In the Conservative party (CPC) Model 2, using manual tone, Conservative tone is positively related to vote shares, while Liberal tone is negatively signed but statistically insignificant. In Model 3, now using automated tone, Conservative tone has no significant effect, but Liberal tone is both negative and significant. The predictive capacity of the models is very similar, however. The  $R^2$  values for the models are .868 and .840; the MAEs are 1.107 and 1.087. Clearly, the automated measure is as valuable a predictor here as the manual measure.

The Liberal models are roughly similar. Liberal tone is positive and significant in both the manual and automated models; Conservative tone is consistently negative and insignificant. In terms of model performance, the manual measure is somewhat stronger, leading to a somewhat higher  $R^2$  value and a lower MAE. Generally speaking, there is consistency in the significance of the media variables and only a slight advantage to using the manual measure here. Overall, we find these results very encouraging.

Figure 4 confirms the close relationship between results using the manual and automated measures of party and leader tone. The panels show poll results alongside the predictions made by Models 2 and 3 in Table 4.<sup>37</sup> Clearly, the automated measure is capturing much of what the manual measure captures—initial evidence of the predictive validity of the LSD, at least in the context of the 2006 electoral campaign in Canada.

## Conclusions

We view this study as a critical starting point for a revived discussion about dictionary construction and validation in sentiment analysis for political communication. As digital



**Figure 4.** Media tone and vote shares in the 2006 Canadian election.

information proliferates and attention to the role of affect and emotion in politics increases, it is incumbent on us to both develop and critically evaluate relevant measurement instruments. Given the relative ease of implementing dictionary-based automation (once a dictionary is constructed), it is inevitable that these tools will become more widely available and broadly used in the near future. At a minimum, we hope to have brought attention

to the range of resources that currently exist. We hope, however, that the results above suggest the potential for the LSD to address some of the challenges of dictionary-based sentiment analysis.

Comparatively speaking, we are generally pleased with the ability of the LSD to establish the overall tone of newspaper articles. Our assessment of how well the LSD works is based on what we see as critical tests of external validity: a test of whether the dictionary produces tone codes that are in line with those produced by (expert) human coders and, moreover, a test of whether it does so more often than other available dictionaries. It appears that, in our sample, it does. This is not to say that other samples would not produce somewhat different results. Our focus here has been on political news stories, and we certainly allow for—indeed, welcome—the possibility that the dictionary will be more or less successful as it is applied to other texts.

Importantly, preprocessing accounts for a small portion of the LSD's performance, and we regard this as promising. Such modules are incredibly labor-intensive to produce (as are the dictionaries), but they will only improve with time. And ours may only have scratched the surface. For instance, we considered only one basic set of negation patterns, but linguistically there are many. In any case, by moving beyond a bag of words we have improved our confidence that even modest amounts of contextual analysis can be quite fruitful. Still, there is much room for improvement.

The dictionary also can be refined and made more efficient. An interesting finding (which we did not note above) is that it was neither the size of the dictionary (the number of words) nor the scope of coverage (the number of “hits” in a text) that drove performance. The LSD accounted for the most variance according to the ANOVA tests, though it fell squarely in the middle on both counts. Dictionaries with too many words (i.e., Roget's) or with coverage that was too vast (i.e., DAL) suffered, arguably because they lack discriminative capacity, as did those with too few words and limited scope (i.e., WNA), which may not have had enough.

The avenues for future research are many. We have tested here very broad categories of sentiment: positive and negative. There is a multitude of subcategories covering all manner of cognitive and affective concepts, however, and we hope our efforts here raise flags for anyone interested in using these more refined categories, especially across dictionaries. Operationalized as word lists, is there a difference between hope and optimism? Fear and anger? To what extent are different dictionaries measuring concepts similarly? Future research should continue to probe the development and validation of such measures so they may be used with greater confidence in scholarly work. Another promising pursuit would be in the area of subdocument lexical analysis—relating tone to particular topics or actors. Indeed, this work has begun. Although rudimentary, the proximity-based measures of actor tone used in the vote share models above make clear that attributing tone to actors using automation is both feasible and effective. Refining this technique would not only improve automated results, it would also greatly broaden the substantive research questions that automation could potentially address. Finally, while we do not directly compare the dictionary-based approach to statistical methods, such an effort would certainly be welcome.

In sum, our goal has been to contribute to the production of reliable, valid, and comparable results in automated sentiment analysis. We welcome improvements to the dictionary, and indeed a main theme of this project has been to produce a dictionary that not only is valid, but also one that is readily available to researchers interested in capturing the sentiment of news content. As we noted above, many of the existing lexicons are not readily open to the kind of improvement that comes with repeated testing by multiple researchers,

nor do they facilitate adaptation to particular research interests or goals. The LSD is, in contrast, freely available for academic use, adjustable, adaptable, and thus readily open to improvement. The current version also, based on our results above, seems to work rather well, and the potential applications of such a dictionary are rather vast. The analysis of sentiment plays a central role in work in political communication. With greater confidence in automated content-analytic measures, we will be better equipped to understand the particular ways in which the sentiment of mass media affects public opinion and behavior.

## Notes

1. On the tone of news content, see discussion in the following section. On negative advertising in particular, see, for example, a meta-analysis by Lau, Sigelman, Heldman, & Babbitt (1999).
2. Affect generally refers to any conscious or unconscious feeling as distinct from a cognitive perception, and it is a necessary component of the more complex experience of emotion, which is generally conceived of as having both affective and cognitive components (Huitt, 2003). For the purposes at hand, we use the terms sentiment and tone to refer broadly to affect or emotion. Practically speaking, automation cannot distinguish between the two, and thus the distinction is adequate. Finer distinctions within these broad categories are made explicit in the text that follows.
3. On the role of emotion in politics generally, see especially Elster (1999); Hall (2002); Marcus (2000); Marcus, Neuman, and MacKuen (2000); Neumann, Marcus, Crigler, & MacKuen (2007); and Walzer (2002). On the processing of media information, see, for example, Detenber and Reeves (1996); Lang, Dhillon, and Dong (1995); and Newhagen (1998). On the primacy of affect in the decision-making process, see, for example, Abelson (1963); Damasio (1995); Huitt (2003); LeDoux (1996); Lodge and Taber (2000); and Zajonc (1984).
4. See, for example, Bloom and Price (1975); Fair (1978); Ito, Larsen, Smith, and Cacioppo (1998); Kahneman and Tversky (1979); Quattrone and Tversky (1988); and Soroka (2006).
5. For examples in computational linguistics, see, for instance, Généreux and Evans (2006); Hatzivassiloglou and McKeown (1997); Joachims (1998); Kim and Hovy (2006); Kushal, Lawrence, and Pennock (2003); Leshed and Kaye (2006); Mishne (2005); Pang et al. (2002); and Wiebe (2000). Political scientists have taken up statistical methods to automate policy positions in party manifestos (Laver, Benoit, & Garry, 2003) and the topic of congressional speeches (Purpura & Hillard, 2006).
6. For examples of UML, see, for example, Quinn, Monroe, Colaresi, and Crespin (2006); Simon and Xenos (2004); and Turney and Littman (2002).
7. There are many other examples, including Pennebaker, Slatcher, and Chung's (2005) use of word counts to derive psychological attributes of political candidates from their tone in natural conversation. Frequency analysis has also been applied to international communications to monitor conflict and various efforts to use word counts in the analysis of international relations (e.g., Doucet & Jehn, 1997; Hogenraad, 2005; Holsti, Brody, & North, 1964; Hopmann & King, 1976;). For work using DICTION specifically, see <http://www.dictionsoftware.com>.
8. Several studies have used proximity-based lexical rules to attribute tone to actors or topics at the subdocument level by measuring the local co-occurrence of dictionary words and a "subject" of interest (see, e.g., Mullen & Collier, 2004; Pang et al., 2002; Tong, 2001; see also research on subjectivity analysis, e.g., Wiebe, 2000). Despite many sophisticated approaches, however, state of the art machine-learning and NLP sentiment analysis techniques cannot readily unravel the topic-specific relationship between presented evidence and speaker opinion (see, e.g., Thomas et al., 2006, p. 2).
9. GI combines Osgood's Semantic Differential Scale and the Lasswell Value Dictionary.
10. The original GI software package was programmed with a set of word sense disambiguation rules that corresponded to various senses annotated in the dictionary. However, neither the program nor the rules are maintained. Thus, most research simply collapses or weighs the carefully annotated multiple word senses, to the dismay of creator Philip Stone, who laments the tendency as "a step backwards in both theory and technique" (1986, p. 76).

11. There are other approaches to word scores as well (see, e.g., Hatzivassiloglou & McKeown, 1997; Kamps, Marx, Mokken, & de Rijke, 2004; Scharl et al., 2003; Thelen & Riloff, 2002; Turney & Littman, 2002, 2003).
12. We were unfortunately unable to include other dictionaries in the construction of the LSD due to proprietary restrictions either on their use, modification, or distribution.
13. A complete list of categories classified as positive or negative is available upon request.
14. Alternately, we could have aggregated by category across dictionaries to calculate word scores. We chose to aggregate per dictionary first to avoid biasing the tone in favor of the dictionary with the most categories.
15. The General Inquirer is a notable exception (see Note 9). Hart's DICTION program also makes modest statistical adjustments by differentially weighing homographs.
16. For more on WordStat, see <http://www.provalisresearch.com/wordstat/Wordstat.html>.
17. Trials were conducted using a subset of subjective words, noted in the literature to improve sentiment analysis (Wiebe, 2000). However, this version did not perform as well as the full LSD.
18. These modules are available alongside the LSD online.
19. By way of example, randomly drawn positive terms include beaming, charity, cognizant, comprehend, credible, curious, dignify, dominance, ecstatic, friend, gain, gentle, justifiably, look up to, meticulous, of note, peace, politeness, reliability, and success; randomly drawn negative terms include admonish, appall, disturbed, fight, flop, grouch, huffy, hypocritical, impurity, irritating, limp, omission, oversight, rancor, relapse, sap, serpent, untimely, worrying, and yawn.
20. Lexicoder was developed by Stuart Soroka and Lori Young, and programmed by Mark Daku. It is available at <http://www.lexicoder.com>
21. We would very much have liked to include DICTION in our study; however, the word list—though it can be inspected—is not exportable.
22. Notably, it is not a measure of journalistic tone, bias, or subjectivity.
23. Note that differences in human codes in our sample are a matter of degree. There is no single case of both positive and negative tone codes for a single story, for instance. Thus, we are confident that differences do reflect genuine ambiguity.
24. Proportions are used to control for the varying length of articles.
25. Note that the  $R^2$  value for unprocessed text is 1.4 points lower than above, and about 4 points lower when inflections are not applied. Full results are available from the authors.
26. It need not be, of course, and that is a first difficulty. Another option is to let the mean net tone of all neutral articles, as determined by coders, be zero. Using that value, .45, does not significantly change the results mentioned in the text.
27. The neutral category represents a real problem in this kind of analysis, since it is not clear in the LSD codes where exactly neutral ends, and the error around the mean in this particular sample is of course a very rough proxy.
28. The lengthy lag of media content (4 days or more) is a consequence of trying to build models that can predict shifts in vote shares. The original models are described in detail in Soroka et al. (2009).
29. Stories were selected by searching for the word "election" or any one of the party leader's names in the story text in searches limited by geography (Canada). Newspapers include the *Globe and Mail*, the *Toronto Star*, the *National Post*, the *Vancouver Sun*, and the *Calgary Herald*. We exclude the two French-language newspapers here, since they cannot be coded using the LSD.
30. Using this matched subsample of articles makes for a stricter test of the LSD in comparison with human coding than would a test using the entire database. This makes sense for the purposes at hand. However, we should note that this approach greatly attenuates one of the main advantages of automated coding—namely, the ability to work with much more data than could feasibly be coded by humans. Given that the reliability of automated tone will increase with sample size, and the ease with which sample size can be increased, we are limiting our automated predictions here rather severely.
31. Note that manual tone in the original study is also a measure of net tone, accounting for the relative weight of positive versus negative coverage toward each actor during the campaign.

32. Note that we include variants of party names such as “Grits” or “Tories” for Liberals and Conservatives, for instance.

33. We could also calculate net tone as a proportion of the total number of words analyzed, to account for differences in the volume of coverage of various actors. We see some advantage to the raw measure we use here, however, since it captures, in part, the consequences of a large versus small amount of negative/positive coverage. In any case, results are not very different when we use the percentage-point measure.

34. This automated measure of tone differs from the manual measure in at least one way. Recall that automated tone is a composite measure of the relative negativity of the events or issues being covered and the opinions and attitudes of newsmakers. In the original study, expert coders were trained to measure the latter. We should accordingly expect to see some differences in their relationship to vote shares. However, as the results demonstrate, any such differences turn out to be minor.

35. On the value of the MAE as a goodness of fit measure in prediction and forecasting, see Krueger and Lewis-Beck (2005).

36. And note that results from these models—relying on just 1,590 of the original manually coded articles—are not very different from the original results in Soroka et al. (2009).

37. Predictions are smoothed using lowess smoothing with a bandwidth of .2.

## References

- Abelson, R. P. (1963). Computer simulation of “hot” cognition. In S. Tomkins & S. Messick (Eds.), *Computer simulation of personality* (pp. 277–298). New York, NY: Wiley.
- Andreevskaia, A., & Bergler, S. (2006). *Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses*. Paper presented at the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy.
- Athanasis, T., Bakamidis, S., Dologlou, I., Cowie, R., Douglas-Cowie, E., & Cox, C. (2005). ASR for emotional speech: Clarifying the issues and enhancing performance. *Neural Networks*, 18, 437–444.
- Bloom, H. S., & Price, H. (1975). Voter response to short-run economic conditions: The asymmetric effect of prosperity and recession. *American Political Science Review*, 69, 1240–1254.
- Bolasco, S., & Ratta-Rinaldi, F. (2004). *Experiments on semantic categorisation of texts: Analysis of positive and negative dimensions*. Paper presented at the 7th International Conference on the Statistical Analysis of Textual Data, Louvain-la-Neuve, Belgium.
- Bradley, M. M., & Lang, P. J. (1999). *Affective Norms for English Words (ANEW): Stimuli, instruction manual and affective ratings*. Gainesville: Center for Research in Psychophysiology, University of Florida.
- Cho, J., Boyle, M. P., Keum, H., Shevy, M. D., McLeod, D. M., Shan, D. V., & Pan, Z. (2003). Media, terrorism, and emotionality: Emotional differences in media content and public reactions to the September 11th terrorist attacks. *Journal of Broadcasting & Electronic Media*, 47, 309–327.
- Damasio, A. (1995). *Descartes’ error: Emotion, reason, and the human brain*. New York, NY: Avon Books.
- Detenber, B. H., & Reeves, B. (1996). A bio-informational theory of emotion: Motion and image size effects on viewers. *Journal of Communication*, 46, 66–84.
- Doucet, L., & Jehn, K. (1997). Analyzing harsh words in a sensitive setting: American expatriates in communist China. *Journal of Organizational Behaviour*, 18, 559–582.
- Edelman, M. (1985). Political language and political reality. *Political Science and Politics*, 18, 10–19.
- Elster, J. (1999). *Alchemies of the mind*. Cambridge, England: Cambridge University Press.
- Eshbaugh-Soha, M. (2010). The tone of local presidential news coverage. *Political Communication*, 27, 121–140.
- Fair, R. C. (1978). The effect of economic events on votes for president. *Review of Economics and Statistics*, 60, 159–173.
- Farnsworth, S. J., & Lichter, S. R. (2010). *The nightly news nightmare: Media coverage of U.S. presidential elections, 1988–2008*. Lanham, MD: Rowman & Littlefield.

- Généreux, M., & Evans, R. (2006). *Towards a validated model for affective classification of texts*. Paper presented at the Workshop of Sentiment and Subjectivity in Text, Association for Computational Linguistics, Sydney, Australia.
- Gentzkow, M., & Shapiro, J. (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, 78, 35–71.
- Grefenstette, G., Qu, Y., Evans, D., & Shanahan, J. (2004). Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes. In Y. Qu, J. Shanahan, & J. Wiebe (Eds.), *Exploring attitude and affect in text: Theories and applications* (pp. 93–107). Palo Alto, CA: Association for the Advancement of Artificial Intelligence.
- Hall, C. (2002). Passions and constraint: The marginalization of passion in liberal political theory. *Philosophy and Social Criticism*, 28, 727–748.
- Hart, R. P. (1984). *Verbal style and the presidency: A computer-based analysis*. New York, NY: Academic Press.
- Hart, R. P. (2001). Redeveloping diction: Theoretical considerations. In M. West (Ed.), *Theory, method, and practice in computer content analysis* (pp. 43–60). Westport, CT: Ablex.
- Hart, R. P. (2000a). *DICTION 5.0: The text analysis program*. Thousand Oaks, CA: Sage-Scolari.
- Hart, R. P. (2000b). *Political keywords: Using language that uses us*. New York, NY: Oxford University Press.
- Hatzivassiloglou, V., & McKeown, K. (1997). *Predicting the semantic orientation of adjectives*. Paper presented at the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain.
- Hogenraad, R. (2005). What the words of war can tell us about the risk of war. *Peace and Conflict: Journal of Peace Psychology*, 11, 137–151.
- Hogenraad, R., McKenzie, D., & Péladeau, N. (2003). Force and influence in content analysis: The production of new social knowledge. *Quality & Quantity*, 37, 221–238.
- Holsti, O. R., Brody, R. A., & North, R. C. (1964). Measuring affect and action in international reaction models: Empirical materials from the 1962 Cuban Crisis. *Journal of Peace Research*, 1(3/4), 170–190.
- Hopmann, P. T., & King, T. (1976). Interactions and perceptions in the test ban negotiations. *International Studies Quarterly*, 20, 105–142.
- Huitt, W. (2003). *The affective system: Educational psychology interactive*. Valdosta, GA: Valdosta State University.
- Ito, T. A., Larsen, J., Smith, K., & Cacioppo, J. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology*, 75, 887–900.
- Iyengar, S. (1996). Framing responsibility for political issues. *Annals of the American Academy of Political and Social Science*, 546, 59–70.
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML-98, 10th European conference on machine learning* (pp. 137–142).
- Johnson-Cartee, K. S. (2005). *News narrative and news framing: Constructing political reality*. Lanham, MD: Rowman & Littlefield.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–292.
- Kamps, J., Marx, M., Mokken, R., & de Rijke, M. (2004). *Using WordNet to measure semantic orientation of adjectives*. Paris, France: European Language Resources Association.
- Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence*, 22, 110–125.
- Kim, S., & Hovy, E. (2006). *Extracting opinions, opinion holders, and topics expressed in online news media text*. Paper presented at the Workshop on Sentiment and Subjectivity in Text, Sydney, Australia.

- Krueger, J. S., & Lewis-Beck, M. (2005). *The place of prediction in politics*. Paper presented at the annual meeting of the American Political Science Association, Washington, DC.
- Kushal, D., Lawrence, S., & Pennock, D. (2003). *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. New York, NY: Association for Computing Machinery.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.
- Lang, A., Dhillon, K., & Dong, Q. (1995). The effects of emotional arousal and valence on television viewers' cognitive capacity and memory. *Journal of Broadcasting & Electronic Media*, 39, 313–327.
- Lau, R. R., Sigelman, L., Heldman, C., & Babbitt, P. (1999). The effects of negative political advertisements: A meta-analytic assessment. *American Political Science Review*, 93, 851–875.
- Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97, 311–331.
- LeDoux, J. E. (1996). *The emotional brain*. New York, NY: Simon & Schuster.
- Leshed, G., & Kaye, J. (2006). *Understanding how bloggers feel: Recognizing affect in blog posts*. New York, NY: Association for Computing Machinery.
- Lodge, M., & Taber, C. (2000). Three steps toward a theory of motivated political reasoning. In A. Lupia, M. McCubbins, & S. Popkin (Eds.), *Elements of reason: Cognition, choice, and the bounds of rationality* (pp. 183–213). London, England: Cambridge University Press.
- Lowry, D. T. (2008). Network TV news framing of good vs. bad economic news under Democrat and Republican presidents: A lexical analysis of political bias. *Journalism & Mass Communication Quarterly*, 85, 483–498.
- Marcus, G. E. (2002). *The sentimental citizen: Emotion in democratic politics*. University Park: Pennsylvania State University Press.
- Marcus, G. E., Neuman, W., & MacKuen, M. (2000). *Affective intelligence and political judgment*. Chicago, IL: University of Chicago Press.
- Martindale, C. (1975). *Romantic progression: The psychology of literary history*. Washington, DC: Hemisphere.
- Martindale, C. (1990). *The clockwork muse: The predictability of artistic change*. New York, NY: Basic Books.
- McComas, K., & Shanahan, J. (1999). Telling stories about global climate change: Measuring the impact of narratives on issue cycles. *Communication Research*, 26, 30–57.
- Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology*, 64, 1306–1315.
- Mergenthaler, E. (2008). Resonating minds: A school-independent theoretical conception and its empirical application to psychotherapeutic processes. *Psychotherapy Research*, 18, 109–126.
- Mishne, G. (2005). *Experiments with mood classification in blog posts*. Paper presented at Style 2005, Stylistic Analysis of Text for Information Access, Salvador, Brazil.
- Mullen, A., & Collier, N. (2004). *Incorporating topic information into sentiment analysis models*. Paper presented at the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain.
- Murphy, C., Bowler, S., Burgess, C., & Johnson, M. (2006). *The rhetorical semantics of state ballot initiative arguments in California, 1980–2004*. Paper presented at the American Political Science Association conference, Philadelphia, PA.
- Nadeau, R., Niemi, R., Fan, D., & Amato, T. (1999). Elite economic forecasts, economic news, mass economic judgments, and presidential approval. *Journal of Politics*, 61, 109–135.
- Neumann, R., Marcus, G., Crigler, A., & MacKuen, M. (2007). *The affect effect: The dynamics of emotion in political thinking and behavior*. Chicago, IL: University of Chicago Press.
- Newhagen, J. E. (1998). TV images that induce anger, fear, and disgust: Effects on approach-avoidance responses and memory. *Journal of Broadcasting & Electronic Media*, 42, 265–276.

- Nussbaum, M. (2004). *Hiding from humanity: Shame, disgust and the law*. Princeton, NJ: Princeton University Press.
- Ottati, V. C., Steenbergen, R., & Riggle, E. (1992). The cognitive and affective components of political attitudes: Measuring the determinants of candidate evaluations. *Political Behavior*, 14, 423–442.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques*. Paper presented at the Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA.
- Pennebaker, J. W., Francis, M., & Booth, R. (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Mahwah, NJ: Erlbaum.
- Pennebaker, J. W., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54, 547–577.
- Pennebaker, J. W., Slatcher, R. B., & Chung, C. K. (2005). Linguistic markers of psychological state through media interviews: John Kerry and John Edwards in 2004, Al Gore in 2000. *Analyses of Social Issues and Public Policy*, 5, 197–204.
- Purpura, S., & Hillard, D. (2006). *Automated classification of congressional legislation*. Paper presented at the 7th Annual International Conference on Digital Government Research, San Diego, CA.
- Quattrone, G., & Tversky, A. (1988). Contrasting rational and psychological analyses of political choice. *American Political Science Review*, 82, 719–736.
- Quinn, K., Monroe, B., Colaresi, M., & Crespin, M. (2006). *An automated method to topic-coding legislative speech over time with application to the 105th–109th U.S. Senate*. Paper presented at the American Political Science Association conference, Philadelphia, PA.
- Roget, P. M. (1911). *Roget's thesaurus of English words and phrases* (supplemented electronic version). Salt Lake City, UT: Project Gutenberg Library Archive Foundation.
- Scharl, A., Pollach, I., & Bauer, C. (2003). Determining the semantic orientation of Web-based corpora. *Lecture Notes in Computer Science*, 2690, 840–849.
- Shenhav, S. R. (2006). Political narratives and political reality. *International Political Science Review*, 27, 245–262.
- Simon, A., & Xenos, M. (2004). Dimensional reduction of word-frequency data as a substitute for intersubjective content analysis. *Political Analysis*, 12, 63–75.
- Soroka, S. N. (2006). Good news and bad news: Asymmetric responses to economic information. *Journal of Politics*, 68, 372–385.
- Soroka, S., Bodet, M., Young, L., & Andrew, B. (2009). Campaign news and vote intentions. *Journal of Elections, Public Opinion and Parties*, 19, 359–376.
- Stone, P. J. (1986). Review of Hart, R. P. 1984 *Verbal style and the presidency: A computer-based analysis*. New York: Academic Press. *Contemporary Sociology*, 15(1), 75–77.
- Stone, P. J., Bales, R., Namenwirth, J., & Ogilvie, D. (1962). The General Inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7, 484–494.
- Stone, P. J., Dumphy, D. C., & Ogilvie, D. M. (1966). *The General Inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.
- Strapparava, C., & Valitutti, A. (2004). *WordNet-Affect: An affective extension of WordNet*. Paper presented at the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal.
- Subasic, P., & Huettner, A. (2001). Affect analysis of text using fuzzy typing. *IEEE Transactions on Fuzzy Systems*, 9, 483–496.
- Tetlock, P., Saar-Tsechansky, M., & Macskassy, S. (2007). *More than words: Quantifying language to measure firms' fundamentals*. Retrieved from <http://ssrn.com/abstract=923911>
- Thelen, M., & Riloff, E. (2002). *A bootstrapping method for learning semantic lexicons using extraction pattern contexts*. Paper presented at the Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA.

- Thomas, M., Pang, B., & Lee, L. (2006). *Get out the vote: Determining support or opposition from congressional floor-debate transcripts*. Paper presented at the Conference on Empirical Methods in Natural Language Processing, Sydney, Australia.
- Tong, R. (2001). *Detecting and tracking opinions in online discussions*. Paper presented at the Workshop on Operational Text Classification, New Orleans, LA.
- Turney, P., & Littman, M. L. (2002). *Unsupervised learning of semantic orientation from a hundred-billion-word corpus*. Ottawa, Ontario, Canada: National Research Council of Canada.
- Turney, P., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21, 315–346.
- Walzer, M. (2002). Passion and politics. *Philosophy and Social Criticism*, 28, 617–633.
- Whissell, C. (1989). The dictionary of affect in language. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory and research* (pp. 113–131). New York, NY: Harcourt Brace.
- Wiebe, J. M. (2000). *Learning subjective adjectives from corpora*. Paper presented at the 17th National Conference on Artificial Intelligence, Austin, TX.
- Wilson, T., Wiebe, J., & Hoffmann, P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis*. Paper presented at the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, British Columbia.
- Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist*, 39, 117–123.

### **Appendix: Media Content and Vote Intentions, 2006 Canadian Election, Control Variables**

	Conservatives			Liberals		
	1	2	3	1	2	3
DV <sub>t-4</sub>	.709*	.523*	.442*	.698*	.537*	.728*
	(.119)	(.126)	(.149)	(.123)	(.083)	(.106)
Strategic Council	−.410	.117	−.458	−1.580	−2.372*	−2.588*
	(.775)	(.752)	(1.205)	(.835)	(.544)	(.771)
SES	.393	1.334	.729	.633	.305	.660
	(.979)	(1.060)	(1.110)	(1.051)	(.762)	(.905)
Decima	1.150	1.196	1.947	−3.381	−2.340*	−3.173*
	(1.417)	(1.351)	(1.637)	(1.474)	(.967)	(1.252)
Ekos	2.080	2.660*	2.832*	−2.006	−2.003*	−.404
	(1.098)	(1.025)	(1.338)	(1.188)	(.767)	(1.081)
Environics	1.080	2.463	2.120	−4.305	−5.590*	.867
	(2.307)	(2.158)	(2.728)	(2.426)	(1.508)	(2.501)
Ipsos-Reid	−.741	−.356	−1.045	.424	−.413	.652
	(.920)	(.928)	(.989)	(.981)	(.669)	(.844)
Léger	−1.992	−1.169	−1.417	.154	−.896	.303
	(1.108)	(1.069)	(1.133)	(1.183)	(.745)	(1.007)
Pollara	−.444	−1.277	.213	1.956	3.276*	4.287*
	(1.746)	(1.848)	(1.871)	(1.846)	(1.309)	(1.674)
Intercept	9.888*	12.340*	18.156*	10.554	22.092*	10.896*
	(3.750)	(3.514)	(5.284)	(4.559)	(3.384)	(3.897)

*Note.* Cells contain OLS coefficients with standard errors in parentheses. \* $p < .05$ .

# Estimating Policy Positions from Political Texts

**Michael Laver** Trinity College Dublin  
**John Garry** Trinity College Dublin

The analysis of policy-based party competition will not make serious progress beyond the constraints of (a) the unitary actor assumption and (b) a static approach to analyzing party competition between elections until a method is available for deriving reliable and valid time-series estimates of the policy positions of large numbers of political actors. Retrospective estimation of these positions in past party systems will require a method for estimating policy positions from political texts.

Previous hand-coding content analysis schemes deal with policy emphasis rather than policy positions. We propose a new hand-coding scheme for policy positions, together with a new English language computer-coding scheme that is compatible with this. We apply both schemes to party manifestos from Britain and Ireland in 1992 and 1997 and cross validate the resulting estimates with those derived from quite independent expert surveys and with previous manifesto analyses.

There is a high degree of cross validation between coding methods, including computer coding. This implies that it is indeed possible to use computer-coded content analysis to derive reliable and valid estimates of policy positions from political texts. This will allow vast volumes of text to be coded, including texts generated by individuals and other internal party actors, allowing the empirical elaboration of dynamic rather than static models of party competition that move beyond the unitary actor assumption.

**D**eriving reliable and valid estimates of the policy positions of key actors is fundamental to the analysis of political competition. Various systematic methods have been used to do this, including surveys of voters, politicians, and political scientists, and the content analysis of policy documents. Each method has advantages and disadvantages but, for both theoretical and pragmatic reasons, policy documents represent a core source of information about the policy positions of political actors.

We explore various ways to extract information about policy positions from political texts. We are particularly interested in using computer-coding techniques to derive reliable and valid estimates of the policy positions of political actors. This is not mere laziness on our part, a lack of stomach for the hard graft of expert coding. If analyses of party competition are to move beyond both static models and a view of political parties as unitary actors, this requires information on the policy positions of actors inside political parties and on the development of these over time and between elections. The laborious expert "hand-coding" of text is simply not a viable method for estimating the policy positions of huge numbers of political actors, for example, all members of a legislature. Any serious attempt to operationalize a model of internal party policy competition, or of dynamic policy-based party competition or coalition government between elections, implies using computer-coding for estimating the policy positions of key political actors.

We first review existing methods for estimating policy positions from political texts. These have for the most part concentrated on the expert coding of party manifestos. We then suggest ways to improve these, dealing with both expert- and computer-coded content analysis. We then explore the impact of our suggestions upon estimates of party policy positions derived from British and Irish manifestos issued during the 1992 and 1997 general elections in each country, positions for which a range of

---

Michael Laver is a Professor of Political Science and Director of the Policy Institute, Trinity College, Dublin 2, Ireland (mlaver@tcd.ie). John Garry is a Ph.D. student of Political Science, Trinity College, Dublin 2, Ireland (garryj@tcd.ie).

Versions of this paper have been presented at the ECPR workshop on "Empirical rational choice theory," Warwick, April 1998 and the ECPR workshop on "Estimating the policy positions of political actors," Mannheim, March 1999. The authors are grateful to Ken Benoit, Ian Budge, Miranda de Vries, Matt Gabel, Daniela Giannetti, John Huber, Jan Kleinnijenhuis, Michael Marsh, Michael McDonald, Leonard Ray, and many other participants at these conferences and workshops, as well as three anonymous journal referees for their helpful and constructive comments.

*American Journal of Political Science*, Vol. 44, No. 3, July 2000, Pp. 619–634

©2000 by the Midwest Political Science Association

independent estimates are available. The results suggest that the computer coding of huge volumes of virgin text may be a viable undertaking, with obvious implications for dynamic analyses of party competition that go beyond the unitary actor assumption. We conclude with suggestions for the refinement of this approach.

## **Analysing Party Manifestos: The Story So Far**

Party manifestos are strategic documents written by politically sophisticated party elites with many different objectives in mind. This leaves considerable scope for debate about whether party manifestos reflect the "real" positions of the parties that publish them. In our view this debate is ultimately fruitless, however, since the "real" policy position of a political actor is a fundamentally elusive, even metaphysical, notion. All we can do in practice is use evidence about policy positions in particular political contexts and make context-specific inferences from this. In this sense we keep our feet on firm ground when we study official party documents published during election campaigns. As an official document, it will be difficult for party members to resile from policies in the party manifesto, while party leaders can be charged with failure to implement published manifesto pledges when given the chance to do so. Furthermore, manifestos are typically issued by each party at each election for most of the post-war period. Manifestos thus provide historical evidence of the movement of party policy positions over time. Regardless of the merits of different methods of estimating *contemporary* party policy, manifestos offer an unparalleled way to retrieve data on party policy *in the past*.

### **The Manifesto Research Group Project**

The Manifesto Research Group (MRG) is by far the biggest show on the road as a source of data on party manifestos. The MRG developed its own coding scheme and used this to analyse nearly all manifestos of nearly all political parties contesting nearly all elections in most post-war parliamentary democracies. This involved expert coders, fluent in the language concerned, reading each manifesto sentence by sentence and allocating each sentence to a category in the coding scheme. The project has been running for about 20 years and has acquired near-monopoly status in the field, for the very obvious reason that any attempt to redo such an analysis seems a truly Herculean task. The original motivation of the MRG,

however, was to operationalize a specific model of party competition, which assumes that parties compete in terms of the *salience* of particular issues in the policy package they put to voters. Whatever subsequent use has been made of their data, MRG researchers set out to measure the *relative emphasis* placed on an issue by a party in a manifesto, not the party's *substantive position* on this issue.

Position and emphasis are quite distinct parameters of party policy. Two parties may have *quite different substantive positions on the same issue, but emphasise this issue to precisely the same extent* in their respective manifestos. Recent expressions of saliency theory do assert a strong relationship between party position on, and party emphasis of, an issue—and even that "emphases equal direction" in a particularly forthright statement of the model (Budge, 1999). This, however, is acknowledged to be an empirical proposition to be tested as part of the evaluation of saliency theory. Testing the proposition, furthermore, requires independent estimates of direction and emphasis, rather than an indicator that conflates the two.

While the analytical distinction between substantive position on a policy dimension and the emphasis given to this might seem clear-cut, the situation is more complicated in practice. The great scarcity of time-series data on party policy has created a determination to squeeze the MRG data until they yield useful information on policy positions. Baron (1991), Schofield (1993), and Warwick (1994), among a wide range of authors, explored models of coalition politics using empirical policy spaces derived from the MRG data, on the clear if implicit assumption that these could be used to estimate party policy positions. Laver and Budge (1992, chapter 2) made a more explicit attempt to extract positional information from of an extensive reanalysis of the MRG data. They used a priori reasoning supplemented by exploratory factor analyses to identify clusters of closely interrelated coding categories which they felt were defined in such a way as to convey information about substantive policy positions. The raw variables making up these apparently more positional clusters of coding categories were then aggregated and used as building blocks in the construction of a general left-right scale that Laver and Budge considered to have good face validity. A different and fundamentally inductive version of this approach has recently been proposed by Gabel and Huber (2000), who do not make any a priori assumption about which policy categories are associated with left-right ideology. They use principal factors analysis on the MRG data to extract the first factor for each country. They interpret this as being, by definition, the main left-right dimension in the country concerned and derive regression scores for each manifesto on this.

The justification for this work is that the MRG data already exist and comprise a huge source of information about at least something to do with party policy. Some MRG coding categories do deal in a unipolar way with positional issues: "nationalisation," for example, or "law and order." In some of these cases emphasis may in practice imply position. Few who speak of "law and order," for example, advocate less law and order. Other MRG categories are bipolar and convey more explicitly positional information. Examples are "social services expansion: positive" and "social services expansion: negative," or "decentralisation: positive" and "decentralisation: negative." The MRG coding scheme does not systematically use bipolar categories, however. As we shall see, and as Gabel and Huber (2000) also show, the existence of some positional categories is why reanalysis of the raw MRG data does retrieve positional information on some aspects of party positions.

### The Party Change Project

The content of party manifestos was used in a quite different way by the researchers of the Harmel-Janda "Party Change Project" (PCP), explicitly designed to extract policy positions from party manifestos. The PCP defined a set of nineteen issues of interest on a priori grounds to the researchers. Manifestos were then used as follows to identify positions on each of these. "After identifying, gathering, and carefully reading all of a manifesto's passages relevant to a given issue, coders then assigned the numerical code [on a +5 to -5 scale] which, in their best judgment, best reflected the overall content of these statements." (Harmel, Janda, and Tan, 1995, 7). In effect, the PCP data generation process was like a highly structured expert survey (see below). Instead of asking many experts to locate parties in general terms on particular policy dimensions, at most three experts were given detailed coding instructions and asked to use a party's manifesto to locate it on each policy dimension.<sup>1</sup>

The data generated by this process are explicitly concerned with party policy positions and enable more valid estimates *on the nineteen policy scales under investigation* than could be constructed from the MRG data on policy emphases. However, we must be content with the nineteen scales defined by the PCP, a problem with all expert survey techniques. The PCP content analyses thus cannot be encyclopaedic descriptions of party policy, nor can they chart the rise to prominence of hitherto unimpor-

<sup>1</sup>The small number of coders and the detailed coding instructions thus make this process look much more like content analysis than an expert survey.

tant issues. In common with expert survey techniques, furthermore, the PCP judgments are more explicitly subjective than the basic coding decision of determining whether a particular sentence is in or out of a particular coding category. It seems likely that the PCP's expert coders would have found it much more difficult to separate their subjective placement of parties on scales from their prior knowledge of policy profiles of the parties concerned.<sup>2</sup> Unlike expert surveys, however, which quite explicitly rely *only* on the prior knowledge of experts, the PCP data are not the result of averaging subjective judgments across number of experts. The estimates derived from the expert surveys used below, for example, are based upon between 30 and 110 expert judgments.

### Expert Surveys

It may seem odd to include expert surveys in a review of methods for analysing party manifestos but, as the PCP illustrates, there is a continuum of techniques based upon expert judgments. At one end, the MRG used expert coders to analyse a manifesto on a sentence by sentence basis. In the middle is the PCP technique of using expert coders to identify substantive party positions at one of eleven points on each of nineteen issue dimensions, having read the manifesto as a whole.<sup>3</sup> At the other end is the expert survey technique of having experts locate parties at substantive positions on one (Castles and Mair, 1984; Huber and Inglehart, 1995) or more (Laver and Hunt, 1992) policy or ideological dimension(s), taking account of everything they think might be relevant. "Everything" presumably includes the direct and indirect impact of manifestos as well as many other things besides. However, it may also include aspects of observed behaviour (for example, coalition formation) that the data are then used to explain.

Expert surveys thus have the virtue, shared with the PCP technique, of generating unequivocally positional estimates of policy on well-defined dimensions. Since the experts are not required to study manifestos and explicitly justify every judgment they make, the expert survey technique imposes far fewer costs, allowing far more people to be consulted. The expert placement on scales "taking everything into consideration" is, however, obviously less explicit than the PCP technique. The big disadvantages of expert surveys relative to text-based coding

<sup>2</sup>Indeed, and almost paradoxically, it might be better for PCP-style analyses to use coders who were political scientists knowing little of the country concerned to read manifestos and allocate parties to scale positions, to ensure that prior knowledge of party positions did not color these judgments.

<sup>3</sup>In effect the text unit to be coded is the entire manifesto.

are, first, that text-based techniques are far more valid for the estimation of a historical party positions and, second, that a given text can typically be located at a precise time point so that a time line of cause and effect can be more confidently established.

## **Expert Coding of Text on Policy Positions**

There are two quite distinct parts of the process of estimating policy positions from political texts; some past confusions have arisen from considering both together. First, there is a process of data reduction in which a large and complex text is reduced in a reliable way to a smaller and simpler set of coded data. This can be done by either expert coders or computers and has three stages: the design of a coding scheme; the definition of a text unit to be coded; and the coding of real text units. Second, there is a data manipulation process, during which raw data are processed into variables that give valid estimates of party policy positions. Data manipulation can equally be applied to data collected using new methods and to the reanalysis of existing data such as those generated by the MRG. We begin by considering data reduction, first in relation to expert coding and then in relation to computer coding. Then we consider the estimation of policy positions from coded texts.

### **A New Expert-Coding Scheme for Party Policy Positions**

In the original MRG analysis, the coding scheme transformed a text into a set of sentence counts for fifty-four coding categories. No matter how long and complex the text, it was reduced in coded form to a case in a dataset with observations on fifty-four variables.<sup>4</sup> In the PCP analysis, data reduction used the project's coding scheme to transform the text into scores on nineteen policy scales. No matter how long and complex the text, it was reduced to a case in a dataset with observations on nineteen variables.<sup>5</sup> Given the complexity of the document being analysed and the skill and effort required from each expert coder, the type of raw data set produced by each approach is very coarse-grained. This is particularly

<sup>4</sup>The bipolar nature of many of the coding categories, alluded to above, meant that far fewer than fifty-four different policy *concerns* were in fact picked up by the MRG analysis.

<sup>5</sup>In neither project was any attention paid to the *sequence* in which references to the various coding categories appeared in the text.

the case for the PCP project, which codes data directly at the level of the scales to be estimated. There is no possibility to disaggregate these and subsequently recombine them into new policy scales. This strongly implies the need for collecting raw data using as fine-grained a coding scheme as is consistent with reliable expert coding, leaving the definition and estimation of specific policy scales explicitly to the data manipulation stage. Data generated by a fine-grained coding scheme are also far more useful for the political science community as a whole, allowing more flexible reanalysis for purposes that go beyond anything the original researchers had in mind.

A further issue is that the MRG coding scheme, as we have seen, does not consistently use bipolar coding categories. It seems to us to be axiomatic that any content analysis coding scheme designed to extract substantive information on policy *positions* should use coding categories that are at least bipolar. It is probably more useful, indeed, to ensure that all policy concerns can be coded in a *tripolar* way. This allows any mention in a manifesto to register some concern for the category involved, while all mentions can be coded into those that are pro some well-defined policy position, those that are con, and those that are neutral on it. Those whose theoretical concerns are with the emphasis attached to policy dimensions rather than positions on them can, of course, easily collapse all codings relating to a given policy category into a single variable. Those whose concerns are with policy positions cannot disaggregate data coded at the level of policy emphasis to retrieve positional information. (Indeed, to extract positional information from data that deal only with policy emphasis it is necessary to *assume* the validity of saliency theory). We therefore developed a new coding scheme for the content analysis of political texts, in which no policy category is defined without defining its antithesis, as well as a neutral position between the two. The substantive policy areas covered by the new scheme build on those of the MRG scheme but are considerably more comprehensive and fine-grained than these.<sup>6</sup>

To ensure coherence and systematic coverage of potential policy spaces, the new scheme is hierarchically structured, something that is also axiomatic in any text-coding scheme designed to extract information about policy. At the highest level in the hierarchy, we defined a set of nodes representing broad policy "domains." These are the economy, the political system, the social system, external relations, and a "general" domain that has to do with the cut and thrust of specific party competition, as

<sup>6</sup>The full version of this scheme can be accessed via <http://www.politics.tcd.ie/personnel/staff/laver.html>.

**TABLE 1** Abridged Section of Revised Manifesto Coding Scheme

1 ECONOMY
Role of state in economy
1 1 ECONOMY/+State+ Increase role of state
1 1 1 ECONOMY/+State+/Budget Budget
1 1 1 1 ECONOMY/+State+/Budget/Spending Increase public spending
1 1 1 1 1 ECONOMY/+State+/Budget/Spending/Health
1 1 1 1 2 ECONOMY/+State+/Budget/Spending/Educ. and training
1 1 1 1 3 ECONOMY/+State+/Budget/Spending/Housing
1 1 1 1 4 ECONOMY/+State+/Budget/Spending/Transport
1 1 1 1 5 ECONOMY/+State+/Budget/Spending/Infrastructure
1 1 1 1 6 ECONOMY/+State+/Budget/Spending/Welfare
1 1 1 1 7 ECONOMY/+State+/Budget/Spending/Police
1 1 1 1 8 ECONOMY/+State+/Budget/Spending/Defense
1 1 1 1 9 ECONOMY/+State+/Budget/Spending/Culture
1 1 1 2 ECONOMY/+State+/Budget/Taxes Increase taxes
1 1 1 2 1 ECONOMY/+State+/Budget/Taxes/Income
1 1 1 2 2 ECONOMY/+State+/Budget/Taxes/Payroll
1 1 1 2 3 ECONOMY/+State+/Budget/Taxes/Company
1 1 1 2 4 ECONOMY/+State+/Budget/Taxes/Sales
1 1 1 2 5 ECONOMY/+State+/Budget/Taxes/Capital
1 1 1 2 6 ECONOMY/+State+/Budget/Taxes/Capital gains
1 1 1 3 ECONOMY/+State+/Budget/Deficit Increase budget deficit
1 1 1 3 1 ECONOMY/+State+/Budget/Deficit/Borrow
1 1 1 3 2 ECONOMY/+State+/Budget/Deficit/Inflation

well as uncodable pap and waffle. Within the economic domain, the coding scheme then has four branches: to increase the role of the state in the economy; to reduce the role of the state in the economy; to be neutral on the role of the state in the economy; and to display a general concern with economic growth. Within each of the three broad policy stances on the role of the state in the economy, the coding scheme branches deal with four very general ways in which the state can intervene in the economy: the state budget, state ownership of industry and services, state regulation, and direct action by the state. Within the state budget, policy could relate to spending, taxation, or the deficit. Taxation policy can re-

late to income taxes, sales taxes, capital taxes, and so on. Table 1 shows an abridged section of part of the new scheme dealing with this area. Other policy domains are spanned hierarchically in the same systematic way.

There is no reason to regard this scheme as being fixed for all time. While deleting branches from its hierarchical structure might cause problems of comparison between newly coded documents and those coded before, adding new branches to suit particular local or temporal circumstances presents no problem at all. The beauty of an hierarchically structured coding scheme is that, if perfect comparability is required between a "parent" coding scheme and one that has been expanded, it is

always easy to collapse the expanded scheme back to its parent.<sup>7</sup>

The coding scheme we propose has over 300 categories. It is thus far more fine-grained than schemes used by either the MRG or the PCP, but even so its hierarchical structure considerably simplifies expert coding. For each text unit, coding involves a sequence of straightforward decisions. Does the text unit deal with the economy, the political system, external relations, etc.? Does it deal with the budget, ownership, or regulation? Does it deal with spending, taxation, or budget deficit? Does it deal with spending on housing, education, or health? Pre-testing did not throw up particular problems for coders using this hierarchical decision-making process, each level of which is actually more straightforward than coding into the less structured fifty-four category MRG coding scheme.

### **Text Units to Be Coded**

The MRG used manifesto "quasi sentences" as its fundamental unit of analysis. A quasi sentence is a word string that is either a complete sentence or a part sentence that could have been a complete sentence if the writer had chosen to make it so. It might seem that complete sentences should be used, since these occur unambiguously between particular punctuation marks. But this would put the analyst at the mercy of the writing style of the manifesto author(s). Comparing two manifestos, the first might appear to give more weight to some topic than the second merely because the author of the former used shorter sentences, triggering more "hits" for a text passage of the same length. The big disadvantage of this approach is that the definition of a "quasi-sentence" might itself be a source of unreliability. Accordingly, we have chosen to use words as the unit of analysis or, more precisely, word strings with an average length of ten words. There are two reasons to do this. At a practical level, it is very time-consuming to code individual words of lengthy texts without giving anything like ten times the payoff of coding word strings with an average length of ten words. At a methodological level, the coder has to read and interpret the text in context, and this is not something reliably done one word at a time.<sup>8</sup>

<sup>7</sup>In addition, provided that appropriate computer software is used to manage the coded policy documents, it is a straightforward matter to identify passages in previously analysed documents that have been coded into nodes in the scheme to which additions have been made. It is then possible to make a decision as to whether or not to recode these passages in the light of modifications to the scheme.

<sup>8</sup> Computer software for assisting expert coders presents each text unit on a separate line. Anyone who has tried to read a lengthy text

### **Coding Conventions**

The MRG coders assigned every text unit to one and only one coding category, a convention that derived from an interest in the saliences of policy concerns, which were estimated as the relative proportions of quasi-sentences assigned to each coding category. Since no text unit was coded into more than one category, these relative proportions always added up to 1.00. We follow the same convention in the present analysis to allow our results to be compared with those of the MRG, but we are not wedded to it as a general principle. Multiple coding of text units might well be appropriate in content analyses designed to extract policy positions on a range of dimensions; nothing intrinsic to our scheme precludes multiple coding.

A further important coding convention is that text units *prima facie* "neutral" on a particular policy concern are coded in context. Thus if a "neutral" text unit is embedded in a paragraph that otherwise expresses a "pro" position on some policy concern, it is coded pro. If it is embedded in a paragraph that otherwise expresses a "con" position, it is coded con.<sup>9</sup> Just as it makes no sense to code every occurrence of the word "the" as being neutral in the grounds that it conveys no information about a policy position, it makes no sense to take a string of ten words out of context and, because they convey no policy meaning as they stand, code them as being neutral.

## **Computer Coding of Text on Party Policy Positions**

### **Comparing Expert and Computer Coding**

A radical alternative to the "qualitative" expert coding of text is to use "quantitative" content analysis. Quantitative techniques use a computer to allocate text units to a coding scheme that is closely analogous to an expert-coding scheme. Expert coding uses the subjective judgment of a human coder to allocate texts units and can therefore take greater account of their substantive con-

in which every word is presented on a separate line will know it is much harder to do this than to make sense of a text with an average of ten words per line. We prepared texts for coding using a word processor to create documents with an average line length of ten words and then input them into the NUD.IST computer package that we used for storing and retrieving text, coding scheme and codings, as well as assisting and managing the coding process.

<sup>9</sup> If a neutral text unit is in a paragraph that otherwise expresses a neutral position on the policy concerns at issue, then it is obviously coded neutral. If a neutral text unit is in a paragraph at a point between a pro and a con text unit dealing with the same policy concern, then it is coded neutral.

text.<sup>10</sup> Most quantitative approaches, in contrast, allocate text units according to mechanical criteria that typically imply taking text units out of any wider political context. This is done by defining a content analysis "dictionary" of words or phrases systematically associated with particular coding categories in relevant texts. The computer then counts the number of words or phrases associated with each coding category.

It is quite possible, using quantitative techniques, to take greater account of the textual context of any unit being coded. For example, rather than including individual words, the dictionary can include phrases or more complex text strings. There is, however, a significant trade-off to be faced if this is done. First, particular phrases and word strings are likely to be repeated far less frequently than individual words in any well-written text, greatly reducing the amount of data generated by the coding process. Second, the use of given phrases and word strings is more stylistically idiosyncratic to a particular author. This raises reliability issues when relating text from different authors to the same underlying policy position. Thus, despite a longstanding gut instinct within the profession that more complex text units should be incorporated into quantitative content analysis, the reality is that much valuable information can be extracted from texts by using individual words as the fundamental unit of analysis. Any alternative faces very serious problems of its own.<sup>11</sup> Accordingly, we use individual words as our units of analysis and do not include longer word strings in our dictionary. Since, as we shall see, the level of cross-validation between our estimates and those derived from completely independent sources was high, we see no reason at this stage to move to a more complex unit of analysis.

People who come to computer-coded text analysis for the first time are often understandably sceptical. They immediately think of words that have several quite different meanings in different contexts—race, state, or class, for example. They then think of words that are often qualified by their context to have contradictory ideological meanings—taxes, spending, or services, for example. In an abstract sense this scepticism seems well justified, but closer familiarity with the technique, combined with actual patterns of word use in real texts, tends to allay at least some of these worries. There is absolutely no need to code all words in the text under investigation. Indeed

<sup>10</sup>As we will shortly argue, however, it is very difficult to confine the context taken into account by a human coder to the text being coded, as opposed to other knowledge the coder has about that text and its author.

<sup>11</sup>For an excellent review of this and many other issues in computer assisted content analysis, see Alexa (1997).

ambiguous words are typically not coded at all, and a good quantitative content analysis dictionary will consist of words with as little ambiguity as possible.<sup>12</sup> Furthermore, once we turn to real texts, there are far fewer ambiguous words than might on the face of things be expected. Of many theoretically possible meanings of a word, in practice one meaning tends to dominate in the texts analyzed. Most uses of the word "taxes" in party manifestos, for example, are in practice associated with arguments in favour of cutting taxes. Far fewer actual uses of the word taxes are found in discussions of the need to raise taxes, despite the fact that, in the abstract, both types of occurrence might seem equally likely. Thus if we assign the word "taxes" to a coding category dealing with cutting taxes, we will not always be right, but the number of times we are right will far outweigh the number of times we are wrong. Assigning the word "taxes" in this way and analysing its occurrence in party manifestos gives us valuable information about the texts under investigation.<sup>13</sup> In practice, however, most of the words used in our dictionary have a relatively unambiguous meaning.

With a well designed coding scheme and its associated dictionary, therefore, quantitative content analysis can give us a lot of information about the substantive content of texts. It does, furthermore, score over expert coding in two important respects. The first and most obvious is reliability. Computer coding is 100 percent reliable, while levels of intercoder reliability among experts, and even the intracoder reliability of the same expert coding the same text at different times, can leave a lot to be desired. Indeed, such is the cost, in terms of time and effort, of expert coding that most studies engage in very little systematic evaluation of either intercoder or intracoder reliability. Thus, while mechanically analysing words out of context may on the face of things seem to have an obvious cost in terms of the *validity* of data generated, this is offset by a very significant gain in their *reliability*.

<sup>12</sup>If the analyst feels that certain crucial words are ambiguous, then these can be "disambiguated" by an expert coder—for example, into "race#1" (as in running) and "race#2" (as in ethnicity). This requires intervention from an expert coder and thus reduces the comparative advantage of computer over expert coding. What results is, in a sense, a hybrid technique.

<sup>13</sup>It is possible to generate a probabilistic dictionary that assigns a probability that a particular word comes from a particular manifesto in a given set of calibration texts. (For an early attempt to do this using Dutch texts, see Kleinnijenhuis and Pennings [1999]). Procedures for doing this are far more complex than those we propose below, which in effect assign probabilities of 1.0 that given words are associated with particular coding categories, and there is no indication as yet that they produce better results. The future development of our proposed technique certainly does not preclude probabilistic dictionaries, however.

Second, even when we consider validity, there are important ways in which expert coding may be less valid than computer coding. An expert coder, by definition, comes to a text with prior knowledge of its context. Knowing a particular text to come from a left-wing party, for example, an expert coder might be more inclined to allocate certain text units to a left-wing coding category. The same coder might have allocated the same words in different way if he or she had known that they came from a right-wing party. Inevitably, expert coding impounds the prior opinions of the coder about the text being coded, as well as its actual content. A computer that codes words out of context comes to a text with no prior opinions—what is coded is the text and nothing else. This suggests, for example, that radical but previously unannounced shifts in policy positions might be more effectively identified in computer coding than by an expert coming to a text with certain expectations about the policy positions under investigation.

### **Designing a Quantitative Content Analysis Coding Scheme and Dictionary**

Clearly the design of a good dictionary is vital to good quantitative content analysis. Unfortunately, existing dictionaries were not well suited to the task of extracting information on substantive policy positions from political texts. We thus designed a procedure for designing an English language dictionary to do this job for us. What we present here is our first attempt at such an enterprise, certainly not the last word on the subject and one we will continue to refine. Most important, given changing political meanings of words over time and space, is the *procedure* for deriving a dictionary rather than the substantive content of any given dictionary. Moreover, as we shall argue in the concluding section, there may be a good case, since we view the enterprise as one of estimating *a priori* policy scales, for reconstructing the dictionary as frequently as we have independent data sources against which to cross-validate it. What remains constant over time is thus the dictionary generation procedure, not the actual word lists in the dictionary. Indeed changes in the word lists associated with policy positions are matters that should be of considerable substantive interest in their own right.<sup>14</sup>

We based the categories used in our dictionary on a collapsed version of the new expert-coding scheme dis-

cussed above to allow us to compare the results of computer coding with those generated by expert coders. This gave us categories for economic policy (pro- and conservative intervention in the economy) and social values (liberal and conservative). We included categories on political reform (radical and defensive), environmental policy, and law and order, as well as a number of other matters that we do not deal with here. We defined our dictionary by allocating words to these categories using a combination of *a priori* and empirical criteria. Empirically, we used the British Conservative and Labour manifestos of 1992 as a pool of key words. These were fat manifestos with lots of words, issued by parties that we expected, *a priori*, to have different policies on many issues. Before selecting any specific word for the dictionary, we looked at all words used in the two manifestos, comparing the relative frequencies with which each used all possible words and focusing on words that one party used more than twice as often as the other. We then set about allocating this subset of words to the dictionary, using a combination of *a priori* reasoning and our empirical knowledge of whether the observed frequencies identified the word in question as a "Labour" or a "Conservative" word. Thus the word "taxes" was used twenty-two times in the 1992 Conservative manifesto and only once in the 1992 Labour manifesto. We thus allocated it to the category: "reduce state involvement in the economy." To take another example of a word that in the abstract can be used in many contexts, the word "choice" appeared thirty-eight times in the 1992 Conservative manifesto and only three times in the Labour manifesto. Thus we also allocated this to the category: "reduce state involvement in the economy." We used similar criteria to allocate words to the full coding dictionary.<sup>15</sup> No word was allocated to a coding category unless it had a clear substantive meaning in terms of that category. And no word was allocated to a category unless it was used by one party twice as often as by the other. Finally, no word was allocated to more than one coding category although, as we mentioned above, we have no principled objection to multiple coding.

It is important to note that our use of the 1992 British Labour and Conservative manifestos in the definition of our dictionary means that any substantive codings of these manifestos reported below contain no new information. The coded manifestos appear very different from each other precisely because they were used in the process of defining the dictionary; indeed the dic-

<sup>14</sup>This method need not be confined to English-language dictionaries; associated researchers have recently used the same technique to derive dictionaries for coding political texts written in Dutch (de Vries, 1999), German (Garry, 1999), Italian (Giannetti, 1999), and Norwegian (Garry, 1999).

<sup>15</sup>The complete version of this dictionary can be downloaded from <http://www.politics.tcd.ie/personnel/staff/laver.html>.

tionary was defined to make them different. The codings have substantive meaning only for the “virgin” texts we analyse below.<sup>16</sup>

## Estimating Policy Positions from “Raw” Codings

How should information extracted from policy documents be used to estimate the positions of political actors? One way of doing this is largely inductive and approaches the text with no prior assumption about the substantive content of any underlying ideological dimension. This is the general style of analysis adopted by the original MRG researchers, who analysed their data by first using factor analysis to summarise patterns in the empirical frequencies of text units in different coding categories. They then used both factor loadings and the factor scores of manifestos on these dimensions to interpret them in substantive terms. A careful and persuasive development of this approach has recently been proposed by Gabel and Huber (2000), who review and compare a number of inductive methods of estimating positions on a general left-right scale, using MRG data.

The essential purpose of the Gabel-Huber method is to estimate a general left-right scale on which political actors can be placed, quite explicitly making no assumption about the policy concerns that might form part of such a scale, and quite explicitly not interpreting the resulting scale in terms of substantive policy concerns. They make no attempt to estimate party positions on any other dimension. Gabel and Huber use the entire fifty-four-category MRG dataset to find the underlying dimension that best accounts for the covariation in party positions on the policy categories. They define this as the main left-right dimension without regard to, or indeed any reporting of, its policy content. For this reason they refer to their method as the “vanilla” method. Since the vanilla method is explicitly mute on the substantive policy content of the left-right dimension it may have everything, or nothing, to do with economic policy, social values, or anything else. The key objective for Gabel and Huber is to use the MRG data to generate the best unidimensional account of covariation in party policy positions. Within this frame of reference, they operationalize and test the vanilla method against alternative methods of extracting a left-right scale from MRG data, measuring success in terms of ability to predict independent estimates of party positions on a

<sup>16</sup>Computer coding was conducted using the “tagcoder” routines in the Textpack content analysis package.

general left-right scale.<sup>17</sup> If what is required is a single, general left-right scale, then their results are very encouraging indeed, though it should never be forgotten that any given implementation of this inductive technique will be entirely dependent upon the choice of cases to analyse. Applying the inductive Gabel-Huber technique to different parties and different time periods will produce left-right scales with different substantive meanings. This means that interpretations of party movements on the scale over time must be made with very great care.

In contrast to this explicitly inductive and unidimensional approach, we prefer a more a priori, substantive, and multidimensional method for estimating party positions. We see ourselves as engaged in the estimation of party positions on the type of exogenously defined policy dimensions that have become part of the currency of “multidimensional” spatial models of party competition. These dimensions give structure and meaning to the strategic moves that parties make and typically deal with policy concerns such as economic policy, social values, foreign policy, and so on.

We thus address the problem of analysing political texts to estimate the position,  $P_C$ , of some party,  $P$ , on some policy concern  $C$ . We do this by attempting to model the process by which actors might extract policy information from political texts. Once a policy document has been read by a political actor (or coded by an analyst), a certain number of text units, possibly zero, will have been read (or coded) as dealing with policy concern  $C$ —say  $P_{C\text{tot}}$ . Of these,  $P_{C\text{pro}}$  units will have been read (or coded) as being pro some substantive position on that issue,  $P_{C\text{neut}}$  will have been read (or coded) as being neutral, and  $P_{C\text{con}}$  as being contra. Thus:

$$P_{C\text{tot}} = P_{C\text{pro}} + P_{C\text{neut}} + P_{C\text{con}}$$

The original MRG approach was to use  $P_{C\text{tot}}$ , expressed as a proportion of the total number of text units in the manifesto, as an indicator of the *emphasis* being given by the manifesto to the policy concern under consideration. This is the number that subsequent analysts have (mis)used as the basis of an estimate of the policy position of the manifesto in question. It is clear, however, that the *position* on some policy issue taken by a reader from the manifesto is much more likely to be some function of the balance of pro and contra mentions of the substantive policy concern under investigation.

<sup>17</sup>It is worth noting that some of the scales with which they compare their own do have substantive policy content and might be better evaluated in terms of ability to predict independent estimates of party policy positions on more substantive scales.

Consider the whole corpus of text in the document dealing with the policy concern under investigation. If every text unit had been coded as pro, it seems reasonable to assume that a reader—whether a member of the public, a journalist, or a party politician—would be in no doubt that the position of the party on the policy concern is pro. If every text unit had been con, it seems reasonable to assume that a reader would be in no doubt that the position is con. But what of policy concerns for which every relevant text unit in the manifesto is not coded in the same way, the situation that typically obtains in practice?

Imagine the reader comes to the text with no prior estimate of  $P_C$ , the position of the party under investigation on the policy concern at stake. (We will return in future work to the interesting possibility that the reader does in fact have some prior estimate of  $P_C$ .) The reader thus estimates the party to be neither pro nor con before reading the manifesto, but updates this estimate after the manifesto has been read. Assume that every time a reader encounters a “pro” text unit this increases the probability that she feels the party to have a pro position, while every time she encounters a “con” text unit, this increases the probability that she feels the party to have a con position. Assume that a “con” text unit negates the updating effect of a “pro” text unit and that a “neutral” text unit has no updating effect on her estimate of whether the party has a pro or a con position. In other words, all pro and con text units have information about the party’s substantive position, and the relative balance of pro and con text units is the basis of how the reader updates her estimate of this position.

In order to use this type of information to estimate a party position on some policy scale, we arbitrarily fix the endpoints of the scale. If the position of party P on concern C is unambiguously pro, let  $P_C = +1$ . If the position of party P on concern C is unambiguously con, let  $P_C = -1$ . If the position of party P on concern C is neutral, defined as being a perfect balance between pro and con, let  $P_C = 0$ . We estimate the updating effect,  $U(P_C)$ , of the manifesto on estimates of the position of party P on concern C as the relative balance of pro and con text units, taken as a proportion of all text units conveying information on this matter. Thus:

$$U(P_C) = (P_{C\text{pro}} - P_{C\text{con}}) / (P_{C\text{pro}} + P_{C\text{con}})$$

On the assumption of no prior knowledge about the party policy position, the updating effect of the manifesto is the *only* information the reader has about the position of party P on concern C. Thus:

$$\begin{aligned} P_C &= U(P_C) \\ &= (P_{C\text{pro}} - P_{C\text{con}}) / (P_{C\text{pro}} + P_{C\text{con}}) \end{aligned}$$

Fortunately, this is a number that can easily be calculated from codings of text units in any policy document. For a given policy concern, it is the number of pro text units, minus the number of con text units, as a proportion of all text units conveying information on this matter, pro or con.

### Substantive Policy Scales

We use this approach to generate two substantive policy scales from party manifestos coded using both the revised expert-coding scheme and the new computer dictionary. The first scale relates to economic policy, the second to social values. An economic left-right scale,  $\text{Econ}_{LR}$ , is defined as:

$$\text{Econ}_{LR} = (\text{Econ}_R - \text{Econ}_L) / (\text{Econ}_R + \text{Econ}_L)$$

This scale was calculated from both expert and computer codings by using the following data from each case to estimate its component parts:

$\text{Econ}_L$  = total text units in category: “increase role of state in the economy”

$\text{Econ}_R$  = total text units in category: “reduce role of state in the economy”

In order to compare our results with those that would have been generated by the original MRG coding scheme, we calculated directly analogous scales from the MRG data.<sup>18</sup> To this end, we concentrated only upon the positional clusters of coding categories identified by Laver and Budge (1992), using these clusters as the building blocks of positional policy scales. We used the Laver-Budge cluster of coding categories “state intervention” as our indicator of  $\text{Econ}_L$  and their “capitalist economics” cluster as our indicator of  $\text{Econ}_R$ .

We also defined a scale estimating party positions on a dimension dealing with liberal vs. conservative social values,  $\text{Soc}_{LC}$ :

$$\text{Soc}_{LC} = (\text{Soc}_C - \text{Soc}_L) / (\text{Soc}_C + \text{Soc}_L)$$

<sup>18</sup>This latter measure is similar in form to one independently arrived at, on the basis of a very different argument, by Kim and Fording (1998). For an evaluation of the relative reliability of ratio-based scales such as this and subtractive scales such as that proposed by Laver and Budge (1992), when applied to the MRG dataset, see McDonald and Mendes (1999).

The following data were used to estimate the component parts of this scale:

$Soc_L$  = total text units in category: "liberal, permissive, or nontraditional social values"

$Soc_C$  = total text units in category: "conservative, restrictive, or traditional social values"

In each case it is important to note that we did not include manifesto codings relating to crime or law and order in our social values scale, since we felt on a priori grounds that these tend to cut across other social values and are thus better treated separately. We used a similar approach in calculating a liberal-conservative social values scale from the original MRG data. We used the Laver-Budge "social conservatism" cluster of coding categories as our indicator of  $Soc_C$  and the "anti-establishment" cluster as our indicator of  $Soc_L$ . The approach we used for deriving policy scales from political texts was thus applied in an identical manner to both the original MRG data and the recoded manifestos.

## Results

Estimates of policy positions based on coding British and Irish party manifestos from the 1992 and 1997 elections in the different ways described above were derived as follows. First, we used expert coders<sup>19</sup> and the computer coding software<sup>20</sup> to calculate the raw counts for expert and computer codings of the manifestos in the aggregate coding categories used to build the "economic left-right" and "liberal-conservative values" scales. Second, we used these data to construct "raw" policy scales for 1992 and 1997, as described in the previous section. We also constructed equivalent raw scales based on the original MRG data for 1992 in Britain and Ireland. All of the scales derived from content analyses are constructed on the same basis from the same text information and can be compared directly with each other. However, our completely independent source of cross-validation for the text-based scales came from expert surveys for Britain in 1989 and 1997 (Laver and Hunt, 1992; Laver, 1998a) and Ireland in

<sup>19</sup>Two coders independently coded the documents. The coders met after completing their independent coding to discuss each text unit on which they disagreed and arrive at an agreed coding.

<sup>20</sup>The tagcoder routines in Textpack.

1992 and 1997 (Laver, 1994; Laver 1998b).<sup>21</sup> The raw scales derived from expert surveys are constructed on a different basis, using quite different information. To allow all scales to be compared directly, each was standardised across the full set of sixteen observations for each scale.<sup>22</sup>

These results are presented in Table 2. A systematic summary of the interrelationship between the various ways of estimating economic policy positions can be found in Table 3, which reports Pearson correlations between the economic policy scores produced by each technique for each party in each election.

Tables 2 and 3 show very encouraging cross-validation between the various economic left-right scales under investigation. The scales based upon expert-coded content analysis, under either the MRG or the revised coding scheme, are very close to each other and to scales derived completely independently from expert surveys. Correlations between pairs of these scales, for both the 1992 and 1997 elections, range from 0.94 to 0.99. The left-right scales based upon computer coding generate very similar positions to those of the other scales.<sup>23</sup> Correlations between computer-coded and other scales range from 0.72 to 0.94.<sup>24</sup>

For the British parties, the correspondence is very good, particularly given the need to treat with great caution the computer estimates for Labour and the Conservatives for 1992, since these manifestos were used to generate the computer dictionary, and will thus "artificially" separate these manifestos.

In the case of Ireland for 1997, the expert survey, revised expert coding, and computer coding all put Democratic Left (DL) firmly on the left, Labour on the centre left and the Progressive Democrats (PDs) on the right. All techniques place Fianna Fáil (FF) and Fine Gael (FG) between Labour and the PDs. The expert survey finds almost no difference between FF and FG. The revised expert coding puts FF to the left of FG. The computer coding places FG towards the centre-left, alongside its coalition partners DL and Labour. In 1992, all techniques have DL and the PDs respectively anchoring left and

<sup>21</sup>No expert judgements are available for Britain in 1992.

<sup>22</sup>That is, eight parties in two elections.

<sup>23</sup>It is important to remember that the 1992 computer-generated positions for the British Labour and Conservative parties should be treated with great care, since these manifestos were used to generate the computer-coding dictionary in the first place. Differences between these manifestos, therefore, were assumed a priori rather than inferred from the data.

<sup>24</sup>Gabel and Huber (2000, footnote 12) regard correlations of the order of 0.88 and 0.94 as indications that policy scales are "measuring the same thing."

**TABLE 2** Standardized Economic "Left-Right" and Social Values "Liberal-Conservative" Scores for 1992 and 1997 British and Irish Party Manifestos and Standardized Scores on Comparable Expert Survey

	Economic Policy				Social Policy			
	Computer	Revised Expert	MRG	Expert Survey	Computer	Revised Expert	MRG	Expert Survey
UK Lab 1992	-1.52	-0.84	-0.99	-1.18	-1.75	-0.02	0.19	-0.69
UK LD 1992	-0.15	-0.68	-0.22	-0.57	-1.19	-1.34	-1.20	-0.61
UK Con1992	2.28	1.34	1.06	1.35	0.96	1.03	1.11	1.44
UK Lab 1997	0.38	0.10		-0.12	0.21	0.98		-0.26
UK LD 1997	-0.38	-0.41		-1.09	-0.88	-0.28		-0.63
UK Con1997	0.81	1.45		0.89	0.96	1.76		0.94
Irl DL 1992	-0.95	-1.34	-1.30	-1.36	-0.90	-1.29	-1.63	-1.42
Irl Lab 1992	0.07	-0.77	-0.93	-0.85	-0.80	-0.74	0.56	-0.82
Irl FF 1992	-0.79	-0.13	0.37	0.48	1.11	0.88	1.11	1.96
Irl FG 1992	0.10	0.34	0.72	0.88	1.58	0.21	0.08	1.05
Irl PD 1992	0.82	0.64	1.28	1.44	0.82	-0.69	-0.23	0.01
Irl DL 1997	-1.38	-1.38		-1.15	-0.79	-1.48		-1.07
Irl Lab 1997	-0.68	-0.88		-0.66	-0.74	-0.75		-0.68
Irl FF 1997	0.22	0.04		0.26	0.50	-0.01		1.01
Irl FG 1997	-0.19	0.56		0.31	0.15	1.19		0.35
Irl PD 1997	1.35	1.96		1.37	0.75	0.54		-0.59

**TABLE 3** Pearson Correlations between Alternative Estimates of Economic Left-Right Scale Positions, Britain and Ireland 1992–97

	Computer Codings	Revised Expert Codings	Original MRG Codings	Expert Surveys
<b>1992</b>				
Computer codings	1.00			
Revised expert codings	0.85	1.00		
Original MRG codings	0.72	0.94	1.00	
Expert surveys	0.75	0.95	0.99	1.00
<b>1997</b>				
Computer codings	1.00			
Revised expert codings	0.94	1.00		
Expert surveys	0.91	0.95	n.a.	1.00

right, but computer coding of FF does appear deviant, placing it to the left of Labour while the other techniques place it to the right.

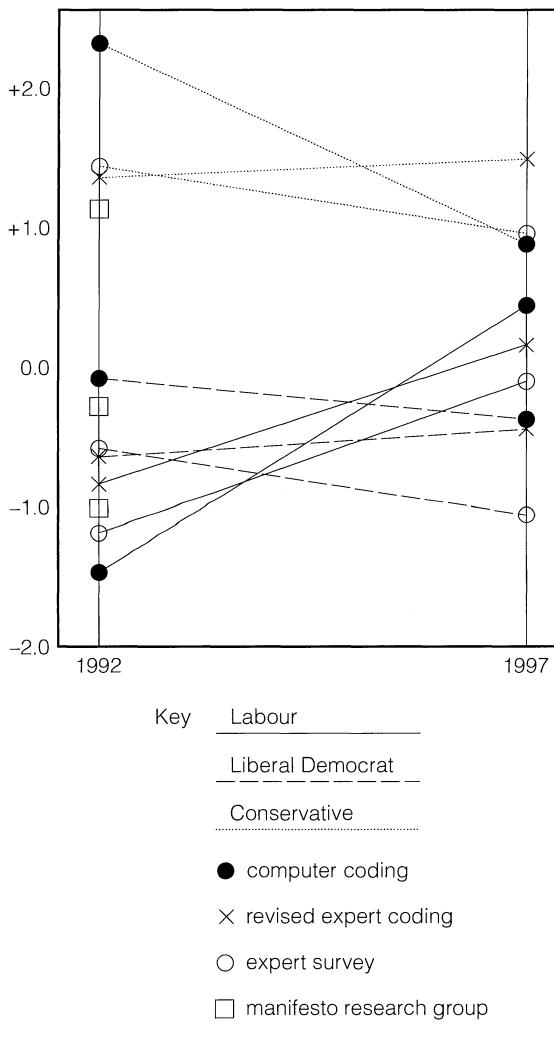
The most striking feature of these results, however, and a clear-cut test of the face validity of the computer coding technique, concerns the widespread informal perception that the British Labour Party shifted sharply towards the centre of the economic policy spectrum in 1997. This is most clearly shown in Figure 1. The expert survey of British party policy positions in 1997 showed Labour making a major move to the centre, with the Liberal Democrats shifting somewhat towards the left. The net result was that the Labour Party was, in 1997, placed by the experts between the Liberal Democrats and the Conservatives on economic policy, rather than to the left of the other two parties as previously.

Table 2 and Figure 1 show that the computer-generated scales did indeed pick up this important shift in British party policy positions. The 1989 and 1997 expert estimates of Labour Party policy are very closely mirrored by the independent computer-generated estimates. The heavy lines in Figure 1 show that the rightwards shift in Labour policy is picked up very clearly by all techniques. The expert surveys and computer coded content analysis also imply a leftwards shift between 1992 and 1997 in the economic policies of both the Liberal Democrats and the Conservatives. The scales based upon expert-coded content analysis of manifestos are more equivocal about this, implying that the Conservatives and Liberal Democrats had more or less remained in the same place.

Whichever technique is used, however, the reversal of the positions of Labour and the Liberal Democrats shows up in a very striking way. Considering that one technique involves averaging the subjective judgments of political scientists, another involves the analysis of party manifestos by expert coders, while another involves the computer counting of key words, the techniques correspond to a remarkable degree.

Turning to party positions on social values, all scales under investigation are in broad agreement for the British parties. The main anomaly is that both the MRG and revised expert manifesto codings place the Liberal Democrats firmly on the liberal side of Labour in 1992, in contrast to the centre-liberal position generated by both the expert judgments and computer-generated scales. A similar pattern is found in Ireland—there is broad agreement between most scales on the positions of all parties except the PDs. The PDs are estimated as a liberal party on social values by the expert text codings, as a centre party on this dimension by the expert judgments, and as a conservative party by the computer-generated

**FIGURE 1** Standardized Expert Survey, Computer Coded and Expert Coded Estimates of Party Policy Positions in Britain 1992–97



scales. The MRG codings, furthermore, identify Labour as having conservative social values, a result that does not on the face of things seem plausible.

A systematic summary of the interrelationship between the various ways of estimating social policy positions can be found in Table 4, which reports Pearson correlations between the various social policy scores. The correlations between the expert survey results and the various manifesto-based estimates are somewhat lower than those for economic policy, but are very respectable nonetheless, ranging from 0.67 to 0.88. For the election

**TABLE 4 Pearson Correlations between Alternative Estimates of Liberal-Conservative Social Values Scale Positions, Britain and Ireland 1992-97**

	Computer Codings	Revised Expert Codings	Original MRG Codings	Expert Surveys
1992				
Computer codings	1.00			
Revised expert codings	0.62	1.00		
Original MRG codings	0.49	0.87	1.00	
Expert surveys	0.84	0.88	0.74	1.00
1997				
Computer codings	1.00			
Revised expert codings	0.80	1.00		
Expert surveys	0.71	0.67	n.a	1.00

for which we have MRG data (1992) it is the MRG codings that are the least well correlated with the expert survey estimates.

Perhaps the most striking overall finding from Tables 2-4 is that there is a high degree of cross-validation between the three quite independent techniques for generating data on party policy. Expert-coded content analysis, computer-coded content analysis and expert surveys all produce consistent results once an explicitly positional approach is used to defining the scales with which to estimate policy positions. No technique stands out as producing deviant results, and all techniques seem very sensitive to the striking shift in the policy position of the British Labour Party that we should expect any valid technique to pick up.

This conclusion also applies the MRG data on economic policy positions, once these data are reanalysed using the new substantive policy scales, and thus offers hope that careful reanalysis of the MRG data on policy *emphases* might yield useful estimates of party *positions* on an economic left-right policy scale. This almost certainly arises from the more explicitly positional nature of at least some of the MRG coding categories relating to economic policy and offers the prospect of putting an existing huge data set to better use than hitherto. The social policy scale generated from the 1992 MRG data was, however, the most deviant. Here we see the impact of the MRG's saliency approach most clearly. While the MRG scheme did cover economic policy from many different angles, it did not code *positions*, as opposed to *emphases*, on a wide range of other matters, including social values. The revised schemes proposed in this paper do set out to do this, and the payoffs become apparent for the main noneconomic policy dimension we consider.

## The Way Forward

Three headline conclusions can be drawn from these results. The first, and for us the most exciting, is that even a very simple form of computer-coded content analysis, one that can be used right now, can generate estimates of policy positions that can be cross-validated against quite independent sources. The second is that parts of the MRG saliency data can be reanalysed using *a priori* positional scales to derive estimates of policy positions that can also be cross-validated. The third is that the new explicitly positional expert-coding scheme works well and may offer clear advantages in noneconomic policy areas for which the original MRG scheme offers fewer quasi-positional categories.

The latter conclusions are important because they imply that the MRG data, in which the profession has already made a huge investment over many years, do represent a valuable source of information on certain substantive policy positions, provided that care is taken to derive estimates in appropriate ways. The *a priori* and substantive approach to building scales used here is based on assumptions about how real people might extract information from real manifestos. Reanalysing the MRG data in this way, we have come close to quite independent estimates of substantive party policy positions derived from expert judgments. Our findings thus echo, for more substantive policy scales, those of Gabel and Huber (2000) concerning a single general left-right dimension.

Real-world research resources are always scarce. These findings imply that, while the new expert-coding scheme we propose offers a methodologically more appropriate way to estimate substantive policy positions, the huge costs of hand-recoding the entire corpus of the

MRG dataset may not justify the benefits. Very serious questions do have to be asked, however, in respect to future work. It does not seem to us to be appropriate to continue the laborious and costly process of hand-coding manifestos using *only* a substantially nonpositional coding scheme designed to estimate the relative salience of different policy concerns.

By far the most radical conclusion to be drawn from the present analysis, however, concerns the way in which computer codings of the manifestos can be used to generate estimates of party positions that are very similar to those estimated using much more resource-hungry expert techniques. This suggests that it may be possible to refine computer-coding techniques to streamline the coding of the vast amount of policy-relevant text now available in machine-readable form. This in turn opens up the possibility of using computer coding to generate systematic estimates of the policy positions of the factions, even the individual members, of political parties, and to chart the development of these positions over time. Such estimates simply do not exist at present, and there is no realistic way they will be generated unless computer-coding techniques, or some reliable and valid working alternative, can be developed. The effective computer coding of political texts will thus represent a significant breakthrough for those concerned with analysing party competition. It will allow the operationalization of models that go beyond the unitary actor assumption, as well as of those that deal with movements in substantive policy positions between elections.

We are acutely aware that we are at an early stage in this research programme. We feel the best way to proceed, since the ideological meaning of key words in the political lexicon clearly changes over space and time, is to formalise the general procedure that we have reported here for generating a computer dictionary. First, party election manifestos can be expert coded using an appropriate positional coding scheme. Scales generated from these codings can also be cross-validated against expert surveys conducted during the same period, although this is not central to the technique. Second, a computer dictionary of key words can be derived from the same manifestos, with words allocated to coding categories in a way that enables computer coding to replicate expert codings as closely as possible. This means, of course, that no new information about *manifestos* will be derived from the computer coding of them. Rather, the computer coding of manifestos for which independent positional estimates are available would be used to "calibrate" the dictionary of key words, adjusting this for changes in the political lexicon between elections.

Having calibrated the computer dictionary in this way for a particular election, this can then be used to computer code vast volumes of "virgin" text in quantities that would be quite beyond the resources of any expert-coding technique. For example, policy statements issued by all members of parliament could be computer-coded to enable the researcher to draw a detailed ideological map of the internal policy spaces of political parties and look at movements in these over time. Alternatively, new party policy statements could be coded in an attempt to track shifts of policy between elections. This process represents the type of interaction between more qualitative expert-coding techniques and more quantitative computer-coding procedures that is now regarded as best practice in the analysis of political texts (Alexa, 1997). The expert coding of manifestos keeps the computer dictionary in touch with realities that may change across space and time, the computer-coding procedure then allows otherwise unmanageable volumes of text to be processed.

An obviously important issue concerns the cross-national extension of this technique, especially to non-English language texts. Early indications in this regard are most encouraging. Researchers associated with this programme have generated computer dictionaries in Dutch (de Vries 1999), German (Garry, 1999), Italian (Giannetti 1999), and Norwegian (Garry, 1999), as well as analysing English language government declarations (Mansergh, 1999). Preliminary findings suggest that computer-generated estimates of party positions on economic policy can be cross-validated against independent estimates at levels of correlation very similar to those reported in Tables 3 and 4 above. These results imply that the technique we propose is relatively robust to changes in political context and does have the potential to be used in cross-national research.

The successful development of this technique could thus have significant implications for the systematic empirical analysis of intraparty politics in particular and party competition in general. The possibilities are both enormous and exciting, once we have at our disposal a technique for the fast, valid, and reliable coding of huge volumes of text containing information on the policy positions of political actors. It will allow empirical accounts of party politics to go beyond the unitary actor assumption. It will also allow them to go beyond static models of party competition that do not recognise any interesting political activity between elections to dynamic models of continuous movement over time, movement driven by an intraparty political game. If the computer coding of political texts can be refined,

applied, and accepted within the profession, the prize is simply enormous.

*Manuscript submitted December 18, 1998.  
Final manuscript received July 19, 1999.*

## References

- Alexa, Melina. 1997. "Computer Assisted Text Analysis Methodology in the Social Sciences." *ZUMA Arbeitsbericht*. 97/07. Mannheim: ZUMA
- Baron, David. 1991. "A Spatial Bargaining Theory of Government Formation in Parliamentary Systems." *American Political Science Review* 85:137–164.
- Budge, Ian. 1999. "Estimating Party Policy Positions: From Ad Hoc Measures to Theoretically Validated Standards." Presented to the workshop "Estimating the Policy Positions of Political Actors." ECPR Joint Sessions, Mannheim.
- Castles, Francis, and Peter Mair. 1984. "Left-Right Political Scales: Some Expert Judgments." *European Journal of Political Research*. 12:73–88.
- DeVries, Miranda. 1999. "Computer-Based Content Analysis of the Party Manifestos of the 1998 Dutch Elections." Presented at the workshop on "Estimating the Policy Positions of Political Actors." ECPR Joint Sessions, Mannheim.
- Gabel, Matt, and John Huber. 2000. "Putting Parties in Their Place: Inferring Party Left-Right Ideological Positions from Party Manifestos Data." *American Journal of Political Science* 44: 94–103.
- Garry, John. 1999. "Computer Coded Content Analysis Of German and Norwegian Manifestos." Presented at the workshop on "Estimating the Policy Positions of Political Actors." ECPR Joint Sessions, Mannheim.
- Giannetti, Daniela. 1999. "The Computer Coding of Italian Political Texts." Presented at the workshop on "Estimating the Policy Positions of Political Actors." ECPR Joint Sessions, Mannheim.
- Harmel, Robert, Kenneth Janda, and Alexander Tan. 1995. "Substance vs. Packaging: An Empirical Analysis of Parties' Issue Profiles." Presented at the Annual Meeting of the American Political Science Association.
- Huber, John, and Ronald Inglehart. 1995. "Expert Interpretations of Party Space and Party Locations in 42 Societies." *Party Politics* 1:73–111.
- Kim, Heemin, and Richard Fording. 1998. "Voter Ideology in Western Democracies, 1946–1989." *European Journal of Political Research* 33:1, 73–97
- Kleinnijenhuis, Jan, and Paul Pennings. 1999. "Measurement of Policy Positions On The Basis Of Party Programmes, Media Coverage and Voter Perceptions." Presented at the workshop on "Estimating the Policy Positions of Political Actors." ECPR Joint Sessions, Mannheim.
- Laver, Michael. 1994. "Party Policy and Cabinet Portfolios in Ireland 1992: Results from an Expert Survey." *Irish Political Studies* 9:157–164.
- Laver, Michael. 1998a. "Party Policy in Britain, 1997: Results from an Expert Survey." *Political Studies* 46:336–347.
- Laver, Michael. 1998b. "Party Policy in Ireland, 1997: Results from an Expert Survey." *Irish Political Studies* 13:159–171.
- Laver, Michael, and Ian Budge (ed.). 1992. *Party Policy and Government Coalitions*. London: St. Martin's Press.
- Laver, Michael, and W. Ben Hunt. 1992. *Policy and Party Competition*. New York: Routledge.
- Mansergh, Lucy. 1999. "Using Computer and Hand Coding to Compare the Policy Positions of Government Declarations and Parties in Government." Presented at the workshop on "Estimating the Policy Positions of Political Actors." ECPR Joint Sessions, Mannheim.
- McDonald, Michael, and Silvia Mendes. 1999. "The Policy Space of Party Manifestos." Presented at the workshop on "Estimating the Policy Positions of Political Actors." ECPR Joint Sessions, Mannheim.
- Schofield, Norman. 1993. "Political Competition and Multi-party Coalition Governments." *European Journal of Political Research*, 23:1–35.
- Warwick, Paul. 1994. *Government Survival in Parliamentary Democracies*. Cambridge: Cambridge University Press.

# Méthodes supervisées



## Extracting Policy Positions from Political Texts Using Words as Data

MICHAEL LAVER and KENNETH BENOIT *Trinity College, University of Dublin*  
 JOHN GARRY *University of Reading*

We present a new way of extracting policy positions from political texts that treats texts not as discourses to be understood and interpreted but rather, as data in the form of words. We compare this approach to previous methods of text analysis and use it to replicate published estimates of the policy positions of political parties in Britain and Ireland, on both economic and social policy dimensions. We "export" the method to a non-English-language environment, analyzing the policy positions of German parties, including the PDS as it entered the former West German party system. Finally, we extend its application beyond the analysis of party manifestos, to the estimation of political positions from legislative speeches. Our "language-blind" word scoring technique successfully replicates published policy estimates without the substantial costs of time and labor that these require. Furthermore, unlike in any previous method for extracting policy positions from political texts, we provide uncertainty measures for our estimates, allowing analysts to make informed judgments of the extent to which differences between two estimated policy positions can be viewed as significant or merely as products of measurement error.

Analyses of many forms of political competition, from a wide range of theoretical perspectives, require systematic information on the policy positions of the key political actors. This information can be derived from a number of sources, including mass, elite, and expert surveys either of the actors themselves or of others who observe them, as well as analyses of behavior in strategic settings, such as legislative roll-call voting. (For reviews of alternative sources of data on party positions, see Laver and Garry 2000 and Laver and Schofield 1998). All of these methods present serious methodological and practical problems. Methodological problems with roll-call analysis and expert surveys concern the direction of causality—"data" on policy positions collected using these techniques are arguably more a product of the political processes under investigation than causally prior to them. Meanwhile, even avid devotees of survey techniques cannot rewind history to conduct new surveys in the past. This vastly restricts the range of cases for which survey methods can be used to estimate the policy positions of key political actors.

An alternative way to locate the policy positions of political actors is to analyze the texts they generate. Political texts are the concrete by-product of strategic political activity and have a widely recognized potential to reveal important information about the policy positions of their authors. Moreover, they can be analyzed, reanalyzed, and reanalyzed again without becoming jaded or uncooperative. Once a text and an

analysis technique are placed in the public domain, furthermore, others can replicate, modify, and improve the estimates involved or can produce completely new analyses using the same tools. Above all, in a world where vast volumes of text are easily, cheaply, and almost instantly available, the systematic analysis of political text has the potential to be immensely liberating for the researcher. Anyone who cares to do so can analyze political texts for a wide range of purposes, using historical texts as well as analyzing material generated earlier in the same day. The texts analyzed can relate to collectivities such as governments or political parties or to individuals such as activists, commentators, candidates, judges, legislators, or cabinet ministers. The data generated from these texts can be used in empirical elaborations of any of the huge number of models that deal with the policies or motivations of political actors. The big obstacle to this process of liberation, however, is that current techniques of systematic text analysis are very resource intensive, typically involving large amounts of highly skilled labor.

One current approach to text analysis is the "hand-coding" of texts using traditional—and highly labor-intensive—techniques of content analysis. For example, an important text-based data resource for political science was generated by the Comparative Manifestos Project (CMP)<sup>1</sup> (Budge, Robertson, and Hearl 1987; Budge et al. 2001; Klingemann, Hofferbert, and Budge 1994; Laver and Budge 1992). This project has been in operation since 1979 and, by the turn of the millennium, had used trained human coders to code 2,347 party manifestos issued by 632 different parties in 52 countries over the postwar era (Volkens 2001, 35). These data have been used by many authors writing on a wide range of subjects in the world's most prestigious journals.<sup>2</sup> Given the immense sunk costs of

Michael Laver's work on this paper was carried out while he was a Government of Ireland Senior Research Fellow in Political Science, Trinity College, University of Dublin, Dublin, Ireland (mlaver@tcd.ie).

Kenneth Benoit's work on this paper was completed while he was a Government of Ireland Research Fellow in Political Science, Trinity College, University of Dublin, Dublin, Ireland (kbenoit@tcd.ie).

John Garry is Lecturer in the Politics Department, University of Reading, White Knights Reading, Berkshire RG6 6AH, UK (j.a.garry@reading.ac.uk).

We thank Raj Chari, Gary King, Michael McDonald, Gail McElroy, and three anonymous reviewers for comments on drafts of this paper.

<sup>1</sup> Formerly the Manifesto Research Group (MRG).

<sup>2</sup> For a sample of such publications, see Adams 2001; Baron 1991, 1993; Blais, Blake, and Dion 1993; Gabel and Huber 2000; Kim and Fording 1998; Schofield and Parks 2000; and Warwick 1994, 2001, 2002.

generating this mammoth data set by hand over a period of more than 20 years, it is easy to see why no other research team has been willing to go behind the very distinctive theoretical assumptions that structure the CMP coding scheme or to take on the task of checking or replicating any of the data.

A second approach to text analysis replaces the hand-coding of texts with computerized coding schemes. Traditional computer-coded content analysis, however, is simply a direct attempt to reproduce the hand-coding of texts, using computer algorithms to match texts to coding dictionaries. With proper dictionaries linking specific words or phrases to predetermined policy positions, traditional techniques for the computer-coding of texts can produce estimates of policy positions that have a high cross-validity when measured against hand-coded content analyses of the same texts, as well as against completely independent data sources (Bara 2001; de Vries, Giannetti, and Mansergh 2001; Kleinnijenhuis and Pennings 2001; Laver and Garry 2000). Paradoxically, however, this approach does not dispense with the need for heavy human input, given the extensive effort needed to develop and test coding dictionaries that are sensitive to the strategic context—both substantive and temporal—of the texts analyzed. Since the generation of a well-crafted coding dictionary appropriate for a particular application is so costly in time and effort, the temptation is to go for large general-purpose dictionaries, which can be quite insensitive to context. Furthermore, heavy human involvement in the generation of coding dictionaries imports some of the methodological disadvantages of traditional techniques based on potentially biased human coders.

Our technique breaks radically from “traditional” techniques of textual content analysis by treating texts not as discourses to be read, understood, and interpreted for meaning—either by a human coder or by a computer program applying a dictionary—but as collections of word data containing information about the position of the texts’ authors on predefined policy dimensions. Given a set of texts about which something is known, our technique extracts data from these in the form of word frequencies and uses this information to estimate the policy positions of texts about which nothing is known. Because it treats words unequivocally as data, our technique not only allows us to estimate policy positions from political texts written in any language but also, uniquely among the methods currently available, allows us to calculate confidence intervals around these point estimates. This in turn allows us to make judgments about whether estimated differences between texts have substantive significance or are merely the result of measurement error. Our method of using words as data also removes the necessity for heavy human intervention and can be implemented quickly and easily using simple computer software that we have made publicly available.

Having described the technique we propose, we set out to cross-validate the policy estimates it generates against existing published results. To do this we reanalyze the text data set used by Laver and Garry

(2000) in their dictionary-based computer-coded content analysis of the manifestos of British and Irish political parties at the times of the 1992 and 1997 elections in each country. We do this to compare our results with published estimates of the policy positions of the authors of these texts generated by dictionary-based computer-coding, hand-coded content analyses, and completely independent expert surveys. Having gained some reassurance from this cross-validation, we go on to apply the technique to additional texts not written in English. Indeed estimating policy positions from documents written in languages unknown to the analyst is a core objective of our approach, which uses computers to minimize human intervention by analyzing text as data, while making no human judgement call about word meanings. Finally, we go on to extend the application of our technique beyond the analysis of party manifestos, to the estimation of legislator positions from parliamentary speeches. If our method can be demonstrated to work well in these various contexts, then we would regard it as an important methodological advance for studies requiring estimates of the policy positions of political actors.

## A MODEL FOR LOCATING POLITICAL TEXTS ON A *PRIORI* POLICY DIMENSIONS

### *A Priori* or Inductive Analyses of Policy Positions?

Two contrasting approaches can be used to estimate the policy positions of political actors. The first sets out to estimate positions on policy dimensions that are defined *a priori*. A familiar example of this approach can be found in expert surveys, which offer policy scales with predetermined meanings to country experts who are asked to locate parties on them (Castles and Mair 1984; Laver and Hunt 1989). Most national election and social surveys also ask respondents to locate both themselves and political parties on predefined scales. Within the realm of text analysis, this approach codes the texts under investigation in a way that allows the estimation of their positions on *a priori* policy dimensions. A recent example of this way of doing things can be seen in the dictionary-based computer-coding technique applied by Laver and Garry (2000), which applies a predefined dictionary to each word in a political text, yielding estimated positions on predefined policy dimensions.

An alternative approach is fundamentally *inductive*. Using content analysis, for example, observed patterns in texts can be used to generate a matrix of similarities and dissimilarities between the texts under investigation. This matrix is then used in some form of dimensional analysis to provide a spatial representation of the texts. The analyst then provides substantive meanings for the underlying policy dimensions of this derived space, and these *a posteriori* dimensions form the basis of subsequent interpretations of policy positions. This is the approach used by the CMP in its hand-coded content analysis of postwar European party manifestos (Budge, Robertson, and Hearl 1987), in which data

analysis is designed to allow inferences to be made about the dimensionality of policy spaces and the substantive meaning of policy dimensions. A forthright recent use of this approach for a single left-right dimension can be found in Gabel and Huber 2000. Warwick (2002) reports a multidimensional inductive analysis of both content analysis and expert survey data.

It should be noted that a *purely* inductive spatial analysis of the policy positions of political texts is impossible. The analyst has no way of interpreting the derived spaces without imposing at least some *a priori* assumptions about their dimensionality and the substantive meaning of the underlying policy dimensions, whether doing this explicitly or implicitly. In this sense, all spatial analyses boil down to the estimation of policy positions on *a priori* policy dimensions. The crucial distinction between the two approaches concerns the point at which the analyst makes the substantive assumptions that allow policy spaces to be interpreted in terms of the real world of politics. What we have called the *a priori* approach makes these assumptions at the outset since the analyst does not regard either the dimensionality of the policy space or the substantive meaning of key policy dimensions as the essential research questions. Using prior knowledge or assumptions about these reduces the problem to an epistemologically straightforward matter of estimating unknown positions on known scales. What we have called the inductive approach does not make prior assumptions about the dimensionality of the space and the meaning of its underlying policy dimensions. This leaves too many degrees of freedom to bring closure to the analysis without making *a posteriori* assumptions that enable the estimated space and its dimensions to be interpreted.

The ultimate methodological price to be paid for the benefits of *a posteriori* interpretation is the lack of any objective criterion for deciding between rival spatial interpretations, in situations in which the precise choice of interpretation can be critical to the purpose at hand. The price for taking the *a priori* route, on the other hand, is the need to accept take-it-or-leave-it propositions about the number and substantive meaning of the policy dimensions under investigation. Using the *a priori* method we introduce here, however, this price can be drastically reduced. This is because, once texts have been processed, it is very easy to reestimate their positions on a new *a priori* dimension in which the analyst might be interested. For this reason we concentrate here on estimating positions on *a priori* policy dimensions. The approach we propose can be adapted for inductive analysis with *a posteriori* interpretation, however, and we intend to return to this in future work.

### **The Essence of Our *A Priori* Approach**

Our approach can be summarized in nontechnical terms as a way of estimating policy positions by comparing two sets of political texts. On one hand is a set of texts whose policy positions on well-defined

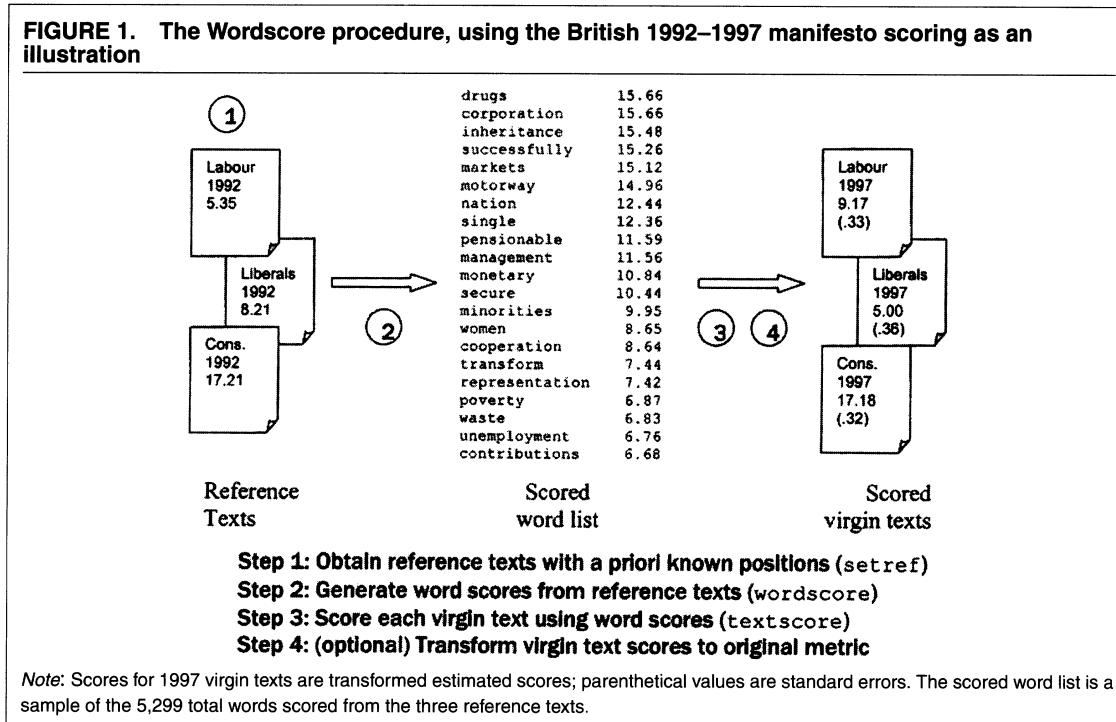
*a priori* dimensions are “known” to the analyst, in the sense that these can be either estimated with confidence from independent sources or assumed uncontroversially. We call these “reference” texts. On the other hand is a set of texts whose policy positions we do not know but want to find out. We call these “virgin” texts. All we do know about the virgin texts is the words we find in them, which we compare to the words we have observed in reference texts with “known” policy positions.

More specifically, we use the relative frequencies we observe for each of the different words in each of the reference texts to calculate the probability that we are reading a particular reference text, given that we are reading a particular word. For a particular *a priori* policy dimension, this allows us to generate a numerical “score” for each word. This score is the expected policy position of any text, given only that we are reading the single word in question. Scoring words in this way replaces the predefined deterministic coding dictionary of traditional computer-coding techniques. It gives words policy scores, not having determined or even considered their meanings in advance but, instead, by treating words purely as data associated with a set of reference texts whose policy positions can be confidently estimated or assumed. In this sense the set of real-world reference texts replaces the “artificial” coding dictionary used by traditional computer-coding techniques.

The value of the set of word scores we generate in this way is not that they tell us anything new about the reference texts with which we are already familiar—indeed they are no more than a particular type of summary of the word data in these texts. Our main research interest is in the virgin texts about which we have no information at all other than the words they contain. We use the word scores we generate from the reference texts to estimate the positions of virgin texts on the policy dimensions in which we are interested. Essentially, each word scored in a virgin text gives us a small amount of information about which of the reference texts the virgin text most closely resembles. This produces a conditional expectation of the virgin text’s policy position, and each scored word in a virgin text adds to this information. Our procedure can thus be thought of as a type of Bayesian reading of the virgin texts, with our estimate of the policy position of any given virgin text being updated each time we read a word that is also found in one of the reference texts. The more scored words we read, the more confident we become in our estimate.

Figure 1 illustrates our procedure, highlighting the key steps involved. The illustration is taken from the data analysis we report below. The reference texts are the 1992 manifestos of the British Labour, Liberal Democrat (LD), and Conservative parties. The research task is to estimate the unknown policy positions revealed by the 1997 manifestos of the same parties, which are thus treated as virgin texts. When performed by computer, this procedure is entirely automatic, following two key decisions by the analyst: the choice of a particular set of reference texts and the identification

**FIGURE 1. The Wordscore procedure, using the British 1992–1997 manifesto scoring as an illustration**



of an estimated or assumed position for each reference text on each policy dimension of interest.

### Selection of Reference Texts

The selection of an appropriate set of reference texts is clearly a crucial aspect of the research design of the type of *a priori* analysis we propose. If inappropriate reference texts are selected, for example, if cookery books are used as reference texts to generate word scores that are then applied to speeches in a legislature, then the estimated positions of these speeches will be invalid. Selecting reference texts thus involves crucial substantive and qualitative decisions by the researcher, equivalent to the decisions made in the design or choice of either a substantive coding scheme for hand-coded content analysis or a coding dictionary for traditional computer-coding. While there are no mechanical procedures for choosing the reference texts for any analysis, we suggest here a number of guidelines as well as one hard-and-fast rule.

The hard-and-fast rule when selecting reference texts is that we must have access to confident estimates of, or assumptions about, their positions on the policy dimensions under investigation. Sometimes such estimates will be easy to come by. In the data analyses that follow, for example, we seek to compare our own estimates of party policy positions with previously published estimates. Thus we replicate other published content analyses of party manifestos, using "reference" party manifestos from one election to estimate the po-

sitions of "virgin" party manifestos in the next election. Our reference scores are taken from published expert surveys of the policy positions of the reference text authors, although this is only one of a number of easily available sources that we could have used with reasonable confidence. While a number of flaws can certainly be identified with expert surveys—some of which we have already mentioned—our purpose here is to compare the word scoring results with a well-known and widely used benchmark. In using these particular reference texts, we are in effect assuming that party manifestos in country  $c$  at election  $t$  are valid points of reference for the analysis of party manifestos at election  $t+1$  in the same country. Now this assumption is unlikely to be 100% correct, since the meaning and usage of words in party manifestos change over time, even over the time period between two elections in one country. But we argue not only that it is likely to be substantially correct, in the sense that word usage does not change very much over this period, but also that there is no better context for interpreting the policy positions of a set of party manifestos at election  $t+1$  than the equivalent set of party manifestos at election  $t$ . Note, furthermore, that any attempt to estimate the policy position of any political text, using any technique whatsoever, must relate this to some external context if the result is to be interpreted in a meaningful way, so that some equivalent assumption must always be made. As two people facing each other quickly discover, any attempt to describe one point as being to the "left" or the "right" of some other point must always have recourse to some external point of reference.

There may be times, however, when it is not easy to obtain simultaneously an authoritative set of reference texts and good estimates of the policy positions of these on all *a priori* dimensions in which the analyst is interested. In such instances it is possible to assume specific values for reference texts representing quintessential expressions of a view or policy whose position is known with a high degree of *a priori* confidence. Later in this paper, we apply our technique to legislative speeches made during a no-confidence debate, *assuming* that the speech of the leader of the government is quintessentially progovernment and that the speech of the leader of the opposition is quintessentially antigovernment.

In other words, what we require for our set of reference texts is a set of estimates of, or assumptions about, policy positions that we are prepared to stand over and use as appropriate points of reference when analyzing the virgin texts in which we are ultimately interested. Explicit decisions of substantive importance have to be made about these, but these are equivalent to the implicit decisions that must always be made when using other techniques for estimating policy positions. We do essentially the same thing when we choose a particular hand-coding scheme or a computer-coding dictionary, for example, both of which can always be deconstructed to reveal an enormous amount of (often hidden) substantive content. The need to choose external points of reference is a universal feature of any attempt to estimate the policy positions of political actors. In our application, the external points of reference are the reference texts.

We offer three further general guidelines in the selection of reference texts. The first is that the reference texts should use the same lexicon, in the same context, as the virgin texts being analyzed. For example, our investigations have (unsurprisingly) revealed very different English-language lexicons for formal written political texts, such as party manifestos, and formal spoken texts, such as speeches in a legislature. This implies that we should resist the temptation to regard party manifestos as appropriate reference texts for analyzing legislative speeches. In what follows, we use party manifestos as reference texts for analyzing other party manifestos and legislative speeches as reference texts for other legislative speeches. The point is that our technique works best when we have a number of "virgin" texts about which we know nothing and want to relate these to a small number of lexically equivalent (or very similar) "reference" texts about which we know, or are prepared to assume, something.

The second guideline is that policy positions of the reference texts should "span" the dimensions in which we are interested. Trivially, if all reference texts have the same policy position on some dimension under investigation, then their content contains no information that can be used to distinguish between other texts on the same policy dimension. An ideal selection of reference texts will contain texts that occupy extreme positions, as well as positions at the center, of the dimensions under investigation. This allows differences in the content of the reference texts to form the basis of inferences about differences in the content of virgin texts.

The third general guideline is that the set of reference texts should contain as many different words as possible. The content of the virgin texts is analyzed in the context of the word universe of the reference texts. The more comprehensive this word universe, and thus the less often we find words in virgin texts that do not appear in any reference text, the better. The party manifestos that we analyze below are relatively long documents. The British manifestos, for example, are between 10,000 and 30,000 words in length, each using between about 2,000 and 4,000 unique words. Most words observed in the virgin texts can be found in the word universe of the reference texts, while those that cannot tend to be used only very occasionally.<sup>3</sup> If the texts in which we are interested are much shorter than this—for example, legislative speeches are typically shorter than party manifestos—then this will tend to restrict the word universe of the reference texts and may reduce our ability to make confident inferences about the policy positions of virgin texts. As we show below when analyzing legislative speeches, the uncertainty of our estimates does increase when texts are short, although it is worth noting that, when other methods of content analysis use short texts, they typically report no estimate at all of the associated increase in uncertainty.<sup>4</sup> The problem of short texts is thus a problem with any form of quantitative content analysis and is not in any way restricted to the technique we propose here. And if the texts in which we are genuinely interested are short, then they are short and we just have to make the best of the situation in which we find ourselves. But the principle remains that it is always better to select longer suitable texts when these are available.

### Generating Word Scores from Reference Texts

We begin with set  $R$  of reference texts, each having a policy position on dimension  $d$  that can be estimated or assumed with confidence. We can think of the estimated or assumed position of reference text  $r$  on dimension  $d$  as being its *a priori* position on this dimension,  $A_{rd}$ . We observe the relative frequency, as a proportion of the total number of words in the text, of each different word  $w$  used in reference text  $r$ .<sup>5</sup> Let this be  $F_{wr}$ . Once

<sup>3</sup> We are more specific about this when discussing particular results below.

<sup>4</sup> We note that in the widely used content analysis data set of the CMP, many of the texts analyzed are very short. Using the CD-ROM distributed with Budge et al. 2001, we find that about one-third of all texts in the data set comprise fewer than 100 quasi-sentences. Generously estimating each quasi-sentence to be about 20 words, this implies that one-third of the CMP texts are about 2,000 words or fewer, while well over half of all texts analyzed are probably fewer than 4,000 words each.

<sup>5</sup> In the analyses reported here, we use the relative frequencies of every single different word in each reference text, even very common words such as prepositions and indefinite articles. We do this for two reasons. First, to do otherwise would require knowledge of the language in which the text under analysis was written, violating our principle of treating words as data and undermining our fundamental objective of being able to analyze texts written in languages we do not understand. Second, where such common words are systematically

we have observed  $F_{wr}$  for each of the reference texts, we have a matrix of relative word frequencies that allows us to calculate an interesting matrix of conditional probabilities. Each element in the latter matrix tells us the probability that we are reading reference text  $r$ , given that we are reading word  $w$ . This quantity is the key to our *a priori* approach. Given a set of reference texts, the probability that an occurrence of word  $w$  implies that we are reading text  $r$  is

$$P_{wr} = \frac{F_{wr}}{\sum_r F_{wr}}. \quad (1)$$

As an example consider two reference texts, A and B. We observe that the word “choice” is used 10 times per 10,000 words in Text A and 30 times per 10,000 words in Text B. If we know simply that we are reading the word “choice” in one of the two reference texts, then there is a 0.25 probability that we are reading Text A and a 0.75 probability that we are reading Text B.

We can then use this matrix  $P_{wr}$  to produce a score for each word  $w$  on dimension  $d$ . This is the expected position on dimension  $d$  of any text we are reading, given only that we are reading word  $w$ , and is defined as

$$S_{wd} = \sum_r (P_{wr} \cdot A_{rd}). \quad (2)$$

In other words,  $S_{wd}$  is an average of the *a priori* reference text scores  $A_{rd}$ , weighted by the probabilities  $P_{wr}$ . Everything on the right-hand side of this expression may be either observed or (in the case of  $A_{rd}$ ) assumed *a priori*. Note that if reference text  $r$  contains occurrences of word  $w$  and no other text contains word  $w$ , then  $P_{wr} = 1$ . If we are reading word  $w$ , then we conclude from this that we are certainly reading text  $r$ . In this event the score of word  $w$  on dimension  $d$  is the position of reference text  $r$  on dimension  $d$ : thus  $S_{wd} = A_{rd}$ . If all reference texts contain occurrences of word  $w$  at precisely equal frequencies, then reading word  $w$  leaves us none the wiser about which text we are reading and  $S_{wd}$  is the mean position of all reference texts.

To continue with our simple example, imagine that Reference Text A is assumed from independent sources to have a position of  $-1.0$  on dimension  $d$ , and Reference Text B is assumed to have a position of  $+1.0$ . The score of the word “choice” is then

$$0.25(-1.0) + 0.75(1.0) = -0.25 + 0.75 = +0.5.$$

Given the pattern of word usage in the reference texts, if we knew only that the word “choice” occurs in some text, then this implies that the text’s expected position on the dimension under investigation is  $+0.5$ . Of course we will update this expectation as we gather more information about the text under investigation by reading more words.

---

used with equal relative frequencies in all reference texts, they convey no useful information, but they do not systematically bias our results. Where such words are systematically used with unequal relative frequencies in reference texts, we assume that this is because they are conveying information about differences between texts.

## Scoring Virgin Texts

Having calculated scores for all words in the word universe of the reference texts, the analysis of any set of virgin texts  $V$  of any size is very straightforward. First, we must compute the relative frequency of each virgin text word, as a proportion of the total number of words in the virgin text. We call this frequency  $F_{vv}$ . The score of any virgin text  $v$  on dimension  $d$ ,  $S_{vd}$ , is then the mean dimension score of all of the scored words that it contains, weighted by the frequency of the scored words:

$$S_{vd} = \sum_w (F_{vv} \cdot S_{wd}). \quad (3)$$

This single numerical score represents the expected position of the virgin text on the *a priori* dimension under investigation. This inference is based on the assumption that the relative frequencies of word usage in the virgin texts are linked to policy positions in the same way as the relative frequencies of word usage in the reference texts. This is why the selection of appropriate reference texts—discussed at some length above—is such an important matter.

## Interpreting Virgin Text Scores

Once raw estimates have been calculated for each virgin text, we need to interpret these in substantive terms, a matter that is not as straightforward as might seem at first sight. Because different texts draw upon the same word universe, relative word frequencies and hence word scores can never distinguish perfectly between texts. Words found in common to all or most of the reference texts hence tend to take as their scores the mean overall scores of the reference texts. The result is that, for any set of virgin texts containing the same set of nondiscriminating words found in the reference texts, the raw virgin text scores tend to be much more clustered together than the reference text scores. While the mean of the virgin scores will have a readily interpretable meaning (relative to the policy positions of the reference texts), the dispersion of the virgin text scores will be on a different scale—one that is much smaller. To compare the virgin scores directly with the reference scores, therefore, we need to transform the scores of the virgin texts so that they have same dispersion metric as the reference texts. For each virgin text  $v$  on a dimension  $d$  (where the total number of virgin texts  $V > 1$ ), this is done as follows:

$$S_{vd}^* = (S_{vd} - \bar{S}_{vd}) \left( \frac{SD_{rd}}{SD_{vd}} \right) + \bar{S}_{vd}, \quad (4)$$

where  $\bar{S}_{vd}$  is the average score of the virgin texts, and the  $SD_{rd}$  and  $SD_{vd}$  are the sample standard deviations of the reference and virgin text scores, respectively. This preserves the mean and relative positions of the virgin scores but sets their variance equal to that of the reference texts. It is very important to note that this particular approach to rescaling is not fundamental to our word-scoring technique but, rather, is a matter of

substantive research design unrelated to the validity of the raw virgin text scores. In our case we wish to express the estimated positions of the virgin texts on the same metric as the policy positions of the reference texts because we wish to compare the two sets of numbers to validate our technique. Further development to interpret raw virgin scores can and should be done, yet the simple transformation (Eq. 4) provides excellent results, as we demonstrate below. Other transformations are of course possible, for example, by analysts who wish to compare estimates derived from text analysis with policy positions estimated by other sources but expressed in some quite different metric. For these reasons we recommend that raw scores always be reported, in addition to any transformed values of virgin scores.

### Estimating the Uncertainty of Text Scores

Our method for scoring a virgin text on some policy dimension generates a precise point estimate, but we have yet to consider any *uncertainty* associated with this estimate. No previous political science work estimating policy positions using quantitative content analysis deals systematically with the uncertainty of any estimate generated. The seminal and widely used CMP content analysis data, for example, are offered as point estimates with no associated measures of uncertainty. There is no way, when comparing the estimated positions of two manifestos using the CMP data, to determine how much the difference between estimates can be attributed to "real" differences and how much to coding unreliability.<sup>6</sup> Notwithstanding this, the time series of party policy positions generated by the CMP data has been seen in the profession as one of its great virtues, and "movements" of parties over time have typically been interpreted as real policy movements rather than as manifestations of coding unreliability.

Here we present a simple method for obtaining uncertainty estimates for our estimates of the policy positions of virgin texts. This allows us for the first time to make systematic judgments about the extent to which differences between the estimated policy positions of two texts are in fact significant.<sup>7</sup> Recall that each virgin text score  $S_{vd}$  is the weighted mean score of the words in

<sup>6</sup> In large part this is because most manifestos in the data set were coded once only by a single coder, making it impossible to provide specific indications of inter- or intracoder reliability. The CMP has not yet published any test of intracoder reliability (Volkens 2001, 39). Intercoder reliability checks have been performed by correlating the frequency distribution of an "official" coding of a single standard text with the codings of hired researchers. The average correlation found for 39 "thoroughly trained" hired coders was 0.72, with correlations running as low as 0.34 (Volkens 2001, 39). Thus we can be certain that there is intercoder unreliability in the CMP data but have no precise way of knowing whether or not the difference between the estimated positions of two texts is statistically significant.

<sup>7</sup> Previous approaches to content analysis typically refer to *reliability*, but that is different from the notion of uncertainty we use here. Reliability refers to the stability of measures across repeated codings, as with the intercoder reliability of hand-coded content analysis. *Uncertainty* in our usage is consistent with the statistical notion of uncertainty, representing confidence that an estimate reflects the true position rather than variation due to chance or other uncontrollable

text  $v$  on dimension  $d$ . If we can compute a mean for any set of quantities, then we can also compute a variance. In this context our interest is in how, for a given text, the scores  $S_{wd}$  of the words in the text vary around this mean. The variance of  $S_{wd}$  for a given text measures how dispersed the individual word scores are around the text's mean score. The less this variance, the more the words in the text all correspond to the final score and hence the lower our uncertainty about that score. Because the text's score  $S_{vd}$  is a weighted average, the variance we compute also needs to be weighted. We therefore compute  $V_{vd}$ , the variance of each word's score around the text's total score, weighted by the frequency of the scored word in the virgin text:

$$V_{vd} = \sum_w F_{wv} (S_{wd} - S_{vd})^2. \quad (5)$$

This measure produces a familiar quantity directly analogous to the unweighted variance, summarizing the "consensus" of the scores of each word in the virgin text.<sup>8</sup> Intuitively, we can think of each scored word in a virgin text as generating an independent prediction of the text's overall policy position. When these predictions are tightly clustered, we are more confident in their consensus than when they are scattered more widely.

As with any variance, we can use the square root of  $V_{vd}$  to produce a standard deviation. This standard deviation can be used in turn, along with the total number of scored virgin words  $N^v$ , to generate a standard error  $\sqrt{V_{vd}}/\sqrt{N^v}$  for each virgin text's score  $S_{vd}$ .<sup>9</sup> As we will see below, this standard error can then be used to perform standard statistical tests, such as the difference between means, to evaluate the significance of any difference in the estimated positions of two texts.<sup>10</sup>

---

factors, since we regard the generation of texts by political actors to be a stochastic process.

<sup>8</sup> Note that while we have employed the weighted formula here because our representation of words thus far has been as frequency distributions, this formula is equivalent to computing a population variance of the score of every (nonunique) word in the text. Each word hence contributes once for each time it occurs.

<sup>9</sup> This standard error applies to the raw virgin scores but not directly to the transformed scores. In the tables that follow (Tables 2–7), we also computed a standard error for the transformed scores along with 95% confidence intervals for the transformed scores, to make more straightforward the task of interpreting the uncertainty of the transformed scores on the original policy metric. The procedure for obtaining the upper and lower bounds of the transformed score confidence interval was straightforward. First, we computed the untransformed 95% confidence interval, calculated as the untransformed score  $S_{vd}$  plus and minus two standard errors (computed as explained in the text). These upper and lower confidence intervals, in the metric of the raw scores, were then transformed using exactly the same rescaling procedure as applied to the raw scores  $S_{vd}$ . The transformed standard error was then taken to be half of the distance between the transformed score and the bounds.

<sup>10</sup> We note that this measure is only one of a number of possible approaches to representing the uncertainty of our estimates of the positions of virgin texts and that numerous alternative measures can be developed to gauge the accuracy and robustness of final scores. In this introductory treatment of the word scoring method, we have deliberately chosen a form that will be familiar to most readers as well as being simple to compute. Diagnostic analysis of the word scoring technique is something to which we will return in future work.

**TABLE 1. Word Scoring Example Applied to Artificial Texts**

Word w	Word Count					Probability of Reading Text r, Given Reading Word w					Score $S_{wd}$	Virgin Score							
	Reference Text					Virgin Text						$P_{w1}$	$P_{w2}$	$P_{w3}$	$P_{w4}$	$P_{w5}$	$F_{wv}$	$F_{wv} * S_{wd}$	$F_{wv}(S_{wd} - S_{vd})^2$
	$r_1$	$r_2$	$r_3$	$r_4$	$r_5$														
A	2	0	0	0	0	0	1.00	0.00	0.00	0.00	0.00	-1.50	0.0000	0.0000	0.0000	0.0000			
B	3	0	0	0	0	0	1.00	0.00	0.00	0.00	0.00	-1.50	0.0000	0.0000	0.0000	0.0000			
C	10	0	0	0	0	0	1.00	0.00	0.00	0.00	0.00	-1.50	0.0000	0.0000	0.0000	0.0000			
D	22	0	0	0	0	0	1.00	0.00	0.00	0.00	0.00	-1.50	0.0000	0.0000	0.0000	0.0000			
E	45	0	0	0	0	0	1.00	0.00	0.00	0.00	0.00	-1.50	0.0000	0.0000	0.0000	0.0000			
F	78	2	0	0	0	0	0.98	0.03	0.00	0.00	0.00	-1.48	0.0000	0.0000	0.0000	0.0000			
G	115	3	0	0	0	0	0.97	0.03	0.00	0.00	0.00	-1.48	0.0000	0.0000	0.0000	0.0000			
H	146	10	0	0	0	2	0.94	0.06	0.00	0.00	0.00	-1.45	0.0020	-0.0029	0.0020				
I	158	22	0	0	0	3	0.88	0.12	0.00	0.00	0.00	-1.41	0.0030	-0.0042	0.0028				
J	146	45	0	0	0	10	0.76	0.24	0.00	0.00	0.00	-1.32	0.0100	-0.0132	0.0077				
K	115	78	2	0	0	22	0.59	0.40	0.01	0.00	0.00	-1.18	0.0220	-0.0261	0.0119				
L	78	115	3	0	0	45	0.40	0.59	0.02	0.00	0.00	-1.04	0.0450	-0.0467	0.0156				
M	45	146	10	0	0	78	0.22	0.73	0.05	0.00	0.00	-0.88	0.0780	-0.0687	0.0146				
N	22	158	22	0	0	115	0.11	0.78	0.11	0.00	0.00	-0.75	0.1150	-0.0863	0.0105				
O	10	146	45	0	0	146	0.05	0.73	0.22	0.00	0.00	-0.62	0.1460	-0.0904	0.0043				
P	3	115	78	2	0	158	0.02	0.58	0.39	0.01	0.00	-0.45	0.1580	-0.0712	0.0000				
Q	2	78	115	3	0	146	0.01	0.39	0.58	0.02	0.00	-0.30	0.1460	-0.0437	0.0032				
R	0	45	146	10	0	115	0.00	0.22	0.73	0.05	0.00	-0.13	0.1150	-0.0150	0.0116				
S	0	22	158	22	0	78	0.00	0.11	0.78	0.11	0.00	0.00	0.0780	0.0000	0.0157				
T	0	10	146	45	0	45	0.00	0.05	0.73	0.22	0.00	0.13	0.0450	0.0059	0.0151				
U	0	3	115	78	2	22	0.00	0.02	0.58	0.39	0.01	0.30	0.0220	0.0066	0.0123				
V	0	2	78	115	3	10	0.00	0.01	0.39	0.58	0.02	0.45	0.0100	0.0045	0.0081				
W	0	0	45	146	10	3	0.00	0.00	0.22	0.73	0.05	0.62	0.0030	0.0019	0.0034				
X	0	0	22	158	22	2	0.00	0.00	0.11	0.78	0.11	0.75	0.0020	0.0015	0.0029				
Y	0	0	10	146	45	0	0.00	0.00	0.05	0.73	0.22	0.88	0.0000	0.0000	0.0000				
Z	0	0	3	115	78	0	0.00	0.00	0.02	0.59	0.40	1.04	0.0000	0.0000	0.0000				
AA	0	0	2	78	115	0	0.00	0.00	0.01	0.40	0.59	1.18	0.0000	0.0000	0.0000				
BB	0	0	0	45	146	0	0.00	0.00	0.00	0.24	0.76	1.32	0.0000	0.0000	0.0000				
CC	0	0	0	22	158	0	0.00	0.00	0.00	0.12	0.88	1.41	0.0000	0.0000	0.0000				
DD	0	0	0	10	146	0	0.00	0.00	0.00	0.06	0.94	1.45	0.0000	0.0000	0.0000				
EE	0	0	0	3	115	0	0.00	0.00	0.00	0.03	0.97	1.48	0.0000	0.0000	0.0000				
FF	0	0	0	2	78	0	0.00	0.00	0.00	0.03	0.98	1.48	0.0000	0.0000	0.0000				
GG	0	0	0	0	45	0	0.00	0.00	0.00	0.00	1.00	1.50	0.0000	0.0000	0.0000				
HH	0	0	0	0	22	0	0.00	0.00	0.00	0.00	1.00	1.50	0.0000	0.0000	0.0000				
II	0	0	0	0	10	0	0.00	0.00	0.00	0.00	1.00	1.50	0.0000	0.0000	0.0000				
JJ	0	0	0	0	3	0	0.00	0.00	0.00	0.00	1.00	1.50	0.0000	0.0000	0.0000				
KK	0	0	0	0	2	0	0.00	0.00	0.00	0.00	1.00	1.50	0.0000	0.0000	0.0000				
Total	1,000	1,000	1,000	1,000	1,000	1,000						1.00	-0.45	0.14					
	-1.50	-0.75	0.00	0.75	1.50	-0.45													

*A priori* positions of reference texts  
Estimated score for virgin text  $S_{vd}$       **-0.45**  
Estimated weighted variance  $V_{vd}$       **0.14**  
Estimated SD  $\sqrt{V_{vd}}$       **0.38**  
Estimated SE  $\sqrt{V_{vd}}/\sqrt{1000}$       **0.018**

### Illustration Using a Sample Text

The method we have outlined can be illustrated by working through the calculation of word scores on an artificial text. Table 1 shows the results of analyzing a very simple hypothetical data set, shown in columns 2–7 in the table (in bold face), containing word counts for 37 different words observed in five reference texts,  $r_1$ – $r_5$ , as well as counts for the same set of words in a hypothetical “virgin” text whose position we wish to

estimate. The policy positions of the reference texts on the dimension under investigation are estimated or assumed *a priori* and are shown at the bottom of the table as ranging between -1.50 and +1.50. Table 1 shows that, in this hypothetical data set, nearly all words can be ranked from left to right in terms of the extent to which they are associated with left- or right-wing parties. Within each individual text, the observed pattern of word frequencies fits a normal distribution. We also indicate the “real” position of the virgin text, which

is unknown to the hypothetical analyst but which we know to be  $-0.45$ . This is the essential quantity to be estimated by comparing the distribution of the word frequencies in the virgin texts with that in the reference texts.

The columns headed  $P_{w1}$ – $P_{w5}$  show the conditional probabilities (Eq. 1) necessary for computing word scores from the reference texts—this is the matrix of probabilities that we are reading reference text  $r$  given that we are reading word  $w$ . Combined with the *a priori* positions of the reference texts, these allow us to calculate scores,  $S_w$ , for each word in the word universe of the reference texts (Eq. 2). These scores are then used to score the virgin text by summing the scores of words used in the virgin text, weighting each score by the relative frequency of the word in question (Eq. 3). The resulting estimate, and its associated uncertainty measure, is provided at the bottom right of Table 1, together with its associated standard error. From this we can see that, in this perfectly behaved data set, our technique perfectly retrieves the position of the virgin text under investigation.

While this simple example illustrates the calculations associated with our technique, it of course in no way shows its efficacy with real-world data, in which there will be much more heavily overlapping patterns of word usage in reference texts, large numbers of very infrequently used words, volumes of words found in virgin texts that do not appear in reference texts and therefore cannot be scored, and so on. The true test of the technique we propose lies in applying it to texts produced by real-world political actors, to see if we can reproduce estimates of their policy positions that have been generated by more traditional means.

### ESTIMATING ECONOMIC POLICY POSITIONS OF BRITISH AND IRISH PARTIES

We now test our technique using real-world texts, by attempting to replicate previously published findings on the policy positions of political parties in Britain and Ireland. We compare our own findings with three sets of independent estimates of the economic policy positions of British and Irish political parties at the time of the 1997 general elections in each country. These are the results of 1997 expert surveys of party policy positions (Laver 1998a, b) and of the hand-coding and deterministic computer-coding of 1997 party manifestos (Laver and Garry 2000).

#### British Party Positions on Economic Policy

The first task is to calculate word scores on the economic policy dimensions for British party manifestos in the 1990s. We selected the 1992 British Labour, Conservative, and LD party manifestos as reference texts. For independent estimates of the economic policy positions of these manifestos, we use the results of an expert survey of the policy positions of the parties that wrote them, on the scale “increase public services vs.

cut taxes,” reported in Laver and Hunt 1992.<sup>11</sup> The first stages in the analysis are to observe frequency counts for all words used in these reference texts<sup>12</sup> and to calculate relative word frequencies from these.<sup>13</sup> Using these relative frequencies and the reference text policy positions, we then calculated a word score on the economic policy dimension for every word used in the reference texts, using the procedures outlined above (Eqs. 1 and 2).

Having calculated word scores on the economic policy dimension for each of the 5,299 words used in the 1992 reference texts, we use these to estimate the positions of three “virgin” texts. These are the Labour, LD, and Conservative manifestos of 1997. Note that this is a tough substantive test for our technique. Most commentators, backed up by a range of independent estimates, suggest that the ordering of the economic policy positions of the British parties changed between the 1992 and the 1997 elections, with Labour and the LDs exchanging places, leaving Labour in the center and the LDs on the left in 1997. This can be seen in 1997 expert survey findings (Laver 1998a) that we set out to replicate using computer word scoring, reported in the third row of the top panel in Table 2. We are particularly interested to see whether our technique can pick up this unusual and significant movement.

We can only score virgin texts on the words that they share with the universe of reference texts. The 1997 British manifestos used a total of 1,573 words that did not appear in the 1992 texts and these could not be scored.<sup>14</sup> We thus applied the word scores derived from

<sup>11</sup> It is very important to note that such expert survey estimates are convenient to use as reference scores in this context but are not in any way intrinsic to our technique. What we require are independent estimates of, or assumptions about, the positions of the reference texts in which we can feel confident. The expert survey scores we use are reported in the first row in the lower half in Table 2. Both in terms of their face validity and because these scores report the mean judgments of a large number of British political scientists, we consider these estimated positions of the reference texts to represent a widely accepted view of the of the British policy space in 1992.

<sup>12</sup> While, for reasons discussed above, we included every single word used in the 1992 manifestos, even common words without substantive political meaning such as “a” and “the,” we did exclude all “non-words,” which we took to be character strings not beginning with letters.

<sup>13</sup> Any computer-coded content analysis software (for example, Textpack) can perform simple word counting. To process large numbers of texts simultaneously and quickly perform all subsequent calculations on the output, however, we wrote our own software. Easy-to-use software—entitled WORDSCORES—for implementing the methods described in this paper is freely available from <http://www.politics.tcd.ie/wordscores/>. A full replication data set for this paper, using the WORDSCORES software, is also available at that web site. Installation or updating of WORDSCORES can be accomplished by any computer connected to the Internet by executing a single command from within the Stata statistical package: `net install http://www.politics.tcd.ie/wordscores/wordscores`. Version information prior to installation can be obtained by executing the Stata command `net describe http://www.politics.tcd.ie/wordscores/wordscores`.

<sup>14</sup> Most of the 1997 words not used in 1992 were used very infrequently, with a median occurrence of 1 and a mean occurrence of between 1.2 and 1.9 (see Table 2). For this reason they would have contributed very little weight to the virgin text scores. Overall for

**TABLE 2. Raw and Standardized Estimated Economic Policy Positions of 1997 British Party Manifestos**

Party	Liberal Democrat	Labour	Conservative	Mean Absolute Difference
<b>Estimates</b>				
1997 transformed virgin text scores	<b>5.00</b>	<b>9.17</b>	<b>17.18</b>	
SE	0.363	0.351	0.325	
1997 expert survey	<b>5.77</b>	<b>10.30</b>	<b>15.05</b>	
SE ( <i>n</i> =117)	0.234	0.229	0.227	
1997 standardized comparison scores				
Word scores	-0.88	-0.21	1.09	<b>0.13</b>
Expert survey	-0.99	-0.02	1.01	—
Hand-coded content analysis	-0.83	-0.28	1.11	<b>0.17</b>
Dictionary-based computer-coding	-1.08	0.18	0.90	<b>0.13</b>
Raw data				
1992 reference texts				
<i>A priori</i> positions	<b>8.21</b>	<b>5.35</b>	<b>17.21</b>	
SE ( <i>n</i> =34)	0.425	0.377	0.396	
Length in words	17,077	11,208	28,391	
No. of unique words	2,911	2,292	3,786	
1997 virgin texts				
Raw mean word scores ( $S_{vd}$ )	10.2181	10.3954	10.7361	
SE	0.015	0.015	0.014	
Length in words	13,709	17,237	20,442	
Unique words scored	1,915	2,211	2,279	
% words scored	94.9	96.2	95.5	
Unique unscorable words	423	697	714	
Mean frequency of unscorable words	1.23	1.26	1.29	

Sources: *A priori* positions 1992 (Laver and Hunt 1992); expert survey scores 1997 (Laver 1998a); hand-coded content analysis and deterministic computer-coding (Laver and Garry 2000).

Note: Standardized scores are reported raw scores for 1997 standardized within each data source. For hand- and deterministic computer-codings, these have been recalculated to facilitate comparison from data presented by Laver and Garry (2000), who standardized their raw score across all observations for Britain and Ireland. The mean absolute difference reports the mean of the absolute differences for the three parties between the standardized party scores for each text-based estimate and the standardized expert survey party score. Standard errors are computed as described in the text.

the 1992 reference texts to the 1997 manifestos, calculating a “raw” score for each of the three manifestos (Eq. 3) and transforming (Eq. 4) it in the way described above. Finally, we calculate the standard errors of our estimates (Eq. 5 and associated discussion).

The key results of this analysis are presented in the top panel in Table 2. The first row reports our estimated positions of the 1997 party manifestos, transformed to the same metric as the 1992 expert survey scores that were used as points of reference. Our first point of comparison is with a set of 1997 expert survey scores, expressed in the same metric, highlighting the shift of the Labour Party to the center of this policy dimension (Laver 1998a). These scores are reported in the third row in Table 2. The comparison is very gratifying. Our word-scored estimates clearly pick up the switch in Labour and LD economic policy positions and are remarkably close, considering that they derive from an utterly independent source, to the expert survey estimates for 1997. Note particularly that the word scores we used were calculated from 1992 reference positions that locate the LDs between Labour and the Conservatives on economic policy, so that it was simply the

changing relative frequencies of word use between the 1992 and the 1997 manifestos that caused the estimated positions of these two parties to reverse, in line with independent estimates.

Table 2 also reports the standard errors associated with our raw estimates, from which we can conclude that differences among the estimated economic policy positions of the three manifestos are statistically significant. Note that this availability of standard errors, allowing such judgments to be made, is unique among published estimates of policy positions based on the content analysis of political texts.

To compare our results with those generated by other content analysis techniques, the last four rows in the top panel in Table 2 report, in addition to our own estimates and those of the 1997 expert survey, two other text-based estimates of the 1997 economic policy positions of the British parties. One of these derives from hand-coded content analysis, and the other from dictionary-based computer-coding, of the 1997 manifestos that we have treated here as virgin texts (both reported in Laver and Garry 2000). Since different published sets of scores had different metrics, all scores have been standardized to facilitate comparison.<sup>15</sup> The main

the 1997 virgin texts, the bottom panel in Table 2 shows that the percentages of virgin words scoreable were 96.2%, 94.9%, and 95.5% for the LDs, Labour, and the Conservatives, respectively.

<sup>15</sup> All sets of standardized estimates in Table 2 have been standardized within country and time period in the tables that follow, to

substantive difference between different estimates of British party positions in 1997 concerns the placement of the Labour Party. All scales locate Labour between the LDs and the Conservatives. The dictionary-based scale places Labour closer to the Conservatives, and the other text-based scales place Labour closer to the LDs, while the independent expert survey locates Labour midway between the two other parties.

As a summary of the fit between the various text-based estimates of party positions and the expert survey, the final column in the top panel in Table 2 reports the mean absolute difference between the estimated positions of the parties on each standardized scale and the positions of the same parties in the expert survey. This confirms our *prima facie* impression that our word-scored estimates are somewhat closer than the hand-coded content analysis to the expert survey estimates (representing the consensus among British political scientists about British party positions in 1997) and are about as close to these as the more traditional dictionary-based computer-coded scale. This is a remarkable achievement considering that, in stark contrast to all other methods, our word scoring technique treats words as data without reading or understanding them in any way, uses no knowledge of English, and does not require a predetermined computer-coding dictionary when analyzing the texts.

### **Irish Party Positions on Economic Policy**

We now report a similar analysis for the Irish party system. As our reference texts for Irish politics in the 1990s, we take the manifestos of the five main parties contesting the 1992 election—Fianna Fáil, Fine Gael, Labour, the Progressive Democrats (PDs), and the Democratic Left (DL). For our independent estimate of the positions of these reference texts, we use an expert survey taken at the time of the 1992 Irish election (Laver 1994). Having used these data in a preliminary analysis to calculate word scores for the economic policy dimension in Ireland in the 1990s, we then analyze 1997 Irish party manifestos as virgin texts. Our aim is once more to replicate independent published estimates of Irish party policy positions in 1997—the results of an expert survey conducted at the time of the 1997 election (Laver 1998b), as well as estimates based on hand-coded content analysis and dictionary-based computer-coding (Laver and Garry 2000). The results of this analysis are listed in Table 3, which has the same format as Table 2.

facilitate comparison of estimates originally reported using different units of analysis. (Thus the 1997 British estimates, for example, are standardized against themselves.) This differs from the practice adopted by Laver and Garry (2000), who standardized across both countries and time periods. This was because they were evaluating the application of a single expert coding scheme and computer-coding dictionary to all observations. In contrast, we use the 1992 manifestos to generate separate sets of words scores for Britain and Ireland and apply these separately to virgin texts taken from subsequent time periods in each country. The standardized figures in Tables 2–5 thus differ from those reported by Laver and Garry (2000) but are calculated directly from them.

Substantively, while nothing as dramatic happened in Ireland between 1992 and 1997 as the vaunted dash to the center by the British Labour Party under Tony Blair, there was a major coalition realignment that we expect to show up in the economic policy positions of the parties. The government that formed immediately after the 1992 election was the first-ever coalition between Fianna Fáil and the Labour Party. As the bottom panel in Table 3 shows, these parties were judged by expert survey respondents in 1992 to be adjacent, though by no means close, on the economic policy dimension. This government fell in 1994 and was replaced without an intervening election by a “rainbow” coalition of Fine Gael, Labour, and DL—so called because of major policy differences among what was essentially a coalition of Fianna Fáil’s opponents. By the time of the 1997 election, the three parties of the Rainbow Coalition presented a common front to the electorate and sought reelection. While promoting independent policy positions, they were nonetheless careful to ensure that their respective party manifestos did not contain major policy differences that would embarrass them on the campaign trail. Confronting the Rainbow Coalition at the election, Fianna Fáil and the PDs formed a pact of their own, promising to go into government together if they received enough support and, also, taking care to clean up any major policy incompatibilities in their respective manifestos that would have been exploited by opponents during the campaign. The 1997 election was thus fought between two rival coalitions—the Fine Gael, Labour, and DL rainbow, on one side, and Fianna Fáil and the PDs, on the other—who published independent but coordinated policy programs.

The top panel in Table 3 shows that the main manifestation of these changes in expert survey data is a collective judgment that Fine Gael Shifted to the left in 1997 as a result of its membership in the Rainbow Coalition with Labour and DL. The experts did not consider Fianna Fáil to have shifted right, despite the fact that the 1997 Fianna Fáil manifesto was designed not to conflict with that of the PDs and that immediately after the election Fianna Fáil agreed to a joint program of government with the right-wing PDs, subsequently governing harmoniously with them for the first full-term coalition government in the history of the Irish state. This is intriguing because, as the last four lines in the top panel in Table 3 show, both expert survey and hand-coded content analyses continue to show Fine Gael to the right of Fianna Fáil in 1997, while both dictionary-based computer-coding and our own word scoring techniques, which proceeded without expert intervention, find Fine Gael to the left of Fianna Fáil. Both sets of computer-coded results reflect the pattern of actual coalitions in the legislature, so we may speculate here that we are seeing signs of experts—whether survey respondents or human text coders—reading between the lines of the published texts and inferring that, in a coalition environment such as this, stated policy positions are not entirely sincere.

Be that as it may, the results in Table 3 show that our approach, while generating results with good face

**TABLE 3. Raw and Standardized Estimated Economic Policy Positions of 1997 Irish Party Manifestos**

	DL	Labour	Fianna Fáil	Fine Gael	PD	Mean Absolute Difference
<b>Estimates</b>						
1997 transformed virgin text scores	<b>3.79</b>	<b>6.78</b>	<b>15.32</b>	<b>13.18</b>	<b>16.44</b>	
SE	1.908	0.503	0.461	0.593	0.797	
1997 expert survey	<b>5.47</b>	<b>7.77</b>	<b>12.07</b>	<b>12.30</b>	<b>17.27</b>	
SE ( <i>n</i> =30)	0.325	0.330	0.398	0.363	0.310	
<b>1997 standardized comparison scores</b>						
Word scores	-1.32	-0.78	0.79	0.37	0.94	<b>0.27</b>
Expert survey	-1.21	-0.70	0.24	0.29	1.38	
Hand-coded content analysis	-1.10	-0.72	-0.02	0.38	1.46	<b>0.11</b>
Deterministic computer-coding	-1.22	-0.52	0.36	-0.06	1.45	<b>0.15</b>
<b>Raw data</b>						
1992 reference texts						
<i>A priori</i> positions	<b>4.50</b>	<b>6.88</b>	<b>13.13</b>	<b>15.00</b>	<b>17.63</b>	
SE ( <i>n</i> =28)	0.40	0.37	0.57	0.47	0.30	
Length in words	6,437	16,373	3,782	3,679	3,523	
No. of unique words	1,763	2,768	1,186	1,019	1,136	
1997 virgin texts						
Raw mean word scores ( $S_{vd}$ )	10.9205	10.9954	11.2087	11.1552	11.2367	
SE	0.048	0.013	0.012	0.015	0.020	
Length in words	2,549	32,171	38,659	24,026	13,922	
Unique words scored	748	2,348	2,609	2,098	1,721	
% words scored	92.4	92.4	89.7	92.1	92.9	
Unique unscorable words	172	1,492	2,203	1,902	991	
Mean frequency of unscorable words	1.13	1.64	1.82	1.59	1.13	

Sources: *A priori* positions 1992 (Laver 1994); expert survey 1997 (Laver 1998b); expert-coded content analysis and deterministic compute coding (Laver and Garry 2000).

Note: See Note to Table 2.

validity in terms of subsequent coalition alignments, does not correspond as well as the other text-based techniques with expert survey. The key difference between our scale and the others is the convergence of Fianna Fáil and the PDs indicated by our technique, followed as we have seen by a coalition between the two parties. While this convergence is substantively plausible, an alternative possibility is that our estimates are less accurate than the others in this case.

One possible source of such a problem is that the 1997 Irish manifestos were on average considerably longer than their short 1992 progenitors, using many words that were not used in 1992. The Fianna Fáil manifesto, in particular, burgeoned dramatically in length. We scored the 4,279 different words in the 1992 manifestos, but a total of 4,188 new words appeared in 1997, albeit many of them only once.<sup>16</sup> There was thus much less overlap than in Britain between the word pools used in 1992 and 1997, leaving more of the 1997 Irish manifestos necessarily unscored. This is reflected in noticeably higher standard errors for our Irish estimates than for the British ones. The short DL manifesto in 1997, for example, generates a word-scored estimated economic policy position of 3.79 on the 1–20 metric of the expert survey with which it is being compared, but

the very high associated standard error tells us that this position might be anything from 0.0 to 7.6 on this scale (its 95% confidence interval). The PD manifesto has a standard error that implies that we cannot statistically distinguish its economic policy position from that of Fianna Fáil. In other words, the standard errors generated by the word scoring technique are telling us that we should not feel as confident with its estimates for Ireland as we feel with those for Britain. We consider this to be an interesting and important result in itself—bearing in mind that all previous content analysis policy estimates of which we are aware report point estimates with no estimate whatsoever of associated error and, thus, are effectively blind to the potential problems arising from short texts we have diagnosed in the Irish case.

### ESTIMATING THE POLICY POSITIONS OF BRITISH AND IRISH PARTIES ON THE LIBERAL–CONSERVATIVE SOCIAL POLICY DIMENSION

A range of techniques has been used to estimate economic policy positions in Britain and Ireland and has been found to have good face validity. When setting out to cross-validate economic policy estimates produced by our word scoring method, therefore, we are working in well-explored territory. We turn now to a more difficult and interesting problem. This is the estimation of policy positions on the “liberal–conservative” dimension of social policy, taken as the second most important

<sup>16</sup> The Fianna Fáil manifesto in 1997 contained more than 10 times as many total words as the 1992 manifesto. Because the pool of reference texts included manifestos from four other parties, however, we were able to score 89.7% of the words in the 1997 manifesto (see Table 3). Results for the other virgin texts were all above 92% words scored.

**TABLE 4. Raw and Standardized Estimated Social Policy Positions of 1997 British Party Manifestos**

	Liberal Democrat	Labour	Conservative	Mean Absolute Difference
<b>Estimates</b>				
1997 transformed virgin text scores	<b>5.17</b>	<b>8.96</b>	<b>15.06</b>	
SE	0.285	0.272	0.254	
1997 expert survey	<b>6.75</b>	<b>8.28</b>	<b>13.26</b>	
SE ( $n=116$ )	0.240	0.228	0.253	
1997 standardized comparison scores				
Word scores	-0.91	-0.15	1.07	<b>0.12</b>
Expert survey	-0.79	-0.34	1.13	
Hand-coded content analysis	-1.07	-0.15	0.91	<b>0.33</b>
Deterministic computer-coding	-1.06	-0.12	0.93	<b>0.31</b>
<b>Raw data</b>				
1992 reference texts				
<i>A priori</i> positions	<b>6.87</b>	<b>6.53</b>	<b>15.34</b>	
SE ( $n=34$ )	0.410	0.358	0.451	
1997 virgin texts				
Raw mean word scores ( $S_{vd}$ )	9.5285	9.6956	9.9649	
SE	0.013	0.012	0.011	

Note: Sources as in Table 2. All statistics for the word counts and frequencies of reference and virgin texts are the same as in Table 2.

dimension of competition in many European party systems, a general perception for which Warwick (2002) found support when extracting common policy spaces from party manifesto and expert survey data.

Traditional techniques of content analysis have been very much less effective at providing reliable and stable estimates of policy positions on this dimension, a conclusion confirmed in a careful study by McDonald and Mendes (2001). Having found a number of economic policy scales to be highly reliable, they found the reliability of content analysis-based social policy scales to be "not so filled with noise as to be completely unreliable" but "below a . . . reliability that we would take as minimally acceptable" (McDonald and Mendes 2001, 111).

In applying our word scoring approach to a new policy dimension, we also reveal one of its chief advantages of flexibility, ease of use, and susceptibility to tests using different *a priori* conditions. Once the reference texts have been converted into the matrix of word probabilities  $P_{wr}$ , it is straightforward to compute word scores for a new dimension  $d'$  simply by changing the *a priori* set of reference scores to  $A_{rd'}$ . We can then very easily apply these new word scores to the virgin texts and thereby estimate their positions on  $d'$ , which in most cases takes under one second of computing time. In contrast to other computer-coding techniques, there is no need for the labor-intensive development and testing of a new coding dictionary for each new policy dimension considered. We demonstrate this by rerunning the analysis for the social policy dimension in Britain and Ireland in a manner identical to that for economic policy, except that the reference scores were taken from expert survey estimates of the social policy positions of the authors of these reference texts (Laver 1994; Laver and Hunt 1992). The social policy positions we estimate are defined *a priori* in terms of promoting liberal policies on matters such as abortion and homosexuality, at one end, and opposing such policies, at the other.

### British Party Positions on Social Policy

The results of rescored of the 1997 virgin texts for Britain are reported in Table 4, which has the same format as Table 2 without repeating raw data unnecessarily. As before, we begin by comparing our estimates with those generated by the completely independent expert survey conducted at the time of the 1997 election. Substantively, the main party movement reported by the expert surveys is a shift from estimates in 1992 that found the social policy positions of Labour and the LDs to be statistically indistinguishable, to one in 1997 in which Labour occupied a statistically distinct position on the conservative side of the LDs. This finding is clearly replicated by our word-scored estimates.

As before, the last four rows in the top panel in Table 4 compare standardized estimates from our word scoring method with those derived from the 1997 expert survey, as well as both hand- and dictionary-based computer-coded content analyses of the 1997 manifestos. These results, summarized by the mean absolute differences, show that computer word scoring performs extraordinarily well in this previously troublesome area, far better than any other content analysis technique. Substantively this is because, according to the expert survey that summarizes the judgments of British political scientists on this matter, the situation in 1997 was one in which Labour and the LDs were relatively close to each other in the more liberal half of the social policy dimension, with the Conservatives firmly on the right. This configuration is retrieved from the 1997 manifestos by our language-blind word scoring technique—it can be seen in the negative standard scores for the Labour Party. The more traditional techniques of content analysis, whether hand- or computer-coded, place Labour much closer to the Conservatives on social policy than to the LDs, a finding that does not seem to have good face validity.

The mean absolute differences between the results of the various content analyses and those of the expert survey show that our word scoring technique did as well on the liberal-conservative dimension in Britain as it did for economic policy. What is striking, however, is that it did distinctly better than more traditional text analysis techniques in what has previously been a very problematic area for content analysis.

### Irish Party Positions on Social Policy

We reran the analysis in the same way to estimate the social policy positions of the 1997 Irish party manifestos, treating these as virgin texts. The results are reported in Table 5. The most important substantive pattern to watch for in the Irish case is the relative position of Fianna Fáil and the PDs. Since the PDs are regarded by many as a classical liberal party, their right-wing economic policy position is widely perceived to be combined with a relatively leftist position on social issues. As Table 5 shows, this received wisdom is reflected in expert survey estimates. Fianna Fáil, in contrast, is typically seen as the guardian of traditional Catholic social values in Ireland. This pattern can be seen clearly in the expert surveys, which place Fianna Fáil very firmly on the right of the liberal-conservative dimension of social policy.

In contrast to the situation in Britain, therefore, the relative positions of parties on the liberal-conservative social policy dimension in Ireland differ in important substantive ways from those on the economic policy dimension. The top row in Table 5 shows that our language-blind word scoring techniques picks this difference up very well, coming close to the 1997 expert survey results in its analysis of the 1997 manifestos as virgin texts. As the last four rows in the top panel in Table 5 show, the more traditional content analysis techniques cannot replicate independent estimates of

the social policy position of the PDs, (mis)placing the PDs, with high positive standard scores, on the conservative side of the social policy dimension at a position much more conservative than that of Fianna Fáil. This neither corresponds to the consensus of political scientists reflected in the expert judgments nor has good face validity.

The mean absolute differences again summarize the relative performance of the three content analysis techniques. These show that our word scoring technique, despite the fact that it uses no knowledge of the English language, performs strikingly better than the other content analysis techniques, performing remarkably well on a dimension that has previously presented content analysts with considerable problems.

### Overall Fit with Expert Surveys

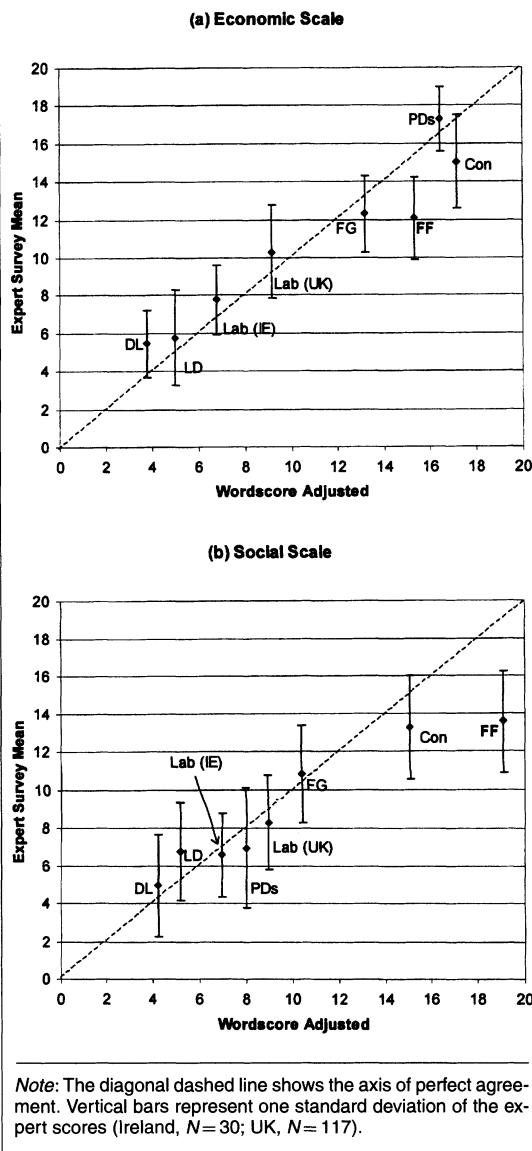
Figure 2 summarizes the fit between independent expert survey findings and our rescaled estimates of the policy positions of virgin texts, using computer word scoring. The X axis gives the word-scored estimates for 1997 virgin texts; the Y axis, expert survey estimates for 1997 of the positions of the authors of those texts. The vertical bars on each point represent a single standard deviation among the expert survey results. These bars may be interpreted as the range within which a single standard deviation of experts ranked the party on each scale. Where this bar crosses the vertical line of perfect correspondence, it indicates that approximately the middle 65% of the experts surveyed could easily have chosen the policy position estimated by the word scoring procedure. Of all of the texts we analyzed, on two policy dimensions, the only text for which word-scored estimates were more than a single standard deviation away from expert survey results was the Fianna Fáil manifesto in Ireland. And this difference, as we have argued, could possibly have been the result of

**TABLE 5. Raw and Standardized Estimated Social Policy Positions of 1997 Irish Party Manifestos**

	DL	Labour	Fianna Fáil	Fine Gael	PD	Mean Absolute Deviation
<b>Estimates</b>						
1997 transformed virgin text scores	<b>4.23</b>	<b>6.96</b>	<b>19.07</b>	<b>10.37</b>	<b>8.01</b>	
SE	1.178	0.319	0.339	0.378	0.474	
1997 expert survey	<b>4.97</b>	<b>6.57</b>	<b>13.55</b>	<b>10.82</b>	<b>6.93</b>	
SE ( $n=30$ )	0.495	0.405	0.491	0.467	0.577	
1997 standardized comparison scores						
Word scores	-0.97	-0.49	1.65	0.12	-0.31	<b>0.21</b>
Expert survey	-1.02	-0.57	1.42	0.64	-0.47	
Hand-coded content analysis	-1.31	-0.62	0.09	1.23	0.62	<b>0.67</b>
Deterministic computer-coding	-1.07	-1.02	0.75	0.25	1.09	<b>0.62</b>
<b>Raw data</b>						
1992 reference texts						
<i>A priori</i> positions	<b>3.50</b>	<b>6.00</b>	<b>17.50</b>	<b>13.71</b>	<b>9.43</b>	
SE ( $n=28$ )	0.416	0.404	0.391	0.554	0.809	
1997 virgin texts						
Raw mean word scores ( $S_{vd}$ )	9.4960	9.6098	10.1157	9.7523	9.6537	
SE	0.049	0.013	0.014	0.016	0.020	

*Note:* Sources as in Table 3. Word statistics and counts as in Table 3.

**FIGURE 2. Agreement Between Word Score Estimates and Expert Survey Results, Ireland and United Kingdom, 1997, for (a) Economic and (b) Social Scales**



contextual judgments made by experts about the “real” position of Fianna Fáil, rather than of error in the computer analysis of the actual text of the party manifesto. Put in a slightly different way, the technique we propose performed, in just about every case, equivalently to a typical expert—which we take to be a clear confirmation of the external validity of our technique’s ability to extract meaningful estimates of policy positions from political texts.

## CODING NON-ENGLISH-LANGUAGE TEXTS

Thus far we have been coding English-language texts, but since our approach is language-blind it should work equally well in other languages. We now apply it to German-language texts, analyzing these using no knowledge of German. Our research design is essentially similar to that we used for Britain and Ireland. As reference texts for Germany in the 1990s, we take the 1990 manifestos of four German political parties—the Greens, Social Democratic Party (SPD), Christian Democrats (CDU), and Free Democrats (FDP). Our estimates of the *a priori* positions of these texts on economic and social policy dimensions derive from an expert survey conducted in 1989 by Laver and Hunt (1992). Having calculated German word scores for both economic and social policy dimensions in precisely the same way as before, we move on to analyze six virgin texts. These are the manifestos of the same four parties in 1994, as well as manifestos for the former Communists (PDS) in both 1990 and 1994. Since no expert survey scores were collected for the PDS in 1990, or for any German party in 1994, we are forced to rely in our evaluation upon the face validity of our estimated policy positions for the virgin texts. However, the corpus of virgin texts presents us with an interesting and taxing new challenge. This is to locate the PDS on both economic and social policy dimensions, even though no PDS reference text was used to calculate the German word scores. We are thus using German word scores, calculated using no knowledge of German, to locate the policy positions of the PDS, using no information whatsoever about the PDS other than the words in its manifestos, which we did not and indeed could not read ourselves. The top panel in Table 6 summarizes the results of our analysis.

The first row in Table 6 reports our rescaled computer estimates of the economic policy positions of the six virgin texts. The main substantive pattern for the economic policy dimension is a drift of all established parties to the right, with a sharp rightward shift by the SPD. Though this party remains between the position of the Greens and that of the CDU, it has moved to a position significantly closer to the CDU. The face validity of this seems very plausible. Our estimated economic policy positions of the 1990 and 1994 PDS manifestos locate these firmly on the left of the manifestos of the other four parties, which has excellent face validity. The rescaled standard errors show that the PDS is indeed significantly to the left of the other parties but that there is no statistically significant difference between the 1990 and the 1994 PDS manifestos. In other words, using only word scores derived from the other four party manifestos in 1990 and no knowledge of German, the manifestos of the former Communists were estimated in both 1990 and 1994 to be on the far left of the German party system. We consider this to be an extraordinarily good result for our technique.

The third row in Table 6 reports our estimates of the social policy positions of the virgin texts. As in the

**TABLE 6. Estimated Economic and Social Policy Positions of German Party Manifestos, 1990–94**

Party	1990 PDS	1994 PDS	1994 Green	1994 SDP	1994 CDU	1994 FDP
<b>Estimates</b>						
1994 transformed economic policy virgin text scores	<b>4.19</b>	<b>3.98</b>	<b>7.47</b>	<b>10.70</b>	<b>13.67</b>	<b>17.15</b>
SE	0.436	0.511	0.259	0.365	0.391	0.220
1994 transformed social policy virgin text scores	<b>1.16</b>	<b>1.93</b>	<b>4.09</b>	<b>11.07</b>	<b>13.65</b>	<b>8.12</b>
SE	0.306	0.421	0.221	0.325	0.368	0.182
<b>Raw data</b>						
1990 reference texts						
Economic policy						
<i>A priori</i> positions	—	—	<b>5.21</b>	<b>6.53</b>	<b>13.53</b>	<b>15.68</b>
SE ( <i>n</i> = 19)	—	—	0.652	0.436	0.544	0.613
Social policy						
<i>A priori</i> positions	—	—	<b>2.90</b>	<b>6.68</b>	<b>14.42</b>	<b>6.84</b>
SE ( <i>n</i> = 19)	—	—	0.908	0.856	0.537	0.603
Length in words	—	—	6,345	9,768	7,322	42,446
No. of unique words	—	—	1,838	2,517	1,987	6,594
1994 virgin texts						
Economic policy						
Raw mean word scores ( $S_{vd}$ )	10.3048	10.2802	10.4459	10.5997	10.7407	10.9059
SE	0.020	0.024	0.012	0.017	0.019	0.010
Social policy						
Raw mean word scores ( $S_{vd}$ )	7.4136	7.5096	7.6076	7.9257	8.0420	7.7909
SE	0.016	0.019	0.010	0.015	0.017	0.008
Length in words	15,296	10,078	36,419	16,341	14,562	50,452
Unique words scored ( $N_v$ )	2,031	1,674	3,455	2,466	2,281	4,168
% words scored	86.7	86.8	86.1	89.8	90.2	87.0
Unique unscorable words	1,294	945	5,064	1,669	1,236	4,707
Mean frequency of unscorable words	1.57	1.41	1.40	1.18	1.16	1.39

Sources: As in previous tables. The rescaled values for PDS 1990 are in the context of the virgin scores for the non-PDS 1994 parties, using the four 1990 texts as references. This procedure has no effect on the value of the raw scores.

In Irish case, an important matter to watch for is whether the word scoring technique can pick up what is widely perceived as the classical liberal position of the FPD—on the right of the economic policy dimension and on the liberal side of the social policy dimension—a perception confirmed by the expert survey results reported in the bottom panel in Table 6. The results again suggest a general conservative shift among the establishment parties, most marked with the SDP. Language-blind word scoring also picks up the liberal positions of the FDP, putting this party on the liberal side of the social policy dimension and the right-wing side of the economic policy dimension. Again providing strong face validity for our general approach, the word-scored estimates place the PDS very firmly at the liberal end of the liberal-conservative dimension of social policy. Again, the standard errors imply that, while the position of the PDS is significantly to the left of the other parties, there is no statistically significant difference between the PDS manifesto of 1990 and that of 1994.

Overall, we take these results to show that our word scoring technique can migrate effectively into a non-English-language environment. They illustrate the enormous payoffs available from using language-blind text coding, since our technique allowed us to analyze very quickly and effectively texts written in a language that we do not speak!

## USING THE WORD SCORING TECHNIQUE TO ANALYZE LEGISLATIVE SPEECHES

Our demonstration of the word scoring technique thus far has been limited to estimating party policy positions from party manifestos. However, computer word scoring also offers the important prospect of moving into completely new areas of text analysis for which the effort involved has previously been prohibitive. Our technique removes this effort, making it possible to analyze the speeches of all members of a given legislature, for example, opening up the prospect of generating policy spaces that locate politicians in a time series and, thereby, the possibility of much more sophisticated analyses of intra- and interparty politics. Moving beyond party politics, there is no reason the technique should not be used to score texts generated by participants in any policy debate of interest, whether these are bureaucratic policy documents, the transcripts of speeches, court opinions, or international treaties and agreements.

To demonstrate the applicability of our computer word scoring technique to texts other than party manifestos, we now use word scoring to analyze legislative speeches. Although most legislatures have long preserved such speeches as part of the written parliamentary record, these speeches have become

**TABLE 7. Mean Raw and Standardized Scores of Speakers in 1991 Confidence Debate on "Pro- versus Antigovernment" Dimension, by Category of TD**

Group	N	Median		Raw		Standardized	
		Total Words	Unique Words	Mean	SD	Mean	SD
<b>Reference texts</b>							
FF Prime Minister Haughey	1	6,711	1,617	1.0000	—	—	—
FG opposition leader Bruton	1	4,375	1,181	-1.0000	—	—	—
DL leader de Rossa	1	6,226	1,536	-1.0000	—	—	—
<b>Virgin texts</b>							
FF ministers	12	3,851	727	-0.2571	0.0383	1.15	0.66
PD minister	1	2,818	593	-0.2947	—	0.50	—
FF	10	1,553	397	-0.2999	0.0721	0.41	1.24
Independent	1	3,314	582	-0.3360	—	-0.21	—
Greens	1	1,445	415	-0.3488	—	-0.43	—
WP	2	2,001	455	-0.3501	0.0423	-0.46	0.73
FG	21	1,611	394	-0.3580	0.0306	-0.59	0.53
Labour	7	2,224	475	-0.3599	0.0220	-0.62	0.38

highly amenable to computerized analysis as they are increasingly published electronically. While the analysis of speeches holds considerable promise, it also raises new challenges for content analysis—whether computerized or traditional—because such speeches differ substantially from party manifestos in several key respects. First, manifestos are typically comprehensive documents addressing a wide range of policy issues, while speeches tend to be much more restricted in focus. Second, manifestos are published in a political context that is fairly well defined. Greater care must be taken in establishing the political context of speeches if we are to justify the comparison of different speeches in the same analysis. Third, because manifestos and speeches use different language registers and lexicons, the analysis of speeches requires types of reference text different from those used in the analysis of manifestos. Finally, political speeches tend to be much shorter than manifestos. With fewer words to analyze, statistical confidence in the results is likely to be reduced. In these respects, the analysis of legislative speeches will be more problematic than the analysis of party manifestos, and therefore we expect this final test of the word scoring technique to be particularly difficult. If successful, however, we would consider it a major confirmation of the ability of our technique to extract political positions from texts using word frequencies as data.

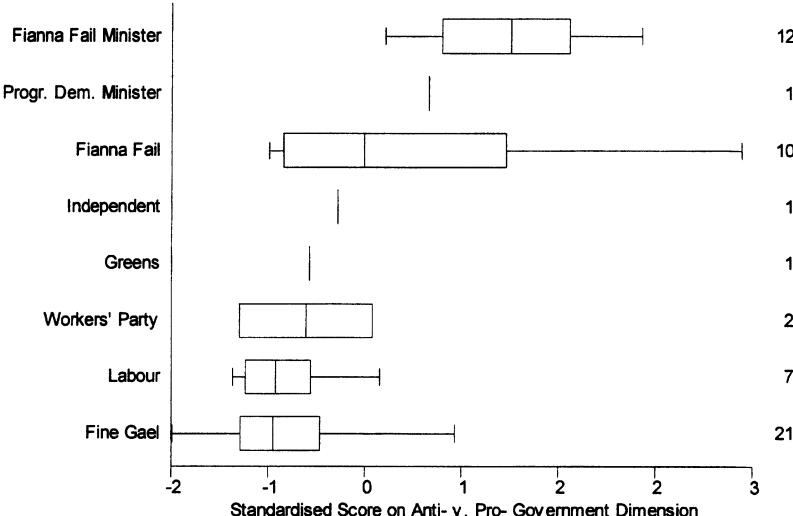
Laver and Benoit (2002) analyzed the acrimonious confidence debate in the Irish Dáil that took place in October 1991 over the future of the incumbent Fianna Fáil–PD coalition government. The matter of interest is the extent to which each individual legislator speaking in this debate was pro- or antigovernment. The texts analyzed were the set of 58 set-piece speeches, extracted from the verbatim transcript of the debate published on the web site of the Irish legislature (<http://www.irlgov.ie/oireachtas>). The tightly structured debate gave each member of the 166-member Dáil a single opportunity to speak. Including speeches by each party leader, the 58 speeches generated a written record of just over 167,000 words.

For reference positions in the debate, we postulated *a priori* the location of certain party leaders on the “pro-versus antigovernment” dimension. The speech of the *Taoiseach* (prime minister) was assumed axiomatically to be progovernment and assigned a reference position of +1.0. The speech of the Fine Gael leader of the day and leader of the opposition, John Bruton, was assumed axiomatically to be antigovernment and assigned a reference position of -1.0, as was the speech of Prionsias de Rossa, then leader of the opposition Workers’ Party. This allowed the calculation of word scores for all different words used in the debate in at least one of the reference texts—a total of 2,856 different words in all. Having calculated word scores from the reference texts, it was then possible to estimate the positions of 55 other speakers on the pro- versus antigovernment dimension, scoring their speeches as virgin texts.

Table 7 presents the results of this analysis, along with descriptive statistics about each text or group of texts from the same party. The three reference speeches were relatively long, as indicated by the median numbers of total and unique words. The virgin texts were typically short, however, with medians of 2,224 total and 508 unique words—much shorter than the typical manifesto analyzed in previous examples, which ranged in length from about 3,000 to 28,000 total words. The bottom half of Table 7 groups legislators by party and, in the case of governing Fianna Fáil and PD parties, by whether or not the legislator was a government minister. Taking the standardized scores as the most interpretable results, our expectation is that members of the coalition parties should be relatively progovernment, that government ministers should be more strongly progovernment than backbenchers, and that opposition legislators should be relatively antigovernment.

The results give us strong encouragement about the possibility of extending the use of our word scoring technique to the analysis of political speeches. Figure 3 shows these results graphically, generating a scale of “support versus opposition” to the government that is readily recognizable by any observer of Irish politics.

**FIGURE 3. Box Plot of Standardized Scores of Speakers in 1991 Confidence Debate on "Pro- versus Antigovernment" Dimension, by Category of Legislator**



Note: Values at the right indicate the number of legislators in each category.

Fianna Fáil ministers were overwhelmingly the most progovernment speakers in the debate, with Fianna Fáil TDs (members of parliament) on average less progovernment in their speeches. At the other end of the scale, Labour, Fine Gael, and Workers' Party TDs were the most systematically antigovernment in their speeches, closely followed by the sole Green TD.

Not only does the word scoring plausibly locate the party groupings, but also it yields interesting information about individual legislators, whose scores may be compared to those of the various groupings. The position of government minister and PD leader Des O'Malley, for instance (the sole PD minister in Table 7), was less staunchly progovernment than that of his typical Fianna Fáil ministerial colleagues. This may be evidence of the impending rift in the coalition, since in 1991 the PDs were shortly to leave the coalition with Fianna Fáil.

We already noted that the word scoring of relatively short speeches may generate estimates of a higher uncertainty than those for relatively longer party manifestos. This is because our approach treats words as data and reflects the greater uncertainty that arises from having fewer data. In the point estimates of the 55 individual speeches we coded as virgin texts (not shown), greater uncertainty about the scoring of a virgin text was directly represented by its associated standard error. For the raw scores (with a minimum of -0.41 and a maximum of -0.25), the standard errors of the estimates derived from speeches ranged from 0.020, for the shortest speech of 625 words, to 0.006, for the longest speech of 6,396 words, delivered by the Labour Party leader Dick Spring. These errors are indeed larger than those arising in our manifesto analy-

ses. However, substantively interesting distinctions between speakers are nonetheless possible on the basis of the resulting confidence intervals. Considering policy differences within Fine Gael, for example, the raw estimates (and 95% confidence intervals) of the positions of former FG Taoiseach Garrett Fitzgerald were -0.283 (-0.294, -0.272), while those of future party leader Enda Kenny were -0.344 (-0.361, -0.327). This allows us to conclude with some confidence that Kenny was setting out a more robustly antigovernment position in the debate than party colleague Fitzgerald. Thus even when speeches are short, our method can detect strong variations in underlying positions and permit discrimination between texts, allowing us to infer how much of the difference between two estimates is due to chance and how much to underlying patterns in the data.

Overall we consider the use of word scoring beyond the analysis of party manifestos to be a considerable success, reproducing party positions in a no-confidence debate using no more than the relative word frequencies in speeches. This also demonstrates three important features of the word scoring technique. First, in a context where independent estimates of reference scores are not available, *assuming* reference text positions using substantive local knowledge may yield promising and sensible results. Second, we demonstrate that our method quickly and effortlessly handles a large number of texts that would have presented a daunting task using traditional methods. Third, we see that the method works even when texts are relatively short and provides estimates of the increased uncertainty arising from having less data.

## CONCLUSIONS AND FUTURE WORK

Our word scoring approach to placing political texts on policy dimensions has been demonstrated to be effective at replicating the results of content analysis techniques based on human- or computer-coding. The scores produced by our technique are both substantively plausible and congruent with independent estimates—even when parties made dramatic moves on policy positions, as with the British Labour party in 1997. Furthermore, it avoids many of the problems of traditional techniques of content analysis. First, it produces policy estimates for texts whose positions are unknown, at a low cost and terrific speed—typically completing the analysis in a matter of seconds. In an analysis reported elsewhere, we were able to estimate the policy positions of the Irish political parties during the 2002 Irish general election, updating the analysis the same day each party released its election manifesto on-line (Benoit and Laver Nd.). Second, unlike traditional methods of content analysis, our technique provides quantitative measures of uncertainty for text-based estimates of policy positions. These allow analysts to make informed judgments, when comparing two estimated policy positions, about whether differences between them can be viewed as significant or merely as products of chance or measurement error—something that has not been possible before. Finally, because it treats words simply as data rather than requiring any knowledge of their meaning as used in the text, our word scoring method works irrespective of the language in which the texts are written. In other words, while our method is designed to analyze the content of a text, it is not necessary for an analyst using the technique to understand, or even read, the texts to which the technique is applied. The primary advantage of this feature is that the technique can be applied to texts in any language.

Given these advantages, the computer word scoring approach to text analysis opens up exciting possibilities for the rapid analysis and reanalysis of large and complex text data sets. As political texts become ever more easily available electronically, for example, it is now possible to analyze party manifestos and other election addresses before the election concerned has even taken place. It is worth reiterating that the great leap forward in efficiency made possible by our computational text analysis approach is made possible by a no less dramatic shift from previous applications of content analysis in political science. Our crucial move is to abandon the notion, which runs throughout most political science content analysis, that the objective of an analyst coding a text is to identify its meaning. Indeed, this notion has been so much taken for granted that it is seldom even recognized as an assumption. It is also why many early attempts to computerize content analysis within political science have in effect attempted to automate tasks otherwise performed by human experts, rather than cashing in on the things computers do really well. The results have been rather like the early robots designed in the 1960s—remarkable more because they could do anything at all than because they

actually did anything better or faster than real people. As with dictionary-based computer-coding applications, these early robots required frequent human intervention, close monitoring, and occasional direct control to make their behavior realistic. Furthermore, neither robots nor computer algorithms to analyze texts can understand meaning “in context,” something easily, if unreliably, performed by humans.<sup>17</sup> Consider an attempt to computer-code the following text: “Some say that I am not averse to the argument that it would be dangerous not to raise taxes. They have every right to say this and nobody would deny them this right, but in this case it is impossible not to conclude they are wrong.” While everyone agrees that this would be a wonderful thing to do, no published work has yet reported success at coding large volumes of political text in context, in this sense. Our approach avoids these pitfalls by circumventing them entirely, by treating individual words simply as data rather than attempting to use computerized algorithms to ascribe meaning to these words in an emulation of a human reader.

Nonetheless, sensitive to the issue of analyzing words in context while retaining our insistence on an essentially statistical method, we intend in future work to extend our approach to allow us to analyze word pairs, triples, and indeed  $n$ -tuples, as a way of taking one step toward a probabilistic analysis of the context in which individual words are located. Two comments are in order here, however. The first is the purely arithmetical point that, in a text with a total of  $m$  words, we must find  $m - 1$  word pairs,  $m - 2$  word triples, and  $m - n$  word  $n$ -tuples. In other words, the number of short word strings in a text is effectively the same as the number of words. But, if there are  $d$  different words in a given text, then there are  $d^2$  different possible word pairs and  $d^n$  different possible  $n$ -tuples. In short, the number of different possible word  $n$ -tuples increases exponentially with  $n$ , meaning that the relative frequencies of even short word strings in a text are likely to be very, very much lower than the relative frequencies of individual words. Much lower relative frequencies will combine with a much higher probability of unscoreable word strings in virgin texts, meaning that our estimates of the policy positions of the virgin texts will be more uncertain when we move from scoring individual words to scoring word  $n$ -tuples. But this will nonetheless be an interesting and important matter to explore.

The second comment on scoring short word strings concerns why our technique appears to work so well without doing this at present. As Laver and Garry (2000) point out when discussing the dictionary-based computer-coding of individual words, this almost certainly has to do with the way that words are used in

<sup>17</sup> Recall the first published reliability tests of the expert coders used by the Comparative Manifestos Project (CMP), Volkens 2001, in which a significant number of coders produced codings that correlated with an “official” coding in the 30–60% range. This is almost certainly the most professionally run and prestigious content analysis project in political science to date. We have seen no other published tests of intercoder reliability in relation to political science content analysis, but we know informally from our own experiences with this that it is a major unspoken problem.

practice in the advocacy of particular policy positions. With regard to our own technique, take the individual word used in our earlier example—"choice." Of course the word "choice" has several meanings, while each meaning can also be qualified with a negative or even a double negative. Someone coming to computational text analysis for the first time might reasonably feel for these reasons that the relative frequency of the word "choice" in a given text does not convey substantive information. This might well be true if our frame of reference were all possible texts written in the English language, read in all possible contexts, but this is very precisely not the frame of reference we propose here. For a given virgin text dealing with a given policy debate in a given political context at a given time—all of these things crucially defined by our selection of a set of reference texts—our approach works because particular words do, empirically, tend to have policy-laden content. Thus, in post-Thatcher Britain, those using the word "choice" in relation to education or health policy, for example, tended to be advocating greater choice of schools or health providers and correspondingly less central control. Those opposing such policies tended, as a matter of empirical observation, not to argue for "no choice" or "less choice" but rather to talk about the benefits of central planning and coordination. This is why the use of the word "choice" in this precise context conveyed substantive information about policy positions. Of course, if the political context changes, the information content of words may well change too—perhaps "citizens now face a stark choice and must sweep out this corrupt administration." If the context changes, however, so must the set of reference texts and hence all word scores—highlighting once more the role of the expert analyst in ensuring that the reference texts reflect in a valid way the political context of the virgin texts to be analyzed. It is patterns in the relative word frequencies observed in the reference texts that define the information content of the words to be analyzed.

In short, our technique works as well as we have shown it to work because, in practice and in a precisely defined context, individual words convey information about policy positions, information revealed in the preliminary analysis of the reference texts. Of course we will almost certainly not always be right when we apply a given word score to a given virgin text. However, provided that we are right more often than we are wrong, a function of choosing good reference texts, and provided that we analyze a large enough number of words, the slender pieces of information we extract by scoring individual words compound to allow us to make what we have shown to be valid estimates.<sup>18</sup>

Computerized word scoring offers the potential for a huge increase in the scope and power of text analysis within political science, but there is still no such thing as a methodological free lunch. While the word scoring technique automates much of the dreary and time-consuming mechanical tasks associated with traditional

text analysis, it in no way dispenses with the need for careful research design by an analyst who is an expert in the field under investigation. The key to our *a priori* approach is the identification of an appropriate set of reference texts for a given research context and the estimation or assumption of policy positions for these reference texts with which everyone can feel comfortable. This is by no means a trivial matter, since the word scores for each policy dimension, and hence all subsequent estimates relating to virgin texts, are conditioned on the selection of reference texts and their *a priori* positions on key policy dimensions. This is thus something to which a considerable amount of careful and well-informed thought must be given before any analysis gets under way. In this, our method shares the "garbage in—garbage out" characteristic of any effective method of data analysis; potential users should, indeed, be comforted by this.<sup>19</sup> The casual or ill-informed choice of reference texts or *a priori* policy positions will result in findings that are unreliable—in the same way as will the choice of inappropriate or poorly worded survey questions or an inappropriate or ambiguously defined content analysis coding scheme. Given a valid set of reference texts, however, and good estimates or assumptions of the policy positions of these, computer word scoring offers the potential to crunch huge volumes of virgin text very rapidly indeed, with an enormous range of intriguing political science applications.

## REFERENCES

- Adams, J. 2001. "A Theory of Spatial Competition with Biased Voters: Party Policies Viewed Temporally and Comparatively." *British Journal of Political Science* 31: 121–58.
- Bara, Judith. 2001. "Tracking Estimates of Public Opinion and Party Policy Intentions in Britain and the USA." In *Estimating the Policy Positions of Political Actors*, ed. Michael Laver. London: Routledge, 217–36.
- Baron, David. 1991. "A Spatial Bargaining Theory of Government Formation in Parliamentary Systems." *American Political Science Review* 85: 137–64.
- Baron, David. 1993. "Government Formation and Endogenous Parties." *American Political Science Review* 87: 34–47.
- Benoit, Kenneth, and Michael Laver Nd. "Estimating Irish Party Positions Using Computer Wordscoring: The 2002 Elections." *Irish Political Studies*.
- Blais, A., D. Blake, and S. Dion. 1993. "Do Parties Make a Difference—Parties and the Size of Government in Liberal Democracies." *American Journal of Political Science* 37: 40–62.
- Budge, Ian, and Denis Farlie. 1977. *Voting and Party Competition*. London: Wiley.
- Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tannenbaum, with Richard Fording, Derek Hearl, Hee Min Kim, Michael McDonald, and Silvia Mendes. 2001. *Mapping Policy Preferences: Parties, Electors and Governments, 1945–1998; Estimates for Parties, Electors and Governments, 1945–1998*. Oxford: Oxford University Press.
- Budge, Ian, David Robertson, and Derek Hearl, eds. 1987. *Ideology, Strategy and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press.
- Castles, Francis and Peter Mair. 1984. "Left-Right Political Scales: Some Expert Judgments." *European Journal of Political Research* 12: 83–88.

<sup>18</sup> Statistically, there is an analogy with the Condorcet Jury Theorem—if we treat individual words as jurors deciding on the policy content of texts.

<sup>19</sup> If they are not, they should consider what they would feel about a method offering "garbage in, gold out."

- de Vries, Miranda, Daniela Giannetti, and Lucy Mansergh. 2001. "Estimating Policy Positions from the Computer Coding of Political Texts: Results from Italy, The Netherlands and Ireland." In *Estimating the Policy Positions of Political Actors*, ed. Michael Laver. London: Routledge, 193–216.
- Gabel, Matthew, and John Huber. 2000. "Putting Parties in Their Place: Inferring Party Left-Right Ideological Positions from Party Manifesto Data." *American Journal of Political Science* 44: 94–103.
- Kim, H. M., and R. C. Fording. 1998. "Voter Ideology in Western Democracies, 1946–1989." *European Journal of Political Research* 33: 73–97.
- Kitchshelt, Herbert. 1994. *The Transformation of European Social Democracy*. Cambridge: Cambridge University Press.
- Kleinmuntz, Jan, and Paul Pennings. 2001. "Measurement of Party Positions on the Basis of Party Programmes, Media Coverage and Voter Perceptions." In *Estimating the Policy Positions of Political Actors*, ed. Michael Laver. London: Routledge, 161–82.
- Klingemann, Hans-Dieter, Richard Hofferbert, and Ian Budge, with Hans Keman, Torbjorn Bergman, François Pétry, and Kaare Strom. 1994. *Parties, Policies and Democracy*. Boulder, CO: Westview Press.
- Laver, Michael. 1994. "Party Policy and Cabinet Portfolios in Ireland 1992: Results from an Expert Survey." *Irish Political Studies* 9: 157–64.
- Laver, Michael. 1998a. "Party Policy in Britain, 1997: Results from an Expert Survey." *Political Studies* 46: 336–47.
- Laver, Michael. 1998b. "Party Policy in Ireland, 1997: Results from an Expert Survey." *Irish Political Studies* 13: 159–71.
- Laver, Michael, and Kenneth Benoit. 2002. "Locating TDs in Policy Spaces: Wordscore Dáil Speeches." *Irish Political Studies* 17 (Summer): 59–73.
- Laver, Michael, and Ian Budge, eds. 1992. *Party Policy and Government Coalitions*. London: Macmillan.
- Laver, Michael, and John Garry. 2000. "Estimating Policy Positions from Political Texts." *American Journal of Political Science* 44: 619–34.
- Laver, Michael, and William Ben Hunt. 1992. *Policy and Party Competition*. London: Routledge.
- Laver, Michael, and Norman Schofield. 1998. *Multiparty Government: The Politics of Coalition in Europe*. Ann Arbor: University of Michigan Press.
- McDonald, Michael D., and Silvia M. Mendes. 2001. "The Policy Space of Party Manifestos." In *Estimating the Policy Positions of Political Actors*, ed. Michael Laver. London: Routledge.
- Muller, Wolfgang, and Kaare Strom, eds. 2000. *Coalition Governments in Western Europe*. Oxford: Oxford University Press.
- Schofield, N., and R. Parks. 2000. "Nash Equilibrium in a Spatial Model of Coalition Bargaining." *Mathematical Social Science* 39: 133–74.
- Volkens, Andrea. 2001. "Manifesto Research Since 1979: From Reliability to Validity." In *Estimating the Policy Positions of Political Actors*, ed. Michael Laver. London: Routledge.
- Warwick, P. 1994. *Government Survival in Parliamentary Democracies*. Cambridge: Cambridge University Press.
- Warwick, Paul. 2001. "Coalition Policy in Parliamentary Democracies—Who Gets How Much and Why." *Comparative Political Studies* 34: 1212–36.
- Warwick, Paul. 2002. "Toward a Common Dimensionality in West European Policy Spaces." *Party Politics* 8: 101–22.

# Understanding Wordscores

Will Lowe

*Methods and Data Institute, School of Politics and International Relations,  
University of Nottingham, Nottingham, NG7 2RD, UK*  
e-mail: will.lowe@nottingham.ac.uk

Wordscores is a widely used procedure for inferring policy positions, or scores, for new documents on the basis of scores for words derived from documents with known scores. It is computationally straightforward, requires no distributional assumptions, but has unresolved practical and theoretical problems. In applications, estimated document scores are on the wrong scale and the theoretical development does not specify a statistical model, so it is unclear what assumptions the method makes about political text and how to tell whether they fit particular text analysis applications. The first part of the paper demonstrates that badly scaled document score estimates reflect deeper problems with the method. The second part shows how to understand Wordscores as an approximation to correspondence analysis which itself approximates a statistical ideal point model for words. Problems with the method are identified with the conditions under which these layers of approximation fail to ensure consistent and unbiased estimation of the parameters of the ideal point model.

## 1 Introduction

Wordscores (Benoit and Laver 2003; Laver et al. 2003) is a pioneering method of automated content analysis that assigns policy positions or “scores” to documents on the basis of word counts and known document scores via the computation of “wordscores.” The method is straightforward to implement, requires no functional or distributional assumptions, and works well in many applications (e.g., Benoit and Laver 2003; Klemmensen et al. 2007). It also has some more troubling features: estimated document scores are not directly interpretable without rescaling, and it is often unclear how best to choose a suitable rescaling method. Wordscores is also expressed directly as an algorithm rather than being derived from an underlying model. In the absence of a statistical model, it is unclear what assumptions Wordscores makes about the relationship between document scores and words, so it is difficult to tell if it will be well suited to particular political text analysis problems.

In the first half of this paper, I introduce the Wordscores algorithm, describe the problem of interpreting estimated document scores and the available rescaling solutions, and argue that no existing rescaling will work by demonstrating several fundamental problems in the method. In the second half of the paper, I show how to understand Wordscores as making classical ideal point assumptions about the relationship between document scores and words. After formulating a statistical ideal point model for words and comparing it to

---

*Author's note:* I would like to thank Ken Benoit, Mik Laver, Cees van der Eijk, and Wijbrandt van Schuur for useful comments and discussion. The remaining errors are my own.

© The Author 2008. Published by Oxford University Press on behalf of the Society for Political Methodology.  
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

existing work in political text analysis, I show how the model's structure and parameterization can avoid the problems identified in the first half of the paper. To support the claim that Wordscores approximates an ideal point model, I show first that Wordscores partly realizes an iterative method for computing a correspondence analysis and second that the parameters computed by correspondence analysis correspond to the word and document score parameters of the ideal point model. Finally, I specify the conditions under which correspondence analysis is a reasonable approximation to the ideal point model and relate problems with Wordscores to violations of particular conditions.

## 2 Wordscores

Given  $R$  documents or “reference texts” with known positions or scores on a policy dimension, Wordscores attempts to estimate the scores of  $L$  out-of-sample documents, the “virgin texts.” To do so the method first estimates scores for each word type occurring in the reference texts and then combines these wordscores into a score for each virgin document. It is important to distinguish the two parts: estimating wordscores and estimating document scores using wordscores because they are, at least in principle, independent parts of the method. There is usually a third and final part of the method that rescales virgin document score estimates, so they can be more easily compared with the reference text scores.

Although the Wordscores algorithm is not explicitly derived from any statistical model of word generation, many aspects of the method can support such interpretations. In particular, the methods for assigning scores to words and documents have a symmetric probabilistic interpretation.

### 2.1 Estimating Scores for Documents

Wordscores computes the estimated score for a document  $d$ ,  $\hat{\theta}_d$ , as the average of the scores,  $\pi_w$ , of the words contained in it<sup>1</sup>. When  $V$  types of words appear in a collection of reference texts and there are  $W$  word tokens in  $d$ ,

$$\hat{\theta}_d = \frac{1}{W} \sum_w^W \hat{\pi}_w \quad (1)$$

$$= \sum_j^V \hat{\pi}_j \hat{P}(w_j|d). \quad (2)$$

The weighting probability is estimated by the proportion of tokens of each word type in the reference documents

$$\hat{P}(w_j|d) = \frac{c(w_j \text{ in } d)}{c(d)}, \quad (3)$$

where  $c(\cdot)$  is the word token counting function. Laver, Benoit, and Garry (LBG) suggest

$$\hat{\sigma}^2(\hat{\theta}) = \sum_j^V [\hat{\pi}_j - \hat{\theta}]^2 \hat{P}(w_j|d) \quad (4)$$

as an estimator for the variance of this document score estimate, although it ignores sampling variation in  $\hat{P}$ .

---

<sup>1</sup>In fact, LBG assume only that virgin document scores are predicted by the average of the scores of their words, but since there is nothing special about virgin documents except our ignorance of their scores, this assumption must apply to all documents or it is impossible to explain why document scoring works.

Equation (1) reflects the assumption that each observed word token provides the same amount of information about the document's score. Equation (2) emphasizes the same point at the level of word types: frequency is assumed to be a direct reflection of a word type's importance in determining a document's score. From this development, it is natural to assume that the true document score is (hats removed) simply

$$\theta_d = \sum_j^V \pi_j P(w_j | d). \quad (5)$$

## 2.2 Estimating Scores for Words

Wordscores computes the score for word  $w$ ,  $\pi_w$ , as an average of document scores, weighted by the posterior probability of each document given that  $w$  occurs within it. When there are  $R$  documents, the word can appear in

$$\hat{\pi}_w = \sum_r^R \theta_r \hat{P}(d_r | w). \quad (6)$$

Posterior probabilities are computed in the following way: The probability of seeing  $w$  given that we are reading document  $i$  is given by equation (3). Assume that the prior probabilities of each reference document are equal, so  $P(d_i) = 1/R$ . The posterior probability of reading  $d_i$  after having seen  $w$  is then estimated as

$$\begin{aligned} \hat{P}(d_i | w) &= \frac{\hat{P}(w | d_i) P(d_i)}{\sum_r^R \hat{P}(w | d_r) P(d_r)} \\ &= \frac{\hat{P}(w | d_i)}{\sum_r^R \hat{P}(w | d_r)} \\ &= \frac{c(w \text{ in } d_i) / c(d_i)}{\sum_r^R c(w \text{ in } d_r) / c(d_r)}. \end{aligned} \quad (7)$$

Note that under this interpretation,  $P(d_i)$  is a *prior* probability, so it would be inappropriate to estimate it from data, for example as  $c(d_i) / \sum_r^R c(d_r)$ . This is not because of Bayesian scruples but because the words in each document are sampled *conditional* on policy position, and for political text we know that the decision about what score to express and therefore what words to generate is not random but strategic and not explicitly modeled.

Although LBG do not make use of information about sampling variation in wordscore estimates, the document scoring procedure suggests that

$$\hat{\sigma}^2(\hat{\pi}) = \sum_r^R [\hat{\theta}_r - \hat{\pi}]^2 \hat{P}(d_r | w) \quad (8)$$

might be a reasonable estimator.

This development suggests that equation (6) should be understood as an estimate of the true wordscore, defined as

$$\pi_w = \sum_r^R \theta_r P(d_r | w). \quad (9)$$

**Table 1** Mapping between LBG's notation and the notation used in this paper

Paper	LBG
$\hat{P}(w d_r)$	$F_{wr}$
$\hat{P}(d_r w)$	$P_{wr}$
$\hat{\pi}_w$	$S_{wd}$
$\hat{\theta}_w$	$S_{vd}$

Table 1 connects LBG's notation to the notation used in this paper. This is necessarily a partial mapping because LBG do not provide a way to distinguish population values from their estimates. Note also that all references to the dimension being scored, for example the  $d$  in  $S_{vd}$ , have been suppressed in order to focus attention on the estimation process. In this paper,  $d$  is used instead to refer to documents.

### 3 Problems with Document Scores

In applications, estimated document scores invariably have a *much* smaller variance than reference document scores and are bunched around  $\bar{\theta}$ , the mean of the reference document scores. For example, in LBG's U.K. party manifestos data the sample variance of the known scores is approximately 500 times larger than the estimated scores. This makes it difficult to compare estimated document scores with reference scores (Laver et al. 2003; Martin and Vanberg 2007), although Benoit and Laver (2007) have countered that raw estimated scores are nevertheless interpretable relative to each other. In an attempt to make the scores of virgin documents interpretable on the same scale as reference texts, two methods have been proposed for rescaling virgin document score estimates.

#### 3.1 Rescaling Document Scores

LBG (2003) transform  $\hat{\theta}$  into the more interpretable  $\tilde{\theta}$  according to

$$\tilde{\theta} = [\hat{\theta} - \bar{\theta}_{vir}]T + \bar{\theta}_{vir},$$

where  $T$  is the ratio of standard deviations of the reference and virgin document scores and  $\bar{\theta}_{vir}$  is the average of estimated virgin document scores. This rescales virgin document scores to have the same variance as the original reference scores. Consequently, it only applies when more than one virgin document is to be scored.

LBG's transformation reflects an implicit assumption that the distribution of estimated document scores has the correct mean but the incorrect variance. This is problematic because in applications, the virgin score predicted mean is invariably close to  $\bar{\theta}$  regardless of which virgin scores are scored, a shrinkage effect that is analyzed in more detail below.

In applications where the mean and variance of document scores can be expected to be approximately constant, the LBG transformation is very natural. LBG's empirical examples are panels of party manifestos, where it may be reasonable to expect policy position variance across elections to be stable. It is nevertheless worth noting the substantive implications of their transformation. Joint rightward or leftward movement of a set of parties relative to their positions in the previous election will be hard to discern because the mean of the virgin score estimates will always be close to that of the reference scores while the variance is not affected. Likewise an expansion of party positions to more extreme locations or increasing polarization in a legislature will also be masked.

Martin and Vanberg (2007) have suggested the alternative rescaling transformation

$$\tilde{\theta} = [\hat{\theta} - \hat{\theta}_a]T + \hat{\theta}_a.$$

Here,  $T = [\hat{\theta}_a - \hat{\theta}_b]/[\theta_a - \theta_b]$ , where  $a$  and  $b$  index the reference documents with the lowest and highest scores, respectively. This transformation ensures that  $\tilde{\theta}_a = \theta_a$  and  $\tilde{\theta}_b = \theta_b$  and is most natural when these two anchoring documents can be identified on substantive grounds.

The transformation of Martin and Vanberg (MV) is valuable because it focuses attention on the important question of consistency. However, their notion of consistency is both very strong— $\theta_a$  and  $\theta_b$  will be recovered exactly—and limited because it holds only for two documents. It may be better understood as enforcing a limited form of unbiasedness: samples of any size will always yield two correct estimates although the bias in remaining document score estimates will be unknown.

Practically, both the LBG and the MV rescaling operations are linear in  $T$ , so the variance of the document score estimates is  $\delta^2(\hat{\theta})T^2$  with a corresponding correction to standard errors. But which rescaling transformation *should* be used?

Both transformations are derived from reasonable and general principles and yet yield different results in applications. In particular, the results of LBG (2003) cannot be replicated using the MV transform. In the light of these difficulties, we might try to avoid a decision by following LBG's suggestion to interpret untransformed scores. But although estimating scores for the reference texts as if they were virgin texts does successfully put all documents in the same scale, their relative positions on this scale do not replicate LBG's original analysis either.

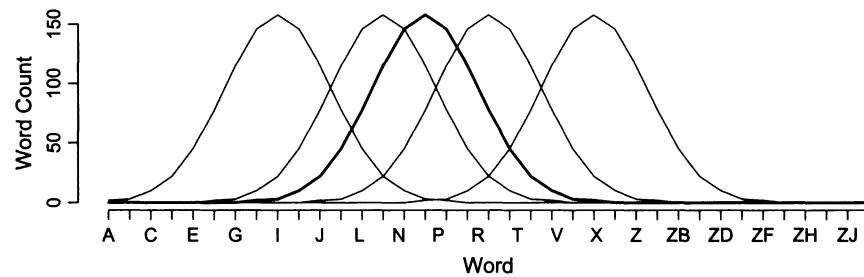
A more basic problem with these transformations is that they shift the sensitivity of Wordscores output to different documents rather than removing it. LBG's transformation is insensitive to the composition of the reference document set but makes an estimated virgin document score depend on the composition of the virgin document set via its sample standard deviation. In contrast, MV's transformation is indifferent to the composition of the virgin document set but sensitive to the choice of anchoring reference texts. To get more insight, it is necessary to look more closely at the component processes of word and document score estimation.

## 4 A Closer Look at Score Estimation

Wordscores consists of two separate processes: the estimation of document scores from wordscores and the estimation of wordscores from document word counts. Introducing rescaling transformations is an attempt to fix their joint output but will not remedy any of the more basic problems with the component processes described below.

### 4.1 Problems Scoring Documents

The method for combining wordscores—averaging—and the method for transforming them into more interpretable forms—rescaling—are both linear, so they could be combined into one process to replace the process of averaging the wordscores of a virgin document with something more complex. But transforming output according to any linear mapping ignores a fundamental issue noted above: averaging wordscores to estimate a document's score implies that each word adds the same amount of information about the document.



**Fig. 1** Distribution of word counts for each document in the example data. These data are taken from Table 1 of Laver et al. (2003). Word count profiles for reference texts R1, R2, R3, R4, and R5 are plotted from left to right with the profile of virgin text V1 superimposed as a darker line.

In real text this is almost certainly false. Words like “taxes” are informative about economic policy in a way that words like “the” are not. However, Wordscores has no way to represent the difference between a genuinely informative politically centrist word—one that is used preferentially by center parties to denote centrist policy positions—and a word that all documents contain in roughly equal numbers for functional linguistic rather than political reasons. The problem is that if document scores are spread evenly across a policy dimension, then centrist words and politically uninformative words will both have wordscores close to  $\bar{\theta}_{\text{ref}}$ . Centrist words get centrist wordscores because  $P(d|w)$  puts more weight on documents with scores near  $\bar{\theta}_{\text{ref}}$  and uninformative words get centrist wordscores because indifference to policy position implies  $P(d|w) \approx 1/R$ , so the resulting wordscore is simply an average of the reference scores.

It is easy to see that the larger the number of word tokens with scores close to  $\bar{\theta}$ , the greater the movement of all estimated document scores toward  $\bar{\theta}$ . This effect is appropriate when all words are equally informative but overstated when there are also words with scores close to  $\bar{\theta}$  simply because  $P(d|w) \approx 1/R$ . No linear transformation of document scores will fix this problem because it is the wordscore averaging process that is at fault. An ideal procedure would have to generate correspondingly extreme scores for just these informative words to offset the bias.

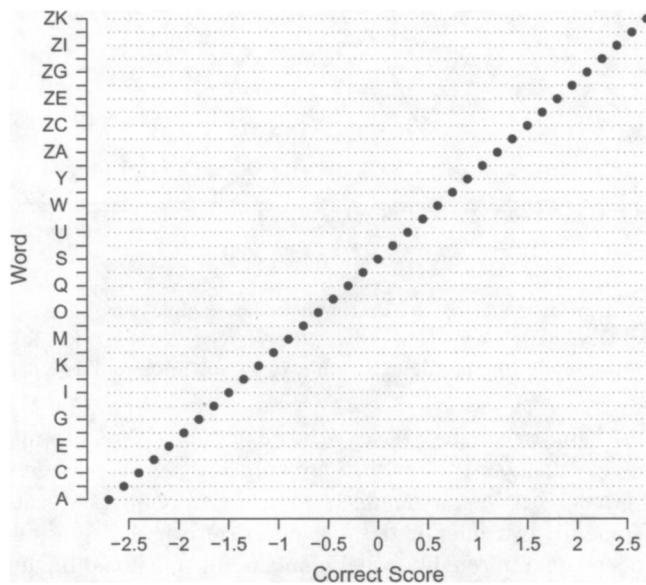
It is theoretically possible that a document could be scored as *more* extreme than any reference document, provided it is similarly constructed but contains a higher proportion of extremely scored words than any of the reference documents. However, in applications this effect will be swamped by the shrinkage due to uninformative words and by the effect of biases in wordscore estimation discussed next.

#### 4.2 Problems Scoring Words

Problems relating to differing word frequency and informativeness will cause some of the shrinkage toward  $\bar{\theta}$  that makes document score estimates so hard to interpret. But there are also problems with the method of assigning scores to words. These are most easily demonstrated using LBG’s example data.

In LBG’s example, there are  $V = 37$  word types available, spread across 6 pseudo-documents containing 1000 word tokens each. These “words,” shown in Fig. 1, are the 26 letters of the alphabet and the first 11 letters of the alphabet prefixed by the letter<sup>2</sup> Z.

<sup>2</sup>This is not quite the same as the paper, but since the words are arbitrary nothing depends upon it.



**Fig. 2** Correct scores for words in the example data.

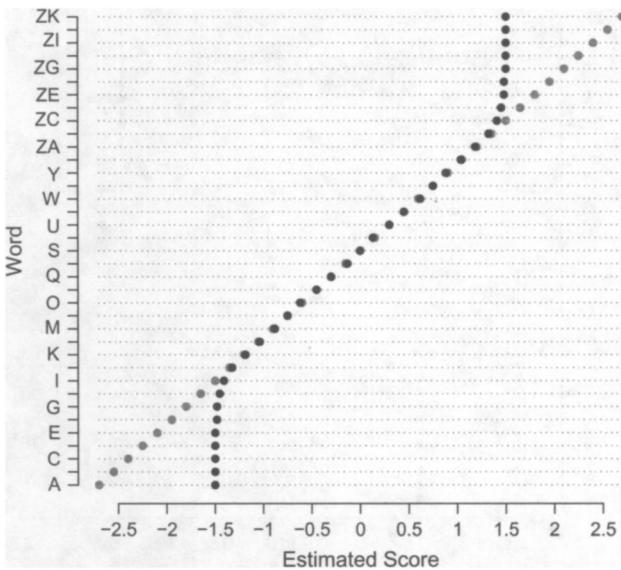
The reference documents R1, R2, R3, R4, and R5 are assigned scores of  $-1.5, -0.75, 0, 0.75$ , and  $1.5$ , respectively. The task is to estimate the score of virgin text V1, whose word count distribution over the vocabulary is shown as a dark line in the figure.

From the word frequencies and document scores in the example, it is straightforward to identify a set of wordscores that fit the data perfectly: that is, scores that not only assign  $\hat{\theta}_{V1} = -0.45$  without transformation but also score the reference documents correctly. Such scores are consistent in MV's sense, except that consistency holds over the complete document set. The scores, shown in Fig. 2, start at  $-2.7$  for word "A" and increase in increments of  $0.15$  until they reach  $2.7$  for word "ZK."

In contrast, Fig. 3 shows scores estimated according to the standard Wordscores method. These agree with the correct scores where there is large amount of overlapping word frequency data but diverge at the edges where overlap decreases. In extreme cases, any word  $w$  that is unique to  $d$  has  $\pi_w = \theta_d$  because  $\hat{P}(d|w) = 1$  leading to strongly biased wordscore estimates. Unfortunately, such words are prevalent in real data. In LBG's 1992 U.K. party manifestos, approximately one half of all word types occurred in only one of the three manifestos.

In Fig. 3, the difference between the Wordscores estimates and correct scores appears only at the edges of the score distribution, but it is easy to show that this is not the only place it can arise. LBG's example uses a relatively large number of reference documents with scores that evenly span the range of possible document scores. Figure 4 shows the result of recomputing wordscores after removing the second and fourth reference documents. Here, only five estimated wordscores agree with the correct set of wordscores shown on the diagonal. The characteristic pattern of estimated scores in Fig. 4 will be familiar to users of the Wordscores because it appears in many data sets. For comparison, Fig. 5 shows the sorted economic dimension wordscores for the 1992 U.K. election manifestos. The vertical bands correspond to words that occur in only one manifesto.

There is nothing inherently wrong with assigning the same score to all the words that occur in only one document. This is because LBG define wordscores implicitly as any



**Fig. 3** Scores estimated according to the Wordscores method for the example data. Correct scores are marked in gray.

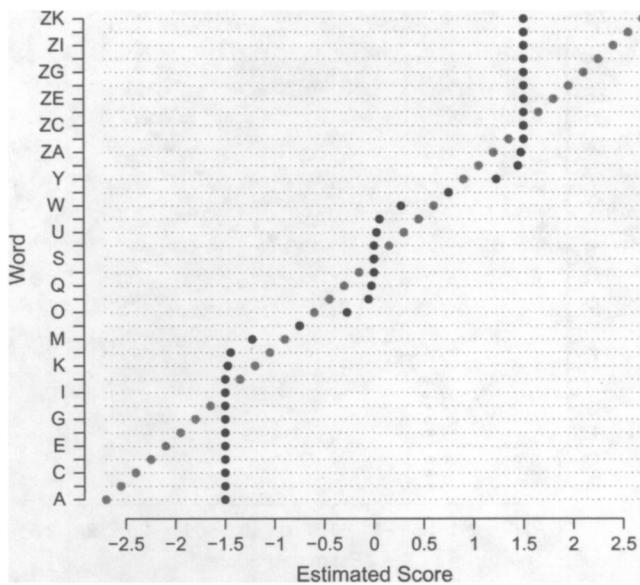
assignment of numbers to words that obeys the  $R$  sum constraints generated by equation (5), the reference document word counts, and their scores. The “correct” wordscores described above are therefore not unique because  $R$  constraints will not typically identify  $V$  wordscores<sup>3</sup>. Among the possible sets of correct wordscores, there will always be assignments where all words unique to a document are given the same score. The problem is that Wordscores will not in general assign these scores correctly. As an example, consider words A to E that are unique to document R1. Manual calculation shows that *any* assignment of wordscores that, when multiplied by the conditional probabilities of these words, generates the value  $-0.181$  will obey the  $R = 5$  sum constraints imposed by equation (5) and thus yield correct virgin and reference score estimates. So, if words A to E are to be assigned a single wordscore  $s$ , it should be the solution to

$$-0.18105 = 0.002s + 0.003s + 0.01s + 0.022s + 0.045s,$$

which is  $-2.208$ . This is not only quite different from the Wordscores estimate of  $-1.5$  but also *larger* than any reference document score. It is therefore an example of the need to assign more extreme wordscores to offset shrinkage toward  $\bar{\theta}$ .

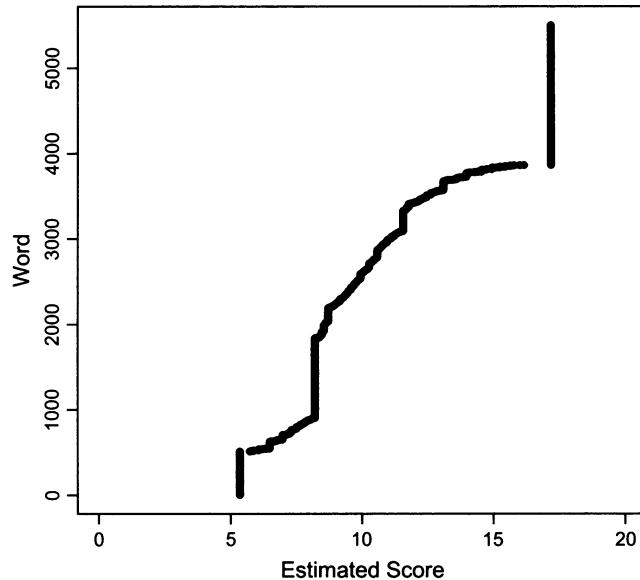
This example points to another fundamental problem with the Wordscores method of estimating scores for words and documents: equation (9) ensures that no wordscore can be *more* extreme than any of the document scores. However, if no wordscore can be more extreme than the lowest (or highest) document score, then that document score *cannot* be the average of the scored words within it as is required by equation (5). The method of generating wordscores is therefore incompatible with the method of scoring documents.

<sup>3</sup>For example, a trivial set of wordscores that estimates all documents correctly is to assign words I, N, P, S, X, and ZC the scores  $-9.49367$ ,  $-4.746835$ ,  $-2.848101$ ,  $0$ ,  $4.746835$ , and  $-9.49367$ , respectively (the scores for documents R1, R2, V1, R3, R4, and R5 multiplied by each word’s conditional probability 0.158) and all other words the score 0.



**Fig. 4** Wordscores computed without reference documents R2 and R4 with scores  $-0.75$  and  $0.75$ .

In the example, the scores of the words appearing in the virgin document are well estimated because they are well within the range of the reference scores, so the virgin document score is itself well estimated. But where overlap is weaker, wordscore estimates can be strikingly biased. This is particularly problematic in applications where there are often



**Fig. 5** Wordscores computed for the 1992 U.K. Labour, Liberal Democrat, and Conservative party manifestos on the economic dimension. Vertical bands of words with identical scores appear at the reference document scores  $5.35$ ,  $8.21$ , and  $17.2$ , respectively. Compare with Fig. 4.

only two available reference texts, chosen to have maximally different policy positions, a choice that minimizes word overlap and risks biased wordscore estimates.

In summary, there are several fundamental problems with Wordscores' method of estimating scores for documents and words: first, the method has no way to distinguish the effects of word frequency and informativeness causing estimates to shrink toward apparently centrist policy positions. Second, it generates systematically biased wordscore estimates when there is insufficient overlap of word distributions across reference documents. Third, the assumptions of document scoring method are incompatible with the assumptions of the wordscore estimation method. That these problems can be demonstrated in example data that contain no error suggests they are basic to the method.

Despite these problems in the method, Wordscores can work well in applications, so at least some of the assumptions built into Wordscores must be appropriate to political text. These assumptions are best made explicit in a statistical model of the word generation process.

## 5 A Probability Model for Wordscores

The basic problem understanding Wordscores is that it appears to make no assumptions about the functional or distributional form of the mechanism by which words are generated from documents with particular scores. Wordscores estimates  $P(w|d_i)$ , irrespective of document scores, and then essentially asserts that the score of  $d_i$  on some policy dimension is  $\theta_i$ . What is needed instead is an explicit form for  $P(w|\theta)$ , parameterized in a way that reflects Wordscores' assumptions about word generation and any *a priori* knowledge about the scores of particular documents.

A parametric form for  $P(w|\theta)$  helps solve problems in both document and word scoring. For document scoring, adding explicit parameters representing how frequent and how informative a word is about policy position allows estimates of document scores to reflect the relative information available in each word count, rather than relying on averaging. Defining  $P(w|\theta)$  also addresses the overlapping word count problem. Wordscores assigns words unique to a document the score of that document because  $P(d_i|w) = 1$ . However, with an explicit and reasonably smooth functional form for  $P(w|\theta)$  the posterior distribution need not collapse to 1 over a single reference document score. Nearby possible scores will also affect the posterior via  $P(w|\theta + \delta)$ , despite the fact that no documents with a score of  $\theta + \delta$  exist in the reference document set. Wordscores has no functional form to lean on between observed data points, so  $P(w|\theta + \delta)$  is undefined.

### 5.1 A Functional Form for Wordscores

The word count distributions shown in Fig. 1 suggest that  $P(w|\theta)$  should be a unimodal function centered on the document score. The Wordscores algorithm also supports this interpretation: Wordscores consists of two symmetrical weighted averaging procedures, the first to estimate scores for words and the second to estimate scores for documents. This averaging process has the effect of moving each  $\pi_w$  toward the center of the distribution of  $\theta$  for the documents that contain  $w$  and to move each  $\theta_d$  to the center of the distribution of  $\pi$  for the words in  $d$ . This process makes sense if Wordscores is in fact a classical ideal point model for words (Enelow and Hinich 1984).

With this intuition, it is possible to formulate a generalized linear latent variable model of the relationship between policy position and word generation (Bartholomew 1984; Elff 2008) in the same statistical framework as roll call voting analysis (Jackman 2001; Baker

and Kim 2004; Clinton et al. 2004). One simple model is that word counts are Poisson distributed with mean

$$\log E[w] = c_w - \frac{1}{2} \frac{(\pi_w - \theta)^2}{\tau_w^2}, \quad (10)$$

where  $c_w$  is the maximum probability of seeing word  $w$  and  $\tau_w > 0$  is an informativeness term representing the rate of decrease in the word probability as  $\pi_w$  moves away from  $\theta$ . This function has a maximum when  $\pi_w = \theta$ . In applications, all the models in this section will also require an offset to control for varying document length, suppressed here for clarity.

Intuitively, words with large  $c_w$  occur more frequently than words with small  $c_w$  in all documents, regardless of their policy position. Words with small  $\tau_w$  are specific to a region of policy space around  $\pi_w$  and tend to appear only when documents express positions in this region. In contrast, the probability of seeing words with large  $\tau_w$  does not depend strongly on the policy position of the document that contains them.  $\tau_w$  therefore distinguishes between words that are frequent in all documents for functional linguistic reasons and those that are only frequent in documents expressing a centrist policy position. Both may have values of  $\pi_w$  toward the center and large  $c_w$ , but the former will have large and the latter small values of  $\tau_w$ .

It is helpful to compare equation (10) with the models put forward by Monroe and Maeda (2004) and Slapin and Proksch (2008). In the simplest of these

$$\log E[w] = c_w - \beta_w \theta, \quad (11)$$

where  $c_w$  represents the word probability when  $\theta = 0$  and  $\beta_w$  represents the sensitivity of  $w$ 's occurrence probability to changes in document policy position. Words that occur often in documents of all positions will have large  $c_w$  but small  $\beta_w$ . Like  $\tau$  in equation (10),  $\beta_w$  distinguishes informative from uninformative words.

In contrast to the previous model, equation (11) has no parameters that can be identified as wordscores; words do not have a preferred position ( $\pi_w$ ) in policy space, only a sensitivity ( $\beta_w$ ) to changes in document policy position. Consequently, it cannot represent words that are used primarily to express centrist policy positions. Whether this is problematic depends on whether there are such words in political language, an empirical question that can only be answered by model comparison using real data.

Equations (10) and (11) are special cases of a quadratic model

$$\log E[w] = b_0 + b_1 \theta + b_2 \theta^2. \quad (12)$$

To translate back to equation (10), let  $\pi_w = b_1/2b_2$ ,  $\tau = 1/\sqrt{-2b_2}$ , and  $b_2 < 0$  (to ensure a peak at  $\theta = \pi_w$ ). To recover equation (11), let  $b_2 = 0$  (ter Braak and Looman 1986).

Equation (10) offers the possibility of correcting at least some of the problems of word frequency, informativeness, and limited word overlap uncovered above. But what reason is there to think of Wordscores in ideal point terms? The next sections show how Wordscores is related to the ideal point model in equation (10) via the method of correspondence analysis.

## 6 Wordscores as Correspondence Analysis

Correspondence analysis (Greenacre 1993) is a method for extracting latent variables from a contingency table that has often been applied to linguistic data (Benzécri 1992). Given

a  $V \times R$  matrix  $\mathbf{C}$  where  $\mathbf{C}_{wd} = c(w \text{ in } d)$ , a one-dimensional correspondence analysis associates a number  $\theta$  with each column and a number  $\pi$  with each row such that the two sets of numbers have maximal correlation<sup>4</sup>. Correspondence analysis is usually presented as the result of an eigen decomposition, but for the purposes of understanding Wordscores it is more useful to note that the same solutions can be found using a simple iterative algorithm known as weighted or reciprocal averaging<sup>5</sup> (Hill 1973, 1974).

### 6.1 Reciprocal Averaging

The reciprocal averaging algorithm for a one-dimensional correspondence analysis starts by randomly choosing  $\hat{\theta}_1, \dots, \hat{\theta}_R$ . It then computes

$$\hat{\pi}_w = \frac{\sum_d^R \mathbf{C}_{wd} \hat{\theta}_d}{\sum_d^R \mathbf{C}_{wd}} \quad (13)$$

for each row,

$$\hat{\theta}_d = \frac{\sum_w^V \mathbf{C}_{wd} \hat{\pi}_w}{\sum_w^V \mathbf{C}_{wd}} \quad (14)$$

for each column, and then normalizes the estimates of  $\theta$ . Normalization is necessary to prevent repeatedly averaged quantities converging to a single value and is typically implemented by fixing the mean and variance of  $\hat{\theta}_1, \dots, \hat{\theta}_R$  at each iteration. Equations (13) and (14) and the normalization step are repeated until parameter changes are sufficiently small (ter Braak and Prentice 2004).

If columns of  $\mathbf{C}$  contain compositional data, then  $\sum_w^V \mathbf{C}_{wd} = 1$  and equation (14) reduces to

$$\hat{\theta}_d = \sum_w^V \mathbf{C}_{wd} \hat{\pi}_d. \quad (15)$$

To see the connection to Wordscores, note that the normalized word counts in equation (3) are compositional data of this type. Equation (15) is therefore identical to the document score estimator in equation (3) and equation (13) is identical to the wordscore estimator in equation (6), with conditional probability estimated as in equation (7). The final normalization step is performed in the Wordscores algorithm either by fixing  $R$  reference document scores and the (virgin) document score variance (the LBG method) or by anchoring with two documents (the MV method). Either is sufficient to identify the model and to prevent a degenerate solution.

Wordscores does not realize exactly this algorithm: initial document scores are not chosen randomly, the virgin texts are treated as out of sample, and Wordscores performs each step of the reciprocal averaging algorithm only once. But these are minor differences compared to the similarities in symmetrical structure. Wordscores is thus a single-step approximation to the reciprocal averaging algorithm for correspondence analysis.

<sup>4</sup>Many other criteria also lead to correspondence analysis as a matrix decomposition. Beh (2004) provides an extensive review.

<sup>5</sup>I would like to thank Wijbrandt van Schuur for suggesting this connection.

Interpreting Wordscores as correspondence analysis connects the method to a well-developed statistical literature and also opens the possibility of extracting more than one latent dimension from data. However, it does not by itself help us understand what assumptions Wordscores makes about word generation since it is defined either algorithmically or in terms of a matrix decomposition rather than as probability model. However, correspondence analysis can be shown to be closely related to the ideal point model in equation (10).

## 7 Correspondence Analysis as Ideal Point Estimation

The relationship between correspondence analysis and ideal point estimation can be illuminated by looking at the maximum likelihood equations of  $\theta_d$  and  $\pi_w$  in equation (10). These are (ter Braak 1985)

$$\theta_d = \sum_w^V \frac{\mathbf{C}_{wd}\pi_w}{\tau_w^2} \Bigg/ \sum_w^V \frac{\mathbf{C}_{wd}}{\tau_w^2} - \left[ \sum_w^V \frac{(\theta_d - \pi_w)\mathbf{E}[w]}{\tau_w^2} \Bigg/ \sum_w^V \frac{\mathbf{C}_{wd}}{\tau_w^2} \right] \quad (16)$$

$$\pi_w = \frac{\sum_d^R \mathbf{C}_{wd}\theta_d}{\sum_d^R \mathbf{C}_{wd}} - \left[ \frac{\sum_d^R (\theta_d - \pi_w)\mathbf{E}[w]}{\sum_d^R \mathbf{C}_{wd}} \right]. \quad (17)$$

If all words are equally informative, then  $\tau$  cancels and the first terms on the right hand sides of equations (16) and (17) are the correspondence analysis recursions in equations (13) and (14).

Even if  $\tau$  is shared, the terms in square brackets distinguish these equations from the correspondence analysis recursions. These terms will be small under two circumstances: when word probabilities are small and when word probabilities decrease in proportion to the distance between  $\theta$  and  $\pi$ , that is, when words are generated according to the ideal point model in equation (10). The correspondence analysis recursions can be therefore be seen as approximations to the maximum likelihood equations for the parameters of an ideal point model for words.

To summarize the argument: Wordscores approximates a correspondence analysis because it performs a single iteration of the reciprocal averaging algorithm. And the reciprocal averaging algorithm for correspondence analysis approximates the maximum likelihood equations for an ideal point model because it ignores the bracketed terms in equations (16) and (17) and assumes a constant  $\tau$ . How good are these approximations?

### 7.1 Inconsistency and Bias

In general, correspondence analysis estimators of word and document scores will be inconsistent. In the maximum likelihood estimation context, ideal point analysis practitioners are familiar with the “incidental parameter problem” that prevents consistent estimation of equation (10) or (11) for latent fixed effects  $\theta$  when  $V$  is fixed and  $R \rightarrow \infty$ . Lynn and McCulloch (2000) prove the stronger result that  $\pi$  cannot be consistently estimated by correspondence analysis estimators in the same limit whether  $\theta$  is treated as fixed or random.

Despite these inconsistency results, correspondence analysis and therefore Wordscores estimates of  $\theta$  and  $\pi$  can under certain conditions correlate highly with their true values in finite samples (Lynn and McCulloch 2000). ter Braak provides a useful summary of the conditions under which bias (the bracketed terms in equations (16) and (17)) is minimized:

1.  $\theta$  are equally spaced and extend over the whole range of  $\pi$ .
2.  $\theta$  are closely spaced relative to  $\tau$ .
3.  $\pi$  are equally spaced and extend past each  $\theta$  in both directions.
4.  $\pi$  are closely spaced relative to  $\tau$ .
5.  $\tau_w$  is the same for all words.
6.  $c_w$  is the same for all words.

Assuming that equation (10) holds, these conditions describe the kind of text analysis problems in which Wordscores can be expected to work well. Conversely, violations correspond to the problems identified in previous sections and summarized below.

Wordscores treats all words as equally informative, providing no way to distinguish politically uninformative from centrist words or discount words that occur more frequently than others for linguistic rather than political reasons. Conditions 5 and 6 will therefore *never* hold for word count data because text exhibits highly skewed word frequency distributions regardless of genre (Zipf 1949; Mandelbrot 1954) and inevitably contains many uninformative words. Indeed, Fig. 2 of Slapin and Proksch (2008) supports the intuition that a word's utility in distinguishing policy position is not only quite variable but also inversely correlated with its frequency.

The second problem is that Wordscores generates biased wordscore estimates when there is insufficient overlap of word distributions between reference documents, as required by conditions 1 and 2. Conditions 3 and 4 remind us of a symmetrical problem in the document score estimation process when wordscores are inappropriately distributed. However, since words are more plentiful than documents this aspect of the insufficient overlap problem has less practical importance.

The third problem is that the assumptions of document and wordscore estimation used in Wordscores are incompatible. There would be no incompatibility if conditions 1 and 3 could hold simultaneously. However, that is impossible for any finite data set. Bias in wordscores, document scores, or both is therefore inevitable if correspondence analysis or Wordscores is used as an estimator.

There are typically many word types in a speech or political manifesto, so we might hope that they may relatively evenly spread out across a policy dimension. Then conditions 3 and 4 may be plausible. When many documents with known scores are also available, for example the speeches of a large number of legislators with varying and evenly distributed policy positions, conditions 1 and 2 might also be well approximated. Word and document scores should then be well estimated, except for those at the edges of the policy space because of the incompatibility problem. This is perhaps the best class of political text analysis problems for Wordscores. When there are very few documents with known scores, for example when analyzing party manifestos, then condition 2 will not hold even approximately. Large bias will therefore appear in wordscore estimates that will compromise new document score estimates.

If the parameters of equation (10) are estimated directly, for example by maximum likelihood or inferred using Bayesian methods rather than via the correspondence analysis or Wordscores approximations, then these biases should disappear. The empirical validation of this assertion is future work.

## 8 Conclusion

I have argued that Wordscores algorithm's computational straightforwardness, apparent absence of functional or distributional assumptions, and empirical effectiveness hide a number

of fundamental problems that are not solved by the rescaling transformations suggested in the literature. In order to address these problems, it is necessary to *understand* Wordscores. Understanding Wordscores involves determining what the method implicitly assumes about political text, particularly about the relationship between document policy positions and words. I have argued that Wordscores reflects an ideal point model for words and justified the claim in two steps. First, by showing that Wordscores is a partial implementation of the reciprocal averaging algorithm for correspondence analysis, and second that there is a close relationship between correspondence analysis and maximum likelihood estimation of some ideal point model parameters. To the extent that words in political text are generated according to an ideal point structure such as equation (10) rather than, for example, a factor structure like equation (11), Wordscores should be an effective method of inferring policy positions from documents. The empirical success of the method suggests that these assumptions may be reasonable. Conversely, I argued that although correspondence analysis, and therefore also Wordscores, will in general be inconsistent as an ideal point estimator, the nature and extent of its approximation to the ideal point model will determine the degree to which it will be biased in applications. In order to determine when Wordscores should work well, I list the conditions under which these biases can be expected to be small.

### Funding

Enterprise Ireland (PC/2003/147).

### References

- Baker, F., and S. H. Kim. 2004. *Item response theory*. 2nd ed. New York: Wiley.
- Bartholomew, D. J. 1984. *Latent variable models and factor analysis*. Vol. 40. London: Charles Griffin and Company Limited.
- Beh, E. J. 2004. Simple correspondence analysis: A bibliographic review. *International Statistical Review* 72:257–84.
- Benoit, K., and M. Laver. 2003. Estimating Irish party positions using computer wordscoring: The 2002 elections. *Irish Political Studies* 17:97–107.
- Benoit, K., and M. Laver. 2008. Compared to what? A comment on “A robust transformation procedure for interpreting political text” by Martin and Vanberg. *Political Analysis* 16:101–11.
- Benzécri, J.-P. 1992. *Correspondence analysis handbook*. New York: Marcel Dekker.
- Clinton, J., S. Jackman, and D. Rivers. 2004. The statistical analysis of roll call voting: A unified approach. *American Journal of Political Science* 98:355–70.
- Elff, M. 2008. *A spatial model of electoral platforms*. Annual meeting of the Political Methodology Society, Ann Arbor, Michigan.
- Enelow, J. M., and M. J. Hinich. 1984. *The spatial theory of voting: An introduction*. New York: Cambridge University Press.
- Greenacre, M. J. 1993. *Correspondence analysis in practice*. London: Academic Press.
- Hill, M. O. 1973. Reciprocal averaging: An eigenvector method of ordination. *Journal of Ecology* 61:237–51.
- Hill, M. O. 1974. Correspondence analysis: A neglected multivariate method. *Applied Statistics* 23:340–54.
- Jackman, S. 2001. Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference and model checking. *Political Analysis* 9:227–41.
- Klemmensen, R., S. B. Hobolt, and M. E. Hansen. 2007. Estimating policy positions using political texts: An evaluation of the wordscores approach. *Electoral Studies* 26:746–55.
- Laver, M., K. Benoit, and J. Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review* 97:311–31.
- Lynn, H. S., and C. E. McCulloch. 2000. Using principal component analysis and correspondence analysis for estimation in latent variable models. *Journal of the American Statistical Association* 95:561–72.
- Mandelbrot, B. 1954. Structure formelle des textes et communication. *Word* 10:1–27.
- Monroe, B. L., and K. Maeda. 2004. *Talk's cheap: Text-based estimation of rhetorical ideal points*. Annual meeting of the Political Methodology Society. Stanford, CA.

- Monroe, B., and K. Maeda. 2004. *Talk's cheap: Text-based estimation of rhetorical ideal-points*. POLMETH Working Paper.
- Slapin, J. B., and S.-O. Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science* 52:705–22.
- ter Braak, C., and I. C. Prentice. 2004. A theory of gradient analysis. *Advances in Ecological Research: Classic Papers* 34:235–82.
- ter Braak, C. J. F. 1985. Correspondence analysis of incidence and abundance data: Properties in terms of a unimodal response model. *Biometrics* 41:859–73.
- ter Braak, C. J. F., and C. W. N. Loaman. 1986. Weighted averaging, logistic regression and the Gaussian response model. *Plant Ecology* 65:3–11.
- Zipf, G. K. 1949. *Human behavior and the principal of least effort*. Reading, MA: Addison Wesley.

WILL LOWE  
*Maastricht University*  
KENNETH BENOIT  
*London School of Economics and Political Science*  
SLAVA MIKHAYLOV  
*University College London*  
MICHAEL LAVER  
*New York University*

## *Scaling Policy Preferences from Coded Political Texts*

Scholars estimating policy positions from political texts typically code words or sentences and then build left-right policy scales based on the relative frequencies of text units coded into different categories. Here we reexamine such scales and propose a theoretically and linguistically superior alternative based on the logarithm of odds-ratios. We contrast this scale with the current approach of the Comparative Manifesto Project (CMP), showing that our proposed logit scale avoids widely acknowledged flaws in previous approaches. We validate the new scale using independent expert surveys. Using existing CMP data, we show how to estimate more distinct policy dimensions, for more years, than has been possible before, and make this dataset publicly available. Finally, we draw some conclusions about the future design of coding schemes for political texts.

Almost anyone interested in party competition, whether this takes place in legislatures, the electoral arena, or government, needs sooner or later to estimate the policy positions of key political actors, whether these be individual legislators or the political parties to which they affiliate. Indeed, “how to best measure the policy preferences of individual legislators and of legislative parties” (Loewenberg 2008, 499) forms one of the central problems of legislative research. This is particularly true for scholars of comparative legislative research. While in the American settings policy preferences of legislators have been conceptualized as individual-level variables, tight party discipline in many non-American contexts makes it difficult to derive

LEGISLATIVE STUDIES QUARTERLY, XXXVI, 1, February 2011 123  
DOI: 10.1111/j.1939-9162.2010.00006.x

[The copyright line for this article was changed on 9 May 2014 after original online publication.]

© 2011 The Authors. *Legislative Studies Quarterly* published by Wiley Periodicals, Inc. on behalf of The Comparative Legislative Research Center of The University of Iowa.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

estimates of legislators' ideal points that are distinct from aggregate policy stances of the parties to which they belong. Over the last two years this journal has devoted particular attention to the problem of measuring the policy preferences of legislators (e.g., Alemán et al. 2009; Carroll et al. 2009; Carrubba, Gabel, and Hug 2008; Clinton and Jackman 2009; Hix and Noury 2009; Saiegh 2009; Schickler and Pearson 2009). Here we contribute to this discussion by focusing on the estimates of policy positions of parties in legislatures on different dimensions over time.

In comparative legislative research, there are many sources of data from which estimates of the policy positions of key political actors—be these legislators or legislative parties—can be derived. These include, among others: mass surveys; expert surveys; political text; roll-call votes; and bill sponsorship (see Benoit and Laver 2006, for a review). By far the most abundant source of data on policy positions, both cross-sectionally and over time, is political text. Text is a direct by-product of political activity by the political actors whose positions we wish to estimate, whether this text takes the form of speeches, debates, written submissions, written rulings, or—by far the most commonly used in the profession for estimating party policy positions—election manifestos issued by political parties. These manifestos outline policies that parties will enact once elected to legislative or executive office and serve as the empirical basis for many models of party competition in legislative and other policymaking settings.

The wide availability of these materials in electronic form has led to a large number of automated and semiautomated methods for scaling positions from political texts based on the statistical analysis of word patterns (e.g., Bara, Weale, and Biquelet 2007; Benoit and Laver 2003; Hilliard, Purpura, and Wilkerson 2007; Hopkins and King 2010; Klemmensen, Hobolt, and Hansen 2007; Laver and Garry 2000; Lowe 2008; Martin and Vanberg 2007; Monroe and Maeda 2004; Pennings and Keman 2002; Quinn et al. 2010; Slapin and Proksch 2008; Yu, Kaufmann, and Diermeier 2008). Despite this growth in automated methods, however, the most common means of analyzing political text remains manual content analysis (Krippendorff 2004; Neuendorf 2002). In a traditional manual content analysis, a predefined categorical coding scheme is applied to segments of text by trained human coders (e.g., Baumgartner, Green-Pedersen, and Jones 2008). The most comprehensive and most frequently used such dataset comes from the Comparative Manifesto Project (Budge et al. 2001; Klingemann et al. 2006, hereafter CMP) which contains the results of coding more than 3,000 election manifestos for more than 650 parties in over 50 countries. CMP data

form the basis for hundreds of published studies by third-party authors and are almost always used to estimate policy positions for political parties on left-right scales. Almost everyone using CMP data does so for the same reason: they want to estimate *positions* of parties on different common policy dimensions. Doing this typically implies assuming that a set of party positions, whether a cross-section or a time series, can be located on some (continuously defined) metric scale. Such a scale allows analysts to make statements to the effect that, for example: party *A* is “moving” towards the left; parties *A* and *B* are “closer” to each other than either is to party *C*; given parties *A*, *B*, and *C*, the “median legislator” in the set of three parties is at *X*; and so on. Spatial theories of policy preferences typically assume that party positions exist on a continuous scale, usually an interval scale, although content coding schemes such as the CMP record only absolute and relative category counts of discrete text units. To convert these observed category counts into points on a continuous policy dimension, therefore, some scaling procedure is required. The CMP data offer several general political scales based on aggregating counts of text categories. The most widely used of these is the CMP’s left-right “Rile” scale, constructed by subtracting the sum of 13 “left”-associated categories from the sum of 13 “right”-oriented categories.<sup>1</sup> There are many different ways to construct such scales, however, and the choice of scaling procedure involves decisions that must be defended on methodological and substantive grounds.

In this article we present a new method for scaling continuous left-right policy positions from political text coded into discrete categories and demonstrate its superiority to current approaches. Comparing our measure to previous scales, we demonstrate that our proposed scale not only better satisfies general political, linguistic, and psychological criteria, but also that it exhibits superior empirical properties when applied to the CMP data. We validate our new scale externally through comparison to independent expert surveys. Not only can our new approach be applied to improve existing policy estimates for the most commonly used CMP scales, it can also be used with existing CMP data to unlock reliable positional estimates on new policy dimensions. These new and improved scales will provide researchers in legislative studies with not only more valid measures of the policy positions than ever before, but also unlock measures for previously unused dimensions of policy that can be used to test empirical models of party competition, legislative coalitions, government formation, and executive-legislative relations. To make the scale immediately useful to applied researchers, we provide a full dataset,

described in an appendix and in Tables 1 and 3, of these newly scaled policy positions with 21 new left-right scales, at least half of which have never before been used in applied, published research. Following the method for estimating uncertainty from political text of Benoit, Laver, and Mikhaylov (2009), we also provide confidence intervals for every new estimate. Finally, by justifying and demonstrating what types of coding categories are best compared to create continuous scales, our findings provide direct lessons for the future design of improved political text coding schemes.

### **How Should Policy Mentions Be Counted?**

The CMP's manual coding process involves several stages. In the first step, a human coder is given a political party manifesto, which he or she then divides into discrete, nonoverlapping text units known as "quasi-sentences." Quasi-sentences are textual units that express a policy proposition and may be either a complete natural sentence or part of one. Once identified, the quasi-sentence is then assigned to one of 56 mutually exclusive policy categories, distributed across seven broad policy domains such as "Political System" or "Economy." CMP data thus take the form of counts of sentences in categories, a unit of analysis that is intermediate between the more holistic analysis offered by an interpretative approach and more detailed syntactic analyses (Popping 2007; van Atteveldt, Kleinnijenhuis, and Ruigrok 2008) and purely lexical approaches (Laver, Benoit, and Garry 2003; Slapin and Proksch 2008). Category counts are then converted to percentages by dividing by the total number of sentences in the manifesto. These category percentages are then either interpreted directly as conveying information about the policy preferences of their authors or may be additively scaled to construct more general indices.

Normalizing counts this way makes sense under three conditions that we will not, for the purposes of this article, dispute: first, the sentence is the fundamental unit of policy assertion; second, different sentences assigned to the same category are exchangeable or independently distributed conditional on their policy category; third, the total number of sentences assigned to any policy category contains no information about the policy preferences that a platform expresses. The precise choice of how to construct a left-right scale from the normalized sentence counts, however, requires decisions to be made in the construction of scales. Scaling category counts, that is, choosing a procedure to transform observed category counts into estimates of unobserved policy positions, means addressing two independent

questions about the content and the form of a scale. These are two fundamental questions to which we will return as we evaluate different methods of scaling left and right policy.

First, how should sentences be counted when constructing a scale for a particular policy domain? Should one category be considered against an absolute standard, or relative to the counts in a different category, or perhaps relative to the entire document? Second, what is the functional form of the relationship between position and counts? In particular, what is the nature of the marginal effect on sentence counts of changes in a party's position in the policy domain linked to the sentence counts? While these two key issues frame a debate that has previously occupied methodologists concerned specifically with scaling policy positions from the CMP data (e.g., Kim and Fording 2002; McDonald and Mendes 2001a), the debate applies much more generally to *any* effort to construct continuous scales from text coded into discrete categories. In what follows, we reexamine both issues from both a substantive political standpoint and also from linguistic and psychological perspectives.

### **Previous Approaches to Scaling Policy Measures**

In the discussion of scaling measures we assume that for each policy dimension there exists a "left" and a "right" direction represented by at least one CMP category.<sup>2</sup> We will denote the *number* of sentences in a manifesto assigned to the "left" and "right" categories constituting a policy issue as  $L$  and  $R$ , respectively, and the total number of sentences in all categories as  $N$ . (There is also an "other" category count  $O$  to completely partition the sentences, such that  $L + R + O = N$ .) For instance, for a policy dimension of more to less protectionism,  $L$  would be the number of sentences coded to "406 Protectionism: Positive," while  $R$  would be the number of sentences coded to "407 Protectionism: Negative," and the corresponding "PER" variables defined as  $\frac{L}{N}100\%$  and  $\frac{R}{N}100\%$ , respectively. The *output* of any scaling procedure is an estimate of the position which we will refer to as  $\theta$ , superscripting to indicate the scaling procedure and subscripting as necessary to indicate the policy dimension.

#### *Previous Scaling Procedures*

The CMP was designed to reflect "saliency theory," a particular view of how parties compete and therefore how they express their

policy preferences, asserting that “all party programmes endorse the same position, with only minor exceptions” (Budge et al. 2001, 82). Parties are assumed to differentiate themselves by emphasising issues on which they have the best reputation with voters (Budge 1994). Because positioning is a matter of emphasis, the answer to the first general methodological question posed above must be that the frequency of quasi-sentences in one policy category should be compared to all other sentences in the manifesto. Budge (1999) suggests that a party’s position according to saliency theory,  $\theta^{(S)}$ , should be defined as

$$\theta^{(S)} = \frac{R - L}{N}.$$

This saliency measure is based on the difference in counts between left and right sentences counts normalized by the total number of sentences in the manifesto on any issue or on none.<sup>3</sup> From this definition it is clear that the answer to the second general question posed above is that each count in  $L$  or  $R$  has the same marginal effect:  $1/N$ . The quantity  $\theta^{(S)}$  is equal to zero when there are exactly the same number of left- as right-coded sentences,  $-1$  when there is only one issue on which the party is perfectly “left,” and  $1$  when there is one issue and the party is perfectly “right.” In practice, however, the extreme values are never reached because party competition almost never occurs on one dimension only. For instance, the distribution of the CMP’s “Rile” left-right index, a measure that encompasses 26 different coding categories, has an empirical range of about  $[-.5, .5]$ .

There is a more subtle constraint on  $\theta^{(S)}$  hidden in this formulation. All theories accept that if an issue becomes less important then a party will devote fewer sentences to it. That is, the relative counts  $R + L$  assigned to the contrasting policy pairs  $R$  and  $L$ , for a specific policy subset of all policy dimensions in a manifesto, will shrink. But because  $R + L$  is also by definition the maximum range of  $R - L$ , then deemphasizing an issue will push  $\theta^{(S)}$  to a more centrist position by moving it closer to 0, even though the proportion of left and right sentences, the raw material for expressing a position, have not changed. For the composite “Rile” scale, this means that counts of the 30 categories *not* in the scale still affect estimated party positions. For instance, a 200-sentence manifesto with 100 right sentences and no left sentences would have a Rile score of  $(50 - 0) = 50$ , but the same manifesto with 50 sentences added that are neither left nor right would change its Rile score to 40 (Benoit and Laver 2007; McDonald and Mendes 2001b; Ray 2007)—suggesting that the party shifted 20%

toward the left. In the CMP, this approach is carried to an extreme by including even uncodeable content in the definition of a manifesto.<sup>4</sup>

Primarily in order to address this problem, Kim and Fording (2002) propose an alternative measure that restricts the difference to sentences from the constituent left and right categories (see also Laver and Garry 2000). This *relative proportional difference* estimate of position is

$$\theta^{(R)} = \frac{R - L}{R + L}.$$

The measure also ranges from  $-1$  to  $1$ , but makes explicit the range constraint hidden in  $\theta^{(S)}$ . Dividing by  $R + L$  decouples the measure from variation in the importance a party assigns to any issue area. The only remaining influence of variable issue importance is that the overall number of sentences available to *express* a position is increased or reduced. To take an extreme case, only three positions are expressible within a budget of two sentences: either both are left, both right, or one is assigned to each category, leading to estimated positions of  $-1/2$ ,  $0$ , or  $1/2$ . Coarse sampling does not necessarily imply anything about the party's actual position on the issue but rather limits the level of nuance and specificity that it can be expressed in a manifesto and the precision that may be inferred from it by readers and researchers. According to spatial theory assumptions the party has a position on the issue dimensions, but has chosen to use its supply of sentences on other dimensions. Finally, unlike  $\theta^{(S)}$  this measure will not necessarily create an apparent move to a more centrist position if the party decides to focus on other policy areas.

In terms of the two methodological questions above,  $\theta^{(R)}$  compares category counts only to counts in the opposing category rather than to counts of all quasi-sentences. The marginal effect of another sentence on the left or right side of the issue is therefore  $1/(R + L)$ .

Although  $\theta^{(R)}$  appears to fix the problem of sentences in unrelated or uncoded categories affecting position estimates, it shares the assumptions embodied in  $\theta^{(S)}$  about the fixed marginal effect of another coded sentence and the existence of fixed endpoints. This has the unfortunate effect of forcing the  $\theta^{(R)}$  to  $-1$  when  $R = 0$  irrespective of the value of  $L$ , or to  $1$  when  $L = 0$  irrespective of the value of  $R$ , leading to spikes at the boundaries of the scale. That the scale has boundaries at all is a basic problem with both procedures that attempt to measure policy positions that are more naturally conceptualized in

an underlying continuum. The essential insight behind  $\theta^{(R)}$  is surely correct—the position of a party on a policy dimension should depend only on  $L$  and  $R$ . The problem is that the nature of the quantity being estimated is not respected in the measure. A different answer to the second general question is needed.

### A Scaling Method Based on Log Odds-Ratios

To motivate a new scaling method, consider the process of reading a party manifesto for changes in policy content, as a voter might do, for example, if trying to identify any change in some party’s policy position on the European Union. If the party’s previous platform contained 50 sentences in favour of increased European integration, and 20 emphasizing its disadvantages, then a new manifesto containing 50 sentences in favor and 21 against would barely register as an indicator of policy change. But if the previous platform had contained 10 and 4 sentences for and against the EU, and the new platform 10 and 5, then a policy change is more plausible. This suggests that the balance between assertions in favour of the EU and against it between platforms is usefully summarized not by the difference between sentence counts, but rather by their ratio. The effect of adding one more sentence in the first case decreases the ratio of pro-to anti-EU sentences by about 5%, and in the second by 20%. By this reasoning, the marginal effect of one more sentence is decreasing in the amount that has already been said on the topic. Proportional or relative emphasis on different topics does indeed determine a reader’s estimate of position, but such changes must be perceivable against the background of existing policy emphasis.

This simple linguistic intuition about reading and writing manifestos can be supported by evidence from psychology. The decreasing marginal effect of an extra unit is a general property of many perceptual quantities such as temperature, heat, or loudness studied by psychophysicists.<sup>5</sup> The Weber-Fechner law (Fechner 1965; Stevens 1957) formalises this observation: the size of the “just perceivable difference” of a subjective quantity is a constant proportion of the quantity already present.<sup>6</sup> Consequently we should operate in proportions, not levels, and work with a logarithmic scale relationship between the underlying quantity and subjective estimations of it. For loudness, this relationship is the familiar decibel scale, which relates perceived loudness as the log of the physical power of the sound.<sup>7</sup> Following this logic it should also be possible to consider the “just perceivable policy

difference,” the proportional change necessary to infer a difference in position on an issue between two party platforms.

### *The Logit Scale of Position*

Our logic suggests that from the point of view of a party manifesto writer wanting to communicate a position effectively, it is important to manipulate not so much the *absolute quantity* of sentences allocated ( $R + L$ ), but rather their relative *balance*, or  $R/L$ . Increasing  $R + L$  allows a wider range of expressible policy positions, but manipulating  $R/L$  expresses the position itself. Furthermore, because we are primarily interested in inferring positions, we view it as most natural to consider proportional changes on a *symmetrical* left-right scale. One natural measure for this purpose is the empirical logit:

$$\theta^{(L)} = \log \frac{R + .5}{L + .5} \quad (1)$$

$$= \log(R + .5) - \log(L + .5). \quad (2)$$

Like  $\theta^{(R)}$ ,  $\theta^{(L)}$  is conditional because it only considers sentences that are assigned to left or right. Unlike  $\theta^{(S)}$  and  $\theta^{(R)}$ , however, the logit scale  $\theta^{(L)}$  has no predefined end points: given enough sentences, it is possible to generate positions of any level of extremity.<sup>8</sup> In this respect,  $\theta^{(L)}$  better reflects spatial politics assumptions about the possible range of ideal points. However, although any real valued policy position can be represented, expressing extreme positions requires exponentially more sentences in  $L$  or  $R$  to move the policy position the same distance left or right as can be seen by considering its alternative formulation (2) as a difference measure.<sup>9</sup>

We should note that although  $\theta^{(L)}$  is defined as a (logged) ratio, it offers *interval* not ratio level measurement. In particular,  $\theta^{(L)} = 0$  should not automatically be identified as a substantively centrist policy position. In the absence of an external anchor, e.g., to policy outcomes, a centrist position would be some function of the mean or median position on an issue of the parties contesting the election. How this position will be expressed in  $R, L$  terms will depend on historically contingent country-level factors.<sup>10</sup>

Using the logit function to transform count data represents a novel approach to scaling left-right policy positions, but logit transformations are found in many inferential models used to estimate latent party positions. Log odds-ratios form the basis of the most

commonly used statistical models of bounded count data (Agresti 1996; Fleiss, Levin, and Paik 2003), item response and unfolding (Elff 2008), and have been studied directly by Monroe, Quinn, and Colaresi (2008).<sup>11</sup> Nevertheless,  $\theta^{(L)}$  is explicitly *not* itself a model of the structure of policy positions but rather a way to measure them that is compatible with several theories of spatial politics. We do not pursue such models here because we are unwilling either to introduce the independence of irrelevant alternatives (IIA) constraint on policy dimensions that would be imposed by logit models or to estimate explicitly the distribution of party positions on multiple dimensions as required by probit models. Consequently we also take no position on important substantive issues such as the underlying dimensionality of the policy space and the correlational structure connecting issue dimensions (Elff 2008; Gabel and Huber 2000) or the dynamics of party positions over time. Our more modest goal here is to improve the future use of the hugely popular CMP dataset, after demonstrating a better way to scale policy positions than the CMP's existing, flawed approach. Furthermore, our confrontational pairing method provides scales for more policy dimensions than ever before used from the CMP dataset. Whether these new positions are comparable over time, or accurately reflect the underlying dimensions of politics, are separate questions that are broader than we can feasibly address here.

Instead, we focus on the scaling procedure that connects basic data of the CMP—counts of sentences in categories—to the policy positions that form the substantive quantities of interest. Even without making model assumptions, we can show that  $\theta^{(L)}$  is a far better predictor of party policy positions than previous measures. We do make one concession towards model structure by adding 0.5 to all counts, a standard statistical practice for the analysis of contingency tables (Agresti 1996) that can also be motivated as a measure to reduce bias when estimating category proportions (Brown, Cai, and Das-Gupta 2001; Firth 1993). This smooths  $\theta^{(L)}$  slightly towards 0 and makes position estimates created from very small counts more stable, while barely affecting those derived from more reasonable numbers of sentences.

#### *A Log Scale of Policy Importance*

In addition to having different *positions* on each of a given set of policy dimensions, political actors may also differ in terms of the relative *importance* they attach to these dimensions. As Laver and Hunt (1992) demonstrated, some issues are simply more important to

some parties than to others, quite independent of their party positions on these dimensions, a distinction long-recognized by other scholars (e.g., Grofman 2004; Riker 1996). We thus expect “green” parties to treat the environmental dimension as the most important policy domain, and indeed this is part of our implicit definition of the set of green parties. Likewise, we expect far-right parties to treat immigration and social values as the most important dimensions. Both liberal and far-right parties might consider social values to be very important, yet take very different positions on this dimension. Scholars concerned with the policies of political actors are typically concerned with both position and importance. Empirical methods often draw this distinction very explicitly, as with the expert surveys of Laver and Hunt (1992), Benoit and Laver (2006), and Hooghe et al. (2008).

Notwithstanding this very clear analytical distinction between the importance, or salience, of a policy dimension and party positions on that dimension, the widely used policy *scales* (as opposed to raw *data*) generated by the CMP are fundamentally grounded in the CMP’s “saliency theory” of party competition (*MPP*, 76). This explicitly conflates party positions on policy dimensions and the relative salience of these dimensions. The core idea of saliency theory is that, in a given setting, parties will endorse only single sides of each issue, such as reducing crime, providing for the national defense, or protecting the environment. Parties differentiate themselves by emphasizing the issues on which their stances are most credible (*MPP*, 7). Consequently, the “taking up of positions is done through emphasizing the importance of certain policy areas compared to others” (Budge 1994, 455).<sup>12</sup> Operationally, “saliency” theory suggests that the relative mention of different policy areas in manifestos provides a direct measure of their importance to the party. Despite this prediction that issues are overwhelmingly one-sided, however, the CMP’s coding scheme makes numerous practical concessions to the fact that many issues are clearly two-sided, such as positions on free trade, on the level of government regulation, or on attitudes toward European integration. The existence of paired categories in the CMP scheme covering opposite sides of the same issue complicates the straightforward assessment of policy salience based on counting relative mentions of a single policy category. Our solution is simple: to group mentions of an issue, whether positive or negative, and to consider their sum as a direct indicator of policy importance. Our scale also follows the psychological and linguistic rationale for logarithmic emphasis as explained previously, however.

Our suggested measure of policy importance is

$$\theta^{(I)} = \log \frac{R+L+1}{N},$$

with a value of 1.0 added to the numerator for consistency with the 0.5 for  $R$  and  $L$  in the position formulation. This measure follows directly from the relative emphasis logic of saliency theory and also conforms to the linguistic model we have already outlined by increasing logarithmically in extremity with additional mentions.

### *Estimating Scale Uncertainty*

It has become widely accepted that text-based measures of policy quantities should come with associated estimates of uncertainty, rather than simply being presented as if they contained no stochastic element or measurement error (see Benoit, Laver, and Mikhaylov 2009). For this reason we also provide a means of computing standard errors and confidence intervals associated with our new scales of position and importance.

If a parametric measure of uncertainty is required, we suggest a simple Bayesian approach: a standard Beta prior over the proportions of  $L$  and  $R$  sentences with parameters  $a_R = a_L = a$  implies a posterior distribution over position that is well approximated as

$$\begin{aligned} \theta^{(L)} | R, L &\sim \text{Normal}(\mu, \sigma^2) \\ \mu &= \log \frac{(R+a)}{(L+a)} \\ \sigma^2 &= (R+a)^{-1} + (L+a)^{-1} \end{aligned}$$

when  $R+L \geq 10$ . Setting  $a = 0.5$  corresponds to a symmetrical invariant Jeffreys prior over party position (Jeffreys 1946). This distribution above suggests the 95% credible interval

$$[\theta^{(L)} - 1.96\sigma, \theta^{(L)} + 1.96\sigma]$$

which corresponds closely to the classical confidence intervals (when they are defined) while being numerically more stable (Newcombe 2001).

Many counts of quasi-sentences representing  $R$  or  $L$ , however, may be zero or close to zero in observed data, implying nonsymmetric bounds that will affect the parametric computation of confidence inter-

vals. An alternative to the parametric estimation that we propose is to use bootstrapping methods (Efron and Tibshirani 1994) to provide nonparametric intervals by resampling  $R$  and  $L$  categories in each policy dimension. In the dataset provided with this article and in the analyses presented here, we compute nonparametric confidence intervals and standard errors for all position and importance scales represented in the article, using the approach outlined by Benoit, Laver, and Mikhaylov (2009).

### New Policy Scales

We have constructed a set of 13 policy scales from the CMP dataset, each representing a distinct dimension of policy on which parties may take positions. These are detailed in Table 1. For each scale, we have identified a pair of CMP categories expressing policy opposites and classified the elements of each pair as either *Right* or *Left*. The pairings in Table 1 are natural and probably closer to what was originally intended by the designers of the CMP's coding scheme, although most are seldom or never used in this way. This alternative to the saliency approach has often been termed the "confrontational" approach to policy (Budge et al. 2001; Gemenis and Dinas 2010) and involves parties declaring competing positions on the same issue. In this view of policy, what matters is not whether each party purports to emphasize the issue or downplay it, but rather what the party's specific policy stances are relative to the extreme positions on any given issue, for instance what degree of permissiveness or restrictiveness regulation it favors regarding the issues of euthanasia, homosexual marriage, and abortion (Laver 2001, 66) or whether a party favours expanding the power of European-level institutions or instead reinforcing national sovereignty. Our logit scale extends and generalizes this logic while applying the notion of *relative* difference that also scales policy extremity in a way that relates to repetition in a nonlinear fashion.

In addition to these natural opposites, there are many categories for which natural policy alternatives could have been identified when the CMP coding scheme was being designed, but which do not in fact exist in the coding scheme. We identify these categories in Table 2. With the sole exception of 408 Economic Goals, these categories all relate to matters of public policy that are inherently positional.

The rationale for the CMP's unwillingness to define polar opposites for these coding categories appears to be that one position seems likely to be almost universally unpopular. Consider corruption or the

TABLE 1  
Paired Policy Dimensions and Corresponding Variable Names in the Dataset

Policy Dimension	“Left” Position	“Right” Position	Variable Name
Foreign Alliances	101 Foreign Special Relationships: Positive	102: Foreign Special Relationships: Negative	foreignalliances
Militarism	105 Military: Negative	104 Military: Positive	militarism
Internationalism	107 Internationalism: Positive	109 Internationalism: Negative	internationalism
EU	108 European Integration: Positive	110 European Integration: Negative	logeu
Constitutionalism	203 Constitutionalism: Positive	204 Constitutionalism: Negative	constitutionalism
Decentralisation	301 Decentralisation: Positive	302 Centralisation: Positive	decentralization
Protectionism	406 Protectionism: Positive	407 Protectionism: Negative	protectionism
Keynesian Policy	409 Keynesian Demand Management: Positive	414 Economic Orthodoxy: Positive	keynesian
Nationalism	602 National Way of Life: Negative	601 National Way of Life: Positive	nationalism
Traditional Morality	604 Traditional Morality: Negative	603 Traditional Morality: Positive	morality
Multiculturalism	607 Multiculturalism: Positive	608 Multiculturalism: Negative	multiculturalism
Labour Policy	701 Labour Groups: Positive	702 Labour Groups: Negative	laborpolicy
Welfare State*	504 Welfare State Expansion: Positive	505 Welfare State Limitation: Positive	welfare
Education spending	506 Educational Provision Expansion: Positive	507 Education Expenditure Limitation: Positive	education

\*This differs from the CMP’s welfare scale in that the CMP’s version is not confrontational (and does not include per504).

TABLE 2  
CMP Scales with No Natural Policy Opposites

Policy Issue	CMP Category
Imperialism	103 Anti-Imperialism: Anti-Colonialism
Peace	106 Peace: Positive
Freedom/Human Rights	201 Freedom and Human Rights: Positive
Democracy	202 Democracy: Positive
Efficiency	303 Governmental and Administrative Efficiency: Positive
Corruption	304 Political Corruption: Negative
Political Authority	305 Political Authority: Positive
(General) Economic Goals	408 Economic Goals
Corporatism	405 Corporatism: Positive
Technology and Infrastructure	411 Technology and Infrastructure: Positive
Cultural Policy	502 Culture: Positive
Social Justice	503 Social Justice: Positive
Law and Order	605 Law and Order: Positive
Social Harmony	606 Social Harmony: Positive
Agricultural Policy	703 Farmers: Positive
Middle Class Policy	704 Middle Class and Professional Groups: Positive
Affirmative Action	705 Underprivileged Minority Groups: Positive

environment: Since no party is likely to support corruption or call for trashing the ecosystem, “saliency” theory assumptions seem plausible for such policy issues. A closer look, however, reveals a more nuanced picture. On environmental policy, for instance, parties do not always produce purely one-sided statements. Many parties do in fact take progrowth stances that contain thinly veiled antienvironmental messages. For instance, the 1988 Danish Liberal Party manifesto contains this statement: *“Environmental policy should not result in Danish companies being worse off than the companies in the countries with which we compete.”*<sup>13</sup> The Danish Liberal Party is clearly not proenvironment, preferring instead to let the natural environment suffer in exchange for the economic benefits that presumably come from easing environmental regulations on firms. This direct preference for industry over the environment is in fact how other schemes for measuring environmental policy have expressed the environmental policy dimension: as contrasting priorities for environmental protection (at the cost of economic growth) versus economic growth (at the cost of environmental damage; Benoit and Laver 2006; Laver and Hunt 1992). We believe that this logic of contrasting extremes applies quite generally.

Not every quantity of end-user interest from the CMP may exist in the form of text units assigned to one of two bipolar categories.

Indeed, most users are only interested in the CMP dataset for its aggregate left-right scale. Fortunately, our measure works equally well for aggregated categories of  $R$  and  $L$  when each  $R$  and  $L$  consists of more than one component category count. Furthermore, as with the “Rile” index that includes quantities such as “305 Political Authority: Positive” that have no opposite category in the CMP coding scheme, many of these measures may denote right-measured positions yet not be usable in any simple, bipolar scale. For multicategory indexes,  $\theta^{(L)}$  is defined the same way after aggregating category counts into a composite  $L$  and  $R$ :

$$\theta_{\text{index}}^{(L)} = \log \frac{\sum_j R_j}{\sum_k L_k}.$$

As with simple scales involving only two categories, the zero point on this scale is not substantively privileged and should not necessarily be identified with a centrist policy position. This is particularly clear when different numbers of categories are used in the numerator and denominator.

In Table 3 we have listed a set of proposed additive indexes that are amenable to use with the logit scale, wherever possible identifying the source where this index was developed. We have also proposed several new scales of our own, such as “Free Market Economy” and “State-provided Services.” Our proposed scale of environmental protection follows the confrontational pairing logic by treating the two proenvironmental categories “Antigrowth Economy: Positive” (416) and of course “Environmental Protection: Positive” (501) together to capture antigrowth politics, ecologism as “left,” and the environmentally opposed paradigm of economic growth is represented in the CMP by the category “Productivity: Positive” (410).<sup>14</sup> In the next section we compare our scale to previous formulations and also compare the scale estimates to independent measures of position and importance from expert surveys.

### **Validating the Logit Policy Scale**

#### *Comparing Scales*

Before turning to validation against experts it is helpful to compare the properties of  $\theta^{(S)}$ ,  $\theta^{(R)}$ , and  $\theta^{(L)}$  as measurements. The problems we have identified with both  $\theta^{(S)}$  and  $\theta^{(R)}$  are fairly easy to

**TABLE 3**  
**Additive Scaled Dimensions and Corresponding Variable Names in the Dataset**

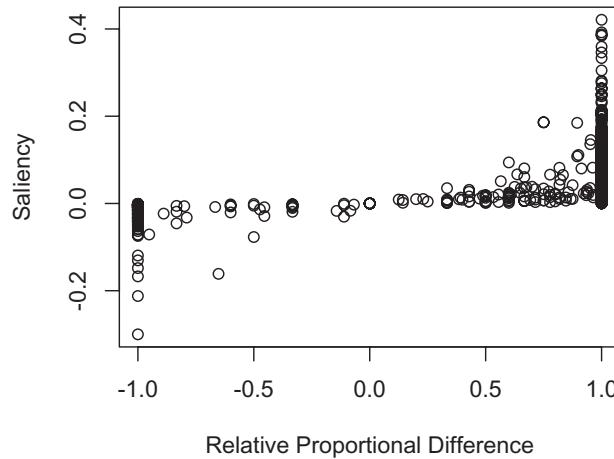
Policy Dimension	“Left” Position	“Right” Position	Source	Variable Name
Free Market Economy	401 Free Enterprise: Positive +	403 Market Regulation: Positive +	(Proposed)	
	402 Incentives: Positive	412 Controlled Economy: Positive +		freetemarket
Environmental Protection	501 Environmental Protection: Positive +	413 Nationalisation: Positive +		
	416 Anti-Growth Economy: Positive	415 Marxist Analysis: Positive		
	403 Market Regulation: Positive +	410 Productivity: Positive	(Proposed)	
	404 Economic Planning: Positive +	401 Free Enterprise: Positive +	Benoit & Laver (2007)	stateconomy
	406 Protectionism: Positive +	402 Incentives: Positive +		
State Involvement in Economy	412 Controlled Economy: Positive +	407 Protectionism: Negative +		
	413 Nationalisation: Positive +	414 Economic Orthodoxy: Positive +		
	504 Welfare State Expansion: Positive +	505 Welfare State Limitation: Positive		
	506 Education Expansion: Positive +	506 Education Expansion: Positive +		
	701 Labour Groups: Positive	505 Welfare State Limitation: Positive +	(Proposed)	
	504 Welfare State Expansion: Positive +	507 Education Limitation: Positive		stateservices
	506 Education Expansion: Positive			

(continued on next page)

TABLE 3 (continued)

Policy Dimension	“Left” Position	“Right” Position	Source	Variable Name
Left-Right: CMP “Rile”	103 Anti-Imperialism: Anti-Colonialism + 105 Military: Negative + 106 Peace: Positive + 107 Internationalism: Positive + 202 Democracy: Positive + 403 Market Regulation: Positive + 404 Economic Planning: Positive + 406 Protectionism: Positive + 412 Controlled Economy: Positive + 413 Nationalisation: Positive + 504 Welfare State Expansion: Positive + 506 Education Expansion: Positive + 701 Labour Groups: Positive Planned v. Market Economy	104 Military: Positive + 201 Freedom and Human Rights: Positive + 203 Constitutionalism: Positive + 305 Political Authority: Positive + 401 Free Enterprise: Positive + 402 Incentives: Positive + 407 Protectionism: Negative + 414 Economic Orthodoxy: Positive + 505 Welfare State Limitation: Positive + 601 National Way of Life: Positive + 603 Traditional Morality: Positive + 605 Law and Order: Positive + 606 Social Harmony: Positive 401 Free Enterprise: Positive + 414 Economic Orthodoxy: Positive	Laver & Budge (1992)	logrie
Social Liberal-Conservative	103 Anti-Imperialism: Anti-Colonialism + 105 Military: Negative + 106 Peace: Positive + 107 Internationalism: Positive + 202 Democracy: Positive	104 Military: Positive + 201 Freedom and Human Rights: Positive + 203 Constitutionalism: Positive + 305 Political Authority: Positive + 601 National Way of Life: Positive + 603 Traditional Morality: Positive + 605 Law and Order: Positive + 606 Social Harmony: Positive	Benoit & Laver (2007)	libcons

FIGURE 1  
Comparing the Relative Proportional Difference to the Saliency Scale for “National Way of Life” (categories 601 and 602)



illustrate by comparing them for a range of values across almost any scale. Notably,  $\theta^{(R)}$  has a problem of reaching its limits for the extremes when  $L > 0$ ,  $R = 0$  or  $R > 0$ ,  $L = 0$ . While the problem with  $\theta^{(S)}$  is that it registers linear changes with each additional extreme-coded text unit in such situations,  $\theta^{(R)}$  registers no changes at all. Hence, a manifesto registering five exclusively left text units would be the same as one registering 500. To demonstrate this using the CMP data, Figure 1 plots the relative proportional difference versus the “saliency” scalings of the confrontational pair for “National Way of Life: Positive/Negative” (categories 601 and 602). The vertically stacked points at the limits of the scale (at  $-1.0$  and  $1.0$ ) show that additional mentions cause linear increases for the saliency scale, but no change for the relative proportional difference measure.

Several other problems with the existing scales also emerge from an inspection of Figure 1. First, because mentions of “national way of life” are relatively low in absolute frequency across manifestos, and because  $(R - L)/N \leq (R + L)/N$ , the low frequency of these statements relative to all other statements severely shrinks  $\theta^{(S)}$  toward zero. The saliency measure is insensitive to changes for policy dimensions with low absolute frequency and misleadingly assigns a difference score close to the zero point. While we have shown this here for only the national way of life categories 601 and 602, it also applies to the CMP’s biggest scale, Rile, encompassing

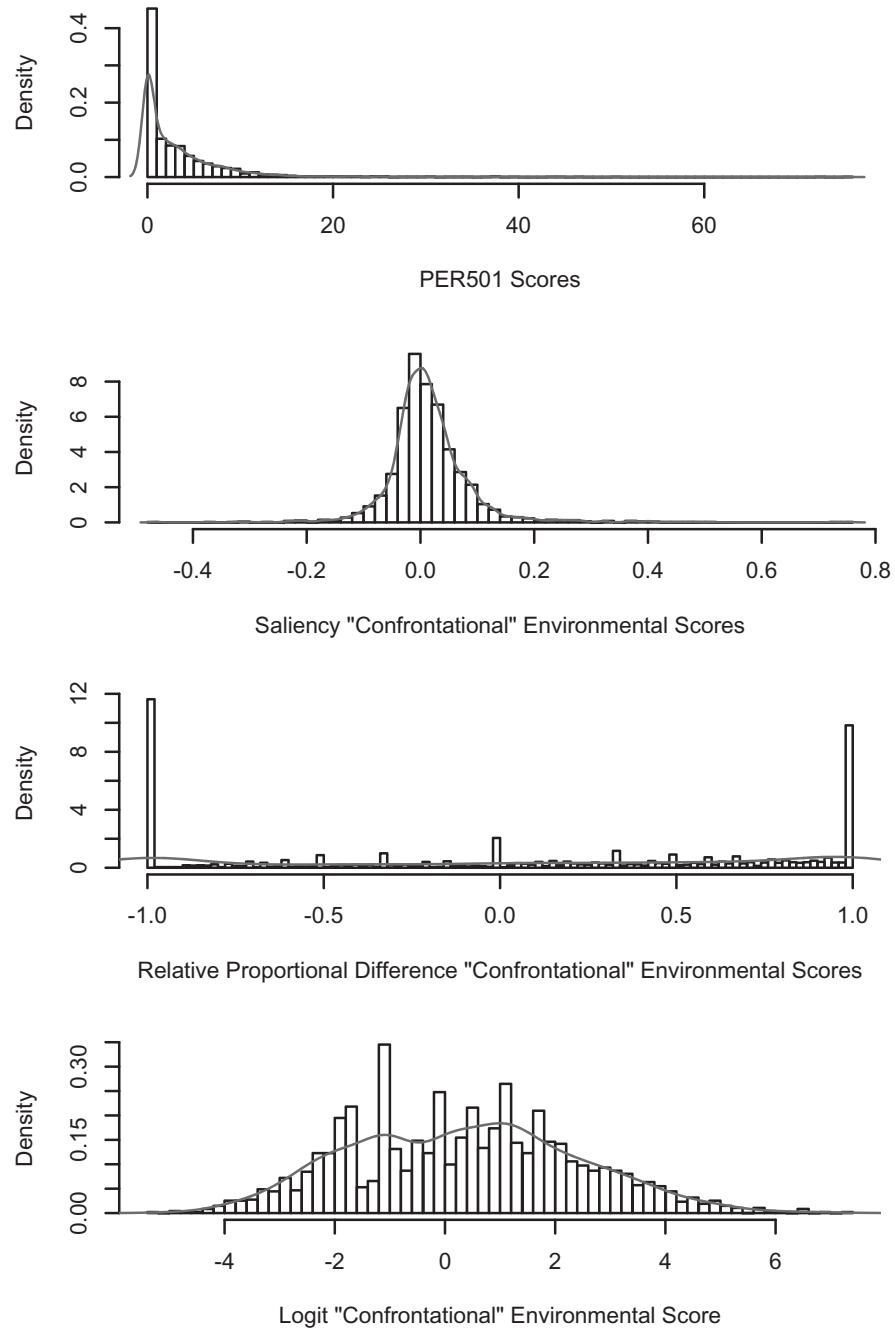
26 categories in all. While in theory this scale runs from  $-1.0$  to  $1.0$  (as a proportion), in practice the range spans only from  $-.5$  to  $.5$  for almost every manifesto measured.

A second problem, again with  $\theta^{(S)}$  can be seen at the extremes defined as  $L > 0$ ,  $R = 0$  for a left extreme, or  $R > 0$ ,  $L = 0$  for a right extreme. Extremity on the saliency measure  $\theta^{(S)}$  increases at a linear rate with each additional text unit in the extreme category. Substantively, the suggestion is that the same change occurs when the extreme-only category text units increase from 5 to 10 units, as when it increases from 105 to 100 units. This assumption of linear change in position given observed text unit counts is neither sensible nor supported by perceptual theory (see our discussion of the Weber-Fechner law above).

For each of these problems we have identified in  $\theta^{(S)}$ , a corresponding problem can also be found in  $\theta^{(R)}$ . The middle-range problem of lack of sensitivity for  $\theta^{(S)}$  is exactly reversed in  $\theta^{(R)}$ : small differences between  $R$  and  $L$  become highly influential on  $\theta^{(R)}$  when these are scaled as ratios of relative content  $R + L$ . An extreme example makes the point: imagine a series of manifestos from a party that had no real interest whatsoever in, and effectively no position on, protectionism. Irrelevant stochastic factors in text generation, or in the coding of the text, could plausibly result in a few essentially random counts of text units into each of the protectionism categories. The effect on  $\theta^{(R)}$  will be drastic in this situation, massively leveraging the error because it is only concerned with relatively proportional content.

If we compare the overall distribution of data for one of our composite scales—the environmental policy scale described above—we can see a fairly stark contrast between the spread of values for different scales that reinforces the patterns we observed in examining the National Way of Life scale. In Figure 2, we compare the distribution of scores for “PER501 Environmental Protection: Positive” to a “confrontational” scale constructed from opposing categories. Our new scale of environmental protection is based on adding the two proenvironmental categories “Antigrowth Economy: Positive” (PER416) and “Environmental Protection: Positive” (PER501) as capturing antigrowth politics and “ecologism” and contrasting this with the environmentally opposed paradigm of economic growth, represented in the CMP by the category “Productivity: Positive” (PER410). Figure 2 not only shows the better dispersion of the logit scale, but also demonstrates anew the problems we have already seen: bunching around zero of the saliency scale, as well as the

FIGURE 2  
Distributions of Scales for Proposed Environmental Scales



bunching around the extremes of the relative proportional difference scale.

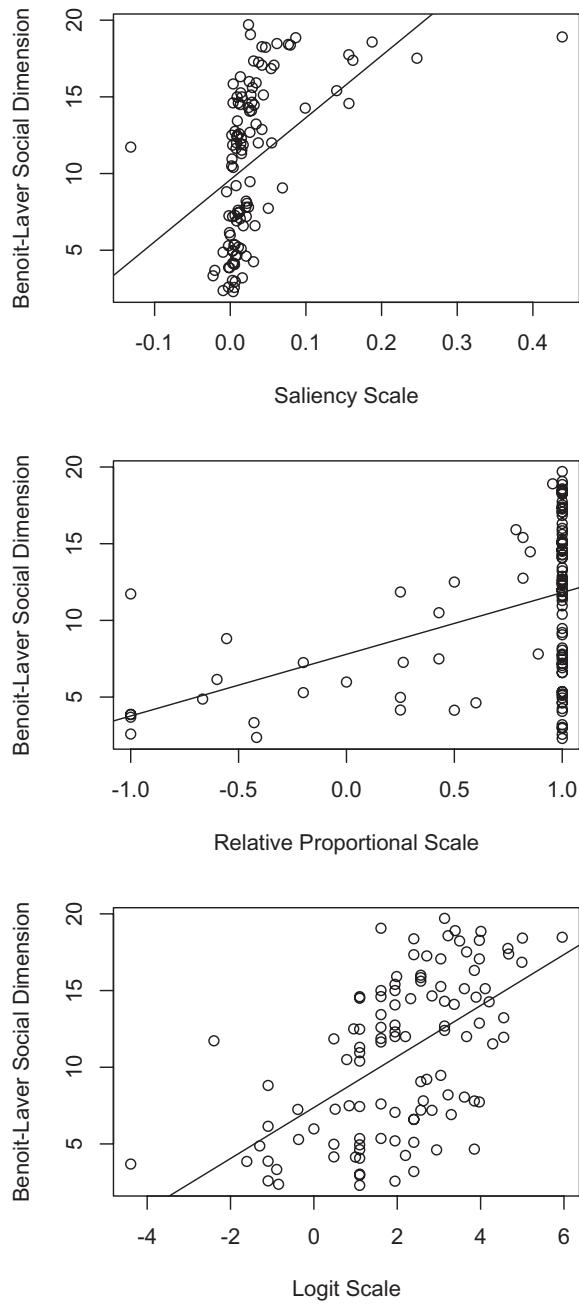
### *Comparisons to Expert Surveys of Policy*

Up to this point we have only compared one scale with another. To judge more conclusively whether a particular scale measures what we hope it measures, we can compare the CMP-based scales to independent, external measures of party positions based on expert surveys. Expert surveys such as Benoit and Laver (2006) have been shown to provide valid and reliable measures of party policy positions, but existing measures are limited in their time frame to the two decades since 1990. Only text-based measures such as the CMP have the potential to provide valid estimates of policy positions going further back in time. Limited comparisons of expert survey estimates to CMP measures were conducted by Benoit and Laver (2007), who tested the saliency-based Rile measures against expert survey ratings of left-right from Benoit and Laver (2006) and found a high correlation and lack of bias between the two measures. Because the large number of categories tends to wash out differences in large additive indexes such as Rile, here we perform the same comparison using smaller, more policy-specific scales.

We have compared the CMP-based indexes to the Benoit and Laver (2006) expert survey estimates of party position on the issue of social liberalism, one of two fundamental axes of political competition (the other being economic left-right) on which they place parties in every country. Some variant of this noneconomic dimension has been identified as a distinct, basic axis of political competition in numerous studies (e.g., Inglehart 1984; Marks, Wilson, and Ray 2002). Figure 3 plots the Benoit and Laver social liberalism dimension scores against each of the three scales based on counts of “604/603 Traditional Morality: Negative/Positive.”

The patterns from the plots are consistent with the interscale comparisons examined earlier. The saliency scale is highly bunched around zero, suppressing variation even when huge differences are identified by the Benoit and Laver scores. The relative proportional scale in the middle panel shows spotty variation in the middle ranges, with a very high proportion of values at the right side where Benoit and Laver indicate a complete range of differences but the relative proportional scale has reached its maximum value. Finally, the logit scale looks approximately linear, has no bunching at the extremes, or dispersed points in the middle. Its scale is centred to

FIGURE 3  
Comparison of CMP Scales of Traditional Morality with Benoit and Laver (2006) Social Liberalism Dimension

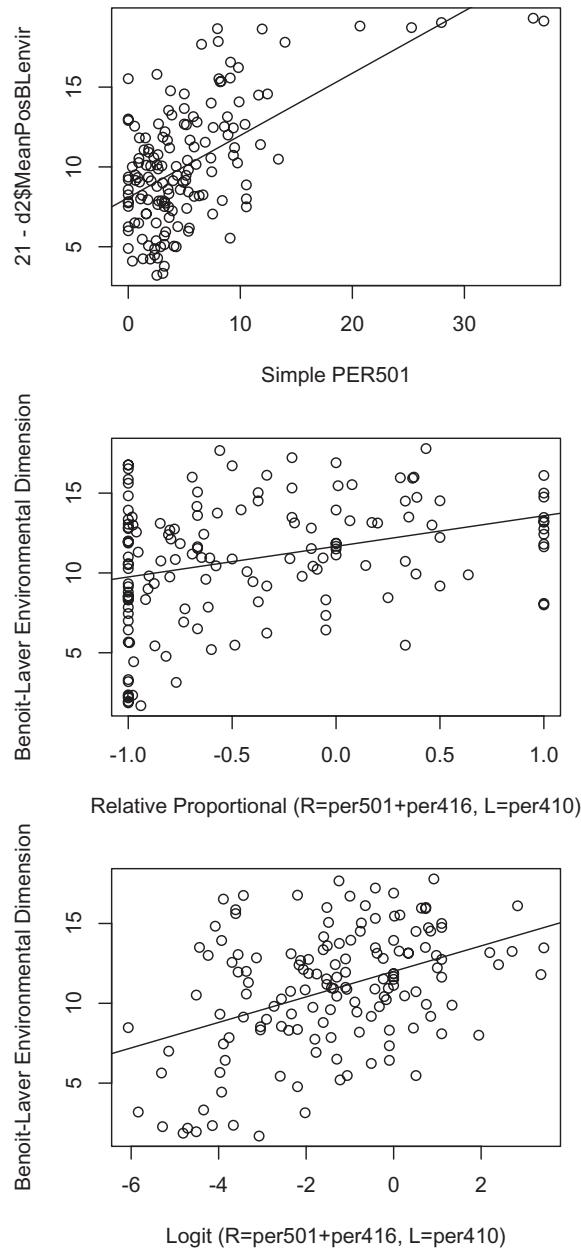


the right of zero, reflecting the higher proportion of text units of “Traditional Morality: Positive” (and many exclusively so), but this does not perturb the scale’s linear relationship with the expert survey scores. Residual analysis suggests that the relationship between expert survey scores and  $\theta^{(L)}$  are both linear and homoskedastic.<sup>15</sup>

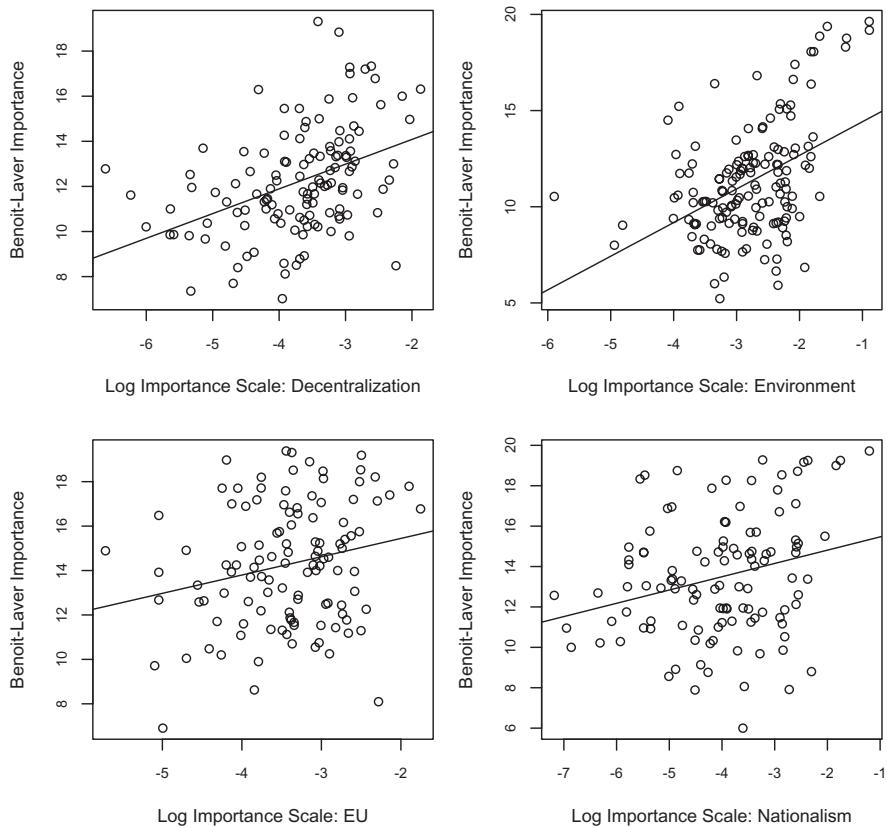
Finally, we perform the comparison with one of our simple additive scales that is not strictly constructed from a bipolar pair. This is the dimension of environmental protection, where the “left” side has two components. This is also an interesting category for comparison since, as previously discussed, the CMP’s saliency approach identifies only one possible side to this issue. In Figure 4 we compare the CMP’s default single-category scale of 501, to the relative absolute proportional difference scale constructed as per Table 3, and to the corresponding logit scale. The two versions of the saliency scale (per501 and the difference scale we have proposed) are not particularly poor, although 501 is clearly bounded to be positive, with several values where the zero boundary suggests antienvironmental policies when in fact the party said nothing at all about the environment. The saliency difference scale (middle plot) is left-skewed with some extreme proenvironmental values skewing the pattern. The new scale (bottom plot), however, shows a much better behaved linear relationship with the Benoit and Laver scores, without the skew from the saliency or relative proportional difference plots. The logit scale suggests the more sensible conclusion that a manifesto containing 40% proenvironmental sentences is not 10 times more proenvironment than a manifesto with just 5%, but rather only about 2.3 times more. And by comparison to the one-sided policy issue approach suggested by pure saliency theory, the comparison in Figure 4 also reinforces our earlier argument that even seemingly one-sided issues perform better when recast in terms of confrontationally opposite categories.

As a final form of external validation, we also compare our suggested importance scale to the separately measured policy importance estimates from Benoit and Laver (2006). These are plotted, for four directly comparable measures, in Figure 5. The positive linear relationship for these scales suggests that the proposed measure of importance based on total mentions on either side of an issue do indeed form a valid indicator of the political importance of this issue to a party. Our proposed importance measure provides a scale for importance that is more valid linguistically, based on logarithmically increasing extremity. It also unlocks a general measure of importance

FIGURE 4  
Comparison of CMP Scale of Environment with Benoit and Laver  
(2006) Environmental Dimension



**FIGURE 5**  
 Comparison of Proposed Log Importance Scale  
 with Benoit and Laver (2006) Importance Measures  
 (only for cases where total  $R$  and  $L$  mentions are non-zero)



from the CMP data that has never before been made systematically available.

### Recommendations

By focusing attention on producing better measures of party policy positions over time, as well as introducing new measures of party policy, our study should contribute to new developments in the field of legislative studies, especially the study of legislatures in multiparty settings. To “understand what a legislature does (and why it

does it) we need to know the policy preferences of its members” (Loewenberg 2009, 415). This need for data becomes all the more interesting at the party level in contexts with multiparty governments, coalitions, and high party discipline.

Our conclusions can be summarized as follows. First, our analysis of the use of the logit scale to estimate left-right positions from counts of textual categories, as well as our demonstration through direct comparison to other scales as well as to independent external data, suggests that the logit scale is superior and should be used in place of the “saliency” and “relative proportional difference” approaches used previously. We recommend using the logit scale for all policy categories and have provided a set of 21 such scales (in Tables 1 and 3) that can be constructed directly from the existing CMP dataset. In addition, we have calculated uncertainty estimates for all quantities using the simulation method proposed by Benoit, Laver, and Mikhaylov (2009). This dataset is available immediately and offers a superior alternative to the estimates supplied by the CMP, estimates that we have shown are based on the inferior saliency-based scales, and with few exceptions not constructed in the confrontational pairing approach we recommend here.

Second, we have proposed a new and separate measure of policy importance that is consistent with our logit scale of position and demonstrated that this proposed scale correlates well with independent, external measures of policy importance from expert surveys. These importance estimates are also provided with accompanying uncertainty measures.

Finally, we have shown that the assumption that individual parties take only one side, and indeed that all parties take the same side, of an issue, is demonstrably false, even given the CMP’s own dataset. For our purposes, this implies a critique of the basic CMP coding scheme, since the existing scheme consists of a mixture of confrontational and saliency-based categories. Our analysis suggests that any revision of the coding scheme would complete the step toward a fully confrontational coding scheme, consisting only of opposing, *pro* and *contra* categories. It would also be possible to go one step further and include a neutral category for each confrontational policy scale, which could be ignored when computing  $\theta^{(L)}$  but counted when considering  $\theta^{(I)}$ . This would address the concerns of McDonald and Mendes (2001b) about the nonreflection of neutral stances in the positional scales, as well as better reflecting overall policy importance based on counting text units.

*Will Lowe <w.lowe@maastrichtuniversity.nl> is Assistant Professor in Research Methods, Department of Political Science, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Kenneth Benoit <kbenoit@lse.ac.uk> is Professor and Director of the Methodology Institute, London School of Economics, Columbia House, Houghton Street, London WC2A 2AE, United Kingdom. Slava Mikhaylov <v.mikhaylov@ucl.ad.uk> is a Lecturer in Political Science, University College London, The Rubin Building, 29/30 Tavistock Square, London, WC1H 9QU, United Kingdom. Michael Laver <michael.laver@nyu.edu> is Professor of Politics, New York University, 19 W. 4th Street, New York, NY 10012.*

## APPENDIX: DATASET DESCRIPTION

The data described in this paper are available for download from <http://www.kenbenoit.net/cmp/scales/>. The dataset contains all of the variables described in Tables 1 and 3, with suffixes indicating whether the variable refers to position, the 95% confidence interval on position, the estimate of importance, and the standard errors of position and importance. For the variable protectionism (for example), the five associated variables are:

---

protectionism	position estimate
protectionism_lo95	95% lower bound for position estimate
protectionism_hi95	95% upper bound for position estimate
protectionism_SE	standard error for position estimate
protectionism_imp	importance estimate
protectionism_impSE	standard error for importance estimate

---

## NOTES

This paper was originally prepared for presentation at the 2008 ECPR General Conference, Potsdam. We thank Thomas Däubler and Jonathan Slapin for comments on multiple drafts of this manuscript. This research was supported in part by the Irish Research Council for Humanities and the Social Sciences.

1. Details may be found in Table 3. We return to this scale later in the text.

2. For the initial development we treat each policy area as defined by one “left” and one “right” CMP category. In fact neutral categories are also possible, and in some cases it is helpful to aggregate more than one CMP category to generate a substantively appropriate left of right count.

3. More precisely, the CMP's saliency-based scale multiplies  $\theta$  by 100 to allow interpretation as a percentage.

4. The percentage of uncodeable content in the average manifesto in the CMP combined dataset is 6.8%, making the inclusion of uncoded content a real worry for many texts.

5. More complex nonlinear marginal effects that decrease the effect of early as well as later mentions have long been suggested (e.g., Jakobovits and Lambert 1963; Jakobovits and Hogenraad 1967) and are intuitively reasonable for manifesto data, perhaps as part of an interaction with issue salience. Investigating this relationship is further work.

6. Later research (Stevens 1957) has established a range of power law relationships between physical and subjective magnitudes in different modalities, not all of which exhibit decreasing marginal effects. Nevertheless we work here with the logarithmic relationship because of its simplicity, its linguistic motivation as sketched above, and most importantly, its excellent fit to policy positions generated independently by experts, considered below.

7. The scale is given units by reference to a barely perceptible reference sound.

8. In practice, however, the logit scales applied to the CMP data ranges from approximately  $-7$  to  $+7$ , since few  $R$  or  $L$  categories (or indeed,  $N$ ) tend to exceed  $\log(1000) = 6.9$ .

9. In applications we follow standard statistical practice in the analysis of contingency tables and add 0.5 to  $L$  and  $R$ , (Agresti 1996). This can be motivated as a means to reduce estimation bias (Firth 1993) or to provide better behaved interval measures of uncertainty (Brown, Cai, and DasGupta 2001).

10. Statistical latent variable modelers (e.g., Clinton, Jackman, and Rivers 2004; Slapin and Proksch 2008) make the same observation by noting that the zero point, direction, and units of the measurements are model identification constraints not substantive assertions about position.

11. In the framework of parametric models,  $\theta^{(L)}$  could be seen as subpart of a multinomial logistic regression model of the category counts  $[R, L, N - (R + L)]$  in party platform, where  $N - (R + L)$  is the number of sentences assigned to other categories or left uncoded. Using  $L$  as a base category,  $\theta^{(L)}$  will, as  $N$  increases, approximate the first linear predictor in such a model.

12. In the saliency theory approach, policy dimensions are assumed to consist of issue areas or clusters of issues (Robertson 1976, 61).

13. In Danish: “Miljøpolitikken [environmental policy] måikke stille danske virksomheder dårligere, end virksomhederne i de lande vi konkurrerer med” (Venstre Manifesto 1988). We thank Jacob Rathlev for this suggestion and Martin Hansen for drawing attention to this example and for help with the translation.

14. This reflects the definition of one of the core four dimensions in the expert survey in Benoit and Laver (2006, 129).

15. Similar patterns are easily observed for numerous other scales, such as Multiculturalism: Positive/Negative against the Benoit and Laver scores for nationalism and immigration.

## REFERENCES

- Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. New York: Wiley.
- Alemán, E., E. Calvo, M.P. Jones, and N. Kaplan. 2009. “Comparing Cosponsorship and Roll-Call Ideal Points.” *Legislative Studies Quarterly* 34 (1): 87–116.
- Bara, J., A. Weale, and A. Biquelet. 2007. “Analysing Parliamentary Debate with Computer Assistance.” *Swiss Political Science Review* 13 (4): 577–605.
- Baumgartner, F.R., C. Green-Pedersen, and B.D. Jones. 2008. *Comparative Studies of Policy Agendas*. London: Routledge.
- Benoit, K., and M. Laver. 2003. “Estimating Irish Party Positions using Computer Wordscore: The 2002 Elections.” *Irish Political Studies* 17 (2): 97–107.
- Benoit, Kenneth, and Michael Laver. 2006. *Party Policy in Modern Democracies*. London: Routledge.
- Benoit, Kenneth, and Michael Laver. 2007. “Estimating Party Policy Positions: Comparing Expert Surveys and Hand Coded Content Analysis.” *Electoral Studies* 26 (1): 90–107.
- Benoit, Kenneth, Michael Laver, and Slava Mikhaylov. 2009. “Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions.” *American Journal of Political Science* 53 (April): 495–513.
- Brown, L.D., T.T. Cai, and A. DasGupta. 2001. “Interval Estimation for a Binomial Proportion (with discussion).” *Statistical Science* 16: 101–33.
- Budge, Ian. 1994. “A New Spatial Theory of Party Competition: Uncertainty, Ideology and Policy Equilibria Viewed Comparatively and Temporally.” *British Journal of Political Science* 24 (4): 443–67.
- Budge, Ian. 1999. *Estimating Party Policy Preferences: From Ad Hoc Measures to Theoretically Validated Standards*. Number 139 in “Essex Papers in Politics and Government.” University of Essex and Department of Government.
- Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tanenbaum. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998*. Oxford: Oxford University Press.
- Carroll, Royce, Jeffrey B. Lewis, James Lo, Keith T. Poole, and Howard Rosenthal. 2009. “Comparing NOMINATE and IDEAL: Points of Difference and Monte Carlo Tests.” *Legislative Studies Quarterly* 34 (4): 555–91.
- Carrubba, C., M. Gabel, and S. Hug. 2008. “Legislative Voting Behavior, Seen and Unseen: A Theory of Roll-Call Vote Selection.” *Legislative Studies Quarterly* 33 (4): 543–72.
- Clinton, J., S. Jackman, and D. Rivers. 2004. “The Statistical Analysis of Roll Call Voting: A Unified Approach.” *American Journal of Political Science* 98 (2): 355–70.
- Clinton, J.D., and S. Jackman. 2009. “To Simulate or NOMINATE?” *Legislative Studies Quarterly* 34 (4): 593–621.
- Efron, Bradley, and Robert Tibshirani. 1994. *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC Hall.
- Elff, M. 2008. “A Spatial Model of Electoral Platforms.” POLMETH Working Paper.

- Fechner, G.T. 1965. Elemente der Psychophysik, Sections VII and XVI. In *A Source Book in the History of Psychology*, ed. R.J. Herrnstein and E.G. Boring. Harvard University Press, 66–75.
- Firth, D. 1993. “Bias Reduction of Maximum Likelihood Estimates.” *Biometrika* 80 (1): 27–38.
- Fleiss, Joseph L., B. Levin, and M.C. Paik. 2003. *Statistical Methods for Rates and Proportions*. 3d ed. New York: John Wiley.
- Gabel, Matthew, and John Huber. 2000. “Putting Parties in Their Place: Inferring Party Left-Right Ideological Positions from Party Manifesto Data.” *American Journal of Political Science* 44: 94–103.
- Gemenis, K., and E. Dinas. 2010. “Confrontation Still? Examining Parties’ Policy Positions in Greece.” *Comparative European Politics* 8: 179–201.
- Grofman, Bernard. 2004. “Downs and Two-Party Convergence.” *Annual Review of Political Science* 7: 25–46.
- Hilliard, D., S.J. Purpura, and S. Wilkerson. 2007. “Computer Assisted Topic Classification for Mixed Methods Social Science Research.” *Journal of Information Technology and Politics* 4 (4): 31–46.
- Hix, S., and A. Noury. 2009. “After Enlargement: Voting Patterns in the Sixth European Parliament.” *Legislative Studies Quarterly* 34 (2): 159–74.
- Hooghe, Liesbet, Ryan Bakker, Anna Brivevich, Catherine de Vries, Erica Edwards, Gary Marks, Jan Rovny, and Marco Steenbergen. 2008. “Reliability and Validity of Measuring Party Positions: The Chapel Hill Expert Surveys of 2002 and 2006.” University of North Carolina, Chapel Hill manuscript.
- Hopkins, D., and G. King. 2010. “A Method of Automated Nonparametric Content Analysis for Social Science.” *American Journal of Political Science* 54 (1): 229–47.
- Inglehart, Ronald. 1984. The Changing Structure of Political Cleavages in Western Society. In *Electoral Change, Realignment and Dealignment in Advanced Industrial Democracies*, ed. Russell Dalton et al. Princeton: Princeton University Press, 25–69.
- Jakobovits, Leon A., and R. Hogenraad. 1967. “Some Suggestive Evidence on the Operation of Semantic Generation and Satiation in Group Discussions.” *Psychological Reports* 20: 1247–50.
- Jakobovits, Leon A., and W.E. Lambert. 1963. The Effects of Repetition in Communication on Meanings and Attitudes. In *Television and Human Behavior*, ed. L. Arons and M.A. May. New York: Appleton-Century-Crofts, 167–76.
- Jeffreys, H. 1946. “An Invariant Form for the Prior Probability in Estimation Problems.” *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 186 (1007): 453–61.
- Kim, Heemin, and Richard C. Fording. 2002. “Government Partisanship in Western Democracies, 1945–1998.” *European Journal of Political Research* 41 (2): 187–206.
- Klemmensen, R., S.B. Hobolt, and M.E. Hansen. 2007. “Estimating Policy Positions using Political Texts: An Evaluation of the Wordscores Approach.” *Electoral Studies* 26 (4): 746–55.

- Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge, and Michael McDonald. 2006. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990–2003*. Oxford: Oxford University Press.
- Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*. 2d ed. Thousand Oaks, CA: Sage.
- Laver, M., and J. Garry. 2000. “Estimating Policy Positions from Political Texts.” *American Journal of Political Science* 44 (3): 619–34.
- Laver, Michael. 2001. *Estimating the Policy Position of Political Actors*. New York: Routledge.
- Laver, Michael, and Ben W. Hunt. 1992. *Policy and Party Competition*. New York: Routledge.
- Laver, Michael, and Ian Budge. 1992. *Party Policy and Government Coalitions*. New York: St. Martin’s Press.
- Laver, Michael, Kenneth Benoit, and John Garry. 2003. “Estimating the Policy Positions of Political Actors using Words as Data.” *American Political Science Review* 97 (2): 311–31.
- Loewenberg, G. 2008. “Introduction.” *Legislative Studies Quarterly* 33 (4): 499–500.
- Loewenberg, G. 2009. “Introduction.” *Legislative Studies Quarterly* 34 (4): 451–53.
- Lowe, W. 2008. “Understanding Wordscores.” *Political Analysis* 16 (4): 356–71.
- Marks, Gary, Carole Wilson, and Leonard Ray. 2002. “National Political Parties and European Integration.” *American Journal of Political Science* 46 (3): 585–94.
- Martin, L.W., and G. Vanberg. 2007. “A Robust Transformation Procedure for Interpreting Political Text.” *Political Analysis* 16 (1): 93–100.
- McDonald, Michael, and Silvia Mendes. 2001a. “Checking the Party Policy Estimates: Convergent Validity.” In *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998*, ed. Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tanenbaum. Oxford University Press, 127–41.
- McDonald, Michael, and Silvia Mendes. 2001b. “The Policy Space of Party Manifestos.” In *Estimating the Policy Position of Political Actors*, ed. Michael Laver. London: Routledge, 90–114.
- Monroe, B., and K. Maeda. 2004. “Talk’s Cheap: Text-Based Estimation of Rhetorical Ideal-Points.” POLMETH Working Paper.
- Monroe, B.L., K.M. Quinn, and M.P. Colaresi. 2008. “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict.” *Political Analysis* 16 (4): 372–403.
- Neuendorf, K.A. 2002. *The Content Analysis Guidebook*. Thousand Oaks CA: Sage.
- Newcombe, R.G. 2001. “Logit Confidence Intervals and the Inverse Sinh Transformation.” *American Statistician* 55: 200–02.
- Pennings, P., and H. Keman. 2002. “Towards a New Methodology of Estimating Party Policy Positions.” *Quality and Quantity* 36 (1): 55–79.
- Popping, Roel. 2007. “Text Analysis for Knowledge Graphs.” *Quality and Quantity* 41: 691–709.

- Quinn, K.M., B. Monroe, M. Colaresi, M. Crespin, and D. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209–28.
- Ray, Leonard. 2007. "Validity of Measured Party Positions on European Integration: Assumptions, Approaches, and a Comparison of Alternative Measures." *Electoral Studies* 26 (1): 11–22.
- Riker, William H. 1996. *The Strategy of Rhetoric: Campaigning for the American Constitution*. New Haven, CT: Yale University Press.
- Robertson, David. 1976. *A Theory of Party Competition*. London and New York: Wiley.
- Saiegh, S. 2009. "Recovering a Basic Space from Elite Surveys: Evidence from Latin America." *Legislative Studies Quarterly* 34 (1): 117–45.
- Schickler, E., and K. Pearson. 2009. "Agenda Control, Majority Party Power, and the House Committee on Rules, 1937–52." *Legislative Studies Quarterly* 34 (4): 455–91.
- Slapin, J.B., and S.-O. Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–22.
- Stevens, S.S. 1957. "On the Psychophysical Law." *Psychological Review* 64 (3): 153–81.
- van Atteveldt, W., J. Kleinnijenhuis, and N. Ruigrok. 2008. "Parsing, Semantic Networks, and Political Authority Using Syntactic Analysis to Extract Semantic Relations from Dutch Newspaper Articles." *Political Analysis* 16 (4): 428–46.
- Yu, B., S. Kaufmann, and D. Diermeier. 2008. "Classifying Party Affiliation from Political Speech." *Journal of Information Technology and Politics* 5 (1): 33–48.

## Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark

**Will Lowe**

*MZES, University of Mannheim*  
*e-mail: will.lowe@uni-mannheim.de*

**Kenneth Benoit**

*Department of Methodology, London School of Economics and the Department of  
Political Science, Trinity College, Dublin*  
*e-mail: kbenoit@lse.ac.uk (corresponding author)*

Edited by R. Michael Alvarez

Automated and statistical methods for estimating latent political traits and classes from textual data hold great promise, because virtually every political act involves the production of text. Statistical models of natural language features, however, are heavily laden with unrealistic assumptions about the process that generates these data, including the stochastic process of text generation, the functional link between political variables and observed text, and the nature of the variables (and dimensions) on which observed text should be conditioned. While acknowledging statistical models of latent traits to be “wrong,” political scientists nonetheless treat their results as sufficiently valid to be useful. In this article, we address the issue of substantive validity in the face of potential model failure, in the context of unsupervised scaling methods of latent traits. We critically examine one popular parametric measurement model of latent traits for text and then compare its results to systematic human judgments of the texts as a benchmark for validity.

A vast amount of effort in political science focuses on estimating characteristics of political actors—parties, legislators, candidates, voters, and so on—that may be estimated, but never directly observed. Whether we call them “ideal points,” policy preferences, topics, or issue emphases, these latent traits and latent classes are not only fundamentally unobservable but also exist in a dimensional space that is fundamentally unknowable.<sup>1</sup> This has hardly prevented political researchers from attempting to identify and estimate such quantities, however, and a variety of such methods are widely used. Many, such as the analysis of roll call votes, suffer from problems of data censorship and selection that produce biased estimates of the quantities desired. Not all actors vote, co-sponsor bills, or return our questionnaires, but there is one activity that always accompanies political action: speech. This simple fact, coupled with a revolution in the availability of vast quantities of recorded text and speech, has spurred the development of a wide range of methods for analyzing textual data, most of which are surveyed in Grimmer and Stewart (2013).

Every statistical model applied to data—textual or otherwise—requires assumptions. As Grimmer and Stewart (2013) point out, such assumptions are always wrong. For textual data, these assumptions concern the process that generates the observed textual data, including the stochastic process of text generation; the functional model linking political variables of interest and observed text; and the nature of the variables (and dimensions) on which observed text should be conditioned. The reality is that even though we know that these assumptions are “wrong,” we have no real benchmark by which to assess the *consequences* of or the degree of wrongness, because

*Authors' note:* Replication materials for this article are available from the *Political Analysis* dataverse at <http://hdl.handle.net/1902.1/20387>. Supplementary materials for this article are available on the *Political Analysis* Web site.

<sup>1</sup> For a good survey with examples of this problem, see Benoit and Laver (2012).

the data-generating process for natural language cannot ever be really “known” and therefore can never be reliably simulated. Moreover, the quantities we seek to estimate from text, such as latent traits for scaling models, or latent classes for topic models, are fundamentally unobservable. They can only be defined as inferences from, and therefore as inherently dependent on, the assumptions of some model. This core problem leaves us in a position of applying models of an unknown conditional data-generating process to estimate latent quantities in unknown dimensional spaces that can never be directly checked.

In what follows, we contend that the key question is not how wrong models of text are, but rather how *useful* these models are in helping us obtain valid measures of real political quantities. Furthermore, we focus on the hardest class of text models to validate, unsupervised scaling models, which we agree with Grimmer and Stewart (2013) have fewer direct methods for validation. Their challenge is that to validate the results from unsupervised methods, “scholars must combine experimental, substantive, and statistical evidence to demonstrate that the measures are as conceptually valid as measures from equivalent supervised methods.” Although in many contexts, this is both appropriate and feasible—as shown in the analysis of credit-claiming, for instance, in Grimmer and Stewart (2013; Fig. 6)—there are many cases where hand coding is practically infeasible. The scaling of continuous latent traits such as ideology, for instance, may be both costly and conceptually difficult for texts of any significant size without resorting to a system of strict coding rules, and existing examples of such schemes have been shown to suffer extreme reliability problems (see Mikhaylov, Laver, and Benoit 2012). Yet, scaling ideology is a core activity in political measurement, because no model of political competition can be tested without reliable and valid information on the relative preferences of political actors.

Scaling ideology is not as simple as comparing supervised models. In general, supervised models assume much *less* about the data-generating process than unsupervised scaling models while assuming more about the quantity estimated (see, e.g., Jordan 1995). Consequently, a successful supervised model confirms that enough information is available in the data to allow the relevant imposed distinctions, whereas a well-validated unsupervised model demonstrates that the assumption of specific additional process structure, here conditional independence, is sufficient to recover it without external imposition. For these models, the core question is instead what Grimmer and Stewart (2013) call “semantic validity”: whether the quantity being scaled reflects the quantity that the analyst intends to measure. Similar to the approach used by Grimmer and King (2011) to assess the semantic validity of unsupervised clustering, we apply experimentally elicited human input to validate an unsupervised scaling of relative ideological preferences. By demonstrating a detailed research design for validating estimates from quantitative text, we not only agree with Grimmer and Stewart (2013) but also show precisely *how* such a validation framework can be deployed. Rather than comparing the unsupervised results to those from a supervised model, we compare the unsupervised scaling results directly to elicited human judgments of the texts. Furthermore, we not only use human perceptions of *relative position* to validate the point estimates from the scaling model but also use human perceptions of *difference* to evaluate the estimates of uncertainty from statistical methods for scaling textual data. This second aspect returns directly to the point on which Grimmer and Stewart (2013) and we agree, namely, that our oversimplified models of the textual data-generating process are wrong, but provides a concrete assessment of *how* wrong these are by comparing the consequences for inference.

Our practical example comes from a set of speeches made and against a historic austerity budget debate in the Irish parliament in late 2009. For testing, we compare the results of the Poisson scaling model of Slapin and Proksch (2008) to a systematic rating of these same texts by twenty human readers. We have chosen the Poisson scaling model because its assumptions about the word data-generating process are the most explicit, and existing validation methods borrowed from the computer science literatures are not directly applicable, as we discuss further below. In addition, scaling latent traits such as policy positions are of central importance to empirical and theoretical political science, since without reliable and valid measurements of these concepts, it is impossible to test models of party competition, representation, or policy outcomes. The validation design we outline is easily understandable and can be easily replicated and extended to almost any research problem involving quantitative text, subject to sufficient human resources.

## 1 A Validation Design for Textual Data Analysis

Text generates unique quantitative data in that prior to being converted into numbers, text can be interpreted directly through a qualitative process of human reading. After all, the central purpose of text is to communicate a message to a reader or listener. Judging the message contained in a text, whether written or spoken, is something that humans do every day. This makes textual data fundamentally different from other forms of data, such as test item responses, the coded responses from completed survey questionnaires, or simple continuous quantitative data such as years of schooling, which do not convey a direct message through a careful construction that includes a rich vocabulary, a meaningful sequence, and a grammar and syntax. Other forms of quantitative data are often most meaningfully interpreted when aggregated and summarized, whereas with text as data, this process is reversed. When disassembled into quantitative information, typically a term-document matrix of word-type frequencies, humans can no longer make sense of textual data as they can in its raw form. Furthermore, few quantitative scales for measuring textual data have a natural metric whose summary interpretation is self-evident. This poses challenges for validating quantitative models of text as data, but also presents unique opportunities.

Our validation framework sets the following expectations: First, *valid positional estimates* from quantitative scaling, for a given dimension, should match a human reader's placement of these texts with respect to identifying relative differences along this dimension. Furthermore, this dimension should be clear to the reader and self-evident from reading of the texts. Second, *meaningful estimates of uncertainty* from a quantitative text scaling model should yield statistical conclusions of similarity or difference that correspond to a human reader's perceptions of difference between pairs of the same set of texts. A model that is "useful" in statistical decision-making will at least roughly correspond to this benchmark of human judgment of similarities and differences.

In the natural language processing, human qualitative interpretation has long formed the benchmark for validating automated content analysis methods. "Supervised" methods that assign complete documents to classes on the basis of qualitative human category assignments have long been validated by specificity and sensitivity (equivalently precision and recall; see Manning, Raghavan, and Schütze 2008). "Unsupervised" versions of these methods that attempt to simultaneously learn categories and their assignment to documents, such as clustering methods, as well as those that attempt to learn categories and to decompose documents into category proportions—known as "topic models"—present greater validation challenges while still using category-oriented methods (Wallach et al. 2009; Grimmer and King 2011). These cannot be applied directly to the validation of "unsupervised" scaling models, however, because we lack any kind of category labeling to work with, and there typically is no "test" set whose values are known. Moreover, because the latent traits that political scientists typically wish to estimate are often fundamentally unobservable—such as left-right ideology or some positive or negative level of affect—although they are selected to correspond to some set of characteristics or categories that humans have defined as meaningful and interesting. For unsupervised methods, human validation is all the more important to establish not only the correctness of the estimates but also the semantic validity of the scales or classes.

## 2 How Text Models Are "Wrong"

Statistical methods for scaling latent traits have received widespread attention in political science in the last decade. Here, we focus on the parametric scaling model formulated by Slapin and Proksch (2008) because it contains explicit, strong assumptions. As we will demonstrate through our validation exercise, while there are many linguistic reasons to recognize the drastically simplifying assumptions of this model's word generation process as wrong, the model nonetheless produces sufficiently valid results to be extremely useful as a measurement model of latent political traits.

In the scaling model—a reparameterization of Goodman’s (1979) row–column (RC) association model—the count of the  $j$ th word in the  $i$ th document,  $C_{ij}$ , is a Poisson process with rate conditional on the document’s position  $\theta_i$ :

$$\begin{aligned} C_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \log \lambda_{ij} &= \alpha_i + \psi_j + \theta_i \beta_j. \end{aligned} \tag{1}$$

Word parameters  $\beta$  and  $\psi$ , incidental document-level parameters  $\alpha$ , and the speaker positions  $\theta$  are jointly estimated by alternating conditional maximum likelihood.<sup>2</sup> Confidence intervals for  $\theta_i$  can be estimated either asymptotically using the information matrix of the likelihood conditional on the word parameter estimates, by using a parametric bootstrap as explained by Slapin and Proksch (2008), or by alternative nonparametric bootstrap methods described below.

This model rests on several strong statistical assumptions to produce valid point estimates and confidence intervals for  $\theta$  that we investigate in what follows.

## 2.1 Unidimensional Latent Trait

One important assumption of the Poisson scaling model in equation (1) is that the set of texts contains word counts generated from a single dimension of relevant variation. The difficulties involved in identifying dimensionality and the consequence of over- and underestimating dimensionality in exploratory factor analysis and related models are well known and no easier when text is involved. One text-specific issue does arise, however: the presence of extra dimensions is confounded with the existence of the kind of topic or frame structure that topic models are designed to extract. For these models, a “topic” consists roughly of a subset of vocabulary with strongly intercorrelated usage. Scaling models will instead treat such correlations as indications of position and scale them accordingly. This implies that the dimensions of difference identified through scaling may differ from the dimensions of difference intended by the researcher, and this is a possibility for which we must be vigilant.

## 2.2 Conditional Independence

The assumption that observed word counts are conditionally independent means that each word is generated independently from others according to a Poisson process with rate specified by the model parameters according to equation (1). Word count variation from causes other than expressed position is assumed to be noise. When this assumption is false, then information from one observed word provides information about the probability that another word is observed. These residual correlations mean that each new word observed provides less information than if it had been independently generated. Consequently, parameter uncertainty, in particular uncertainty about  $\theta$ , will be underestimated.

There are two related problems with respect to conditional independence: unmodeled lexical associations that cause serial dependence and contemporaneous correlation in the form of a hierarchical document structure.

First, to the extent that text scaling models (at least those that treat only words as data) do not account for lexical associations, for example, collocations, compound nouns, and names, these sorts of conditional independence failure are to be expected. In defense of the conditional independence assumption, Laver, Benoit, and Garry (2003) point out that some single words *do* have strong directional associations—the word “tax” and its variants, for instance, is used almost exclusively by more right-leaning parties (who prefer to cut taxes). However, this fails to distinguish all sorts of politically interesting differences among taxes, such as “income taxes,” “taxes on banks,” “carbon taxes,” “inheritance taxes,” and “capital gains taxes.” Not all words, however, tend to have such

---

<sup>2</sup> The model is also straightforward to implement in a Bayesian-MCMC framework using random effects for all parameters, and indeed was fit for two dimensions this way by Monroe and Maeda (2004).

stable associations. The word “free,” for instance, is used for both “free enterprise” (a right-leaning phrase) as well as “GMO-free” and “free health care” (left-leaning phrases). From this perspective, it may seem remarkable that text scaling models based purely on the relative frequencies of atomic words—what linguists call the “bag of words” approach—work at all.<sup>3</sup>

### 2.3 Words as a Poisson Process

The assumption that word-count rates are conditionally distributed as Poisson implies that the variance of this rate is equivalent to the expected rate—a strong assumption that may not hold in natural language word counts. This may be because variation in  $\theta$  may be present in manifestos as different ideological wings of a party add their own sections of text, a fact modeled explicitly by Lo, Proksch, and Slapin (2011). Unmodeled variation in the rate of word occurrence for fixed  $\theta$  may also result from linguistic features. Interestingly, this may be over- or underdispersion. For example, in English, each sentence contains on average about one instance of the word “the.” This regularity is very strong: in the Irish budget debate speeches we examine in more depth later, the rate per hundred words of the word “the” is 7.28 with variance 0.75, about ten times smaller than even a Poisson model with no covariates would predict. Structural zeros are another frequently encountered feature of term-frequency matrixes, caused when a word has *no* chance of occurring in some documents, for example, the term “European Union” prior to the 1980s (when the EU was still called the European Economic Community), or in the party manifestos of Australia where EU policy was simply never a feature of the political discourse. The counterpart of structural absences is when informative words that occur may trigger additional occurrences of the same words. This dependency may be either the result of a real dependency between nearby observations—“true contagion” (or “burstiness”; see Church and Gale 1995)—or when there is merely “apparent contagion” due to variation or serial correlation in values of  $\theta$  (see Cameron and Trivedi 1998, 106 for a review).

To conclude, the problems with conditional independence point to a fundamental observation about applying measurement models to the text scaling task: We have very little idea about what the functional form of the relationship between  $Y$  and  $\theta$  is. The best we can do is identify the model assumptions that fail, be realistic (but sanguine) about the limits of our models given these assumption failures, and—if possible—seek ways to correct them. The dilemma is that we simply have no “true” benchmark to compare the point estimates and confidence intervals from scaling models’ estimates from natural language texts as data. Furthermore, standard techniques designed to test estimator assumption failure, such as Monte Carlo simulation, offer little respite because we have no realistic model for simulating natural language text. We can simulate quantitative data that looks like the quantitative form of natural language text, but only by relying on the very assumptions whose appropriateness we would like to test. For what appears to be a very useful model, in other words, we have no means through standard methods of validating the estimates it produces or of assessing how confidence about the estimates. This points to the need for some other form of validating the strong assumptions of quantitative models of text as data: one based on a human rating of the texts in their original, qualitative forms.

## 3 Data and Methods

Parliamentary speech has been analyzed previously with an aim to locating legislators’ policy preferences (e.g., Monroe and Maeda 2004; Proksch and Slapin 2010), but the dimensions of policy measured in these applications have been less clear.<sup>4</sup> Such problems point to a need to choose texts where the topic is plausibly limited to a single dimension of difference, and where a

<sup>3</sup> Zhang (2004) argues that tight collocations like the ones above need not compromise inferences about  $\theta$  in models with strong conditional independence assumptions, although he admits they will still bias uncertainty estimates downward.

<sup>4</sup> In Monroe and Maeda (2004), for instance, the primary dimension that emerged from a two-dimensional scaling model of US Senate speeches was labeled the “workhorse/showhorse” dimension, for want of a better interpretation. Proksch and Slapin (2010) had to interpret their single estimated dimension from the European Parliament by resorting to correlations with roll-call vote analysis and independent expert surveys.

lot of external information exists on speaker positions that can be used to assess the validity of the text scaling results. For political texts, this suggests a debate where the format and content of text are limited to a single topic: in our case, a budget debate dominated by a single dimension of willingness to accept the burden of austerity measures.

### 3.1 Texts: Legislative Debates over the 2010 Irish Budget

The set of texts we use for comparing human to Poisson-scaled estimates comes from the debate following the presentation of the Irish budget of 2010, taking place in December 2009 in the Irish *Dáil*, the lower house of the Irish parliament. At the time, this budget was widely acknowledged to be the harshest budget in Irish history. In a total of fourteen speeches by key members of each of five political parties, speakers urged either adoption of the harsh fiscal measures as a necessary measure to get the economy back on track, or rejection of the budget as unfair, unnecessary, or unworkable, along with a rejection of the government proposing it. On the government side, speeches by the *Taoiseach* (Prime Minister) Brian Cowen of the governing Fianna Fáil party, and Finance Minister Brian Lenihan of the same party, represented the most pro-budget positions. Three speeches from Green party ministers (Gormley, Cuffe, and Ryan) provided support for the budget, but somewhat more reluctantly, as many in the Green party regretted the austerity measures but felt bound to support the budget by the terms of their party's coalition agreement with Fianna Fáil. On the opposition side, the leaders of the Fine Gael and Labour parties—in addition to two deputies from the opposition, anti-system Sinn Féin party—showed the greatest opposition to the budget, and had allied in an opposition pact to replace the governing coalition in the next election. In all, the budget debates provide a good example of text expressing positions that plausibly reflect a single dimension of relative preference for fiscal austerity versus social protection, and also directly relate to the approval or rejection of specific legislation.

Table 1 lists the fourteen texts we analyze, by speaker and political party, along with the number of total words (tokens) and unique words (types). The median text had 3629 total words, although the shortest contained just 919, and 361 unique words (total 1644). Overall, the corpus contained 49,019 tokens and 4840 different word types. Our construction of the term-document matrix did not exclude any words, such as “stop words” thought *a priori* to be politically uninformative or very low-frequency words such as the sixty-seven hapaxes found in the corpus. Nor did we apply

**Table 1** Quantitative summary of 2010 budget debate texts

Speaker	Party	Tokens	Types
Brian Cowen	FF	5842	1466
Brian Lenihan	FF	7737	1644
Ciaran Cuffe	Green	1141	421
John Gormley	Green	919	361
Eamon Ryan	Green	1513	481
Richard Bruton	FG	4043	947
Enda Kenny	FG	3863	1055
Kieran O'Donnell	FG	2054	609
Joan Burton	LAB	5728	1471
Eamon Gilmore	LAB	3780	1082
Michael Higgins	LAB	1139	437
Ruairí Quinn	LAB	1182	413
Arthur Morgan	SF	6448	1452
Caoimhghin Ó Caoláin	SF	3629	1035
All texts		49,019	4840
Minimum		919	361
Maximum		7737	1644
Median		3704	991
Hapaxes		67	

a stemmer to the words, although tests showed virtually identical results when applied to the set of stemmed words.<sup>5</sup>

Working with these fourteen texts, we selected a panel of human raters to read and evaluate the position expressed in each text, an exercise described next.

### 3.2 *The Qualitative Coding Exercise*

The objective of the statistical scaling model is to estimate the latent positions  $\theta_i$  from a term-document matrix, along a single dimension of difference. The challenges for such a model lie in knowing whether the dimension of difference on which positions are estimated in the scaling model, in fact, correspond to the dimension expected by the researcher, and whether the estimated positions driven by the relative differences in term frequencies accurately reflect each text's position on this dimension. Related to this is the question of meaningfully measuring our *uncertainty* about these point estimates: whether statistical estimates of uncertainty reflect human confidence in perceived differences between texts. Our validation methodology targets both position and uncertainty by asking human readers to assess both, by reading the original texts and answering a series of structured questions about texts and pairs of texts. This takes human judgment beyond the realm of post hoc face validity, into generating independent judgments against which statistical estimates of textual traits may be compared directly.

Our experiment consisted of printing a collection of the budget debates as an eighty-four-page booklet, prefaced by detailed instructions (with full details and examples available in the supplementary materials to this article). The first speech was clearly identified as the introduction of the budget by Finance Minister Lenihan, abridged to remove from the middle of the speech some of the (tedious and specific) technical details of the budget measures. Prime Minister Cowen's speech followed, also slightly abridged to reduce the length caused by technical detail in the middle section. The remaining twelve speeches were referred to only by number. The instructions encouraged readers to make notes in or on the margins of each text, or in a box for notes provided at the end of each text. Most readers made moderate to extensive notes, as judged by the booklets returned to us for coding.

Each speech was followed by at least one question asking the reader to compare the speech just read to a previous speech. We did not ask readers to attempt all ninety-one possible pairwise comparisons, but were able to gauge their perception of differences on twenty-five of these pairs, including all pairs of party leaders, many key intraparty pairs, and several additional pairs. If the reader did perceive one of the two texts in the pairwise comparison to be more pro-budget than the other, then the questionnaire asked the reader to assess his or her confidence in there being a difference, on a scale from one ("Not at all confident") to ten ("Very confident").

At the end of the qualitative exercise, readers were asked to use their notes to place each text on a scale of 0–100, with zero indicating complete support for ("lack of opposition to") the budget, and one hundred indicating complete opposition to the budget. Speeches 1 and 2 from Lenihan and Cowen, respectively, were fixed to zero. Readers had a choice of indicating a point estimate, drawing an interval, or doing both. Most chose to use intervals or some combination of intervals and point estimates, indicating that they understood the instructions and examples and were using the intervals to approximate a confidence region in which they felt the speaker's true position lay.

A total of nineteen readers completed the questionnaire, in exchange for a modest (fixed) honorarium. The readers consisted of five government postgraduate students from a course in qualitative text analysis from the London School of Economics, one post-doctoral fellow from the London School of Economics (LSE) associated with text analysis, nine PhD students and

---

<sup>5</sup> One reason for not excluding common words ("stop words") is that this presumes that we know which words are uninformative. For instance, the pronouns "he" and "she" emerge as highly discriminant words, with Poisson scaling parameter estimates of  $\beta = -0.54$  and  $\beta = -1.68$ , respectively. This is similar to the results of Monroe, Quinn, and Colaresi (2008), who found that uncorrected partisan association measures for female pronouns (of the kind that the Poisson scaling model uses) indicated that they were Democrat words.

three MSc students in political science from Trinity College Dublin, and one PhD student from New York University.<sup>6</sup>

#### 4 Results: Human Qualitative Placements of the Texts

We present the results of the human judgment exercise first because it forms the benchmark by which we will judge the validity of the automated scaling model's estimates of the position of each text.

In Fig. 1a, we plot the mean positions according to the direct human placement of each budget speech along the 0–100 scale, after reflecting the scale (so that one hundred indicates complete support rather than the original zero) and rescaled the means to the unit normal. The bars indicate the 95% confidence intervals assuming the distribution sampling means are *t*-distributed (and the Fianna Fáil deputies Cowen and Lenihan have no intervals because their positions were fixed in the questionnaire at zero). These positions represent our confidence that the point estimate of the speaker's position, on the scale we have asked respondents to place each speech, is contained in this interval. By aggregating the judgment of each human expert into a mean across experts, we have averaged out individual differences in scale usage as well as perceptions of difference to form a “consensus” position. Similar methods have been advocated and defended at length for the use of expert ratings of party positions by Benoit and Laver (2006).

The results shown in Fig. 1a confirm what almost any knowledgeable observer of Irish politics would expect. The degree of antipathy toward the budget, according to the human expert readers, neatly separates opposition parties on the left from government parties on the right, indicated in Fig. 1a by the dashed red line. Also, in accord with expectations, we see strong similarities in intraparty positions, where speakers of the same party tended to cluster together in their positions. Taking far, pro-budget positions are the *Fianna Fáil* deputies Brian Cowen and Brian Lenihan, Taoiseach and Finance Minister respectively. Slightly less pro-budget were the Greens, a position we would expect given their reluctant identification with the governing *Fianna Fáil* whose economic policies were widely seen to be the cause of the financial crisis necessitating the austerity budget being debated. On the opposition side were Fine Gael and Labour, two would-be governing partners whose speakers' positions overlapped, although the median Labour speaker was more opposed to the budget than the median Fine Gael speaker, in line with expectations. Since they reject not only the government and its budgets but also the system itself, it is also entirely plausible that Sinn Féin would be the most opposed to the government's austerity budget, a result also affirmed by the human placements.

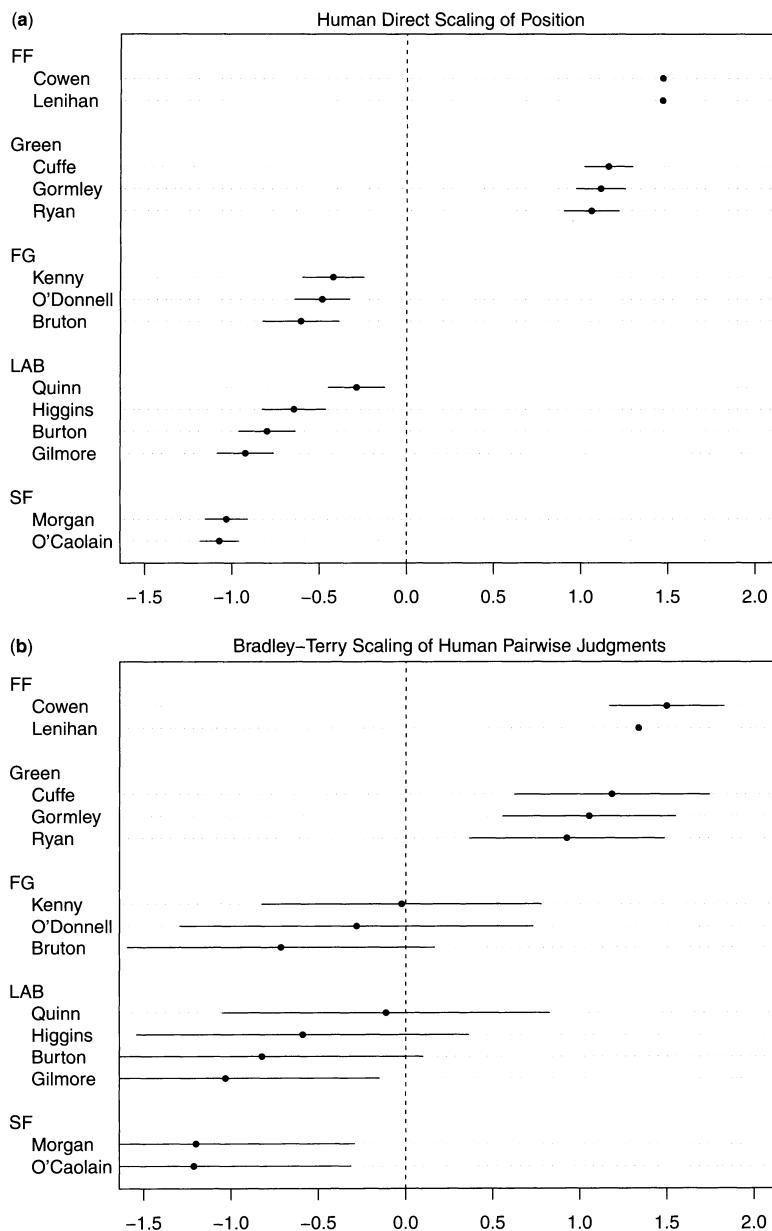
Since we will lean heavily on the human results, we also checked their internal consistency by comparing explicit positions to pairwise judgments. Figure 1b shows the result of applying a Bradley-Terry model for the paired comparisons.<sup>7</sup> The Bradley-Terry model provides an alternative scaling method of the budget positions. Using  $>$  to mean “is more pro-budget than” and denoting two speakers to be compared by  $a$  and  $b$ , this model assumes that  $\text{logit } P(a > b) = \theta_a - \theta_b$ , with  $\theta$  suitably normalized for identification.<sup>8</sup> The results indicate that the human judgments themselves are very consistent across the two types of questions we asked regarding the positions of the speeches, as the placement of people and parties is nearly exactly the same as the direct scaling discussed previously.<sup>9</sup>

<sup>6</sup> In order to test for any effect of prior knowledge of Irish politics, coder education, or prior coding experience, we also compared the results according to a variety of coder-specific variables, but found no statistically significant differences in any factors.

<sup>7</sup> For estimation, we used the R package BradleyTerry2 (Firth 2005).

<sup>8</sup> Following Firth's (2005) suggestions, responses indicating a pair of speeches that could not be distinguished were distributed evenly between the other two responses. Removing pair judgments marked as uncertain perfectly preserved the ordering of speakers.

<sup>9</sup> The uncertainty estimates in Fig. 1b are wide for three reasons. First, we asked a rather small subset of possible comparisons to reduce respondent fatigue, second because pairwise comparisons are inherently less informative than, for example, scalar judgments, and third because we do not make use of the uncertainty information that respondents provided.



**Fig. 1** Human placement results on anti- versus pro-budget scale, human direct scaling (a) and Bradley-Terry scaling from pairwise comparisons (b). Human scaling results are means of the 0–100 placements, with 95% confidence intervals assuming a standard normal sampling distribution of means. The Bradley-Terry results are rescaled to the unit normal for comparison, and were estimated using Lenihan as an anchor point (at zero, prior to rescaling).

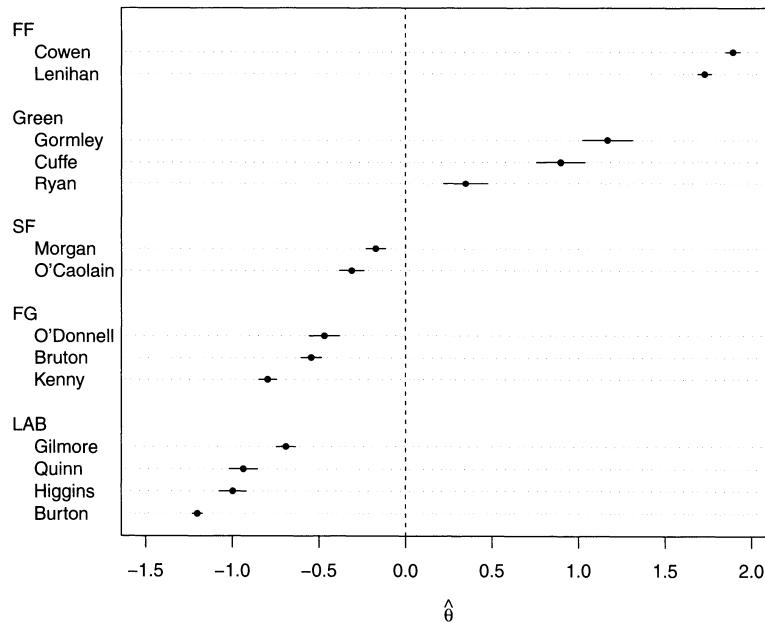
## 5 Results: Validating the Quantitative Text Scaling Results

Having established the benchmark for assessing the placement of each speech on a pro- versus anti-budget divide, and having confirmed the validity of these placements, we now turn to the comparison of the quantitative scaling results against the human ratings to illustrate how validation of quantitative text scaling can be performed.

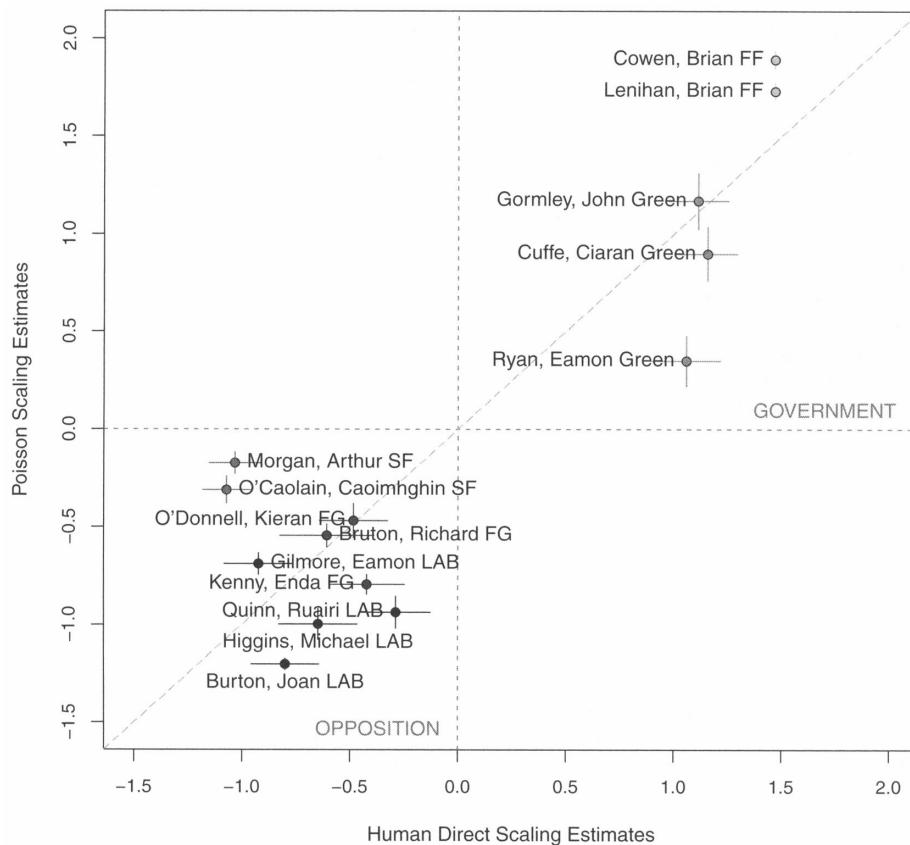
### 5.1 Poisson Scaling Results

Figure 2 presents the results of the Poisson scaling model on the unmodified speeches, along with 95% confidence intervals computed from the maximum likelihood estimation. (The details of this method for computing the confidence intervals, which rely on the model assumptions being correct, are discussed below.) As it turns out, the results correspond to the human placements very closely, correctly separating government and opposition (in the plot, by the dashed red line) and clustering each set of speakers by party. As in the human ratings, the opposition parties of Labour and Fine Gael opposed the budget in the same order as placed by the human raters. On the government side, the Greens and Fianna Fáil argued strongly for the government position. Slightly less pro-budget were the Greens, as expected. On the opposition side were Fine Gael and Labour, two would-be governing partners whose speakers' positions overlapped, although the median Labour speaker was more opposed to the budget than the median Fine Gael speaker, in line with expectations. Using “face validity” as a criterion for model assessment, the Poisson scaling gets at least a B+ for these results.

To compare the methods directly, we have plotted in Fig. 3 the statistical estimates against the human-scaled results, along with the standard parametric 95% confidence intervals from the Poisson scaling model and the standard errors of the mean from the human-scaled results. The dashed cross-hair lines (in red) divide government and opposition for both sets of estimates, as expected. Most parties cluster in similar groups with only minor differences. In the Poisson-scaled estimates, for example, Fine Gael opposition leader Enda Kenny was the most critical of the



**Fig. 2** Poisson scaling results.



**Fig. 3** Direct comparison of human versus Wordfish estimates.

budget, although in the human placements, the Fine Gael positions were indistinguishable. Among Labour speakers, party leader Eamon Gilmore's speech was the most pro-budget in the statistical results, but was rated by human readers as the most opposed to the budget, although they could not distinguish it from Labour Deputy Burton's. For the Green party, the statistical results found significant differences between the three Green ministers, whereas the qualitative placement found their positions indistinguishable.

All in all, most estimates of the speaker positions now lie along the 45° axis of agreement (shown by the long dashed gray lines), except for one party: Sinn Féin. The relatively middle position of Sinn Féin as on the opposition side of the budget, yet still between the main opposition and government party blocs, is incongruent with the human placements who clearly regarded the Sinn Féin speeches as the most extreme expressions of an anti-budget position. The Poisson scaling results for Sinn Féin thus results contradict our expectations based on the knowledge of Irish politics, as well as the results established from the human placements and pairwise comparisons. They suggest a model failure or that some deeper look at the textual data is needed.

## 5.2 Explaining the Incongruent Position of Sinn Féin

Among the main political parties in Irish politics, Sinn Féin is unusual in that one of its key political goals is to redefine the Irish state to reunite the Republic of Ireland with the northern counties currently part of the United Kingdom. As a left-wing party committed to democratic socialism, Sinn Féin considers itself the champion of the poor, minorities, migrants, and the working

class—all groups who would feel most keenly the bite of the austerity measures called for in the government’s budget. Sinn Féin is also Euro-skeptical, having campaigned for a “No” vote in the Lisbon referendum held in 2008, and vociferously opposed both the creation of the National Asset Management Agency in 2009 to take over the bad loans of Ireland’s failing banks, and the acceptance in the same year of an €85 billion rescue package from the European Union and the International Monetary Fund (IMF).

Prior to the budget debate, Sinn Féin publicized its own budget plan, an alternative to both the government budget announced by Minister Lenihan and the Fine Gael proposal articulated by Finance Spokesperson Richard Bruton. This plan called for a nearly €4 billion economic stimulus package, to be offset by raising €3.7 billion in revenue mainly through increased taxation, especially on the wealthiest and highest earners.<sup>10</sup> In the Dáil debates, Sinn Féin Deputy Arthur Morgan expressed not only his rejection of the budget in no uncertain terms but also his rejection of the Fine Gael–Labour alternative:

Replacing Fianna Fáil and the Green Party with Fine Gael and the Labour Party will make no difference to economic recovery... Fine Gael and the Labour Party would implement the same policies in a different package, with the same bad results for the economy. While the establishment parties close ranks and display a disturbing uniformity in their policies... we are unique because we are the only party with an alternative analysis of the situation.

This statement suggests a possible second dimension to the budget debate, not captured here, in the one-dimensional scaling result. To the extent that the opposition to the budget is also opposition to the government, the speeches made by Sinn Féin Deputy Morgan and Dáil party leader Caoimhghín Ó Caoláin are expressing a rejection not only of the budget but also of both “system” party alternatives, including Fine Gael and Labour. The unidimensional scaling results, however, are unable to detect this difference, and estimate the Sinn Féin positions as lying in the middle of the axis of opposition to and support for the budget. Furthermore, there is no easy “fix” to this problem such as trimming a section of irrelevant text from the Sinn Féin speeches, because the anti-establishment language is interwoven with the Sinn Féin commentary on the budget.<sup>11</sup> Although there may be more sophisticated methods for “fixing” the Sinn Féin’s speeches, these are neither simple nor generalizable.

## 6 Results: Evaluating the Uncertainty Estimates

As we saw from the confidence intervals in the direct comparison presented in Fig. 3, many differences *within* party suggested by the Poisson scaling results are judged indistinguishable by humans. This suggests either that the statistical scaling model is capable of detecting more nuanced differences than the human readers or—far more probably—that the standard errors from the statistical estimates are unrealistically small. Having used the human placements as a benchmark to validate the point estimates, we now draw on the aggregate pairwise judgments of difference to validate, or at least to pass some judgment on, the approximate validity of the standard errors produced by the Poisson scaling model.<sup>12</sup> For these differences to be meaningful for text analysis, we focus only on judgments of *difference*: substantively meaningful confidence intervals for positional estimates should assess differences in positions in a manner that corresponds to human perceptions of difference between two speeches, from a qualitative reading. If humans cannot detect a difference in the position of two texts, then a statistical model that declares them to different is underestimating the true level of uncertainty.

<sup>10</sup> *The Road to Recovery: Sinn Féin Pre-Budget 2010 Submission*, [http://www.sinnfein.ie/files/2009/Pre-Budget2010\\_small.pdf](http://www.sinnfein.ie/files/2009/Pre-Budget2010_small.pdf).

<sup>11</sup> In supplementary materials to this article, we also show how Green Deputy John Gormley’s position is wrongly estimated when a section of off-topic speech is included within his text.

<sup>12</sup> This approach holds strong similarities to the pairwise comparisons asked of human raters in Grimmer and King (2011), where a three-scale judgment of similarity from pairwise comparisons of documents was used to evaluate the performance of automated clustering methods on different texts.

The asymptotic maximum likelihood method we use to quantify uncertainty in the Poisson scaling model relies on three assumptions: that the model is correctly specified, that there are enough data available for the curvature of the log likelihood to be approximately quadratic, and that the word parameters  $\psi$  and  $\beta$  are sufficiently well estimated that they can be treated as known. By explicitly conditioning on document lengths, we remove the  $\alpha$  parameters that decouple the likelihoods for each position parameter (see Lowe and Benoit 2011 for further details). If the first assumption is maintained, but we are less confident about the second and third, then we can instead use a bootstrap method (Davison and Hinkley 1997). Slapin and Proksch (2008) recommend a parametric bootstrap to maintain the assumption of a correctly specified model.

Both this and the asymptotic method share a key assumption: the model is correctly specified in all respects. Therefore, we attempt a less model-dependent method by nonparametric bootstrapping directly from the data itself. This involves extending a method that is now well established for nontextual data to resampling from the text itself, *prior to* conversion into the quantitative matrix required for the application of the statistical model. Here, we make a deliberately simple first step, albeit a very significant one, away from the restrictive assumptions of the parametric Poisson scaling model, by resampling texts from their constituent words. More elaborate alternatives are possible, and while we have explored them in other work (Lowe and Benoit 2011), here we focus only on the simplest of these methods, by bootstrapping texts from their words.<sup>13</sup>

To assess the judgments of difference from the human ratings, we draw on the pairwise comparison questions following each text, in which a human reader compared the text just read to other texts previously read. In Table 2, we report the results of pairwise comparisons testing the “null hypothesis” of no difference between the pair of texts indicated in each row. In the “Qualitative/Pairwise” column, we report the results of the aggregated (modal) judgment of the human raters whether there was a detectable difference between the two texts. The “Qualitative/Scaling” column is based on a test of  $H_0 : \theta_a = \theta_b$  for the pair of texts indicated, applying a *t*-test of the difference in sampling means of the human (0–100) placements. The “Poisson/Analytical” and “Poisson/Parametric Bootstrap” columns report similar tests using the estimated standard errors of each  $\hat{\theta}_i$ . In the “NP BS” (nonparametric bootstrap) column, we report the test that the middle 95% region of bootstrap replicates of  $\hat{\theta}_a - \hat{\theta}_b$ , from one hundred replicates, included zero.

Similar to the comparisons of errors, the most conservative measure (in terms of estimating uncertainty) was the nonparametric bootstrap method of Poisson scaling, which found only sixteen of the twenty-five pairwise comparisons to be different.

The asymptotic method and the parametric bootstrap lead to numerically very similar measures of uncertainty, although the bootstrap offers persistently slightly larger uncertainty measures for positions in this setting. This is presumably because for small numbers of documents, the word parameters are *not* particularly well estimated.<sup>14</sup> Nevertheless, both methods are clearly overcertain relative to our respondents’ judgments.

Although it would hardly be novel or controversial to apply a nonparametric bootstrap to any other form of quantitative data, our use of this technique to avoid over reliance on unverifiable model assumptions for the Poisson scaling model has, as far as we are aware, not been applied to the quantitative analysis of textual data in this fashion. There is a tradition of using the bootstrap in correspondence analyses involving text, for example, as discussed in Greenacre (2007, chap. 25). But the focus there is on stability rather than estimating sampling variation.<sup>15</sup>

All the nonparametric bootstrap methods resample from the observed data matrix rather than reconstructing textual data itself, *prior to* converting these data into quantitative form. Our bootstrapping method thus simulates the process of stochastic production of text as in Benoit, Laver, and Mikhaylov (2009): it produces texts that might have been observed, by resampling and

<sup>13</sup> Alternatives to resampling texts from words would involve *block bootstrapping* methods. In block bootstrapping, consecutive blocks of observations of length  $K$  are resampled from the original time series, in either fixed blocks (Carlstein 1986) or overlapping blocks (Künsch 1989).

<sup>14</sup> Empirically, we find that with larger numbers of documents the two methods converge.

<sup>15</sup> Also there is no model reestimation—a single model is fitted and resampled rows of the data matrix are projected onto the space of positions. This effectively assumes known word parameters.

**Table 2** Pairwise comparisons of difference

	Qualitative		Quantitative ( <i>Poisson</i> ) scaling		
	Pairwise	Scaling	Analytical	Parametric bootstrap	NP BS word
<b>Party leader comparisons</b>					
Gilmore LAB versus Cowen FF	DIFF	DIFF	DIFF	DIFF	DIFF
Gormley Green versus Cowen FF	DIFF	DIFF	DIFF	DIFF	DIFF
Gormley Green versus Gilmore LAB	DIFF	DIFF	DIFF	DIFF	DIFF
Gormley Green versus Kenny FG	DIFF	DIFF	DIFF	DIFF	DIFF
Gormley Green versus O'Caolain SF	DIFF	DIFF	DIFF	DIFF	DIFF
Kenny FG versus Cowen FF	DIFF	DIFF	DIFF	DIFF	DIFF
Kenny FG versus Gilmore LAB	DIFF	DIFF	DIFF	DIFF	—
O'Caolain SF versus Cowen FF	DIFF	DIFF	DIFF	DIFF	DIFF
O'Caolain SF versus Gilmore LAB	—	—	DIFF	DIFF	DIFF
O'Caolain SF versus Kenny FG	DIFF	DIFF	DIFF	DIFF	DIFF
<b>Within-party comparisons</b>					
Cowen FF versus Lenihan FF	—	—	DIFF	DIFF	DIFF
Bruton FG versus Kenny FG	DIFF	—	DIFF	DIFF	DIFF
Cuffe Green versus Gormley Green	—	—	DIFF	DIFF	—
Ryan Green versus Cuffe Green	—	—	DIFF	DIFF	DIFF
Ryan Green versus Gormley Green	DIFF	—	DIFF	DIFF	DIFF
Burton LAB versus Quinn LAB	DIFF	DIFF	DIFF	DIFF	—
Burton LAB versus Higgins LAB	DIFF	—	DIFF	DIFF	—
Higgins LAB versus Quinn LAB	DIFF	DIFF	—	—	—
O'Caolain SF versus Morgan SF	—	—	DIFF	DIFF	—
<b>Other comparisons</b>					
Bruton FG versus Gilmore LAB	DIFF	DIFF	DIFF	DIFF	—
O'Donnell FG versus Burton LAB	DIFF	DIFF	DIFF	DIFF	DIFF
Ryan Green versus Morgan SF	DIFF	DIFF	DIFF	DIFF	—
Burton LAB versus Bruton FG	—	—	DIFF	DIFF	DIFF
Quinn LAB versus Bruton FG	DIFF	DIFF	DIFF	DIFF	—
Morgan SF versus Gilmore LAB	DIFF	—	DIFF	DIFF	DIFF
Total observed differences (max 25)	19	15	24	24	16

reconstructing texts that could have been generated from the same data-generating process, and then converts these into quantitative data. The advantages of this method are that it can be easily adapted to preserve any essential features of the textual data-generating process, such as the sequence, grammar, and syntax, simply by redefining the resampling units. Lowe and Benoit (2011), for instance, tested different forms of block bootstrap resampling, although here (primarily to keep the discussion focused) we use only a word-level bootstrap. A second advantage is its generality: There is no method of analyzing text as data to which a text-level bootstrapping cannot be applied. Bootstrapping by resampling units from original texts offers a plausible means of approximating the sampling distributions for just about any form of quantitative text analysis.

## 7 Discussion

From our analysis, comparing qualitative judgments from a panel of nineteen human readers about pairwise differences, uncertainty regarding those differences, and placements of each speech overall on an “anti-/pro-budget” dimension to results to those from a statistical scaling model of relative word frequencies, emerge two novel contributions. First, we have demonstrated a complete research design to assess the semantic validity of unsupervised text scaling models by benchmarking the results to a human cognitive level. By using a panel of readers and aggregating their perceptions,

we have a high degree of confidence in the human judgments. This knowledge combined with direct qualitative judgments, we argue, provides the most *meaningful* benchmark against which to assess quantitative scaling results, in terms of both positional estimates and confidence about these estimates. Although our application has focused on a unidimensional latent trait model, the approach is quite general and could be applied more generally to almost any class of statistical model that treats text as quantitative data.

Second, our validation exercise has demonstrated how even “wrong” models can be useful. Despite numerous political and linguistic reasons why the strict assumptions of the Poisson scaling model are violated—a simplistic model of word counts as a multilevel Poisson process conditional on an unobserved  $\theta_i$ —we have shown that unsupervised text scaling can produce semantically valid results, benchmarked according to independent human judgments of the texts. In comparing the positional estimates from the qualitative and quantitative placements of each speaker, we found a high degree of correspondence, with two exceptions. First, human readers found the positions of the Green party speakers to be indistinguishable, while all methods of Poisson scaling except the nonparametric bootstrapping method found their positions to be different. Second, and more significantly, the Poisson scaled results placed the two Sinn Féin speakers in a middle position, whereas the human readers judged these two speeches to be the most anti-budget of all. This difference may reflect an additional dimension of opposition to both the government and the system-party opposition of Fine Gael and Labour, a dimension not picked up by the unidimensional statistical scaling model.

Our comparison also attempted to benchmark the uncertainty estimates of the parametric scaling model through both direct comparison of the interval sizes and pairwise judgments from human rating and statistical decision making, and we also observed marked differences. Although the standard parametric approaches to the Poisson scaling model judged almost all pairwise speakers’ positions to be significantly different from one another (twenty-four out of twenty-five compared), methods based on human reading produced far fewer perceptions of speakers as different: nineteen from direct questions of difference and fifteen from the aggregated scale placements. The bootstrap method produced results much closer to these perceptions of difference, suggesting that this method offers a far less assumption-reliant, and hence more meaningful, method for estimating confidence intervals from quantitative scaling models of textual data. Wrong model assumptions do a reasonable job of recovering estimates of position but will, unless corrected for, vastly overstate our confidence in them. This result offers a valuable lesson for using “wrong” models, since it shows that while positional estimates may be largely correct, an over-reliance on wrong assumptions will lead us to overstate our confidence in these estimates.

A third contribution lies in the demonstration of text-level bootstrapping. In applying nonparametric bootstrapping methods to textual data, we have taken the treatment of “text as data” to a new level. Bootstrapping methods have been applied previously to coded text units from qualitative content analysis (e.g., Benoit, Laver, and Mikhaylov 2009), but not for the purposes of purely quantitative approaches using text as data. Our nonparametric approach contrasts with both model-dependent asymptotic error computation and parametric bootstrapping methods, by avoiding assumptions that are unrealistic for natural language text. Here, we have applied text bootstrapping to the Poisson scaling model, but this approach is completely general and can be used for any scaling or statistical model drawing on text as data, including classifiers, other scaling methods, topic models, or even descriptive textual statistics, such as vocabulary diversity, readability, or keyword difference measures.

### Funding

European Research Council (ERC-2011-StG 283794-QUANTESS).

### References

- Benoit, K., and M. Laver. 2006. *Party policy in modern democracies*. London: Routledge.  
 ———. 2012. The dimensionality of political space: Epistemological and methodological considerations. *European Union Politics* 13:194–218.

- Benoit, K., M. Laver, and S. Mikhaylov. 2009. Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science* 53(2):495–513.
- Cameron, A. C., and P. K. Trivedi. 1998. *Regression analysis of count data*. Cambridge, UK: Cambridge University Press.
- Carlstein, E. 1986. The use of subseries methods for estimating the variance of a general statistic from a stationary time series. *Annals of Statistics* 14:1171–79.
- Church, K., and W. Gale. 1995. Poisson mixtures. *Natural Language Engineering* 1:163–90.
- Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.
- Firth, D. 2005. Bradley-Terry models in R. *Journal of Statistical Software* 12:1–12.
- Goodman, L. A. 1979. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association* 74(367):537–52.
- Greenacre, M. 2007. *Correspondence analysis in practice*, 2nd ed. London: Chapman and Hall.
- Grimmer, J., and G. King. 2011. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences* 108(7):2643–50.
- Grimmer, J., and B. M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3):267–97.
- Jordan, M. I. 1995. *Why the logistic function? A tutorial discussion on probabilities and neural networks*. Computational Cognitive Science Report 9503, MIT.
- Künsch, H. R. 1989. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* 17:1217–41.
- Laver, M., K. Benoit, and J. Garry. 2003. Estimating the policy positions of political actors using words as data. *American Political Science Review* 97:311–31.
- Lo, J., S.-O. Proksch, and J. B. Slapin. 2011. Party ideology and intra-party cohesion: A theory and measure of election manifestos. Paper presented at MPSA 2011.
- Lowe, W., and K. R. Benoit. 2011. Practical issues in text scaling models: Estimating legislator ideal points in multi-party systems using speeches. Paper presented at MPSA 2011.
- Manning, C. D., P. Raghavan, and H. Schütze. 2008. *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press.
- Mikhaylov, S., M. Laver, and K. Benoit. 2012. Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis* 20:78–91.
- Monroe, B., and K. Maeda. 2004. *Talk's cheap: Text-based estimation of rhetorical ideal-points*. Working paper, Michigan State University.
- Monroe, B. L., K. M. Quinn, and M. P. Colaresi. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16:372–403.
- Proksch, S.-O., and J. B. Slapin. 2010. Position taking in the European Parliament speeches. *British Journal of Political Science* 40(3):587–611.
- Slapin, J. B., and S.-O. Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science* 52(3):705–22.
- Wallach, H. M., I. Murray, R. Salakhutdinov, and D. Mimno. 2009. Evaluation methods for topic models. Proceedings of the 26th International Workshop on Machine Learning, New York, NY.
- Zhang, H. 2004. The optimality of Naïve Bayes. In *FLAIRS Conference*, eds. V. Barr and Z. Markov. Menlo Park, CA: AAAI Press.

# Computer-Assisted Topic Classification for Mixed-Methods Social Science Research

Dustin Hillard  
Stephen Purpura  
John Wilkerson

**ABSTRACT.** Social scientists interested in mixed-methods research have traditionally turned to human annotators to classify the documents or events used in their analyses. The rapid growth of digitized government documents in recent years presents new opportunities for research but also new challenges. With more and more data coming online, relying on human annotators becomes prohibitively expensive for many tasks. For researchers interested in saving time and money while maintaining confidence in their results, we show how a particular supervised learning system can provide estimates of the class of each document (or event). This system maintains high classification accuracy and provides accurate estimates of document proportions, while achieving reliability levels associated with human efforts. We estimate that it lowers the costs of classifying large numbers of complex documents by 80% or more.

**KEYWORDS.** Topic classification, data mining, machine learning, content analysis, information retrieval, text annotation, Congress, legislation

Technological advances are making vast amounts of data on government activity newly available, but often in formats that are of limited value to researchers as well as citizens. In this article, we investigate one approach to transforming these data into useful information. “Topic classification” refers to the process of assigning individual documents (or parts of documents) to a limited set of categories. It is widely used to facilitate search as well as in the study of patterns and trends. To pick an example

of interest to political scientists, a user of the Library of Congress’ THOMAS Web site (<http://thomas.loc.gov>) can use its Legislative Indexing Vocabulary (LIV) to search for congressional legislation on a given topic. Similarly, a user of a commercial Internet service turns to a topic classification system when searching, for example, Yahoo! Flickr for photos of cars or Yahoo! Personals for postings by men seeking women.

Topic classification is valued for its ability to limit search results to documents that closely

---

Dustin Hillard is a Ph.D. candidate in the Department of Electrical Engineering, University of Washington. Stephen Purpura is a Ph.D. student in Information Science, Cornell University.

John Wilkerson is Associate Professor of Political Science at the University of Washington.

This project was made possible with support of NSF grants SES-0429452, SES-00880066, SES-0111443, and SES-00880061. An earlier version of the paper was presented at the Coding Across the Disciplines Workshop (NSF grant SES-0620673). The views expressed are those of the authors and not the National Science Foundation. We thank Micah Altman, Frank Baumgartner, Matthew Baum, Jamie Callan, Claire Cardie, Kevin Esterling, Eduard Hovy, Aleks Jakulin, Thorsten Joachims, Bryan Jones, David King, David Lazer, Lillian Lee, Michael Neblo, James Purpura, Julianna Rigg, Jesse Shapiro, and Richard Zeckhauser for their helpful comments.

Address correspondence to: John Wilkerson, Box 353530, University of Washington, Seattle, WA 98195 (E-mail: [jwilker@u.washington.edu](mailto:jwilker@u.washington.edu)).

Journal of Information Technology & Politics, Vol. 4(4) 2007

Available online at <http://jitp.haworthpress.com>

© 2007 by The Haworth Press. All rights reserved.

doi:10.1080/19331680801975367

31

match the user's interests, when compared to less selective keyword-based approaches. However, a central drawback of these systems is their high costs. Humans—who must be trained and supervised—traditionally do the labeling. Although human annotators become somewhat more efficient with time and experience, the marginal cost of coding each document does not really decline as the scope of the project expands. This has led many researchers to question the value of such labor-intensive approaches, especially given the availability of computational approaches that require much less human intervention.

Yet there are also good reasons to cling to a proven approach. For the task of topic classification, computational approaches are useful only to the extent that they "see" the patterns that interest humans. A computer can quickly detect patterns in data, such as the number of E's in a record. It can then very quickly organize a dataset according to those patterns. But computers do not necessarily detect the patterns that interest researchers. If those patterns are easy to objectify (e.g., any document that mentions George W. Bush), then machines will work well. The problem, of course, is that many of the phenomena that interest people defy simple definitions. "Bad" can mean good—or bad—depending on the context in which it is used. Humans are simply better at recognizing such distinctions, although computerized methods are closing the gap.

Technology becomes increasingly attractive as the size and complexity of a classification task increase. But what do we give up in terms of accuracy and reliability when we adopt a particular automated approach? In this article, we begin to investigate this accuracy/efficiency tradeoff in a particular context. We begin by describing the ideal topic classification system where the needs of social science researchers are concerned. We then review existing applications of computer-assisted methods in political science before turning our attention to a method that has generated limited attention within political science to date: supervised learning systems.

The Congressional Bills Project (<http://www.congressionalbills.org>) currently includes

approximately 379,000 congressional bill titles that trained human researchers have assigned to one of 20 major topic and 226 subtopic categories, with high levels of inter-annotator reliability.<sup>1</sup> We draw on this corpus to test several supervised learning algorithms that use case-based<sup>2</sup> or "learning by example" methods to replicate the work of human annotators. We find that some algorithms perform our particular task better than others. However, combining results from individual machine learning methods increases accuracy beyond that of any single method, and provides key signals of confidence regarding the assigned topic for each document. We then show how this simple confidence estimate can be employed to achieve additional classification accuracy more efficiently than would otherwise be possible.

### **TOPIC CLASSIFICATION FOR SOCIAL SCIENCE DOCUMENT RETRIEVAL**

Social scientists are interested in topic classification for two related reasons: retrieving individual documents and tracing patterns and trends in issue-related activity. Mixed-method studies that combine pattern analyses with case-level investigations are becoming standard, and linked examples are often critical to persuading readers to accept statistical findings (King, Keohane, & Verba, 1994). In *Soft News Goes to War*, for example, Baum (2003) draws on diverse corpora to analyze media coverage of war (e.g., transcripts of "Entertainment Tonight," the jokes of John Stewart's "The Daily Show," and network news programs).

Keyword searches are fast and may be effective for the right applications, but effective keyword searches can also be difficult to construct without knowing what is actually in the data. A search that is too narrow in scope (e.g., "renewable energy") will omit relevant documents, while one that is too broad (e.g., "solar") will generate unwanted false positives. In fact, most modern search engines, such as Google, consciously reject producing a reasonably

comprehensive list of results related to a topic as a design criterion.<sup>3</sup>

Many political scientists rely on existing databases where humans have classified events (decisions, votes, media attention, legislation) according to a predetermined topic system (e.g., Jones & Baumgartner, 2005; Poole & Rosenthal, 1997; Rohde, 2004; Segal & Spaeth, 2002). In addition to enabling scholars to study trends and compare patterns of activity, reliable topic classification can save considerable research time. For example, Adler and Wilkerson (2008) wanted to use the Congressional Bills Project database to study the impact of congressional reforms. To do this, they needed to trace how alterations in congressional committee jurisdictions affected bill referrals. The fact that every bill during the years of interest had already been annotated for topic allowed them to reduce the number of bills that had to be individually inspected from about 100,000 to "just" 8,000.

Topic classification systems are also widely used in the private sector and in government. However, a topic classification system created for one purpose is not necessarily suitable for another. Well-known document retrieval systems such as the Legislative Indexing Vocabulary of the Library of Congress' THOMAS Web site allow researchers to search for documents using preconstructed topics (<http://thomas.loc.gov/liv/livtoc.html>), but the THOMAS Legislative Indexing Vocabulary is primarily designed to help users (congressional staff, lobbyists, lawyers) track down contemporary legislation. This contemporary focus creates the potential for "topic drift," whereby similar documents are classified differently over time as users' conceptions of what they are looking for change.<sup>4</sup>

For example, "women's rights" did not exist as a category in the THOMAS system until sometime after 1994. The new category likely was created to serve current users better, but earlier legislation related to women's rights was not reclassified to ensure intertemporal comparability. Topic drift may be of little concern where contemporary search is concerned, but it is a problem for researchers hoping to compare legislative activity or attention across time. If the topic categories are changing, researchers risk confusing shifts in the substance of legisla-

tive attention with shifts in coding protocol (Baumgartner, Jones, & Wilkerson, 2002).

So, what type of topic classification system best serves the needs of social scientists? If the goals are to facilitate trend tracing and document search, an ideal system possesses the following characteristics. First, it should be discriminating. By this we mean that the topic categories are mutually exclusive and span the entire agenda of topics. The search requires that the system indicate what each document is primarily about, while trend tracing is made more difficult if the same document is assigned to multiple categories. Second, it should be accurate. The assigned topic should reflect the document's content, and there should be a systematic way of assessing accuracy. Third, the ideal system should be reliable. Pattern and trend tracing require that similar documents be classified similarly from one period to the next, even if the terminology used to describe those documents is changing. For example, civil rights issues have been framed very differently from one decade to the next. If the goal is to compare civil rights attention over time, then the classification system must accurately capture attention despite these changing frames. Fourth, it should be probabilistic. In addition to discriminating a document's primary topic, a valuable topic system for search should also identify those documents that address the topic even though they are not primarily about that topic. Finally, it should be efficient. The less costly the system is to implement, the greater its value.

Human-centered approaches are attractive because they meet most of these standards. Humans can be trained to discriminate the main purpose of a document, and their performance can be monitored until acceptable levels of accuracy and reliability are achieved. However, human annotation is also costly. In this article, we ask whether supervised machine learning methods can achieve similar levels of accuracy and reliability while improving efficiency.

We begin by contrasting our approach to several computer-assisted categorization methods currently used in political science research. Only supervised learning systems have the potential to address the five goals of topic classification described above.

### **COMPUTER ASSISTED CONTENT ANALYSIS IN POLITICAL SCIENCE**

Content-analysis methods center on extracting meaning from documents. Applications of computer-assisted content analysis methods have developed slowly in political science over the past four decades, with each innovation adding a layer of complexity to the information gleaned from the method. Here we focus on a selected set of noteworthy projects that serve as examples of some of these important developments (see Table 1).

Data comparison, or keyword matching, was the first search method ever employed on digital data. Keyword searches identify documents that contain specific words or word sequences. Within political science, one of the most sophisticated is KEDS/TABARI (Schrodt, Davis, & Weddle, 1994; Schrodt & Gerner, 1994). TABARI turns to humans to create a set of computational rules for isolating text and associating it with a particular event category. Researchers use the resulting system to analyze changing attention in the international media or other venues.

Systems based on keyword searching can meet the requirements for a solid topic classification system. Keyword search systems such as TABARI can be highly accurate and reliable because the system simply replicates coding decisions originally made by humans.<sup>5</sup> If the system encounters text for which it has not been trained, it does not classify that text. Keyword search systems can also be discriminating, because only documents that include the search terms are labeled. The system can also be probabilistic, by using rules to establish which documents are related to a topic area.

However, for nonbinary classification tasks, achieving the ability to be both discriminating and probabilistic can be expensive, because the system requires explicit rules of discrimination for the many situations where the text touches on more than one topic. For example, the topic “elderly health care issue” includes subjects that are of concern to non-seniors (e.g., health insurance, prevention). Effective keyword searches must account for these contextual factors, and at some point other methods may prove to be more efficient for the same level of accuracy.

TABLE 1. Criteria for Topic Classification and the Appropriateness of Different Computer-Assisted Content Analysis Methods

Criteria	Method				
	Unsupervised Learning (without human intervention)	Keds/Tabari	Wordscores	Hopkins and King, 2007	Supervised Learning
System for topic classification	Partial	Yes	No	Partial	Yes
Discriminates the primary subject of a document?	Yes	No	No	No	Yes
Document level accuracy is assessed	No	No	Yes	No	Yes
Document level reliability is assessed?	No	No	Yes	No	Yes
Indicate secondary topics?	No	No	No	No	Yes
Efficient to implement	Yes, integrating document level accuracy and reliability checks makes the process similar to supervised learning	Yes, but costs rise with scope of task	Yes, costs decline with scope of task	Yes, costs decline with scope of task	Yes, costs decline with scope of task

Unsupervised approaches, such as factor analysis or agglomerative clustering, have been used for decades as an alternative to keyword searching. They are often used as a first step to uncovering patterns in data including document content (Hand, Mannila, & Smyth, 2001). In a recent political science application, Quinn, Monroe, Colaresi, Crespin, and Radev (2006) have used this approach to cluster rhetorical arguments in congressional floor speeches.

Unsupervised approaches are efficient because typically they do not require human guidance, in contrast to data comparison or keyword methods. They also can be discriminating and/or probabilistic, because they can produce mutually exclusive and/or ranked observations. Consider the simplest case of unsupervised learning using agglomerative clustering—near-exact duplicate detection. As a researcher, if you know that 30% of the documents in a data set are near-exact duplicates (99.8% of text content is equivalent) and each has the same topic assigned to it, it would be inefficient to ask humans to label all of these documents. Instead, the research would use an unsupervised approach to find all of the clusters of duplicates, label just one document in the cluster, and then trust the labeling approach to assign labels to the near-exact duplicate documents.<sup>6</sup>

But, to assess the accuracy and reliability of unsupervised methods on more complex content analysis questions, humans must examine the data and decide relevance. And once researchers begin to leverage information from human experts to improve accuracy and reliability (achieve a higher match with human-scored relevance) in the data generation process, the method essentially evolves into a hybrid of the supervised learning method we focus on here.

Another semi-automated method, similar to the method proposed in this article, is the supervised use of word frequencies in Wordscores ([www.wordscores.com](http://www.wordscores.com)). With Wordscores, researchers select model training “cases” that are deemed to be representative of opposite ends of a spectrum (Laver, Benoit, & Garry, 2003). The software then orders other documents along this spectrum based on their similarities and differences to the word frequencies of the end-point model documents. Wordscores

has been used to locate party manifestos and other political documents of multiple nations along the same ideological continuum.

This method can be efficient because it requires only the human intervention required to select training documents and conduct validation. Wordscores is also probabilistic, because it can produce ranked observations. And the method has been shown to be accurate and reliable. However, its accuracy and reliability are application dependent, in that the ranks Wordscores assigns to documents will make sense only if the training documents closely approximate the user’s information retrieval goal. Its small number of training documents limits the expression of the user’s information needs. That is, Wordscores was not designed to place events in discrete categories.

Application-independent methods for conducting algorithmic content analysis do not exist. The goal of such a system would be to generate discriminative and reliable results efficiently and accurately for any content analysis question that might be posed by a researcher. There is a very active community of computer scientists interested in this problem, but, to date, humans must still select the proper method for their application. Many Natural Language Processing (NLP) researchers believe that an application-independent method will never be developed (Kleinberg, 2002).<sup>7</sup>

As a part of this search for a more general method, Hopkins and King recently have developed a supervised learning method that gives, as output, “approximately unbiased and statistically consistent estimates of the proportion of all documents in each category” (Hopkins & King, 2007, p.2). They note that “accurate estimates of these *document category proportions* have not been a goal of most work in the classification literature, which has focused instead on increasing the accuracy of *individual document classification*” (*ibid*). For example, a classifier that correctly estimates 90% of the documents belonging to a class must estimate incorrectly that 10% of those documents belong to other classes. These errors can bias estimates of class proportions (e.g., the proportion of all media coverage devoted to different candidates), depending on how they are distributed.

Like previous work (Purpura & Hillard, 2006), the method developed by Hopkins and King begins with documents labeled by humans, and then statistically analyzes word features to generate an efficient, discriminative, multiclass classification. However, their approach of estimating proportions is not appropriate for the researchers interested in mixed-methods research requiring the ability to analyze documents within a class. Despite this limitation, mixed-methods researchers may still want to use the Hopkins and King method to validate estimates from alternative supervised learning systems. Because it is the only other method (in the political science literature) of those mentioned that relies on human-labeled training samples, it does offer a unique opportunity to compare the prediction accuracy of our supervised learning approach in our problem domain to another approach (though the comparison must be restricted to proportions).

### **SUPERVISED LEARNING APPROACHES**

Supervised learning (or machine learning classification) systems, have been the focus of more than 1,000 scholarly publications in the computational linguistics literature in recent years (Breiman, 2001; Hand, 2006; Mann, Mimno, & McCallum, 2006; Mitchell, 1997; Yang & Liu, 1999). These systems have been used for many different text annotation purposes but have been rarely used for this purpose in political science.

In this discussion, we focus on supervised learning systems that statistically analyze terms within documents of a corpus to create rules for classifying those documents into classes. To uncover the relevant statistical patterns, annotators mark a subset of the documents in the corpus as being members of a class. The researcher then develops a “document representation” that draws on this “training set” to accurately machine annotate previously unseen documents in the corpus referred to as the “test set.”

Practically, a document representation can be any numerical summary of a document in

the corpus. Examples might include a binary indicator variable, which specifies whether the document contains a picture, a vector containing “term weights” for each word in the document, or a real number in the interval (i.e., 0, infinity) that represents the cost of producing the document. Typically, a critical selection criterion is empirical system performance. If a human can separate all of the documents in a corpus perfectly by asking, for example, whether a key e-mail address appears, then a useful document representation would be a binary indicator variable specifying whether the e-mail address appears in each document. However, for classification tasks that are more complex, simplicity, calculation cost, and theoretical justifications are also relevant selection criteria.

Our document representation consists of a vector of term weights, also known as feature representation, as documented in Joachims (1998). For the term weights, we use both tf\*idf (term frequency multiplied by inverse document frequency) and a mutual information weight (Purpura & Hillard, 2006). The most typical feature representation first applies Porter stemming to reduce word variants to a common form (Porter, 1980), before computing term frequency in a sample divided by the inverse document frequency (to capture how often a word occurs across all documents in the corpus) (Papineni, 2001).<sup>8</sup> A list of common words (stop words) also may be omitted from each text sample.

Feature representation is an important research topic in itself, because different approaches yield different results depending on the task at hand. Stemming can be replaced by a myriad of methods that perform a similar task—capturing the signal in the transformation of raw text to numeric representation—but with differing results. In future research, we hope to demonstrate how alternative methods of preprocessing and feature generation can improve the performance of our system.

For topic classification, a relatively comprehensive analysis (Yang & Liu, 1999) finds that support vector machines (SVMs) are usually the best performing model. Purpura and Hillard

(2006) applied a SVM model to the corpus studied here with high fidelity results. We are particularly interested in whether combining the decisions of multiple supervised learning systems can improve results. This combined approach is known as ensemble learning (Brill & Wu, 1998; Curran, 2002; Dietterich, 2000). Research indicates that ensemble approaches yield the greatest improvements over a single classifier when the individual classifiers perform with similar accuracy but make different types of mistakes.

### **Algorithms**

We will test the performance of four alternatives: Naïve Bayes, SVM, Boostexter, and MaxEnt.

#### *Naïve Bayes*

Our Naïve Bayes Classifier uses a decision rule and a Bayes probability model with strong assumptions of independence of the features ( $tf \cdot idf$ ). Our decision rule is based on MAP (maximum a posteriori), and it attempts to select a label for each document by selecting the class that is most probable. Our implementation of the Naïve Bayes comes from the rainbow toolkit (McCallum, 1996).

#### *The SVM Model*

The SVM system builds on binary pairwise classifiers between each pair of categories, and chooses the one that is selected most often as the final category (Joachims, 1998). Other approaches are also common (such as a committee of classifiers that test one vs. the rest), but we have found that the initial approach is more time efficient with equal or greater performance. We use a linear kernel, Porter stemming, and a feature value (mutual information) that is slightly more detailed than the typical inverse document frequency feature. In addition, we prune those words in each bill that occur less often than the corpus average. Further details and results of the system are described in Purpura and Hillard (2006).

#### *Boostexter Model*

The Boostexter tool allows for features of a similar form to the SVM, where a word can be associated with a score for each particular text example (Schapire & Singer, 2000). We use the same feature computation as for the SVM model, and likewise remove those words that occur less often than the corpus average. Under this scenario, the weak learner for each iteration of AdaBoost training consists of a simple question that asks whether the score for a particular word is above or below a certain threshold. The Boostexter model can accommodate multiclass tasks easily, so only one model need be learned.

#### *The MaxEnt Model*

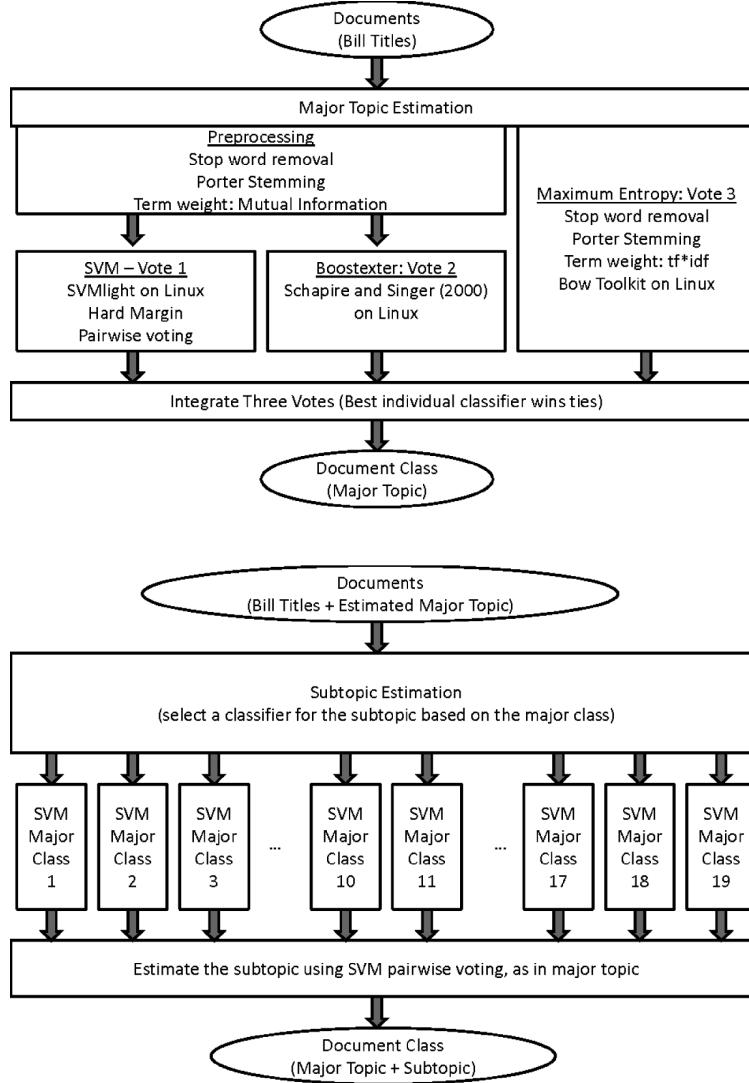
The MaxEnt classifier assigns a document to a class by converging toward a model that is as uniform as possible around the feature set. In our case, the model is most uniform when it has maximal entropy. We use the rainbow toolkit (McCallum, 1996). This toolkit provides a cross validation feature that allows us to select the optimal number of iterations. We provide just the raw words to rainbow, and let it run word stemming and compute the feature values.

Figure 1 summarizes how we apply this system to the task of classifying congressional bills based on the word features of their titles. The task consists of two stages. In the first, we employ the ensemble approach developed here to predict each bill's major topic class. Elsewhere, we have demonstrated that the probability of correctly predicting the subtopic of a bill, given a correct prediction of its major topic, exceeds .90 (Hillard, Purpura, & Wilkerson, 2007; Purpura & Hillard, 2006). We leverage this valuable information about likely subtopic class in the second stage by developing unique subtopic document representations (using the three algorithms) for each major topic.<sup>9</sup>

#### *Performance Assessment*

We assess the performance of our automated methods against trained human annotators. Although we report raw agreement between

FIGURE 1. Topic classifying Congressional bill titles.



human and machine for simplicity, we also discount this agreement for confusion, or the probability of that that the human and machine might agree by chance. Chance agreement is of little concern when the number of topics is large. However, in other contexts, chance agreement may be a more relevant concern.

Cohen's Kappa statistic is a standard metric used to assess interannotator reliability between two sets of results while controlling for chance agreement (Cohen, 1968). Usually, this technique assesses agreement between two human annotators, but the computational linguistics field also uses it to assess agreement between

human and machine annotators. Cohen's Kappa statistic is defined as:

$$\kappa = \frac{p(A) - p(E)}{1 - p(E)}$$

In the equation,  $p(A)$  is the probability of the observed agreement between the two assessments:

$$p(A) = \frac{1}{N} \sum_{n=1}^N I(Human_n == Computer_n)$$

where  $N$  is the number of examples, and  $I()$  is an indicator function that is equal to 1 when the two annotations (human and computer) agree on a particular example.  $P(E)$  is the probability of the agreement expected by chance:

$$p(E) = \frac{1}{N^2} \sum_{c=1}^C (HumanTotal_c \times ComputerTotal_c)$$

where  $N$  is again the total number of examples and the argument of the sum is a multiplication of the marginal totals for each category. For example, for category 3—health—the argument would be the total number of bills a human annotator marked as category 3 multiplied by the total number of bills the computer system marked as category 3. This multiplication is computed for each category, summed, and then normalized by  $N^2$ .

Due to bias under certain constraint conditions, computational linguists also use another standard metric, the AC1 statistic, to assess interannotator reliability (Gwet, 2002). The AC1 statistic corrects for the bias of Cohen's Kappa by calculating the agreement by chance in a different manner. It has a similar form:

$$AC1 = \frac{p(A) - p(E)}{1 - p(E)}$$

but the  $p(E)$  component is calculated differently:

$$p(E) = \frac{1}{C-1} \sum_{c=1}^C (\pi_c (1 - \pi_c))$$

where  $C$  is the number of categories, and  $\pi_c$  is the approximate chance that a bill is classified as category  $c$ :

$$\pi_c = \frac{(HumanTotal_c + ComputerTotal_c)/2}{N}$$

In this study we report just AC1 because there is no meaningful difference between Kappa and AC1.<sup>10</sup> For annotation tasks of this level of complexity, a Cohen's Kappa or AC1 statistic of 0.70 or higher is considered to be very good agreement between annotators (Carletta, 1996).

### **Corpus: The Congressional Bills Project**

The Congressional Bills Project (<http://www.congressionalbills.org>) archives information about federal public and private bills introduced since 1947. Currently the database includes approximately 379,000 bills. Researchers use this database to study legislative trends over time as well as to explore finer questions such as the substance of environmental bills introduced in 1968, or the characteristics of the sponsors of environmental legislation.

Human annotators have labeled each bill's title (1973–98) or short description (1947–72) as primarily about one of 226 subtopics originally developed for the Policy Agendas Project (<http://www.policyagendas.org>). These subtopics are further aggregated into 20 major topics (see Table 2). For example, the major topic of environment includes 12 subtopics corresponding to longstanding environmental issues, including species and forest protection, recycling, and drinking water safety, among others. Additional details can be found online at <http://www.policyagendas.org/codebooks/topicindex.html>.

The students (graduate and undergraduate) who do the annotation train for approximately three months as part of a year-long

TABLE 2. Major Topics of the Congressional Bills Project

1 Macroeconomics
2 Civil Rights, Minority Issues, Civil Liberties
3 Health
4 Agriculture
5 Labor, Employment, and Immigration
6 Education
7 Environment
8 Energy
10 Transportation
12 Law, Crime, and Family Issues
13 Social Welfare
14 Community Development and Housing Issues
15 Banking, Finance, Domestic Commerce
16 Defense
17 Space, Science, Technology, Communications
18 Foreign Trade
19 International Affairs and Foreign Aid
20 Government Operations
21 Public Lands and Water Management
99 Private Legislation

commitment. Typically, each student annotates 200 bills per week during the training process. To maintain quality, interannotator agreement statistics are regularly calculated. Annotators do not begin annotation in earnest until interannotator reliability (including a master annotator) approaches 90% at the major topic level and 80% at the subtopic level.<sup>11</sup> Most bills are annotated by just one person, so the dataset undoubtedly includes annotation errors.

However, it is important to recognize that interannotator disagreements are usually legitimate differences of opinion about what a bill is primarily about. For example, one annotator might place a bill to prohibit the use of live

rabbits in dog racing in the sports and gambling regulation category (1526), while another might legitimately conclude that it is primarily about species and forest protection (709). The fact that interannotator reliability is generally high, despite the large number of topic categories, suggests that the annotators typically agree on where a bill should be assigned. In a review of a small sample, we found that the distribution between legitimate disagreements and actual annotation errors was about 50/50.

## EXPERIMENTS AND FINDINGS

The main purpose of automated text classification is to replicate the performance of human labelers. In this case, the classification task consists of either 20 or 226 topic categories. We exploit the natural hierarchy of the categories by first building a classification system to determine the major category, and then building a child system for each of the major categories that decides among the subcategories within that major class, as advocated by Koller and Sahami (1997).

We begin by performing a simple random split on the entire corpus and use the first subset for training and the second for testing. Thus, one set of about 190,000 labeled samples is used to predict labels on about 190,000 other cases.

Table 3 shows the results produced when using our text preprocessing methods and four off-the-shelf computer algorithms. With 20 major topics and 226 subtopics, a random assignment of bills to topics and subtopics can be expected to yield very low levels of

TABLE 3. Bill Title Interannotator Agreement for Five Model Types

	SVM	MaxEnt	Boostexter	Naïve Bayes	Ensemble
Major topic $N = 20$	88.7% (.881)	86.5% (.859)	85.6% (.849)	81.4% (.805)	89.0% (.884)
Subtopic $N = 226$	81.0% (.800)	78.3% (.771)	73.6% (.722)	71.9% (.705)	81.0% (.800)

Note. Results are based on using approximately 187,000 human-labeled cases to train the classifier to predict approximately 187,000 other cases (that were also labeled by humans but not used for training). Agreement is computed by comparing the machine's prediction to the human assigned labels. (AC1 measure presented in parentheses).

accuracy. It is therefore very encouraging to find high levels of prediction accuracy across the different algorithms. This is indicative of a feature representation—the mapping of text to numbers for analysis by the machine—that reasonably matches the application.

The ensemble learning voting algorithm combines the best of the four (SVM, MaxEnt, and Boostexter) marginally improves interannotator agreement (compared to SVM alone) by 0.3% (508 bills). However, combining information from three algorithms yields important additional information that can be exploited to lower the costs of improving accuracy. When the three algorithms predict the same major topic for a bill, the prediction of the machine matches the human-assigned category 94% of the time (see Table 4). When the three algorithms disagree by predicting different major topics, collectively the machine predictions match the human annotation team only 61% of the time. The AC1 measure closely tracks the simple accuracy measure, so for brevity we present only accuracy results in the remaining experiments.

### **PREDICTING TO THE FUTURE: WHEN AND WHERE SHOULD HUMANS INTERVENE?**

A central goal of the Congressional Bills Project (as well as many other projects) is to turn to automated systems to lower the costs of labeling new bills (or other events) as opposed to labeling events of the distant past. The previous experiments shed limited light on the value of the method for this task. How different are our results if we train on annotated bills from previous Congresses to predict the topics of bills of future Congresses?

From past research we know that topic drift across years can be a significant problem. Although we want to minimize the amount of time that the annotation team devotes to examining bills, we also need a system that approaches 90% accuracy. To address these concerns, we adopt two key design modifications. First, we implement a partial memory learning system. For example, to predict class labels for the bills

of the 100th Congress (1987–1988), we only use information from the 99th Congress, plus whatever data the human annotation team has generated for the 100th as part of an active learning process. We find that this approach yields results equal to, or better than, what can be achieved using all available previous training data.

The second key design decision is that we only want to accept machine-generated class labels for bills when the system has high confidence in the prediction. In other cases, we wish to have humans annotate the bills, because we have found that this catches cases of topic drift and it minimizes mistakes. One implication of Table 4 is that the annotation team may be able to trust the algorithms' prediction when all three methods agree and limit its attention to the cases of disagreement where they disagree. But we need to confirm that the results are comparable when we use a partial memory learning system.

For the purposes of these experiments, we will focus on predicting the topics of bills from the 100th Congress to the 105th Congress using only the bill topics from the previous Congress as training data. This is the best approximation of the “real world” case that we are able to construct, because (a) these congressional sessions have the lowest computer/human agreement of all of the sessions in the data set, (b) the 105th Congress is the last human-annotated session, and (c) the first production experiment with live data will use the 105th Congress’ data to predict the class labels for the bills of the 106th Congress. The results reported in Table 5 are at

TABLE 4. Prediction Success for 20 Topic Categories When Machine Learning Ensemble Agrees and Disagrees

	Methods Agree	Methods Disagree
correct	94%	61%
incorrect	6%	39%
cases	85%	15%
(N of Bills)	(158,762)	(28,997)

Note. Based on using 50% of the sample to train the systems to predict to the other 50%.

the major topic only. As mentioned, the probability of correctly predicting the subtopic of a bill, given a correct prediction of major topic class, exceeds 0.90 (Hillard et al., 2007; Purpura & Hillard, 2006).

Several results in Table 5 stand out. Overall, we find that when we train on a previous Congress to predict the class labels of the next Congress, the system correctly predicts the major topic about 78.5% of the time without any sort of human intervention. This is approximately 12% below what we would like to see, but we have not spent any money on human annotation yet.

How might we strategically invest human annotation efforts to best improve the performance of the system? To investigate this question we will begin by using the major topic class labels of bills in the 99th Congress to predict the major topic class labels of the bills in the 100th Congress. Table 6 reports the percentage of cases that agree between the

machine and the human team in three situations: when the three algorithms agree, when two of them agree, and when none of them agree. When all three agree, only 10.3% of their predictions differ from those assigned by the human annotators. But when only two agree, 39.8% of the predictions are wrong by this standard, and most (58.5%) are wrong when the three algorithms disagree.

Of particular note is how this ensemble approach can guide future efforts to improve overall accuracy. Suppose that only a small amount of resources were available to pay human annotators to review the automated system's prediction for the purpose of improving overall accuracy. (Remember that in an applied situation, we would not know which assignments were correct and which were wrong.) With an expected overall accuracy rate of about 78%, 78% of the annotator's efforts would be wasted on inspecting correctly labeled cases.

TABLE 5. Prediction Success When the Ensemble Agrees and Disagrees

Congress	Train	Test	Bills in Test Set (N)	Ensemble Methods Agree (%)	Correctly Predicts Major Topic (%)			
					When 3 Methods Agree	When Methods Disagree	Combined Agree and Disagree	Best Individual Classifier
99th	100th	8508	61.5	89.7	59.3	78.0	78.3	
100th	101st	9248	62.1	93.0	61.5	81.1	80.8	
101st	102nd	9602	62.4	90.3	61.1	79.3	79.3	
102nd	103rd	7879	64.8	90.1	60.2	79.6	79.5	
103rd	104th	6543	62.4	89.0	57.5	77.1	76.6	
104th	105th	7529	60.0	87.4	58.9	76.0	75.6	
	Mean	8218	62.2	89.9	59.7	78.5	78.4	

Note. The "best individual classifier" is usually the SVM system.

TABLE 6. Prediction Success When the Ensemble Agrees and Disagrees

	3 Methods Agree	2 Methods Agree	No Agreement	Overall
Correct	89.7%	64.2%	41.5%	78.0%
Incorrect	10.3%	36.8%	58.5%	22.0%
Share of incorrect cases	28.8%	50.8%	20.2%	-----
All cases	61.5%	30.8%	7.7%	100.0%
(N of Bills)	(5233)	(2617)	(655)	(8508)

Note. Training on bills of the 99th Congress to predict bills of the 100th Congress.

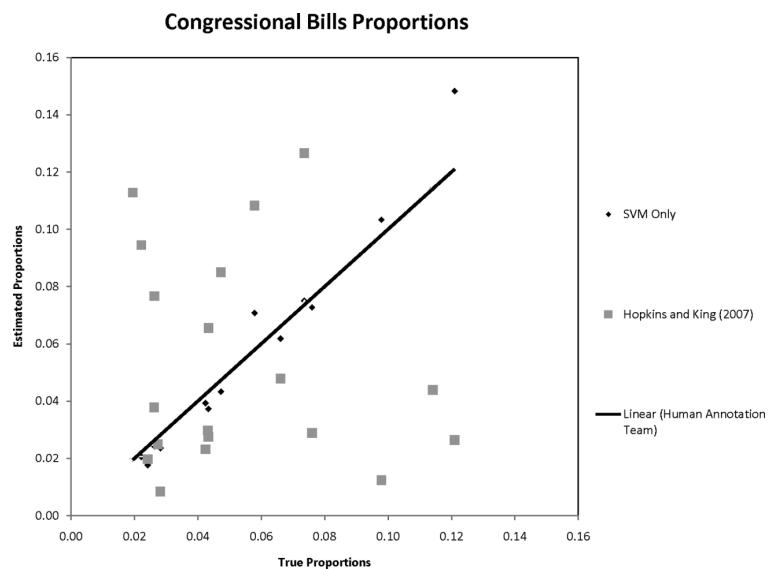
A review of just 655 bills where the three methods disagree (i.e., less than 8% of the sample) can be expected to reduce overall annotation errors by 20%. In contrast, inspecting the same number of cases when the three methods agree would reduce overall annotation errors by just 3.5%. If there are resources to classify twice as many bills (just 1,310 bills, or about 15% of the cases), overall error can be reduced by 32%, bumping overall accuracy from 78% to 85%. Coding 20% of all bills according to this strategy increases overall accuracy to 87%.

In the political science literature, the most appropriate alternative approach for validating the methods presented here is the one recently advocated by Hopkins and King (2007). While their method, discussed earlier, does not predict to the case level and is therefore inadequate for the goals we have established in this work. We can compare estimates of proportions by applying our software and the ReadMe software made available by Hopkins and King (<http://gking.harvard.edu/readme/>) to the same dataset. We trained the ReadMe algorithm and the best performing algorithm of our ensemble (SVM) on the human-assigned topics of bills of the

104th Congress (1995–96), and then predicted the proportion of bills falling into each of 20 major topics of the 105th Congress.

In Figure 2, an estimate that lies along the diagonal is perfectly predicting the actual proportion of bills falling into that topic category. The further the estimate strays from the diagonal, the less accurate the prediction. Thus, Figure 2 indicates that the SVM algorithm—which labels cases in addition to predicting proportions—is performing as well and sometimes much better than the ReadMe algorithm. These findings buttress our belief that estimation bias is just one of the considerations that should affect methods selections. In this case, the greater document level classification accuracy of the SVM discriminative approach translated into greater accuracy of the estimated proportions. In practice, mixed method researchers using our approach gain the ability to inspect individual documents in a class (because they know the classification of each document) while still having confidence that the estimates of the proportions of the documents assigned to each class are reasonable. The technique for bias reduction proposed by Hopkins and King

FIGURE 2. Predicting document proportions via two methods.



could then be used as a postprocessing step—potentially further improving proportion predictions.

## CONCLUSIONS

Topic classification is a central component of many social science data collection projects. Scholars rely on such systems to isolate relevant events, to study patterns and trends, and to validate statistically derived conclusions. Advances in information technology are creating new research databases that require more efficient methods for topic classifying large numbers of documents. We investigated one of these databases to find that a supervised learning system can accurately estimate a document's class as well as document proportions, while achieving the high inter-annotator reliability levels associated with human efforts. Moreover, compared to human annotation alone, this system lowers costs by 80% or more.

We have also found that combining information from multiple algorithms (ensemble learning) increases accuracy beyond that of any single algorithm, and also provides key signals of confidence regarding the assigned topic for each document. We then showed how this simple confidence estimate could be employed to achieve additional classification accuracy.

Although the ensemble learning method alone offers a viable strategy for improving classification accuracy, additional gains can be achieved through other active learning interventions. In Hillard et al. (2007), we show how confusion tables that report classification errors by topic can be used to target follow-up interventions more efficiently. One of the conclusions of that research was that stratified sampling approaches can be more efficient than random sampling, especially where smaller training samples are concerned. In addition, much of the computational linguistics literature focuses on feature representations and demonstrates that experimentation in this area is also likely to lead to improvements. The Congressional Bills Project is in the public domain (<http://www.congressionalbills.org>).

[congressionalbills.org](http://www.congressionalbills.org)). We hope that this work inspires others to improve upon it.

We appreciate the attraction of less costly approaches such as keyword searches, clustering methodologies, and reliance on existing indexing systems designed for other purposes. Supervised learning systems require high-quality training samples and active human intervention to mitigate concerns such as topic drift as they are applied to new domains (e.g., new time periods), but it is also important to appreciate where other methods fall short as far as the goals of social science research are concerned. For accurately, reliably, and efficiently classifying large numbers of complex individual events, supervised learning systems are currently the best option.

## NOTES

1. [http://www.congressionalbills.org/Bills\\_Reliability.pdf](http://www.congressionalbills.org/Bills_Reliability.pdf)

2. We use *case-based* to mean that *cases* (examples marked with a class) are used to train the system. This is conceptually similar to the way that reference cases are used to train law school students.

3. Google specifically rejects use of *recall* as a design criterion in their design documents, available at <http://www-db.stanford.edu/~backrub/google.html>

4. Researchers at the Congressional Research Service are very aware of this limitation of their system, which now includes more than 5,000 subject terms. However, we have been reminded on more than one occasion that THOMAS's primary customer (and the entity that pays the bills) is the U.S. Congress.

5. TABARI does more than classify, but at its heart it is an event classification system just as Google (at its heart) is an information retrieval system.

6. If we were starting from scratch, we would employ this method. Instead, we use it to check whether near-exact duplicates are labeled identically by both humans and software. Unfortunately, humans make the mistake of mislabeling near-exact duplicates more times than we care to dwell upon, and we are glad that we now have computerized methods to check them.

7. Many modern popular algorithmic NLP text classification approaches convert a document into a mathematical representation using the *bag of words* method. This method reduces the contextual information available to the machine. Different corpus domains and applications require more contextual information to increase effectiveness. Variation in the document pre-processing (including morphological transformation) is one of the key methods

for increasing effectiveness. See Manning and Shütze (1999) for a helpful introduction to this subject.

8. Although the use of features similar to tf\*idf (term frequency multiplied by inverse document frequency) dates back to the 1970's, we cite Papineni's literature review of the area.

9. Figure 1 depicts the final voting system used to predict the major and subtopics of each Congressional Bill. The SVM system, as the best performing classifier, is used alone for the subtopic prediction system. However, when results are reported for individual classifier types (SVM, Boostexter, MaxEnt, and Naïve Bayes), the same classifier system is used to predict both major and subtopics.

10. Cohen's Kappa, AC1, Krippendorf's Alpha, and simple percentage comparisons of accuracy are all reasonable approximations for the performance of our system because the number of data points and the number of categories are large.

11. See <http://www.congressionalbills.org/BillsReliability.pdf>

## REFERENCES

- Adler, E. S., & Wilkerson, J. (2008). Intended consequences? Committee reform and jurisdictional change in the House of Representatives. *Legislative Studies Quarterly*, 33(1), 85–112.
- Baum, M. (2003). *Soft news goes to war: Public opinion and American foreign policy in the new media age*. Princeton NJ: Princeton University Press.
- Baumgartner, F., Jones, B. D., & Wilkerson, J. (2002). Studying policy dynamics. In F. Baumgartner & B. D. Jones (Eds.), *Policy dynamics* (pp. 37–56). Chicago: University of Chicago Press.
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231.
- Brill, E. & Wu, J. (1998). Classifier combination for improved lexical disambiguation. In *Proceedings of the COLING-ACL '98* (pp. 191–195). Philadelphia, PA: Association for Computational Linguistics.
- Carletta, J. (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2), 249–254.
- Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220.
- Curran, J. (2002). Ensemble methods for automatic thesaurus extraction. *Proceedings of the conference on empirical methods in natural language processing* (pp. 222–229). Morristown, NJ: Association for Computational Linguistics.
- Dietterich, T. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857, 1–15.
- Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-rater Reliability Assessment*, 1(April), 1–6.
- Hand, D. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1), 1–14.
- Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining (adaptive computation and machine learning)*. Cambridge, MA: MIT Press.
- Hillard, D., Purpura, S., & Wilkerson, J. (2007). An active learning framework for classifying political text. Presented at the annual meetings of the Midwest Political Science Association, Chicago, April.
- Hopkins, D., & King, G. (2007). *Extracting systematic social science meaning from text*. Unpublished manuscript (Sept. 15, 2007), Center for Basic Research, Harvard University.
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European conference on machine learning* (pp. 137–142). Chemnitz, Germany: Springer.
- Jones, B. D., & Baumgartner, F. B. (2005). *The politics of attention: How government prioritizes problems*. Chicago: University of Chicago Press.
- King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.
- Kleinberg, J. (2002). An impossibility theorem for clustering. *Proceedings of the 2002 Conference on Advances in Neural Information Processing Systems*, 15, 463–470.
- Koller, D., & Sahami, M. (1997). Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 170–78. San Francisco, CA: D. H. Fisher, & E. Morgan Publishers.
- Laver, M., Benoit, K., & Garry, J. (2003). Estimating the policy positions of political actors using words as data. *American Political Science Review*, 97(2), 311–331.
- Mann, G., Mimno, D., & McCallum, A. (2006). Bibliometric impact measures and leveraging topic analysis. *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 65–74). New York: Association for Computer Machinery.
- Manning, C., & Shütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- McCallum, A. (1996). *Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering*. Available at <http://www.cs.cmu.edu/mccallum/bow>
- Mitchell, T. (1997). *Machine learning*. New York: McGraw-Hill.
- Papineni, K. (2001). Why inverse document frequency? In *Proceedings of the North American Chapter of the Association for Computational Linguistics* (pp. 1–8). Morristown, NJ: Association for Computing Machinery.

- Poole, K., & Rosenthal, H. (1997). *Congress: A political-economic history of roll call voting*. New York: Oxford University Press.
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 16*(3), 130–137.
- Purpura, S., & Hillard, D. (2006). Automated classification of congressional legislation. In *Proceedings of the International Conference on Digital Government Research* (pp. 219–225). New York: Association for Computer Machines.
- Quinn, K., Monroe, B., Colaresi, M., Crespin, M., & Radev, D. (2006). *An automated method of topic-coding legislative speech over time with application to the 105th–108th U.S. Senate*. Presented at the Annual Meetings of the Society for Political Methodology, Seattle, WA.
- Rohde, D. W. (2004). *Roll call voting data for the United States House of Representatives, 1953–2004*. Compiled by the Political Institutions and Public Choice Program, Michigan State University, East Lansing, MI. Available at <http://crespin.myweb.uga.edu/pipcdta.htm>
- Schapire, R. E., & Singer, Y. (2000). Boostexter: A boosting based system for text categorization. *Machine Learning, 39*(2/3), 135–168.
- Schrodt, P., Davis, S., & Weddle, J. (1994). Political Science: KEDS—a program for the machine coding of event data. *Social Science Computer Review, 12*(4), 561–587.
- Schrodt, P., & Gerner, D. (1994). Validity assessment of a machine-coded event data set for the Middle East, 1982–92. *American Journal of Political Science, 38*(3), 825–844.
- Segal, J., & Spaeth, H. (2002). *The Supreme Court and the attitudinal model revisited*. Cambridge, England: Cambridge University Press.
- Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval*. San Francisco, New York: Association for Computer Machinery.

# A Method of Automated Nonparametric Content Analysis for Social Science

**Daniel J. Hopkins** Georgetown University  
**Gary King** Harvard University

*The increasing availability of digitized text presents enormous opportunities for social scientists. Yet hand coding many blogs, speeches, government records, newspapers, or other sources of unstructured text is infeasible. Although computer scientists have methods for automated content analysis, most are optimized to classify individual documents, whereas social scientists instead want generalizations about the population of documents, such as the proportion in a given category. Unfortunately, even a method with a high percent of individual documents correctly classified can be hugely biased when estimating category proportions. By directly optimizing for this social science goal, we develop a method that gives approximately unbiased estimates of category proportions even when the optimal classifier performs poorly. We illustrate with diverse data sets, including the daily expressed opinions of thousands of people about the U.S. presidency. We also make available software that implements our methods and large corpora of text for further analysis.*

Efforts to systematically categorize text documents date to the late 1600s, when the Church tracked the proportion of printed texts which were non-religious (Krippendorff 2004). Similar techniques were used by earlier generations of social scientists, including Waples, Berelson, and Bradshaw (1940, which apparently includes the first use of the term “content analysis”) and Berelson and de Grazia (1947). Content analyses like these have spread to a vast array of fields, with automated methods now joining projects based on hand coding, and have increased at least sixfold from 1980 to 2002 (Neuendorf 2002). The recent explosive increase in web pages, blogs, emails, digitized books and articles, transcripts, and elec-

tronic versions of government documents (Lyman and Varian 2003) suggests the potential for many new applications. Given the infeasibility of much larger scale human-based coding, the need for automated methods is growing fast. Indeed, large-scale projects based solely on hand coding have stopped altogether in some fields (King and Lowe 2003, 618).

This article introduces new methods of automated content analysis designed to estimate the primary quantity of interest in many social science applications. These new methods take as data a potentially large set of text documents, of which a small subset is hand coded into an investigator-chosen set of mutually exclusive and

---

Daniel J. Hopkins is Assistant Professor of Government, Georgetown University, 681 Intercultural Center, Washington, DC 20057 (dhopkins@iq.harvard.edu, <http://www.danhopkins.org>). Gary King is Albert J. Weatherhead III University Professor, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge St., Cambridge, MA 02138 (king@harvard.edu, <http://gking.harvard.edu>).

Replication materials are available at Hopkins and King (2009); see <http://hdl.handle.net/1902.1/12898>. Our special thanks to our indefatigable undergraduate coders Sam Caporal, Katie Colton, Nicholas Hayes, Grace Kim, Matthew Knowles, Katherine McCabe, Andrew Prokop, and Keneshia Washington. Each coded numerous blogs, dealt with the unending changes we made to our coding schemes, and made many important suggestions that greatly improved our work. Matthew Knowles also helped us track down and understand the many scholarly literatures that intersected with our work, and Steven Melendez provided invaluable computer science wizardry; both are coauthors of the open source and free computer program that implements the methods described herein (ReadMe: Software for Automated Content Analysis; see <http://gking.harvard.edu/readme>). We thank Ying Lu for her wisdom and advice, Stuart Shieber for introducing us to the relevant computer science literature, and <http://Blogpulse.com> for getting us started with more than a million blog URLs. Thanks to Ken Benoit, Doug Bond, Justin Grimmer, Matt Hindman, Dan Ho, Pranam Kolari, Mark Kantrowitz, Lillian Lee, Will Lowe, Andrew Martin, Burt Monroe, Stephen Purpura, Phil Schrodte, Stuart Shulman, and Kevin Quinn for helpful suggestions or data. Thanks also to the Library of Congress (PA#NDP03-1), the Center for the Study of American Politics at Yale University, the Multidisciplinary Program on Inequality and Social Policy, and the Institute for Quantitative Social Science for research support.

*American Journal of Political Science*, Vol. 54, No. 1, January 2010, Pp. 229–247

©2010, Midwest Political Science Association

ISSN 0092-5853

229

exhaustive categories.<sup>1</sup> As output, the methods give approximately unbiased and statistically consistent estimates of the proportion of all documents in each category. Accurate estimates of these *document category proportions* have not been a goal of most work in the classification literature, which has focused instead on increasing the accuracy of *classification into individual document categories*. Unfortunately, methods tuned to maximize the percent of documents correctly classified can still produce substantial biases in the aggregate proportion of documents within each category. This poses no problem for the task for which these methods were designed, but it suggests that a new approach may be of use for many social science applications.

When social scientists use formal content analysis, it is typically to make generalizations using document category proportions. Consider examples as far-ranging as Mayhew (1991, chap. 3), Gamson (1992, chaps. 3, 6, 7, and 9), Zaller (1992, chap. 9), Gerring (1998, chaps. 3–7), Mutz (1998, chap. 8), Gilens (1999, chap. 5), Mendlberg (2001, chap. 5), Rudalevige (2002, chap. 4), Kellstedt (2003, chap. 2), Jones and Baumgartner (2005, chaps. 3–10), and Hillygus and Shields (2008, chap. 6). In all these cases and many others, researchers conducted content analyses to learn about the distribution of classifications in a population, not to assert the classification of any particular document (which would be easy to do through a close reading of the document in question). For example, the manager of a congressional office would find useful an automated method of sorting individual constituent letters by policy area so they can be routed to the most informed staffer to draft a response. In contrast, political scientists would be interested primarily in tracking the proportion of mail (and thus constituent concerns) in each policy area. Policy makers or computer scientists may be interested in finding the needle in the haystack

<sup>1</sup>Although some excellent content analysis methods are able to delegate to the computer both the choice of the categorization scheme and the classification of documents into the chosen categories, our applications require methods where the social scientist chooses the questions and the data provide the answers. The former so-called “unsupervised learning methods” are versions of cluster analysis and have the great advantage of requiring fewer startup costs, since no theoretical choices about categories are necessary *ex ante* and no hand coding is required (Quinn et al. 2009; Simon and Xeons 2004). In contrast, the latter so-called “supervised learning methods,” which require a choice of categories and a sample of hand-coded documents, have the advantage of letting the social scientist, rather than the computer program, determine the most theoretically interesting questions (Kolari, Finin, and Joshi 2006; Laver, Benoit, and Garry 2003; Pang, Lee, and Vaithyanathan 2002). These approaches, and others such as dictionary-based methods (Gerner et al. 1994; King and Lowe 2003), accomplish somewhat different tasks and so can often be productively used together, such as for discovering a relevant set of categories in part from the data.

(such as a potential terrorist threat or the right web page to display from a search), but social scientists are more commonly interested in characterizing the haystack. Certainly, individual document classifications, when available, provide additional information to social scientists, since they enable one to aggregate in unanticipated ways, serve as variables in regression-type analyses, and help guide deeper qualitative inquiries into the nature of specific documents. But they do not usually (as in Benoit and Laver 2003) constitute the ultimate quantities of interest.

Automated content analysis is a new field and is newer still within political science. We thus begin in the second section with a concrete example to help fix ideas and define key concepts, including an analysis of expressed opinion through blog posts about Senator John Kerry. We next explain how to represent unstructured text as structured variables amenable to statistical analysis. The following section discusses problems with existing methods. We introduce our methods in the fifth section along with empirical verification from several data sets in the sixth section. The last section concludes. The appendix provides intercoder reliability statistics and offers a method for coping with errors in hand-coded documents.

## Measuring Political Opinions in Blogs: A Running Example

Although our methodology works for any unstructured text, we use blogs as our running example. Blogs (or “web logs”) are periodic web postings usually listed in reverse chronological order.<sup>2</sup> For present purposes, we define our inferential target as expressed sentiment about each candidate in the 2008 American presidential election. Measuring the national conversation in this way is not the only way to define the population of interest, but it seems to be of considerable public interest and may also be of interest to political scientists studying activists (Verba, Schlozman, and Brady 1995), the media (Drezner and Farrell 2004), public opinion (Gamson 1992), social networks (Adamic and Glance 2005; Huckfeldt and Sprague 1995), or elite influence (Grindle 2005; Hindman, Tsoutsouliklis, and Johnson 2003; Zaller 1992). We attempted to collect all English-language blog posts from highly political people who blog about politics all the time, as

<sup>2</sup>Eight percent of U.S. Internet users (about 12 million people), claim to have their own blog (Lenhart and Fox 2006). The growth worldwide has been explosive, from essentially none in 2000 to estimates today that range up to 185.62 million worldwide. Blogs are a remarkably democratic technology, with 72.82 million in China and at least 700,000 in Iran (Helmond 2008).

well as others who normally blog about gardening or their love lives, but choose to join the national conversation about the presidency for one or more posts. Bloggers' opinions get counted when they post and not otherwise, just as in earlier centuries when public opinion was synonymous with visible public expressions rather than attitudes and nonattitudes expressed in survey responses (Ginsberg 1986).<sup>3</sup>

Our specific goal is to compute the proportion of blogs each day or week in each of seven categories, including extremely negative (-2), negative (-1), neutral (0), positive (1), extremely positive (2), no opinion (NA), and not a blog (NB).<sup>4</sup> Although the first five categories are logically ordered, the set of all seven categories is not (which rules out innovative approaches like Word-scores, which presently requires a single dimension; Laver, Benoit, and Garry 2003). Bloggers write to express opinions and so category 0 is not common, although it and NA occur commonly if the blogger is writing primarily about something other than our subject of study. Category NB ensures that the category list is exhaustive. This coding scheme represents a difficult test case because of the mixed data types, because "sentiment categorization is more difficult than topic classification" (Pang, Lee, and Vaithyanathan 2002, 83), and because the language used ranges from the Queen's English to "my crunchy gf thinks dubya hid the wmd's, :)!!"<sup>5</sup>

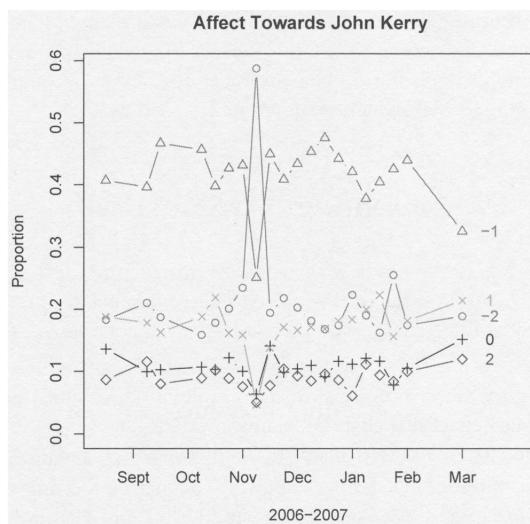
We now preview the type of empirical results we seek. To do this, we apply the nonparametric method described below to blogosphere opinions about John Kerry before,

<sup>3</sup>We obtained our list of blogs by beginning with eight public blog directories and two other sources we obtained privately, including [www.globeofblogs.com](http://www.globeofblogs.com), <http://truthlaidbear.com>, [www.nycbloggers.com](http://www.nycbloggers.com), [http://dir.yahoo.com/Computers\\_and\\_Internet/Internet/](http://dir.yahoo.com/Computers_and_Internet/Internet/), [www.bloghop.com/highrating.htm](http://www.bloghop.com/highrating.htm), <http://www.blogrolling.com/top.phtml>, a list of blogs provided by blogrolling.com, and 1.3 million additional blogs made available to us by Blogpulse.com. We then continuously crawl out from the links or "blogroll" on each of these blogs, adding seeds along the way from Google and other sources, to identify our target population.

<sup>4</sup>Our specific instructions to coders read as follows: "Below is one entry in our database of blog posts. Please read the entire entry. Then, answer the questions at the bottom of this page: (1) indicate whether this entry is in fact a blog posting that contains an opinion about a national political figure. If an opinion is being expressed (2) use the scale from -2 (extremely negative) to 2 (extremely positive) to summarize the opinion of the blog's author about the figure."

<sup>5</sup>Using hand coding to track opinion change in the blogosphere in real time is infeasible and even after the fact would be an enormously expensive task. Using unsupervised learning methods to answer the questions posed is also usually infeasible. Applied to blogs, these methods often pick up topics rather than sentiment or irrelevant features such as the informality of the text.

**FIGURE 1** Blogosphere Responses to Kerry's Botched Joke



*Notes:* Each line gives a time series of estimates of the proportion of all English-language blog posts in categories ranging from -2 (extremely negative, colored red) to 2 (extremely positive, colored blue). The spike in the -2 category immediately followed Kerry's joke. Results were estimated with our nonparametric method.

during, and after the botched joke in the 2006 election cycle, which was said to have caused him to not enter the 2008 contest ("You know, education—if you make the most of it . . . you can do well. If you don't, you get stuck in Iraq"). Figure 1 gives a time-series plot of the proportion of blog posts in each of the opinion categories over time. The sharp increase in the extremely negative (-2) category occurred immediately following Kerry's joke. Note also the concomitant drop in other categories occurred primarily from the -1 category, but even the proportion in the positive categories dropped to some degree. Although the media portrayed this joke as his motivation for not entering the race, this figure suggests that his high negatives before and after this event may have been even more relevant.

These results come from an analysis of word patterns in 10,000 blog posts, of which only 442 from five days in early November were actually read and hand coded by the researchers. In other words, the method outlined in this article recovers a highly plausible pattern for several months using word patterns contained in a small, nonrandom subset of just a few days when anti-Kerry sentiment was at its peak. This was one incident in the

run-up to the 2008 campaign, but it gives a sense of the widespread applicability of the methods. Although we do not offer these in this article, one could easily imagine many similar analyses of political or social events where scale or resource constraints make it impossible to continuously read and manually categorize texts. We offer more formal validation of our methods below.

## Representing Text Statistically

We now explain how to represent unstructured text as structured variables amenable to statistical analysis, first by coding variables and then via statistical notation.

### Coding Variables

To analyze text statistically, we represent natural language as numerical variables following standard procedures (Joachims 1998; Kolari, Finin, and Joshi 2006; Manning and Schütze 1999; Pang, Lee, and Vaithyanathan 2002). For example, for our key variable, we summarize a document (a blog post) with its category. Other variables are computed from the text in three additional steps, each of which works without human input, and all of which are designed to reduce the complexity of text.

First, we drop non-English-language blogs (Cavnar and Trenkle 1994), as well as spam blogs (with a technology we do not share publicly; for another, see Kolari, Finin, and Joshi 2006). For the purposes of this article, we focus on blog posts about President George W. Bush (which we define as those that use the terms “Bush,” “George W.”, “Duby,” or “King George”) and similarly for each of the 2008 presidential candidates. We develop specific filters for each person of interest, enabling us to exclude others with similar names, such as to avoid confusing Bill and Hillary Clinton. For our present methodological purposes, we focus on 4,303 blog posts about President Bush collected February 1–5, 2006, and 6,468 posts about Senator Hillary Clinton collected August 26–30, 2006. Our method works without filtering (and in foreign languages), but filters help focus the limited time of human coders on the categories of interest.

Second, we preprocess the text within each document by converting to lowercase, removing all punctuation, and stemming by, for example, reducing “consist,” “consisted,” “consistency,” “consistent,” “consistently,” “consisting,” and “consists” to their stem, which is “consist.” Preprocessing text strips out information, in addition to reducing complexity, but experience in this liter-

ature is that the trade-off is well worth it (Porter 1980; Quinn et al. 2009).

Finally, we summarize the preprocessed text as dichotomous variables, one type for the presence or absence of each word stem (or “unigram”), a second type for each word pair (or “bigram”), a third type for each word triplet (or “trigram”), and so on to all “n-grams.” This definition is not limited to dictionary words. In our application, we measure only the presence or absence of stems rather than counts (the second time the word “awful” appears in a blog post does not provide as much information as the first). Even so, the number of variables remaining is enormous. For example, our sample of 10,771 blog posts about President Bush and Senator Clinton includes 201,676 unique unigrams, 2,392,027 unique bigrams, and 5,761,979 unique trigrams. The usual choice to simplify further is to consider only dichotomous stemmed unigram indicator variables (the presence or absence of each of a list of word stems), which we have found to work well. We also delete stemmed unigrams appearing in fewer than 1% or greater than 99% of all documents, which results in 3,672 variables. These procedures effectively group the infinite range of possible blog posts to “only”  $2^{3,672}$  distinct types. This makes the problem feasible but still represents a huge number (larger than the number of elementary particles in the universe).

Researchers interested in similar problems in computer science commonly find that “bag of words” simplifications like this are highly effective (e.g., Pang, Lee, and Vaithyanathan 2002; Sebastiani 2002), and our analysis reinforces that finding. This seems counterintuitive at first, since it is easy to write text whose meaning is lost when word order is discarded (e.g., “I hate Clinton. I love Obama”). But empirically, most text sources make the same point in enough different ways that representing the needed information abstractly is usually sufficient. As an analogy, when channel surfing for something to watch on television, pausing for only a few hundred milliseconds on a channel is typically sufficient; similarly, the negative content of a vitriolic post about President Bush is usually easy to spot after only a sentence or two. When the bag of words approach is not a sufficient representation, many procedures are available: we can code where each word stem appears in a document, tag each word with its part of speech, or include selective bigrams, such as by replacing “white house” with “white\_house” (Das and Chen 2001). We can also use counts of variables or code variables to represent meta-data, such as the URL, title, blogroll, or whether the post links to known liberal or conservative sites (Thomas, Pang, and Lee 2006). Many other similar tricks suggested in the computer science

literature may be useful for some problems (Pang and Lee 2008), and all can be included in the methodology described below, but we have not found them necessary for the many applications we have tried to date.

### Notation and Quantities of Interest

Our procedures require two sets of text documents. The first is a small *labeled set*, for which each document  $i$  ( $i = 1, \dots, n$ ) is labeled with one of the given categories, usually by reading and hand coding (we discuss how large  $n$  needs to be in the sixth section, and what to do if hand coders are not sufficiently reliable in the appendix). We denote the *Document category* variable as  $D_i$ , which in general takes on the value  $D_i = j$ , for possible categories  $j = 1, \dots, J$ .<sup>6</sup> (In our running example,  $D_i$  takes on the potential values  $\{-2, -1, 1, 0, 1, 2, \text{NA}, \text{NB}\}$ .) We denote the second, larger *population set* of documents as the inferential target, and in which each document  $\ell$  (for  $\ell = 1, \dots, L$ ) has an unobserved classification  $D_\ell$ . Sometimes the labeled set is a sample from the population and so the two overlap; more often it is a nonrandom sample from a different source than the population, such as from earlier in time.

All other information is computed directly from the documents. To define these variables for the labeled set denote  $S_{ik}$  as equal to 1 if word Stem  $k$  ( $k = 1, \dots, K$ ) is used at least once in document  $i$  (for  $i = 1, \dots, n$ ) and 0 otherwise (and similarly for the population set, substituting index  $i$  with index  $\ell$ ). This makes our abstract summary of the text of document  $i$  the set of these variables,  $\{S_{i1}, \dots, S_{iK}\}$ , which we summarize as the  $K \times 1$  vector of word stem variables  $S_i$ . We refer to  $S_i$  as a *word stem profile* since it provides a summary of all the word stems (or other information) used in a document.

The quantity of interest in most of the supervised learning literature is the set of individual classifications for all documents in the population,  $\{D_1, \dots, D_L\}$ . In contrast, the quantity of interest for most content analyses in social science is the aggregate proportion of all (or a subset of all) of these population documents that fall into each category:  $P(D) = \{P(D = 1), \dots, P(D = J)\}'$  where  $P(D)$  is a  $J \times 1$  vector, each element of which is a proportion computed by direct tabulation:

$$P(D = j) = \frac{1}{L} \sum_{\ell=1}^L \mathbf{1}(D_\ell = j), \quad (1)$$

<sup>6</sup>This notation is from King and Lu (2008), who use related methods applied to unrelated substantive applications that do not involve coding text, and different mnemonic associations.

where  $\mathbf{1}(a) = 1$  if  $a$  is true and 0 otherwise. Document category  $D_i$  is one variable with many possible values, whereas word profile  $S_i$  constitutes a set of dichotomous variables. This means that  $P(D)$  is a multinomial distribution with  $J$  possible values and  $P(S)$  is a multinomial distribution with  $2^K$  possible values, each of which is a word stem profile.

### Issues with Existing Approaches

This section discusses problems with two common methods that arise when they are used to estimate social aggregates rather than individual classifications.

#### Existing Approaches

A simple way of estimating  $P(D)$  is *direct sampling*: identify a well-defined population of interest, draw a random sample from the population, hand code all the documents in the sample, and count the documents in each category. This method requires basic sampling theory, no abstract numerical summaries of any text, and no classifications of individual documents in the unlabeled population.

The second approach to estimating  $P(D)$ , the *aggregation of individual document classifications*, is standard in the supervised learning literature. The idea is to first use the labeled sample to estimate a functional relationship between document category  $D$  and word features  $S$ . Typically,  $D$  serves as a multicategory dependent variable and is predicted with a set of explanatory variables  $\{S_{i1}, \dots, S_{iK}\}$ , using some statistical, machine learning, or rule-based method (such as multinomial logit, regression, discriminant analysis, radial basis functions, CART, random forests, neural networks, support vector machines, maximum entropy, or others). Then the coefficients of the model are estimated, and both the coefficients and the data-generating process are assumed the same in the labeled sample as in the population. The coefficients are then used with the features measured in the population,  $S_\ell$ , to predict the classification for each population document  $D_\ell$ . Social scientists then aggregate the individual classifications via equation (1) to estimate their quantity of interest,  $P(D)$ .

### Problems

Unfortunately, as Hand (2006) points out, the standard supervised learning approach to individual document

classification will fail in two circumstances, both of which appear common in practice. (And even if classification succeeds with high or optimal accuracy, the next subsection shows that estimating population proportions can still be biased.)

First, when the labeled set is not a random sample from the population, both methods fail. Yet “in many, perhaps most real classification problems the data points in the [labeled] design set are not, in fact, randomly drawn from the same distribution as the data points to which the classifier will be applied. . . . It goes without saying that statements about classifier accuracy based on a false assumption about the identity of the [labeled] design set distribution and the distribution of future points may well be inaccurate” (Hand 2006, 2). Deviations from randomness may occur due to “population drift,” which occurs when the labeled set is collected at one point and meant to apply to a population collected over time (as with blogs), or for other reasons. The burdens of hand coding become especially apparent when considering the typical analysis within subgroups and the need for a separate random sample within each.

Second, the data-generation process assumed by the standard supervised learning approach predicts  $D$  with  $S$ , modeling  $P(D | S)$ , but the world works in reverse. For our running example, bloggers do not start writing and only afterwards discover their affect toward the president: they start with a view, which we abstract as a document category, and then set it out in words. That is, the right data-generation process is the inverse of what is being modeled, where we should be predicting  $S$  with  $D$ , and inferring  $P(S | D)$ . The consequence of using  $P(D | S)$  instead (and without Bayes Theorem, which is not very helpful in this case) is the requirement of two assumptions needed to generalize from the labeled sample to the population. The first is that  $S$  “spans the space of all predictors” of  $D$  (Hand 2006, 9), which means that once one controls for the measured variables, there exists no other variable that could improve predictive power. In problems involving human language, this assumption is not met, since  $S$  is intentionally an abstraction and so by definition does not represent all existing information in the predictors. The other assumption is that the class of models chosen for  $P(D | S)$  includes the “true” model. This is a more familiar assumption to social scientists, but it is no easier to meet. In this case, finding even the best model or a good model, much less the “true one,” is a difficult and time-consuming task given the huge number of potential explanatory variables coded from text and potential models to run. As Hand writes, “Of course, it would be a brave person who could confidently assert that these two conditions held” (2006, 9).

## Optimizing for a Different Goal

Here we show that even optimal individual document classification that meets all the assumptions of the last section can lead to biased estimates of the document category proportions. The criterion for success in the classification literature, the percent correctly classified in a test set, is obviously appropriate for individual-level classification, but it can be seriously misleading when characterizing document populations. For example, of the 23 models estimated by Pang, Lee, and Vaithyanathan (2002), the percent correctly predicted ranges from 77% to 83%. This is an excellent classification performance for sentiment analysis, but suppose that all the misclassifications were in a particular direction for one or more categories. In that situation, the statistical *bias* (the average difference between the true and estimated proportion of documents in a category) in using this method to estimate the aggregate quantities of interest could be as high as 17 to 23 percentage points. This does not matter for the authors, since their goal was classification, but it could matter for researchers interested in category proportions.

Unfortunately, except at the extremes, there exists no necessary connection between low misclassification rates and low bias: it is easy to construct examples of learning methods that achieve a high percent of individual documents correctly predicted and large biases for estimating the aggregate document proportions, or other methods that have a low percent correctly predicted but nevertheless produce relatively unbiased estimates of the aggregate quantities. For example, flipping a coin is a bad predictor of which party will win a presidential election, but it does happen to provide an unbiased estimate of the percentage of Democratic presidential victories since 1912. Since the goal of this literature is individual classification, it does not often report the bias in estimating the aggregates. As such, the bulk of the otherwise impressive supervised learning classification literature offers little indication of whether the methods proposed would work well for those with different goals.

## Statistically Consistent Estimates of Social Aggregates

We now introduce a method optimized for estimating document category proportions. To simplify the exposition, we first show how to correct aggregations of any existing classification method and after offer our stand-alone procedure, not requiring (or producing) a method of individual document classification.

## Corrected Aggregations of Individual Classifications

*Intuition.* Consider multinomial logit or any other method which can generate individual classifications. Fit this model to the labeled set, use it to classify each of the unlabeled documents in the population of interest, and aggregate the classifications to obtain a raw, uncorrected estimate of the proportion of documents in each category. Next, estimate misclassification probabilities by first dividing the labeled set of documents into a training set and a test set (ignoring the unlabeled population set). Then apply the same classification method to the training set alone and make predictions for the test set,  $\hat{D}_i$  (ignoring the test set's labels). Then use the test set's labels to calculate the specific misclassification probabilities between each pair of actual classifications given each true value,  $P(\hat{D}_i = j | D_i = j')$ . These misclassification probabilities do not tell us which documents are misclassified, but they can be used to correct the raw estimate of the document category proportions.

For example, suppose we learn, in predicting test set proportions from the training set, that 17% of the documents our method classified as  $D = 1$  really should have been classified as  $D = 3$ . For any one individual classification in the population, this fact is of no help. But for document category proportions, it is easy to use: subtract 17% from the raw estimate of the category 1 proportion in the population,  $P(D = 1)$ , and add it to category 3,  $P(D = 3)$ . Even if the raw estimate was badly biased, which can occur despite optimal individual document classification, the resulting corrected estimate would be unbiased so long as the population misclassification errors were estimated well enough from the labeled set (a condition we discuss below). Even if the percent correctly predicted is low, this corrected method can give unbiased estimates of the category frequencies.

*Formalization for Two Categories.* For the special case where  $D$  is dichotomous, the misclassification correction above is well known in epidemiology—an area of science directly analogous to the social sciences, where much data are at the individual level, but the quantities of interest are often at the population level. To see this, consider a dichotomous  $D$ , with values 1 or 2, a raw estimate of the proportion of documents in category 1 from some method of classification,  $P(\hat{D} = 1)$ , and the true proportion (corrected for misclassification),  $P(D = 1)$ .<sup>7</sup>

<sup>7</sup>The raw estimate  $P(\hat{D} = 1)$  can be based on the proportion of individual documents classified into category 1. However, a better estimate for classifiers that give probabilistic classifications is to sum

Then define two forms of correct classification as “sensitivity,”  $\text{sens} \equiv P(\hat{D} = 1 | D = 1)$  (sometimes known as “recall”), and “specificity,” or  $\text{spec} \equiv P(\hat{D} = 2 | D = 2)$ . For example, sensitivity is the proportion of documents predicted to be in category 1 among those actually in category 1.

Then we note that the proportion of documents estimated to be in category 1 must come from only one of two sources: documents actually in category 1 that were correctly classified and documents actually in category 2 but misclassified into category 1. We represent this accounting identity, known as the Law of Total Probability, as

$$P(\hat{D} = 1) = (\text{sens})P(D = 1) + (1 - \text{spec})P(D = 2). \quad (2)$$

Since equation (2) is one equation with only one unknown [since  $P(D = 1) = 1 - P(D = 2)$ ], it is easy to solve. As Levy and Kass (1970) first showed, the solution is

$$P(D = 1) = \frac{P(\hat{D} = 1) - (1 - \text{spec})}{\text{sens} - (1 - \text{spec})}. \quad (3)$$

This expression can be used in practice by estimating sensitivity and specificity in the first-stage analysis (separating the labeled set into training and test sets as discussed above or more formally by cross-validation) and using the entire labeled set to predict the (unlabeled) population set to give  $P(\hat{D} = 1)$ . Plugging these values in the right side of (3) gives a corrected, and statistically consistent, estimate of the true proportion of documents in category 1.

*Generalization to Any Number of Categories.* The applications in epidemiology for which these expressions were developed are completely different than our problems, but the methods developed there are directly relevant. This connection enables us to use for our application the generalizations developed by King and Lu (2008).<sup>8</sup>

the estimated probability that each document is in the category for all documents. For example, if 100 documents each have a 0.52 probability of being in category 1, then all individual classifications are in this category. However, since we would only expect 52% of documents to actually be in category 1, a better estimate is  $P(\hat{D} = 1) = 0.52$ .

<sup>8</sup>King and Lu's (2008) article contributed to the field in epidemiology called “verbal autopsies.” The goal of this field is to estimate the distribution of the causes of death in populations without medical death certification. This information is crucial for directing international health policy and research efforts. Data come from two sources. One is a sample of deaths from the population, where a relative of each deceased is asked a long (50–100 item) list of usually dichotomous questions about symptoms the deceased may

Thus, we first generalize equation (2) to include any number of categories by substituting  $j$  for 1, and summing over all categories instead of just 2:

$$P(\hat{D} = j) = \sum_{j'=1}^J P(\hat{D} = j | D = j') P(D = j'). \quad (4)$$

Given  $P(\hat{D})$  and the misclassification probabilities  $P(\hat{D} = j | D = j')$ , which generalize sensitivity and specificity to multiple categories, this expression represents a set of  $J$  equations (i.e., defined for  $j = 1, \dots, J$ ) that can be solved for the  $J$  elements in  $P(D)$ . This is aided by the fact that the equations include only  $J - 1$  unknowns since elements of  $P(D)$  must sum to 1.

*Interpretation.* The section entitled “Optimizing for a Different Goal” shows that a method meeting all the assumptions required for optimal classification performance can still give biased estimates of the document category proportions. We therefore offer here statistically consistent estimates of document category proportions, without having to improve individual classification accuracy and with no assumptions beyond those already made by the individual document classifier. In particular, classifiers require that the labeled set be a random sample from the population. Our method only requires a special case of the random selection assumption: that the misclassification probabilities (sensitivity and specificity with 2 categories or  $P(\hat{D} = j | D = j')$  for all  $j$  and  $j'$  in equation (4)) estimated with data from the labeled set also hold in the unlabeled population set. This assumption may be wrong, but if it is, then the assumptions necessary for the original classifier to work are also wrong and will not necessarily even give accurate individual classifications. More importantly, our approach will also work with a biased classifier.

## Document Category Proportions Without Individual Classifications

We now offer an approach that requires no parametric statistical modeling, individual document classification, or random sampling from the target population. It also

have suffered prior to death ( $S_i$ ). The other source of data is deaths in a nearby hospital, where the same data collection of symptoms from relatives is collected ( $S_i$ ) and also where medical death certification is available ( $D_i$ ). Their method produces approximately unbiased and consistent estimates, considerably better than the existing approaches, which included expensive and unreliable physician reviews (where three physicians spend 20 minutes with the answers to the symptom questions from each deceased to decide on the cause of death), reliable but inaccurate expert rule-based algorithms, or model-dependent parametric statistical models.

correctly treats  $S$  as a consequence rather than cause of  $D$ .

*The Approach.* This method requires only one additional step beyond that in the previous section: instead of using  $S$  and  $D$  to estimate  $P(\hat{D} = j)$ , and then separately correcting via equation (4), we avoid having to compute  $\hat{D}$  by using  $S$  in place of  $\hat{D}$  in that same equation. That is, any observable implication of  $D$  can be used in place of  $\hat{D}$  in equation (4); because  $\hat{D}$  is a function of  $S$ —since the words chosen are by definition a function of the document category—it is simplest to use it directly. Thus, we have

$$P(S = s) = \sum_{j=1}^J P(S = s | D = j) P(D = j). \quad (5)$$

To simplify this expression, we rewrite equation (5) as an equivalent matrix expression:

$$P(S) = \begin{matrix} P(S | D) \\ 2^K \times 1 \end{matrix} \begin{matrix} P(D) \\ 2^K \times J \\ J \times 1 \end{matrix} \quad (6)$$

where, as indicated,  $P(S)$  is the probability of each of the  $2^K$  possible word stem profiles occurring,<sup>9</sup>  $P(S | D)$  is the probability of each of the  $2^K$  possible word stem profiles occurring within the documents in category  $D$  (columns of  $P(S | D)$  corresponding to values of  $D$ ), and  $P(D)$  is our  $J$ -vector quantity of interest.

*Estimation.* Elements of  $P(S)$  can be estimated by direct tabulation from the target population, without parametric assumptions: we merely compute the proportion of documents observed with each pattern of word profiles. Since  $D$  is not observed in the population, we cannot estimate  $P(S | D)$  directly. Instead, we make the crucial assumption that its value in the labeled, hand-coded sample,  $P^h(S | D)$ , is the same as that in the population,

$$P^h(S | D) = P(S | D), \quad (7)$$

and use the labeled sample to estimate this matrix (we discuss this assumption below). We avoid parametric assumptions here too, by using direct tabulation to compute the proportion of documents observed to have each specific word profile among those in each document category.

In principle, we could estimate  $P(D)$  in equation (6) assuming only the veracity of equation (7) and the accuracy of our estimates of  $P(S)$  and  $P(S | D)$ , by solving equation (6) via standard regression algebra. That is, if we

<sup>9</sup>For example, if we ran the method with only  $K = 3$  word stems,  $P(S)$  would contain the probabilities of each of these ( $2^3 = 8$ ) patterns occurring in the set of documents: 000 (i.e., none of the three words were used), 001, 010, 011, 100, 101, 110, and 111.

think of  $P(D)$  as the unknown “regression coefficients”  $\beta$ ,  $P(S | D)$  as the “explanatory variables” matrix  $X$ , and  $P(S)$  as the “dependent variable”  $Y$ , then equation (6) becomes  $Y = X\beta$  (with no error term). This happens to be a linear expression but not because of any assumption imposed on the problem that could be wrong. The result is that we can solve for  $P(D)$  via the usual regression calculation:  $\beta = (X'X)^{-1}X'y$  (or via standard constrained least squares to ensure that elements of  $P(D)$  are each in  $[0,1]$  and collectively sum to 1). A key point is that this calculation does *not* require classifying individual documents into categories and then aggregating; it estimates the aggregate proportions directly.

This simple approach poses two difficulties in our application. First,  $K$  is typically very large and so  $2^K$  is far larger than any standard computer could handle. Second is a sparseness problem since the number of observations available for estimating  $P(S)$  and  $P(S | D)$  is much smaller than the number of potential word profiles ( $n << 2^K$ ). To avoid both of these issues, we adapt results from King and Lu (2008) and randomly choose subsets of between approximately 5 and 25 words. The optimal number of words to use per subset is application-specific, but can be determined empirically through cross-validation within the labeled set. Although the estimator remains approximately unbiased regardless of subset size, in practice we find that setting the number of words per subset too high can lead to inefficiency. The reason is that as the number of words per subset increases, the number of unique subsets increases, reducing the number of common subsets that appear in both the labeled and unlabeled data sets. In addition, in the applications below, the words included in each subset are chosen randomly with equal probabilities, although in some applications, performance may improve by weighting words unequally.

Once we determine the optimal number of subsets through cross-validation, we solve for  $P(D)$  in each, and average the results across the subsets. Because  $S$  is treated as a consequence of  $D$ , using subsets of  $S$  introduces no new assumptions. This simple subsetting procedure turns out to be equivalent to a version of the standard approach of smoothing sparse matrices via kernel densities, although, unlike the typical use of this procedure, its application here reduces bias. (Standard errors and confidence intervals are computed via standard bootstrapping procedures.)

*Interpretation.* A key advantage of estimating  $P(D)$  without the intermediate step of computing the individual classifications is that the required assumptions are much less restrictive. They can still be wrong, and as a result our estimates can be biased, but the dramatic re-

duction in their restrictiveness means that under the new approach we have a fighting chance to get something close to the right answer in many applications where valid inferences were not previously likely.

Unlike direct sampling or standard supervised learning approaches, our strategy allows the distribution of documents across word-stem profiles,  $P(S)$ , and the distribution of documents across the categories,  $P(D)$ , to each be completely different in the labeled set and population set of documents. So for example, if a word or pattern of words becomes more popular between the time the labeled set was hand coded and the population documents were collected, no biases would emerge. Similarly, if documents in certain categories are more prevalent in the population than labeled set, no biases would result. In our running example, no bias would be induced if the labeled set includes a majority of conservative Republicans who defend everything President Bush does and the target population has a supermajority of liberal Democrats who want nothing more than to end the Bush presidency. In contrast, changes in either  $P(D)$  or  $P(S)$  between the labeled and population sets would be sufficient to doom existing classification-based approaches. For example, so long as “idiot” remains an insult, our method can make appropriate use of that information, even if the word becomes less common (a change in  $P(S)$ ) or if there are fewer people who think politicians deserve it (a change in  $P(D)$ ).

The key theoretical assumption is equation (7)—that the documents in the hand-coded set contain sufficient good examples of the language used for each document category in the population. To be more specific, among all documents in a given category, the prevalence of particular word profiles in the labeled set should be the same in expectation as in the population set. For example, the language bloggers use to describe an “extremely negative” view of Hillary Clinton in the labeled set must be at least a subset of the way she is described in the target population. They do not need to write literally the same blog posts, but rather need to have the same probabilities of using similar word profiles so that  $P^h(S | D = -2) = P(S | D = -2)$ . This assumption can be violated due to population drift or for other reasons, but we can always hand code some additional cases in the population set to verify that it holds sufficiently well. And as discussed above, the proportion of examples of each document category and of each word profile can differ between the two document sets.

The methodology is also considerably easier to use in practice. Applying the standard supervised learning approach is difficult, even if we meet its assumptions. Even if we forget about choosing the “true” model, merely

finding a “good” specification with thousands of explanatory variables to choose from can be extraordinarily time consuming. One needs to fit numerous statistical models, consider many specifications within each model type, run cross-validation tests, and check various fit statistics. Social scientists have a lot of experience with specification searches, but all the explanatory variables mean that even one run would take considerable tuning and many runs would need to be conducted.

The problem is further complicated by the fact that social scientists are accustomed to choosing their statistical specifications on the basis of prior theoretical expectations and results from past research, whereas the overwhelming experience in the information extraction literature is that radically empirical approaches work best for a given amount of effort. For example, we might think we could carefully choose words or phrases to characterize particular document categories (e.g., “awful,” “irresponsible,” “impeach,” etc., to describe negative views about President Bush), and indeed this approach will often work to some degree. Yet, a raw empirical search for the best specification, ignoring these theoretically chosen words, will typically turn up predictive patterns we would not have thought of ex ante. Indeed, methods based on highly detailed parsing of the grammar and sentence structure in each document can also work exceptionally well (e.g., King and Lowe 2003), but the strong impression from the literature is that the extensive, tedious work that goes into adapting these approaches for each application is more productively put into collecting more hand-coded ex-

amples and then using an automatic specification search routine.

## The Method in Practice

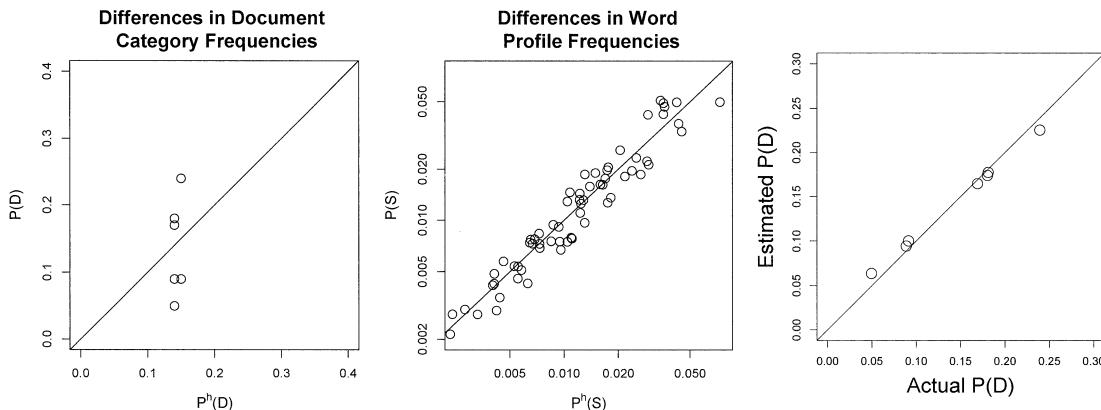
We begin here with a simple simulated example, proceed to real examples of different types of documents, and study how many documents one needs to hand code. We also compare our approach to existing methods and discuss what can go wrong. Readers can replicate and modify any of these analyses using the replication files made available with this article.

### Monte Carlo Simulations

We begin with a simulated data set of 10 words and thus  $2^{10} = 1,024$  possible word-stem profiles. We set the elements of  $P^h(D)$  to be the same across the seven categories, and then set the population document category frequencies,  $P(D)$ , to very different values. We then draw a value  $\tilde{D}$  from  $P^h(D)$ , insert the simulation into  $P^h(S | \tilde{D})$ , which we set to that from the population, and then draw the simulated matrix  $\tilde{S}$  from this density. We repeat the procedure 1,000 times to produce the labeled data set, and analogously for the population.

The left two panels of Figure 2 summarize the sharp differences between the hand-coded and population

**FIGURE 2 Accurate Estimates Despite Differences Between Labeled and Population Sets**



*Notes:* For both  $P(D)$  on the left and  $P(S)$  in the center, the distributions differ considerably. The direct sampling estimator,  $P^h(D)$ , is therefore highly biased. Yet, the right panel shows that our nonparametric estimator remains unbiased.

distributions in these data. The left graph plots  $P^h(D)$  horizontally by  $P(D)$  vertically, where the seven circles represent the category proportions. If the proportions were equal, they would all fall on the  $45^\circ$  line. If one used the labeled, hand-coded sample in this case via direct sampling to estimate the document category frequencies in the population, the result would not even be positively correlated with the truth.

The differences between the two distributions of word frequency profiles appear in the middle graph (where for clarity the axes, but not labels, are on the log scale). Each circle in this graph represents the proportion of documents with a specific word profile. Again, if the two distributions were the same, all the circles would appear on the diagonal line, but again many of the circles fall off the line, indicating differences between the two samples.

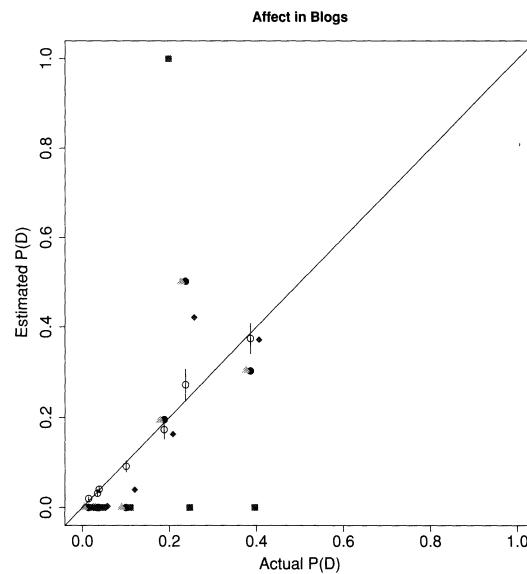
Despite the considerable differences between the labeled data set and the population, and the bias in the direct sampling estimator, our approach still produces accurate estimates. The right panel of the figure presents these results. The actual  $P(D)$  is on the horizontal axis and the estimated version is on the vertical axis, with each of the seven circles representing one of the document frequency categories. Estimates that are accurate fall on the  $45^\circ$  line. In fact, the points are all huddled close to this equality line, with even the maximum distance from the line for any point being quite small.

## Empirical Evidence

We now offer several out-of-sample tests of our nonparametric approach in different types of real data. Our first test includes the 4,303 blog posts that mention George W. Bush. (For levels of intercoder reliability for this task, see the appendix.) These posts include 47,726 unique words and 3,165 unique word stems. We randomly divide the data set in half between the training set and test set and, to make the task more difficult, then randomly delete half (422) of the posts coded –2 in the test set. Our test set therefore intentionally selects on (what would be considered, in standard supervised learning approaches) the dependent variable. The results from our nonparametric estimator appear in Figure 3 as one open circle for each of the seven categories, with 95% confidence intervals appearing as a vertical line. Clearly the points are close to the  $45^\circ$  line, indicating approximately unbiased estimates, and all are within the 95% confidence intervals.

Also plotted on the same graph are the document category proportions aggregated up from individual clas-

**FIGURE 3 Out-of-Sample Validation**



*Notes:* The plot gives the estimated document category frequencies (vertically) by the actual frequencies (horizontally). Our nonparametric approach is represented with black open circles, with 95% confidence intervals as vertical lines. Aggregated optimized SVM analyses also appear for radial basis (black dots), linear (green triangles), polynomial (blue diamonds), and sigmoid kernels (red squares). Estimates closer to the  $45^\circ$  line are more accurate.

sifications given by four separately optimized support vector machine (SVM) classifiers, the most widely used (and arguably the best) of the existing methods. These include SVMs using a radial basis function (black dots), linear kernel (green triangles), polynomial kernel (blue diamonds), and sigmoid kernel (red squares) (Brank et al. 2002; Hastie, Tibshirani, and Friedman 2001; Hsu, Chang, and Lin 2003; Joachims 1998). As can be seen in the graph, these results vary wildly and none do as well as our approach. They are plotted without confidence intervals since SVM is not a statistical method and has no probabilistic foundation. An additional difficulty of using individual classifiers is the highly time-intensive tuning required. Whereas the results from our approach represent only a single run, we followed the advice of the SVM literature and chose the final four SVMs to present in Figure 3 by optimizing over a total of 19,090 separate SVM runs, including cross-validation tests on 10 separate subsets of the labeled set. One run of our nonparametric estimator took 60 seconds of computer time, or a total of five hours for 300 bootstrapped runs. The SVM runs

**TABLE 1 Performance of Our Nonparametric Approach and Four Support Vector Machine Analyses**

	Percent of Blog Posts Correctly Classified			
	In-Sample Fit	In-Sample Cross-Validation	Out-of-Sample Prediction	Mean Absolute Proportion Error
Nonparametric	—	—	—	1.2
Linear	67.6	55.2	49.3	7.7
Radial	67.6	54.2	49.1	7.7
Polynomial	99.7	48.9	47.8	5.3
Sigmoid	15.6	15.6	18.2	23.2

*Notes:* Each row is the optimal choice over numerous individual runs given a specific kernel. Leaving aside the sigmoid kernel, individual classification performance in the first three columns does not correlate with mean absolute error in the document category proportions in the last column.

took approximately 8.7 days (running 24 hours/day) on a powerful server and much more in human time.

We give an alternative view of these results in Table 1. The first three numerical columns report individual classification performance whereas the last gives the mean absolute error in the document category proportions. The last column confirms the overall impression from Figure 3 that the nonparametric method has much lower error in estimating the document category proportions. Leaving aside the sigmoid kernel, which did not work well in these data, the SVM results have the familiar patterns for individual classifiers: the models fit best to the in-sample data, followed next by in-sample cross-validation, and lastly by the true out-of-sample predictions. The key result in this analysis is that, even among the SVM analyses, the best individual classifier (the linear kernel) is different from the best choice for minimizing the mean absolute error in the document category proportions (the polynomial kernel). Of course, nothing is wrong with SVM when applied to the individual classification goal for which it was designed.

### Examples from Other Textual Data Sources

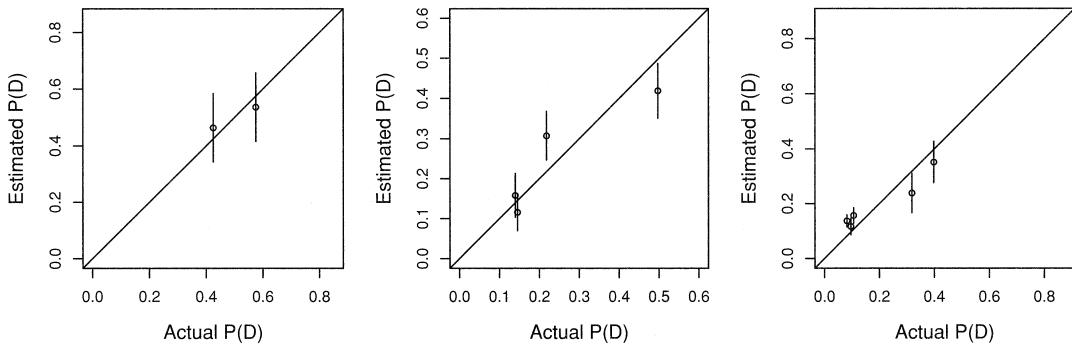
We now give three brief examples applying our method to different sources of unstructured text. The first is from a corpus of congressional speeches used in the computer science literature to evaluate supervised learning methods (Thomas, Pang, and Lee 2006). Researchers selected 3,838 speeches given in the House of Representatives between January 4th and May 12th, 2005, during “contentious” debates, defined as those where more than 20% of the speeches were in opposition. To simulate how a resource-conscious researcher might proceed, we used the 1,887

speeches appearing on even-numbered pages of the congressional record as a training set, and then estimated the distribution of supportive speeches in the test set of 1,951 speeches on odd-numbered pages. The results using the nonparametric estimator appear in the top-left graph in Figure 4 and are again highly accurate.

Another example comes from a data set of 462 immigration editorials that we compiled using Factiva. The editorials appeared in major newspapers between April 1st and July 15th, 2007, and were coded into four nonordered categories indicating editorials supporting the Senate’s immigration bill, those opposing it, and two categories that capture letters to the editor and other miscellaneous articles. Here, the training set includes the 283 editorials prior to June 12th, while the test set includes the 179 editorials on or after that date. Deviations from the 45° line are due to slight violations of the assumption in equation (7). This is quite a hard test, since some categories have as few as 40 examples. The small discrepancy can also be fixed easily if this were a real application by adding to the hand-coded set a small number of documents collected over time.

Our final example comes from 1,726 emails sent by Enron employees and classified into five nonordered categories: company business, personal communications, logistic arrangements, employment arrangements, and document editing.<sup>10</sup> To make the task more difficult, we first created a skewed test set of 600 emails that was more uniformly distributed than the training set, with no category accounting for less than 12% or more than 39% of the observations. We then used the remaining 1,126 emails as a mutually exclusive training set where the comparable bounds were 4% and 50%. The results are quite

<sup>10</sup>See <http://www.cs.cmu.edu/~enron/>.

**FIGURE 4 Additional Out-of-Sample Validation**

*Notes:* The left graph displays the accuracy of the nonparametric method in recovering the distribution of supporting versus opposing speeches in Congress. The center graph shows the same for a categorization of newspaper editorials on immigration, and the right graph shows the distribution across categories of emails sent by enron employees. As before, 95% confidence intervals are represented by vertical lines, and estimates closer to the 45° line are more accurate.

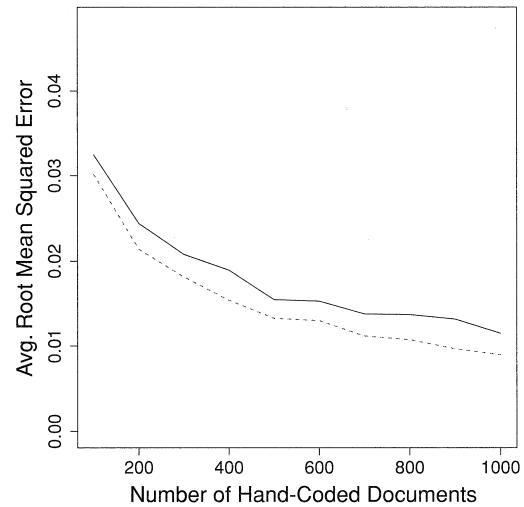
accurate, especially given the paucity of information in many (short) emails, and are displayed in the right panel of Figure 4.

### How Many Documents Need to Be Hand Coded?

Any remaining bias in our estimator is primarily a function of the assumption in equation (7). In contrast, efficiency, as well as confidence intervals and standard errors, are primarily a function of how many documents are hand coded and so are entirely under the control of the investigator. But how many is enough? Hand coding is expensive and time consuming and so we would want to limit its use as much as possible, subject to acceptable uncertainty intervals.

To study this question, we set aside bias by randomly sampling the labeled set directly from the population and plotting in Figure 5 the root mean square error (RMSE) averaged across the categories vertically by the number of hand-coded documents horizontally for our estimator (solid line) and the direct sampling estimator (dashed line). RMSE is lower for the direct estimator, of course, since this sample was drawn directly from the population and little computation is required, although the difference between the two is only about two-tenths of a percentage point.

For our estimator, the RMSE drops quickly as the number of hand-coded documents increases. Even the highest RMSE, with only 100 documents in the labeled

**FIGURE 5 Average Root Mean Square Error by Number of Hand-Coded Documents**

set, is only slightly higher than 3 percentage points, which would be acceptable for some applications. (For example, most national surveys have a margin of error of at least 4 percentage points, even when assuming random sampling and excluding all other sources of error.) At about 500 documents, the advantage of more hand coding begins to

suffer diminishing returns. In part this is because there is little more error to eliminate as our estimator then has an average RMSE of only about 1.5 percentage points.

The conclusion here is clear: coding more than about 500 documents to estimate a specific quantity of interest is probably not necessary, unless one is interested in much more narrow confidence intervals than is common or in specific categories that happen to be rare. For some applications, as few as 100 documents may even be sufficient.

### What Can Go Wrong?

We now discuss five problems that can arise with our methods. If they do arise, and steps are not taken to avoid or ameliorate them, they can cause our estimator to be biased or inefficient. We also discuss what to do to ameliorate these problems.

First, and most importantly, our procedure cannot work without reliable information. This requires that the original documents contain the information needed, the hand codings are reliable enough to extract the information from the documents, and the quantitative summary of the document (in  $S$ ) is a sufficiently accurate representation and sufficient to estimate the quantities of interest. Each of these steps requires careful study. Documents that do not contain the information needed cannot be used to estimate quantities of interest. If humans cannot code these documents into well-defined categories with some reasonable level of reliability, then automated procedures are unlikely to succeed at the same task. And many choices are available in producing abstract numerical summaries of written text documents. Although we have found that stemmed unigrams are a sufficient representation to achieve approximately unbiased inferences in our examples, researchers may have to use some of the other tricks discussed in the section entitled “coding variables” for different applications.

Second, a key issue is the assumption in equation (7) that  $P(S | D)$  is the same in the labeled and population document sets. We thus have much less restrictive assumptions than prior methods, but we still assume a particular type of connection between the two document sets. If we are studying documents over a long time period, where the language used to characterize certain categories is likely to change, it would not be advisable to select the labeled test set only from the start of the period. Checking whether this assumption holds is not difficult and merely requires hand coding some additional documents closer to the quantity presently being estimated and using them as a validation test set. If the data are collected

over time, one can either hand code several data sets from different time periods or gradually add hand-coded documents collected over time. In our running example, we are attempting to track opinions over a single presidential campaign. As such, only one hand-coded data set at the start may be sufficient, but we have tested this assumption, and will continue to do so by periodically hand coding small numbers of blogs.<sup>11</sup>

Third, each category of  $D$  should be defined so as to be mutually exclusive, exhaustive, and relatively homogeneous. To confront cases where the categories are not mutually exclusive, one can define an additional “both” category. Categories that require many examples to define may be too broad for effective estimation as may occur for residual or catch-all categories. Consider the “NB” category in our data as one example. There are innumerable types of web sites that are not blogs, each with very different language; yet this category was essential since our blog search algorithm was not perfect. In fact, we do find slightly more bias in estimating category NB than the others in our categorization, but not so much as to cause a problem for our applications. Given our experiences, the identification of an effective set of categories in  $D$  is an important issue and should involve careful iteration between improving concepts, validation in hand coding tests, and searching for new possibilities in example documents. Intercoder reliability is a crucial metric as well. If human coders cannot agree on a classification, automated approaches are not likely to return sensible results either.

Fourth, our approach requires the choice of the number of word stems to use in each randomly chosen subset. While choosing the number of random subsets is easy (the more the better, and so like any simulation method the number should be chosen based on available computer time and the precision needed), the number of word stems to use in each random subset must be chosen more carefully. Choosing too few or too many will leave  $P(S)$  and  $P(S | D)$  too sparse or too short and may result in attenuation bias due to measurement error in  $P(S | D)$ , which serve as the “explanatory variables” in the estimation equation. To make this choice in practice, we use standard automated cross-validation techniques, such as by randomly dividing the labeled set into training and test sets and then checking what works in those data.

<sup>11</sup>To generate a clear example of where this assumption is violated, we divided a test set into subsets based on the sophistication of the language using Flesch-Kincaid scores, which attempt to measure the grade level needed to read a text. We then tried to estimate the document category frequencies from a labeled set that made no such distinctions. Since the language sophistication is computed directly from the document text, equation (7) was violated and our estimates were biased as a result.

In practice, the number of word stems to choose to avoid sparseness bias mainly seems to be a function of the number of unique word stems in the documents. Fixing any problem that may arise via these types of cross-validation tests is not difficult. Given the other recommendations discussed above—stability in  $P(S | D)$ , coding categories that are homogeneous and clearly defined—the choice of the optimal number of subsets can account for many of the performance problems we observe in practice. In some applications, researchers may find it helpful to weight the word stems unevenly, so that words likely to have more information (such as based on their “mutual information”) appear more frequently in the subsets, although we have not found this necessary.

Finally, we require a reasonable number of documents in each category of  $D$  to be hand coded. Although we studied the efficiency of our procedure as a function of the number of hand-coded documents above, these results would change if by chance some categories had very few hand-coded documents and we cared about small differences in the proportions in these population categories. This makes sense, of course, since the method requires examples from which to generalize. Discovering too few examples for one or more categories can be dealt with in several ways. Most commonly, one can alter the definition of the categories or can change the coding rules.

However, even if examples of some categories are rare, they may be sufficiently well represented in the much larger population set to be of interest to social scientists. To deal with situations like this, we would need to find more examples from these relatively rare categories. Doing so by merely increasing the size of the hand-coded data set would be wasteful given that we would wind up with many more coded documents in the more prevalent categories. Still, it may be possible to use available meta-data to find the needed documents with higher probability. In our blogs data, we could find blog posts of certain types via links from other already hand-coded posts or from popular directories of certain types of blogs. Fortunately, the labeled set is assumed to be generated conditional on the categories, and so no bias is induced if we add extra examples via this “case-control” approach (cf. King and Zeng 2002).

Throughout all these potential problems, the best approach seems to be the radically empirical procedure suggested in the supervised learning literature. If the procedure you choose works, it works; if it doesn’t, it doesn’t. And so one should verify that the procedures work by subdividing the labeled set into training and (truly out of sample) test sets and then directly testing hypotheses about the success of the procedure. Ideally, this should

then be repeated with different types of labeled test sets. The more we make ourselves vulnerable to being wrong, using rigorous scientific procedures, the more we learn. Fortunately, the tools we make available here would seem to make it possible to learn enough to produce a reliable procedure in many applications.

Relatedly, standard errors and confidence intervals take a very different role in this type of research than the typical observational social science work. For most methods, the only way to shrink confidence intervals is to collect more data. For the method introduced here, all a researcher needs to do is to hand code additional documents (selected randomly or randomly conditional on  $D$ ) and rerun the algorithm. As long as no data are discarded along the way, continuing to hand code until one’s confidence intervals are small enough induces no bias, since our methodology (like direct sampling) is invariant to sampling plans (Thompson 2002, 286ff). A reasonably general approach is to hand code roughly 200 documents and run the algorithm. If uncertainty is more than desired, then hand code 100 more randomly selected documents, add them to the first set, reestimate, and continue until the uncertainty is small enough.

Given the many possible applications of this method, it is difficult to provide general guidelines about how time-intensive the entire process is likely to be. However, our experience is that identifying clear categories that humans are consistently able to differentiate takes far longer than the automated analyses we propose. Once users have clearly defined categories hand coded for a few hundred documents, they can often estimate the document category proportions for far larger corpora a few hours later.

## Concluding Remarks

Existing supervised methods of analyzing textual data come primarily from the tremendously productive computer science literature. This literature has been focused on optimizing the goals of computer science, which for the most part involve maximizing the percent of documents correctly classified into a given set of categories. We do not offer a way to improve on the computer scientists’ goals. Instead of seeking to classify any individual document, most social science literature that has hand-(or computer-) coded text is primarily interested in broad characterizations about the whole set of documents, such as unbiased estimates of the proportion of documents in given categories. Unfortunately, since they are optimized for a different purpose, computer science methods often produce biased estimates of these category proportions.

By developing methods for analyzing textual data that optimize social science goals directly, we are able to considerably outperform standard computer science methods developed for a different purpose. In addition, our approach requires no modeling assumptions, no modeling choices, and no complicated statistical approaches, and lets the social scientist pose the theoretical question to be answered. It also requires far less work than projects based entirely on hand coding, and much less work than most computer science methods of individual classification; it is both fast and can be used in real time. Individual-level classification is not a result of this method, and so it is not useful for all tasks, but numerous quantities of interest, from separate subdivisions of the population or different populations, can be estimated. As with all supervised learning methods, our approach does require careful efforts to properly define categories and to hand code a small sample of texts.

Although we have included only a few applications in this article, the methods offered here would seem applicable to many possible analyses that may not have been feasible previously. With the explosion of numerous types and huge quantities of text available to researchers on the web and elsewhere, we hope social scientists will begin to use these methods, and develop others, to harvest this new information and to improve our knowledge of the political, social, cultural, and economic worlds.

## Appendix Correcting for Lack of Intercoder Reliability

Hand coding is often an error-prone task. Intercoder reliability is measured in many different ways in the literature, but the rates tend to be lower with more categories and more theoretically interesting coding schemes and are almost never perfectly reliable. Unfortunately, “the classical supervised classification paradigm is based on the assumption that there are no errors in the true class labels” (Hand 2006, 9). The problem may be due to “conceptual stretching” (Collier and Mahon 1993) or “concept drift” (Widmer and Kubat 1996) that could in principle be fixed with a more disciplined study of the categories or coder training, but in practice some error is always left. In current practice, scholars typically report some reliability statistics and then use methods that assume no misclassification. Here, we propose to address misclassification via simulation-extrapolation (SIMEX; Cook and Stefanski 1994; Küchenhoff, Mwalili, and Lassaffre 2006).

As an example, before we developed our methods, we had at least two coders categorize each of 4,169 blog

**TABLE 2 Intercoder Reliability**

	-2	-1	0	1	2	NA	NB	$P(D_1)$
-2	.70	.10	.01	.01	.00	.02	.16	.28
-1	.33	.25	.04	.02	.01	.01	.35	.08
0	.13	.17	.13	.11	.05	.02	.40	.02
1	.07	.06	.08	.20	.25	.01	.34	.03
2	.03	.03	.03	.22	.43	.01	.25	.03
NA	.04	.01	.00	.00	.00	.81	.14	.12
NB	.10	.07	.02	.02	.02	.04	.75	.45

*Notes:* This table presents conditional probabilities for coder 2's classification (in a set of column entries) given a code assigned by coder 1 (corresponding to a particular row), or  $P(D_2 | D_1)$ . For instance, when coder 1 chooses category -2, coder 2 will choose the same category 70% of the time, category -1 10% of the time, and so on across the first row. This matrix is estimated from all 4,169 coding pairs from five coders. The final column denotes the marginal probability that coder 1 placed the blog in each category.

posts. In these data, our coders agreed on the classification of 66.5% of the blog posts; they agreed on 71.3% of blog posts among those when both coders agreed the post contained an opinion; and they agreed on 92% of the posts for an aggregated classification of negative, neutral, or positive opinions among posts with opinions. Table 2 gives more detailed information. For any two coders, arbitrarily named 1 and 2, each row gives the probability of coder 2's classification given a particular classification  $d$  chosen by coder 1,  $P(D_2 | D_1 = d)$ , with the marginal probability for coder 1 appearing in the last column,  $P(D_1)$ . The “misclassification” (or “confusion”) matrix in this table includes information from all combinations of observed ordered coder pairs.

*Intuition.* For intuition, we illustrate our approach by an analogy to what might occur during a highly funded research project as a coding scheme becomes clearer, the coding rules improve, and coder training gets better. For clarity, imagine that through five successive rounds, we have different, more highly trained coders classifying the same set of documents with improved coding rules. If we do well, the results of each round will have higher rates of intercoder reliability than the last. The final round will be best, but still not perfect. If we could continue this process indefinitely, we might imagine that we would remove all misclassification.

Now suppose our estimate of the percent of documents in category 2 is 5% in the first round, 11% in the second, 14% in the third, 19% in the fourth, and 23% in the last round. Following all previously published content analyses, our estimate of the proportion in category 2 would be 23%. This is not unreasonable, but it appears

to leave some information on the table. In particular, if the proportion of documents in category 2 is increasing steadily as the level of intercoder reliability at each round improves, then we might reasonably extrapolate this proportion to the point where intercoder agreement is perfect. We might thus conclude that the true proportion in category 2 is actually somewhat larger than 23%. We might even formalize this idea by building some type of regression model to predict the category 2 proportion with the level of intercoder reliability and extrapolate to the unobserved point where reliability is perfect. Since this procedure involves extrapolation, it is inherently model dependent and so uncertainty from its inferences will exceed the nominal levels (King and Zeng 2006). However, a crucial point is that even using the figure from the final round and doing no subsequent processing still involves an extrapolation; it is just that the extrapolation ignores the information from previous rounds of coding. So using 23% as our estimate and ignoring this problem is no safer.

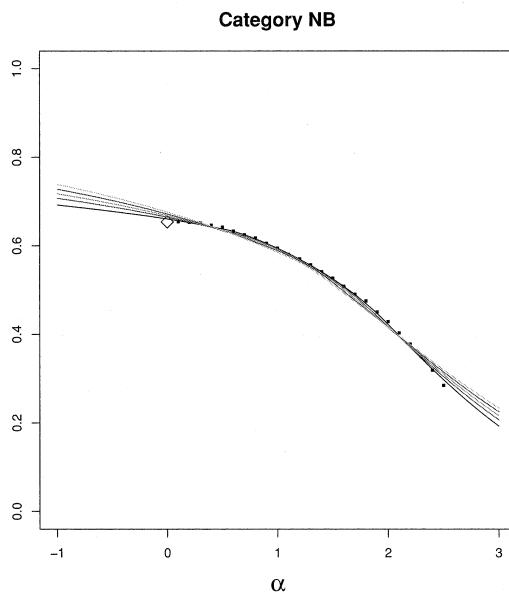
*Formalization.* Following the intuition outlined above, we make use of the misclassifications estimated from a single round of coding with more than one coder, simulate what would have happened to the document category proportions if there were even lower levels of intercoder reliability, and extrapolate back to the point of no misclassification.

To formalize this SIMEX procedure, begin with our estimation method, which would give statistically consistent answers if it were applied to data with no misclassification. The same method applied to error-prone data is presumably biased. However, in this problem, the type of misclassification is easy to characterize, as we do in Table 2. Then we follow five steps: (1) Take each observed data point  $D_i$  in the labeled set and simulate  $M$  error-inflated pseudo-data points, using the misclassification matrix in Table 2. We do this by drawing  $M$  values of  $\tilde{D}_i$  from the probability density  $P(\tilde{D}_i | D_i)$  (given the observed data point  $D_i$ ) which appears in the corresponding row of the table. This step creates  $M$  simulated data sets with twice the amount of measurement error, of the same type as in our observed data, to these pseudo-data. We then repeat this procedure starting with these pseudo-data to produce  $M$  pseudo-data sets with three times the measurement error as in the original data. Then again with four times the amount of measurement error, etc. (2) We apply our estimator to each of the simulated pseudo-data sets and average over the  $M$  results for each level of added error. This leads to a sequence of averaged results from each of the pseudo-estimators, with a different level of intercoder reliability. (3) We transform

these data using the multivariate logistic transformation to keep them constrained to the simplex, and then (4) fit a relationship between the transformed average proportion of observations estimated to be in each category from the error-inflated pseudo-data sets and the amount of added error in each. We then (5) extrapolate back to the unobserved point of zero measurement error, and transform the results.

*Illustration.* Figure 6 gives an example of this procedure for one category from our blogs data. The vertical axis of this graph is the proportion of observations in category NB. The horizontal axis, labeled  $\alpha$ , gives the number of additional units of misclassification error we have added to the original data, with the observed data at value 0. The estimate of  $P(D = \{\text{NB}\})$  from the original data (corresponding to the last round of coding from the earlier example) is denoted with a diamond above the value of zero. A value of  $\alpha$  of 1 means that the original data went through the misclassification matrix in Table 2 once; 2 means twice, etc. Some noninteger values are also included. In the application, it seems likely that the

**FIGURE 6** SIMEX Analysis of the Proportion of Documents in Category NB (Not a Blog)



*Notes:* The estimate from the observed data appears above 0 marked with a diamond; other points are simulated. The goal is to decide on the proportion in category NB at a horizontal axis value of -1.

proportion of documents we would have estimated to be in category NB, if our coders had perfect rates of intercoder reliability, would be higher than the proportion from our actual observed data.

All applications begin with the point estimated from the observed data at zero (marked by a diamond in the figure) and extrapolate it over to the horizontal axis value of  $-1$ , which denotes the data with no misclassification error. The implicit extrapolation used in prior content analysis research occurs by effectively drawing a flat line from the diamond to the vertical axis on the left. Instead, in Figure 6, estimates from the error-inflated data also appear, as well as several alternative (LOESS-based) models used to form possible extrapolations to the left axis where our estimates appear. In all cases, estimates appear somewhat higher than the nominal (flat line) extrapolation. Differences among the lines indicate uncertainty due to extrapolation-induced model dependence.

## References

- Adamic, L. A., and N. Glance. 2005. "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog." *Proceedings of the 3rd International Workshop on Link Discovery*.
- Benoit, Kenneth, and Michael Laver. 2003. "Estimating Irish Party Policy Positions Using Computer Wordscore: The 2002 Election - A Research Note." *Irish Political Studies* 18(1): 97–107.
- Berelson, B., and S. de Grazia. 1947. "Detecting Collaboration in Propaganda." *Public Opinion Quarterly* 11(2): 244–53.
- Brank, Janez, Marko Grobelnik, Natasa Milic-Frayling, and Dunja Mladenic. 2002. "Feature Selection Using Linear Support Vector Machines." Technical report, Microsoft Research.
- Cavnar, W. B., and J. M. Trenkle. 1994. "N-Gram-Based Text Categorization." *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*.
- Collier, David, and James E. Mahon, Jr. 1993. "Conceptual 'Stretching' Revisited." *American Political Science Review* 87(4): 845–55.
- Cook, J., and L. Stefanski. 1994. "Simulation-Extrapolation Estimation in Parametric Measurement Error Models." *Journal of the American Statistical Association* 89: 1314–28.
- Das, Sanjiv R., and Mike Y. Chen. 2001. "Yahoo! for Amazon: Opinion Extraction from Small Talk on the Web." Unpublished manuscript, Santa Clara University.
- Drezner, Daniel W., and Henry Farrell. 2004. "The Power and Politics of Blogs." Paper presented at the annual meeting of the American Political Science Association.
- Gamson, William A. 1992. *Talking Politics*. New York: Cambridge University Press.
- Gerner, Deborah J., Philip A. Schrodt, Ronald A. Francisco, and Judith L. Weddle. 1994. "Machine Coding of Event Data Using Regional and International Sources." *International Studies Quarterly* 38(1): 91–119.
- Gerring, John. 1998. *Party Ideologies in America, 1828–1996*. New York: Cambridge University Press.
- Gilens, Martin. 1999. *Why Americans Hate Welfare*. Chicago: University of Chicago Press.
- Ginsberg, Benjamin. 1986. *The Captive Public: How Mass Opinion Promotes State Power*. New York: Basic Books.
- Grindle, Merilee S. 2005. "Going Local: Decentralization, Democratization, and the Promise of Good Governance." Cambridge, MA: Kennedy School of Government, Harvard University.
- Hand, David J. 2006. "Classifier Technology and the Illusion of Progress." *Statistical Science* 21(1): 1–14.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
- Helmond, Anne. 2008. "How Many Blogs Are There? Is Someone Still Counting?" *The Blog Herald* (2/11). <http://www.blogherald.com/2008/02/11/how-many-blogs-are-there-is-someone-still-counting/>.
- Hillygus, Sunshine, and Todd G. Shields. 2008. *The Persuadable Voter: Wedge Issues in Presidential Campaigns*. Princeton, NJ: Princeton University Press.
- Hindman, Matthew, Kostas Tsoutsoulikis, and Judy A. Johnson. 2003. "Googlearchy: How a Few Heavily-Linked Sites Dominate Politics on the Web." Paper presented at the annual meeting of the Midwest Political Science Association.
- Hopkins, Daniel, and Gary King. 2009. "Replication Data for: A Method of Automated Nonparametric Content Analysis for Social Science." UNF:3:xlE5stLgKvpeMvxzLxzEQ== <hdl:1902.1/12898> Murray Research Archive [Distributor].
- Hsu, C. W., C. C. Chang, and C. J. Lin. 2003. "A Practical Guide to Support Vector Classification." Technical report, National Taiwan University.
- Huckfeldt, R. Robert, and John Sprague. 1995. *Citizens, Politics, and Social Communication*. New York: Cambridge University Press.
- Joachims, Thorsten. 1998. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." In *Machine Learning ECML-98*, ed. Claire Nédellec and Céline Rouvierol. Vol. 1398. New York: Springer, 127–42.
- Jones, Bryan D., and Frank R. Baumgartner. 2005. *The Politics of Attention: How Government Prioritizes Problems*. Chicago: University of Chicago Press.
- Kellstedt, Paul M. 2003. *The Mass Media and the Dynamics of American Racial Attitudes*. New York: Cambridge University Press.
- King, Gary, and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57(3): 617–42. <http://gking.harvard.edu/files/abs/infoex-abs.shtml>.
- King, Gary, and Ying Lu. 2008. "Verbal Autopsy Methods with Multiple Causes of Death." *Statistical Science* 23(1): 78–91. <http://gking.harvard.edu/files/abs/vamc-abs.shtml>.
- King, Gary, and Langche Zeng. 2002. "Estimating Risk and Rate Levels, Ratios, and Differences in Case-Control Studies." *Statistics in Medicine* 21: 1409–27.
- King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2): 131–59. <http://gking.harvard.edu/files/abs/counterft-abs.shtml>.

- Kolari, Pranam, Tim Finin, and Anupam Joshi. 2006. "SVMs for the Blogosphere: Blog Identification and Splog Detection." American Association for Artificial Intelligence Spring Symposium on Computational Approaches to Analyzing Weblogs.
- Krippendorff, D. K. 2004. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage.
- Küchenhoff, Helmut, Samuel M. Mwalili, and Emmanuel Las-saffre. 2006. "A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX." *Biometrics* 62 (March): 85–96.
- Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(2): 311–31.
- Lenhart, Amanda, and Susannah Fox. 2006. "Bloggers: A Portrait of the Internet's New Storytellers." Technical report, Pew Internet and American Life Project. <http://207.21.232.103/pdfs/PIP%20Bloggers%20Report%20July%202019%202006.pdf>.
- Levy, P. S., and E. H. Kass. 1970. "A Three Population Model for Sequential Screening for Bacteriuria." *American Journal of Epidemiology* 91: 148–54.
- Lyman, Peter, and Hal R. Varian. 2003. "How Much Information 2003." Technical report, University of California. <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/>.
- Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: Massachusetts Institute of Technology.
- Mayhew, David R. 1991. *Divided We Govern: Party Control, Law-making and Investigations*. New Haven, CT: Yale University Press.
- Mendelberg, Tali. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton, NJ: Princeton University Press.
- Mutz, Diana C. 1998. *Impersonal Influence: How Perceptions of Mass Collectives Affect Political Attitudes*. New York: Cambridge University Press.
- Neuendorf, K. A. 2002. *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage.
- Pang, Bo, and Lillian Lee. 2008. "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval* 2(1): 1–135.
- Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Porter, M. F. 1980. "An Algorithm for Suffix Stripping." *Program* 14(3): 130–37.
- Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. 2009. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54(1): 209–28.
- Rudalevige, Andrew. 2002. *Managing the President's Program*. Princeton, NJ: Princeton University Press.
- Sebastiani, Fabrizio. 2002. "Machine Learning in Automated Text Categorization." *ACM Computing Surveys (CSUR)* 34(1): 1–47.
- Simon, Adam F., and Michael Xeons. 2004. "Dimensional Reduction of Word-Frequency Data as a Substitute for Intersubjective Content Analysis." *Political Analysis* 12(1): 63–75.
- Thomas, Matt, Bo Pang, and Lillian Lee. 2006. "Get Out the Vote: Determining Support or Opposition from Congressional Floor-Debate Transcripts." *Proceedings of EMNLP*. <http://www.cs.cornell.edu/home/llee/papers/tpl-convote.home.html>.
- Thompson, Steven K. 2002. *Sampling*. New York: John Wiley and Sons.
- Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and Equality: Civic Volunteerism in American Politics*. Cambridge, MA: Harvard University Press.
- Waples, D., B. Berelson, and F. R. Bradshaw. 1940. *What Reading Does to People: A Summary of Evidence on the Social Effects of Reading and a Statement of Problems for Research*. Chicago: University of Chicago Press.
- Widmer, G., and M. Kubat. 1996. "Learning in the Presence of Concept Drift and Hidden Contexts." *Machine Learning* 23(1): 69–101.
- Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

## How Censorship in China Allows Government Criticism but Silences Collective Expression

GARY KING *Harvard University*

JENNIFER PAN *Harvard University*

MARGARET E. ROBERTS *Harvard University*

**W**e offer the first large scale, multiple source analysis of the outcome of what may be the most extensive effort to selectively censor human expression ever implemented. To do this, we have devised a system to locate, download, and analyze the content of millions of social media posts originating from nearly 1,400 different social media services all over China before the Chinese government is able to find, evaluate, and censor (i.e., remove from the Internet) the subset they deem objectionable. Using modern computer-assisted text analytic methods that we adapt to and validate in the Chinese language, we compare the substantive content of posts censored to those not censored over time in each of 85 topic areas. Contrary to previous understandings, posts with negative, even vitriolic, criticism of the state, its leaders, and its policies are not more likely to be censored. Instead, we show that the censorship program is aimed at curtailing collective action by silencing comments that represent, reinforce, or spur social mobilization, regardless of content. Censorship is oriented toward attempting to forestall collective activities that are occurring now or may occur in the future—and, as such, seem to clearly expose government intent.

### INTRODUCTION

The size and sophistication of the Chinese government's program to selectively censor the expressed views of the Chinese people is unprecedented in recorded world history. Unlike in the U.S., where social media is centralized through a few providers, in China it is fractured across hundreds of local sites. Much of the responsibility for censorship is devolved to these Internet content providers, who may be fined or shut down if they fail to comply with government censorship guidelines. To comply with the government, each individual site privately employs up to 1,000 censors. Additionally, approximately 20,000–50,000 Internet police (*wang jing*) and Internet monitors (*wang guanban*) as well as an estimated 250,000–300,000 “50 cent party members” (*wumao dang*) at all levels of government—central, provincial, and local—participate in this huge effort (Chen and

Ang 2011, and our interviews with informants, granted anonymity). China overall is tied with Burma at 187th of 197 countries on a scale of press freedom (Freedom House 2012), but the Chinese censorship effort is by far the largest.

In this article, we show that this program, designed to limit freedom of speech of the Chinese people, paradoxically also exposes an extraordinarily rich source of information about the Chinese government's interests, intentions, and goals—a subject of long-standing interest to the scholarly and policy communities. The information we unearth is available in continuous time, rather than the usual sporadic media reports of the leaders' sometimes visible actions. We use this new information to develop a theory of the overall purpose of the censorship program, and thus to reveal some of the most basic goals of the Chinese leadership that until now have been the subject of intense speculation but necessarily little empirical analysis. This information is also a treasure trove that can be used for many other scholarly (and practical) purposes.

Our central theoretical finding is that, contrary to much research and commentary, the purpose of the censorship program is *not* to suppress criticism of the state or the Communist Party. Indeed, despite widespread censorship of social media, we find that when the Chinese people write scathing criticisms of their government and its leaders, the probability that their post will be censored does not increase. Instead, we find that the purpose of the censorship program is to reduce the probability of collective action by clipping social ties whenever any collective movements are in evidence or expected. We demonstrate these points and then discuss their far-reaching implications for many research areas within the study of Chinese politics and comparative politics.

In the sections below, we begin by defining two theories of Chinese censorship. We then describe our unique data source and the unusual challenges involved in gathering it. We then lay out our strategy for analysis,

Gary King is Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138 (<http://GKing.harvard.edu>, king@harvard.edu) (617) 500-7570.

Jennifer Pan is Ph.D. Candidate, Department of Government, 1737 Cambridge Street, Harvard University, Cambridge MA 02138 (<http://people.fas.harvard.edu/~jjpan/>) (917) 740-5726.

Margaret E. Roberts is Ph.D. Candidate, Department of Government, 1737 Cambridge Street, Harvard University, Cambridge MA 02138 (<http://scholar.harvard.edu/mroberts/home>).

Our thanks to Peter Bol, John Carey, Justin Grimmer, Navid Hassaniour, Iain Johnston, Bill Kirby, Peter Lorentzen, Jean Oi, Liz Perry, Bob Putnam, Susan Shirk, Noah Smith, Lynn Vavreck, Andy Walder, Barry Weingast, and Chen Xi for many helpful comments and suggestions, and to our insightful and indefatigable undergraduate research associates, Wanxin Cheng, Jennifer Sun, Hannah Waight, Yifan Wu, and Min Yu, for much help along the way. Our thanks also to Larry Summers and John Deutch for helping us to ensure that we satisfy the sometimes competing goals of national security and academic freedom. For help with a wide array of data and technical issues, we are especially grateful to the incredible teams, and for the unparalleled infrastructure, at Crimson Hexagon ([CrimsonHexagon.com](http://CrimsonHexagon.com)) and the Institute for Quantitative Social Science at Harvard University ([iq.harvard.edu](http://iq.harvard.edu)).

give our results, and conclude. Appendixes include coding details, our automated Chinese text analysis methods, and hints about how censorship behavior presages government action outside the Internet.

## GOVERNMENT INTENTIONS AND THE PURPOSE OF CENSORSHIP

**Previous Indicators of Government Intent.** Deciphering the opaque intentions and goals of the leaders of the Chinese regime was once the central focus of scholarly research on elite politics in China, where Western researchers used Kremlinology—or Pekingology—as a methodological strategy (Chang 1983; Charles 1966; Hinton 1955; MacFarquhar 1974, 1983; Schurmann 1966; Teiwes 1979). With the Cultural Revolution and with China's economic opening, more sources of data became available to researchers, and scholars shifted their focus to areas where information was more accessible. Studies of China today rely on government statistics, public opinion surveys, interviews with local officials, as well as measures of the visible actions of government officials and the government as a whole (Guo 2009; Kung and Chen 2011; Shih 2008; Tsai 2007a, b). These sources are well suited to answer other important political science questions, but in gauging government intent, they are widely known to be indirect, very sparsely sampled, and often of dubious value. For example, government statistics, such as the number of "mass incidents", could offer a view of government interests, but only if we could somehow separate true numbers from government manipulation. Similarly, sample surveys can be informative, but the government obviously keeps information from ordinary citizens, and even when respondents have the information researchers are seeking they may not be willing to express themselves freely. In situations where direct interviews with officials are possible, researchers are in the position of having to read tea leaves to ascertain what their informants really believe.

Measuring intent is all the more difficult with the sparse information coming from existing methods because the Chinese government is not a monolithic entity. In fact, in those instances when different agencies, leaders, or levels of government work at cross purposes, even the concept of a unitary intent or motivation may be difficult to define, much less measure. We cannot solve all these problems, but by providing more information about the state's revealed preferences through its censorship behavior, we may be somewhat better able to produce useful measures of intent.

**Theories of Censorship.** We attempt to complement the important work on how censorship is conducted, and how the Internet may increase the space for public discourse (Duan 2007; Edmond 2012; Egorov, Guriev, and Sonin 2009; Esarey and Xiao 2008, 2011; Herold 2011; Lindtner and Szablewicz 2011; MacKinnon 2012; Yang 2009; Xiao 2011), by beginning to build an empirically documented theory of why the government censors and what it is trying to achieve through this extensive program. While current scholarship draws the reasonable but broad conclusion that Chinese govern-

ment censorship is aimed at maintaining the status quo for the current regime, we focus on what specifically the government believes is critical, and what actions it takes, to accomplish this goal.

To do this, we distinguish two theories of what constitutes the goals of the Chinese regime as implemented in their censorship program, each reflecting a different perspective on what threatens the stability of the regime. First is a *state critique* theory, which posits that the goal of the Chinese leadership is to suppress dissent, and to prune human expression that finds fault with elements of the Chinese state, its policies, or its leaders. The result is to make the sum total of available public expression more favorable to those in power. Many types of state critique are included in this idea, such as poor government performance.

Second is what we call the theory of *collective action potential*: the target of censorship is people who join together to express themselves collectively, stimulated by someone other than the government, and seem to have the potential to generate collective action. In this view, collective expression—many people communicating on social media on the same subject—regarding actual collective action, such as protests, as well as those about events that seem likely to generate collective action but have not yet done so, are likely to be censored. Whether social media posts with collective action potential find fault with or assign praise to the state, or are about subjects unrelated to the state, is unrelated to this theory.

An alternative way to describe what we call "collective action potential" is the apparent perspective of the Chinese government, where collective expression organized outside of governmental control equals factionalism and ultimately chaos and disorder. For example, on the eve of Communist Party's 90th birthday, the state-run Xinhua news agency issued an opinion that western-style parliamentary democracy would lead to a repetition of the turbulent factionalism of China's Cultural Revolution (<http://j.mp/McRDXk>). Similarly, at the Fourth Session of the 11th National People's Congress in March of 2011, Wu Bangguo, member of the Politburo Standing Committee and Chairman of the Standing Committee of the National People's Congress, said that "On the basis of China's conditions...we'll not employ a system of multiple parties holding office in rotation" in order to avoid "an abyss of internal disorder" (<http://j.mp/Ldhp25>). China observers have often noted the emphasis placed by the Chinese government on maintaining stability (Shirk 2007; Whyte 2010; Zhang et al. 2002), as well as the government's desire to limit collective action by clipping social ties (Perry 2002, 2008). The Chinese regime encounters a great deal of contention and collective action; according to Sun Liping, a professor of Sociology at Tsinghua University, China experienced 180,000 "mass incidents" in 2010 (<http://j.mp/McQeji>). Because the government encounters collective action frequently, it influences the actions and perceptions of the regime. The stated perspective of the Chinese government is that limitations on horizontal communications is a legitimate and effective action designed to protect its people (Perry 2010)—in other words, a

paternalistic strategy to avoid chaos and disorder, given the conditions of Chinese society.

Current scholarship has not been able to differentiate empirically between the two theories we offer. Marolt (2011) writes that online postings are censored when they “either criticize China’s party state and its policies directly or advocate collective political action.” MacKinnon (2012) argues that during the Wenzhou high speed rail crash, Internet content providers were asked to “track and censor critical postings.” Esarey and Xiao (2008) find that Chinese bloggers use satire to convey criticism of the state in order to avoid harsh repression. Esarey and Xiao (2011) write that party leaders are most fearful of “Concerted efforts by influential netizens to pressure the government to change policy,” but identify these pressures as criticism of the state. Shirk (2011) argues that the aim of censorship is to constrain the mobilization of political opposition, but her examples suggest that critical viewpoints are those that are suppressed.

Collective action in the form of protests is often thought to be the death knell of authoritarian regimes. Protests in East Germany, Eastern Europe, and most recently the Middle East have all preceded regime change (Ash 2002; Lohmann 1994; Przeworski et al. 2000). A great deal of scholarship on China has focused on what leads people to protest and their tactics (Blecher 2002; Cai 2002; Chen 2000; Lee 2007; O’Brien and Li 2006; Perry 2002, 2008). The Chinese state seems focused on preventing protest at all costs—and, indeed, the prevalence of collective action is part of the formal evaluation criteria for local officials (Edin 2003). However, several recent works argue that authoritarian regimes may expect and welcome substantively narrow protests as a way of enhancing regime stability by identifying, and then dealing with, discontented communities (Dimitrov 2008; Lorentzen 2010; Chen 2012). Chen (2012) argues that small, isolated protests have a long tradition in China and are an expected part of government.

**Outline of Results.** The nature of the two theories means that either or both could be correct or incorrect. Here, we offer evidence that, with few exceptions, the answer is simple: state critique theory is incorrect and the theory of collective action potential is correct. Our data show that the Chinese censorship program allows for a wide variety of criticisms of the Chinese government, its officials, and its policies. As it turns out, censorship is primarily aimed at restricting the spread of information that may lead to collective action, regardless of whether or not the expression is in direct opposition to the state and whether or not it is related to government policies. Large increases in online volume are good predictors of censorship when these increases are associated with events related to collective action, e.g., protests on the ground. In addition, we measure sentiment within each of these events and show that during these events, the government censors views that are both supportive and critical of the state. These results reveal that the Chinese regime believes suppressing social media posts with collective

action potential, rather than suppression of criticism, is crucial to maintaining power.

## DATA

We describe here the challenges involved in collecting large quantities of detailed information that the Chinese government does not want anyone to see and goes to great lengths to prevent anyone from accessing. We discuss the types of censorship we study, our data collection process, the limitations of this study, and ways we organize the data for subsequent analyses.

### Types of Censorship

Human expression is censored in Chinese social media in at least three ways, the last of which is the focus of our study. First is “The Great Firewall of China,” which disallows certain entire Web sites from operating in the country. The Great Firewall is an obvious problem for foreign Internet firms, and for the Chinese people interacting with others outside of China on these services, but it does little to limit the expressive power of Chinese people who can find other sites to express themselves in similar ways. For example, Facebook is blocked in China but RenRen is a close substitute; similarly Sina Weibo is a popular Chinese clone of Twitter, which is also unavailable.

Second is “keyword blocking” which stops a user from posting text that contain banned words or phrases. This has limited effect on freedom of speech, since netizens do not find it difficult to outwit automated programs. To do so, they use analogies, metaphors, satire, and other evasions. The Chinese language offers novel evasions, such as substituting characters for those banned with others that have unrelated meanings but sound alike (“homophones”) or look similar (“homographs”). An example of a homograph is 目田, which has the nonsensical literal meaning of “eye field” but is used by World of Warcraft players to substitute for the banned but similarly shaped 自由 which means freedom. As an example of a homophone, the sound “hexie” is often written as 河蟹, which means “river crab,” but is used to refer to 和谐, which is the official state policy of a “harmonious society.”

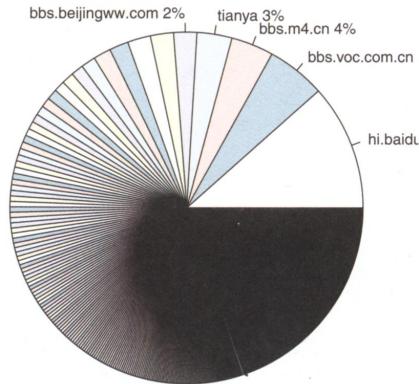
Once past the first two barriers to freedom of speech, the text gets posted on the Web and the censors read and remove those they find objectionable. As nearly as we can tell from the literature, observers, private conversations with those inside several governments, and an examination of the data, content filtering is in large part a manual effort—censors read post by hand. Automated methods appear to be an auxiliary part of this effort. Unlike The Great Firewall and keyword blocking, hand censoring cannot be evaded by clever phrasing. Thus, it is this last and most extensive form of censoring that we focus on in this article.

### Collection

We begin with social media blogs in which it is at least possible for writers to express themselves fully, prior to

**Figure 1. The Fractured Structure of the Chinese Social Media Landscape**

(a) Sample of Sites



(b) All Sites excluding Sina

All tables and figures appear in color in the online version. This version can be found at <http://j.mp/LdVXqN>.

possible censorship, and leaving to other research social media services that constrain authors to very short Twitter-like (*weibo*) posts (e.g., Bamman, O'Connor, and Smith 2012). In many countries, such as the U.S., almost all blog posts appear on a few large sites (Facebook, Google's blogspot, Tumblr, etc.); China does have some big sites such as *sina.com*, but a large portion of its social media landscape is finely distributed over numerous individual sites, e.g., local bbs forums. This difference poses a considerable logistical challenge for data collection—with different Web addresses, different software interfaces, different companies and local authorities monitoring those accessing the sites, different network reliabilities, access speeds, terms of use, and censorship modalities, and different ways of potentially hindering or stopping our data collection. Fortunately, the structure of Chinese social media also turns out to pose a special opportunity for studying localized control of collective expression, since the numerous local sites provide considerable information about the geolocation of posts, much more than is available even in the U.S.

The most complicated engineering challenges in our data collection process involves locating, accessing, and downloading posts from many Web sites before Internet content providers or the government reads and censors those that are deemed by authorities as objectionable;<sup>1</sup> revisiting each post frequently enough to learn if and when it was censored; and proceeding with data collection in so many places in China without affecting the system we were studying or being prevented from studying it. The reason we are able to accomplish this is because our data collection methods

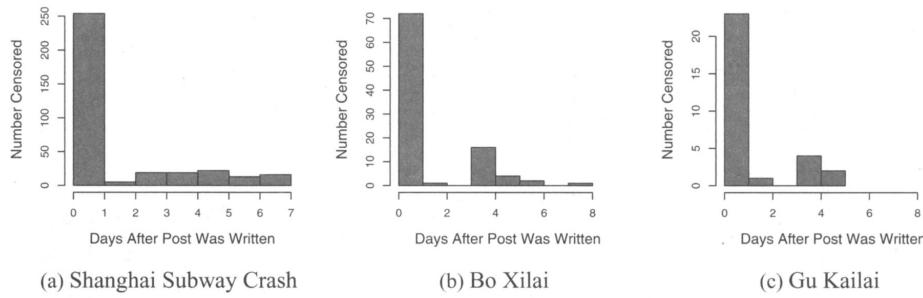
are highly automated whereas Chinese censorship entails manual effort. Our extensive engineering effort, which we do not detail here for obvious reasons, is executed at many locations around the world, including inside China.

Ultimately, we were able to locate, obtain access to, and download social media posts from 1,382 Chinese Web sites during the first half of 2011. The most striking feature of the structure of Chinese social media is its extremely long (power-law like) tail. Figure 1 gives a sample of the sites and their logos in Chinese (in panel (a)) and a pie chart of the number of posts that illustrate this long tail (in panel (b)). The largest sources of posts include blog.sina (with 59% of posts), hi.baidu, bbs.voc, bbs.m4, and tianya, but the tail keeps going.<sup>2</sup>

Social media posts cover such a huge range of topics that a random sampling strategy attempting to cover everything is rarely informative about any individual topic of interest. Thus, we begin with a stratified random sampling design, organized hierarchically. We first choose eighty-five separate topic areas within three categories of hypothesized political sensitivity, ranging from "high" (such as Ai Weiwei) to "medium" (such as the one child policy) to "low" (such as a popular online video game). We chose the specific topics within these categories by reviewing prior literature, consulting with China specialists, and studying current events. Appendix A gives a complete list. Then, within each topic area, defined by a set of keywords, we collected all social media posts over a six-month period. We examined the posts in each area, removed spam, and explored the content with the tool for computer-assisted reading (Crosas et al. 2012; Grimmer and King

<sup>1</sup> See MacKinnon (2012) for additional information on the censorship process.

<sup>2</sup> See <http://blog.sina.com.cn/>, <http://hi.baidu.com/>, <http://voc.com.cn/>, <http://bbs.m4.cn/>, and <http://tianya.cn/>.

**Figure 2. The Speed of Censorship, Monitored in Real-Time**

2011). With this procedure we collected 3,674,698 posts, with 127,283 randomly selected for further analysis. (We repeated this procedure for other time periods, and in some cases in more depth for some issue areas, and overall collected and analyzed 11,382,221 posts.) All posts originating from sites in China were written in Chinese, and excluded those from Hong Kong and Taiwan.<sup>3</sup> For each post, we examined its content, placed it on a timeline according to topic area, and revisited the Web site from which it came repeatedly thereafter to determine whether it was censored. We supplemented this information with other specific data collections as needed.

The censors are not shy, and so we found it straightforward to distinguish (intentional) censorship from sporadic outages or transient time-out errors. The censored Web sites include notes such as "Sorry, the host you were looking for does not exist, has been deleted, or is being investigated" (抱歉,指定的主题不存在或已被删除或正在被审核) and are sometimes even adorned with pictures of Jingjing and

Chacha, Internet police cartoon characters (警察).

Although our methods are faster than the Chinese censors, the censors nevertheless appear highly expert at their task. We illustrate this with analyses of random samples of posts surrounding the 9/27/2011 Shanghai Subway crash, and posts collected between 4/10/2012 and 4/12/2012 about Bo Xilai, a recently deposed member of the Chinese elite, and a separate collection of posts about his wife, Gu Kailai, who was accused and convicted of murder. We monitored each of the posts in these three areas continuously in near real time for nine days. (Censorship in other areas follow the same basic pattern.) Histograms of the time until censorship appear in Figure 2. For all three, the vast majority of censorship activity occurs within 24 hours of the original posting, although a few deletions occur longer than five days later. This is a remarkable organizational accomplishment, requiring large scale military-like precision: The many leaders at different

levels of government and at different Internet content providers first need to come to a decision (by agreement, direct order, or compromise) about what to censor in each situation; they need to communicate it to tens or hundreds of thousands of individuals; and then they must all complete execution of the plan within roughly 24 hours. As Edmond (2012) points out, the proliferation of information sources on social media makes information more difficult to control; however, the Chinese government has overcome these obstacles on a national scale. Given the normal human difficulties of coming to agreement with many others, and the usual difficulty of achieving high levels of intercoder reliability on interpreting text (e.g., Hopkins and King 2010, Appendix B), the effort the government puts into its censorship program is large, and highly professional. We have found some evidence of disagreements within this large and multifarious bureaucracy, such as at different levels of government, but we have not yet studied these differences in detail.

### Limitations

As we show below, our methodology reveals a great deal about the goals of the Chinese leadership, but it misses self-censorship and censorship that may occur before we are able to obtain the post in the first place; it also does not quantify the direct effects of The Great Firewall, keyword blocking, or search filtering in finding what others say. We have also not studied the effect of physical violence, such as the arrest of bloggers, or threats of the same. Although many officials and levels of government have a hand in the decisions about what and when to censor, our data only sometimes enable us to distinguish among these sources.

We are of course unable to determine the consequences of these limitations, although it is reasonable to expect that the most important of these are physical violence, threats, and the resulting self-censorship. Although the social media data we analyze include expressions by millions of Chinese and cover an extremely wide range of topics and speech behavior, the presumably much smaller number of discussions we cannot observe are likely to be those of the most (or most urgent) interest to the Chinese government.

<sup>3</sup> We identified posts as originating from mainland China by sending out a DNS query using the root url of the post and identifying the host IP.

Finally, in the past, studies of Internet behavior were judged based on how well their measures approximated “real world” behavior; subsequently, online behavior has become such a large and important part of human life that the expressions observed in social media is now important in its own right, regardless of whether it is a good measure of non-Internet freedoms and behaviors. But either way, we offer little evidence here of connections between what we learn in social media and press freedom or other types of human expression in China.

## ANALYSIS STRATEGY

Overall, an average of approximately 13% of all social media posts are censored. This average level is quite stable over time when aggregating over all posts in all areas, but masks enormous changes in volume of posts and censorship efforts. Our first hint of what might (not) be driving censorship rates is a surprisingly low correlation between our *ex ante* measure of political sensitivity and censorship: Censorship behavior in the low and medium categories was essentially the same (16% and 17%, respectively) and only marginally lower than the high category (24%).<sup>4</sup> Clearly something else is going on. To convey what this is, we now discuss our coding rules, our central hypothesis, and the exact operational procedures the Chinese government may use to censor.

### Coding Rules

We discuss our coding rules in five steps. First, we begin with social media posts organized into the eighty-five topic areas defined by keywords from our stratified random sampling plan. Although we have conducted extensive checks that these are accurate (by reading large numbers and also via modern computer-assisted reading technology), our topic areas will inevitably (with any machine or human classification technology) include some posts that do not belong. We take the conservative approach of first drawing conclusions even when affected by this error. Afterward, we then do numerous checks (via the same techniques) after the fact to ensure we are not missing anything important. We report below the few patterns that could be construed as a systematic error; each one turns out to strengthen our conclusions.

Second, conversation in social media in almost all topic areas (and countries) is well known to be highly “bursty,” that is, with periods of stability punctuated by occasional sharp spikes in volume around specific subjects (Ratkiewicz et al. 2010). We also found that with only two exceptions—pornography and criticisms of the censors, described below—censorship effort is often especially intense within *volume bursts*. Thus, we organize our data around these volume bursts. We

<sup>4</sup> That all three figures are higher than the average level of 13% reflects the fact that the topic areas we picked *ex ante* had generated at least some public discussion and included posts about events with collective action potential.

think of each of the eighty-five topic areas as a six-month time series of daily volume and detect bursts using the weights calculated from robust regression techniques to identify outlying observations from the rest of the time series (Huber 1964; Rousseeuw and Leroy 1987). In our data, this sophisticated burst detection algorithm is almost identical to using time periods with volume more than three standard deviations greater than the rest of the six-month period. With this procedure, we detected eighty-seven distinct volume bursts within sixty-seven of the eighty-five topic areas.<sup>5</sup>

Third, we examined the posts in each volume burst and identified the real world *event* associated with the online conversation. This was easy and the results unambiguous.

Fourth, we classified each event into one of five content areas: (1) collective action potential, (2) criticism of the censors, (3) pornography, (4) government policies, and (5) other news. As with topic areas, each of these categories may include posts that are critical or not critical of the state, its leaders, and its policies. We define collective action as the pursuit of goals by more than one person controlled or spurred by actors other than government officials or their agents. Our theoretical category of “collective action potential” involves any event that has the potential to cause collective action, but to be conservative, and to ensure clear and replicable coding rules, we limit this category to events which (a) involve protest or organized crowd formation outside the Internet; (b) relate to individuals who have organized or incited collective action on the ground in the past; or (c) relate to nationalism or nationalist sentiment that have incited protest or collective action in the past. (Nationalism is treated separately because of its frequently demonstrated high potential to generate collective action and also to constrain foreign policy, an area which has long been viewed as a special prerogative of the government; Reilly 2012.)

Events are categorized as criticism of censors if they pertain to government or nongovernment entities with control over censorship, including individuals and firms. Pornography includes advertisements and news about movies, Web sites, and other media containing pornographic or explicitly sexual content. Policies refer to government statements or reports of government activities pertaining to domestic or foreign policy. And “other news” refers to reporting on events, other than those which fall into one of the other four categories.

Finally, we conducted a study to verify the reliability of our event coding rules. To do this, we gave our rules above to two people familiar with Chinese politics and asked them to code each of the eighty-seven events (each associated with a volume burst) into one of the five categories. The coders worked independently and classified each of the events on their own. Decisions by the two coders agreed in 98.9% (i.e., eighty-six of eighty-seven) of the events. The only event with divergent codes was the pelting of Fang Binxing (the

<sup>5</sup> We attempted to identify duplicate posts, the Chinese equivalent of “retweets,” sblogs (spam blogs), and the like. Excluding these posts had no noticeable effect on our results.

architect of China's Great Firewall) with shoes and eggs. This event included criticism of the censors and to some extent collective action because several people were working together to throw things at Fang. We broke the tie by counting this event as an example of criticism of the censors, but however this event is coded does not affect our results since we predict both will be censored.

### Central Hypothesis

Our central hypothesis is that the government censors *all* posts in topic areas during volume bursts that discuss events with collective action potential. That is, the censors do not judge whether individual posts have collective action potential, perhaps in part because rates of intercoder reliability would likely be very low. In fact, Kuran (1989) and Lohmann (2002) show that it is information about a collective action event that propels collective action and so distinguishing this from explicit calls for collective action would be difficult if not impossible. Instead, we hypothesize that the censors make the much easier judgment, about whether the posts are on topics associated with events that have collective action potential, and they do it regardless of whether or not the posts criticize the state.

The censors also attempt to censor all posts in the categories of pornography and criticism of the censors, but not posts within event categories of government policies and news.

### The Government's Operational Procedures

The exact operational procedures by which the Chinese government censors is of course not observed. But based on conversations with individuals within and close to the Chinese censorship apparatus, we believe our coding rules can be viewed as an approximation to them. (In fact, after a draft of our article was written and made public, we received communications confirming our story.) We define topic areas by hand, sort social media posts into topic areas by keywords, and detect volume bursts automatically via statistical methods for time series data on post volume. (These steps might be combined by the government to detect topics automatically based on spikes in posts with high similarity, but this would likely involve considerable error given inadequacies in known fully automated clustering technologies.) In some cases, identifying the real world event might occur before the burst, such as if the censors are secretly warned about an upcoming event (such as the imminent arrest of a dissident) that could spark collective action. Identifying events from bursts that were observed first would need to be implemented at least mostly by hand, perhaps with some help from algorithms that identify statistically improbable phrases. Finally, the actual decision to censor an individual post—which, according to our hypothesis, involves checking whether it is associated with a particular event—is almost surely accomplished largely by hand, since no known statistical or machine learning technology can achieve a level of accuracy anywhere

near that which we observe in the Chinese censorship program. Here, censors may begin with keyword searches on the event identified, but will need to manually read through the resulting posts to censor those which are related to the triggering event. For example, when censors identified protests in Zengcheng as precipitating online discussion, they may have conducted a keyword search among posts for Zengcheng, but they would have had to read through these posts by hand to separate posts about protests from posts talking about Zengcheng in other contexts, say Zengcheng's lychee harvest.

## RESULTS

We now offer three increasingly specific tests of our hypotheses. These tests are based on (1) post volume, (2) the nature of the event generating each volume burst, and (3) the specific content of the censored posts. Additionally, Appendix C gives some evidence that government censorship behavior paradoxically reveals the Chinese government's intent to act outside the Internet.

### Post Volume

If the goal of censorship is to stop discussions with collective action potential, then we would expect more censorship during volume bursts than at other times. We also expect some bursts—those with collective action potential—to have much higher levels of censorship.

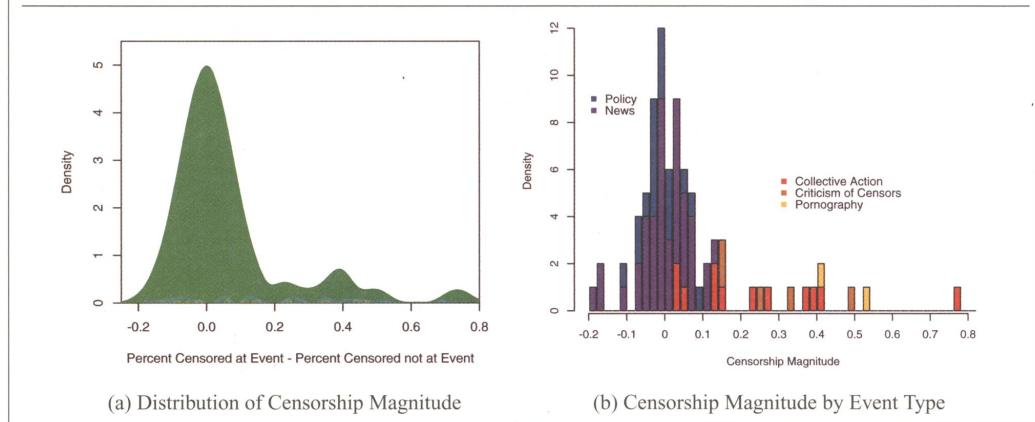
To begin to study this pattern, we define *censorship magnitude* for a topic area as the percent censored within a volume burst minus the percent censored outside all bursts. (The base rates, which vary very little across issue areas and which we present in detail in graphs below, do not impose empirically relevant ceiling or floor effects on this measure.) This is a stringent measure of the interests of the Chinese government because censoring during a volume burst is obviously more difficult owing to there being more posts to evaluate, less time to do it in, and little or no warning of when the event will take place.

Panel (a) in Figure 3 gives a histogram with results that appear to support our hypotheses. The results show that the bulk of volume bursts have a censorship magnitude centered around zero, but with an exceptionally long right tail (and no corresponding long left tail). Clearly volume bursts are often associated with dramatically higher levels of censorship even compared to the baseline during the rest of the six months for which we observe a topic area.

### The Nature of Events Generating Volume Bursts

We now show that volume bursts generated by events pertaining to collective action, criticism of censors, and pornography are censored, albeit as we show in different ways, while post volume generated by discussion of government policy and other news are not.

**Figure 3. “Censorship Magnitude,” The Percent of Posts Censored Inside a Volume Burst Minus Outside Volume Bursts.**



We discuss the state critique hypothesis in the next subsection. Here, we offer three separate, and increasingly detailed, views of our present results.

First, consider panel (b) of Figure 3, which takes the same distribution of censorship magnitude as in panel (a) and displays it by event type. The result is dramatic: events related to collective action, criticism of the censors, and pornography (in red, orange, and yellow) fall largely to the right, indicating high levels of censorship magnitude, while events related to policies and news fall to the left (in blue and purple). On average, censorship magnitude is 27% for collective action, but  $-1\%$  and  $-4\%$  for policy and news.<sup>6</sup>

Second, we list the specific events with the highest and lowest levels of censorship magnitude. These appear, using the same color scheme, in Figure 4. The events with the highest collective action potential include protests in Inner Mongolia precipitated by the death of an ethnic Mongol herder by a coal truck driver, riots in Zengcheng by migrant workers over an altercation between a pregnant woman and security personnel, the arrest of artist/political dissident Ai Weiwei, and the bombings over land claims in Fuzhou. Notably, one of the highest “collective action potential” events was not political at all: following the Japanese earthquake and subsequent meltdown of the nuclear plant in Fukushima, a rumor spread through Zhejiang province that the iodine in salt would protect people from radiation exposure, and a mad rush to buy salt ensued. The rumor was biologically false, and had nothing to do with the state one way or the other, but it was highly censored; the reason appears to be because of the localized control of collective expression by actors other than the government. Indeed, we find that salt

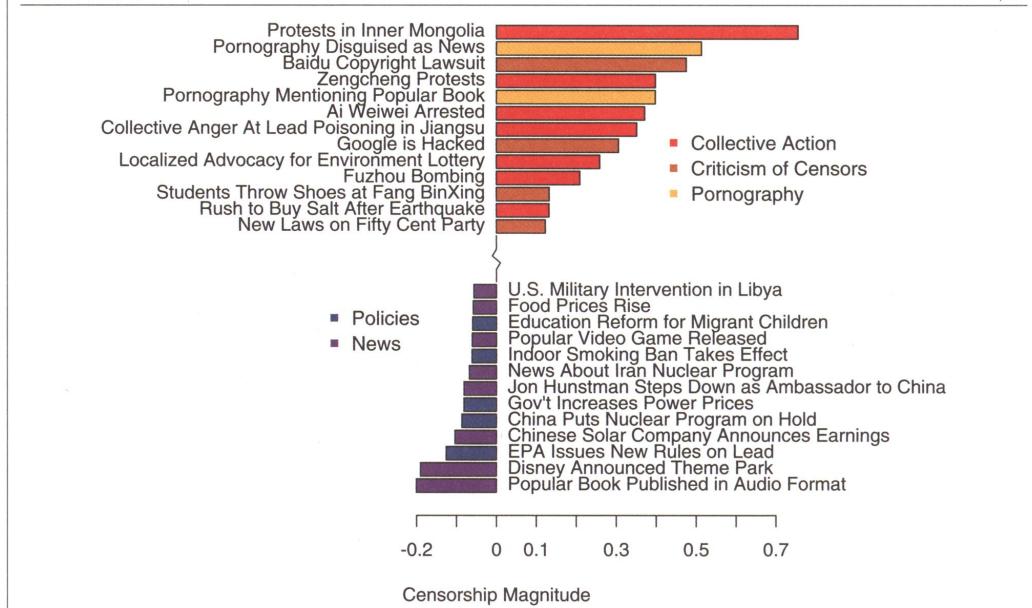
rumors on local Web sites are much more likely to be censored than salt rumors on national Web sites.<sup>7</sup>

Consistent with our theory of collective action potential, some of the most highly censored events are not criticisms or even discussions of national policies, but rather highly localized collective expressions that represent or threaten group formation. One such example is posts on a local Wenzhou Web site expressing support for Chen Fei, a environmental activist who supported an environmental lottery to help local environmental protection. Even though Chen Fei is supported by the central government, all posts supporting him on the local Web site are censored, likely because of his record of organizing collective action. In the mid-2000s, Chen founded an environmental NGO (色环保志愿者协会) with more than 400 registered members who created China’s first “no-plastic-bag village,” which eventually led to legislation on use of plastic bags. Another example is a heavily censored group of posts expressing collective anger about lead poisoning in Jiangxi Province’s Suyang County from battery factories. These posts talk about children sickened by pollution from lead acid battery factories in Zhejiang province belonging to the Tianneng Group (天能集团), and report that hospitals refused to release results of lead tests to patients. In January 2011, villagers from Suyang gathered at the factory to demand answers. Such collective organization is not tolerated by the censors, regardless of whether it supports the government or criticizes it.

In all events categorized as having collective action potential, censorship within the event is more frequent than censorship outside the event. In addition, these events are, on average, considerably more censored than other types of events. These facts are consistent

<sup>6</sup> The baseline (the percent censorship outside of volume bursts) is typically very small, 3-5% and varies relatively little across topic areas.

<sup>7</sup> As in the two relevant events in Figure 4, pornography often appears in social media in association with the discussion of some other popular news or discussion, to attract viewers.

**Figure 4. Events with Highest and Lowest Censorship Magnitude**

with our theory that the censors are intentionally searching for and taking down posts related to events with collective action potential. However, we add to these tests one based on an examination of what might lead to different levels of censorship among events within this category: Although we have developed a quantitative measure, some of the events in this category clearly have more collective action potential than others. By studying the specific events, it is easy to see that events with the lowest levels of censorship magnitude generally have less collective action potential than the very highly censored cases, as consistent with our theory.

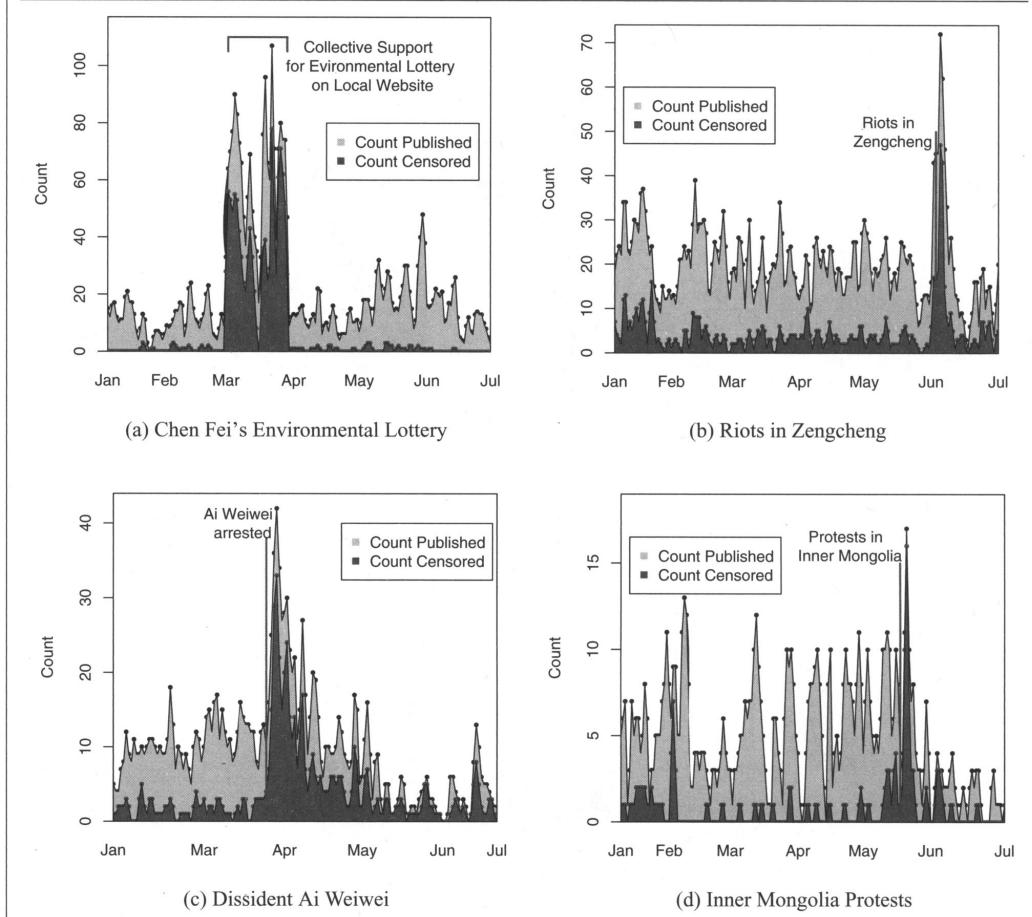
To see this, consider the few events classified as collective action potential with the lowest levels of censorship magnitude. These include a volume burst associated with protests about ethnic stereotypes in the animated children's movie *Kungfu Panda 2*, which was properly classified as a collective action event, but its potential for future protests is obviously highly limited. Another example is Qian Yunhui, a village leader in Zhejiang, who led villagers to petition local governments for compensation for land seized and was then (supposedly accidentally) crushed to death by a truck. These two events involving Qian had high collective action potential, but both were before our observation period. In our period, there was an event that led to a volume burst around the much narrower and far less incendiary issue of how much money his family was given as a reparation payment for his death.

Finally, we give some more detailed information of a few examples of three types of events, each based

on a random sample of posts in one topic area. First, Figure 5 gives four time series plots that initially involve low levels of censorship, followed by a volume spike during which we witness very high levels of censorship. Censorship in these examples are high in terms of the absolute number of censored posts and the percent of posts that are censored. The pattern in all four graphs (and others we do not show) is evident: the Chinese authorities disproportionately focus considerable censorship efforts during volume bursts.

We also went further and analyzed (by hand and via computer-assisted methods described in Grimmer and King 2011) the smaller number of uncensored posts during volume bursts associated with events that have collective action potential, such as in panel (a) of Figure 5 where the red area does not entirely cover the gray during the volume burst. In this event, and the vast majority of cases like this one, uncensored posts are not about the event, but just happen to have the keywords we used to identify the topic area. Again we find that the censors are highly accurate and aimed at increasing censorship magnitude. Automated methods of individual classification are not capable of this high a level of accuracy.

Second, we offer four time-series plots of random samples of posts in Figure 6 which illustrate topic areas with one or more volume bursts but without censorship. These cover important, controversial, and potentially incendiary topics—including policies involving the one child policy, education policy, and corruption, as well as news about power prices—but none of the volume bursts where associated with any localized

**Figure 5. High Censorship During Collective Action Events (in 2011)**

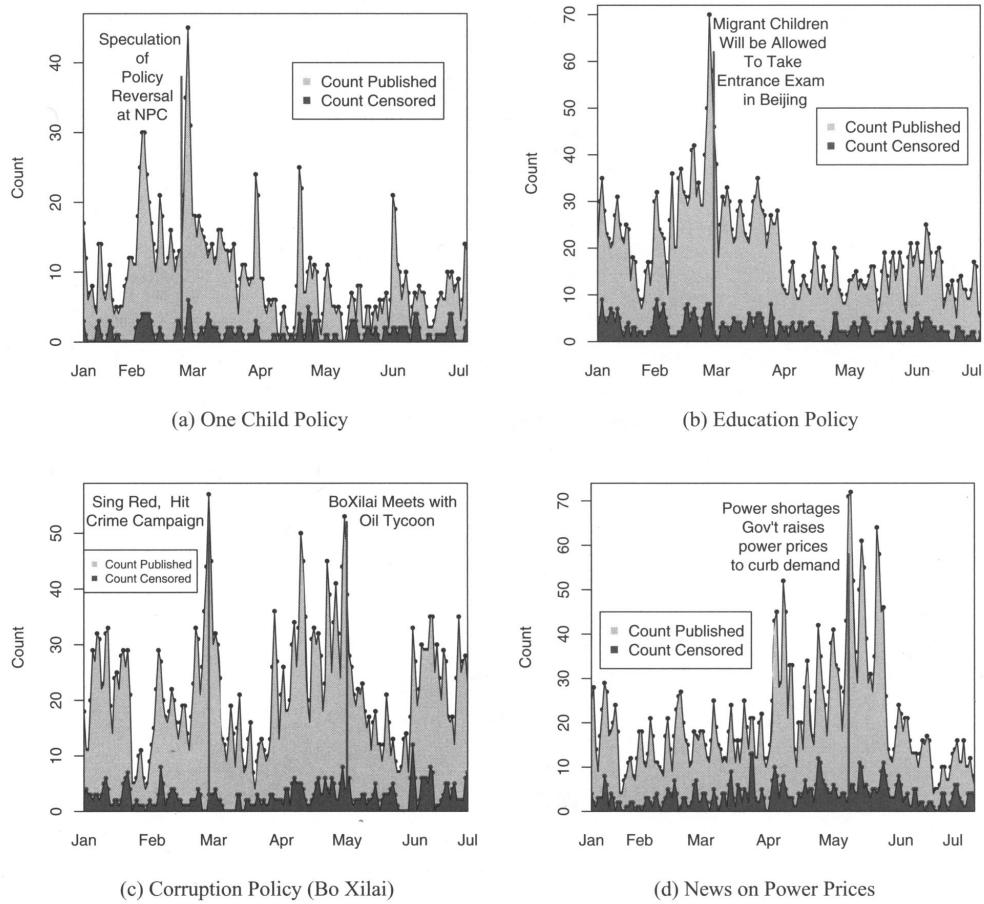
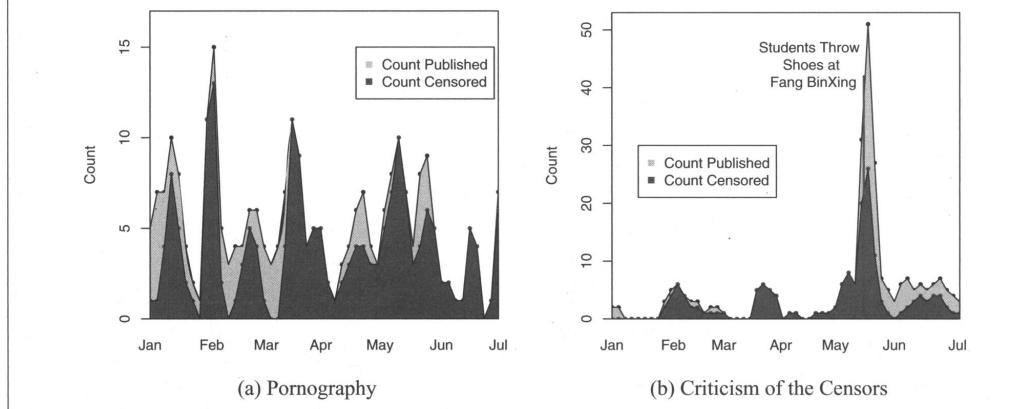
collective expression, and so censorship remains consistently low.

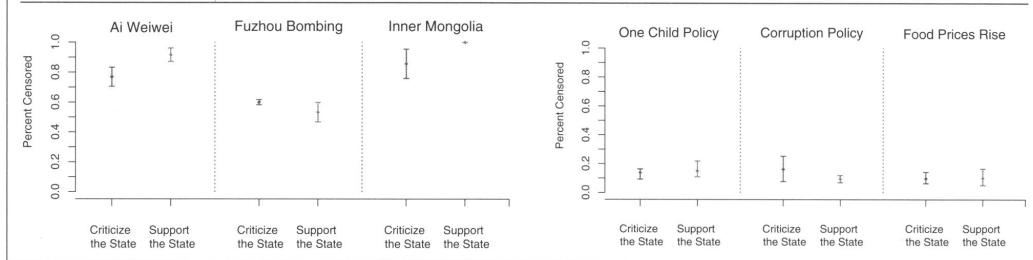
Finally, we found that almost all of the topic areas exhibit censorship patterns portrayed by Figures 5 and 6. The two with divergent patterns can be seen in Figure 7. These topics involve analyses of random samples of posts in the areas of pornography (panel (a)) and criticism of the censors (panel (b)). What is distinctive about these topics compared to the remaining we studied is that censorship levels remain high consistently in the entire six-month period and, consequently, do not increase further during volume bursts. Similar to American politicians who talk about pornography as undercutting the "moral fiber" of the country, Chinese leaders describe it as violating public morality and damaging the health of young people, as well as promoting disorder and chaos; regardless, censorship in one form or another is often the consequence.

More striking is an oddly "inappropriate" behavior of the censors: They offer freedom to the Chinese people to criticize every political leader except for themselves, every policy except the one they implement, and every program except the one they run. Even within the strained logic the Chinese state uses to justify censorship, Figure 7 (panel (b))—which reveals consistently high levels of censored posts that involve criticisms of the censors—is remarkable.

#### Content of Censored and Uncensored Posts

Our final test involves comparing the content of censored and uncensored posts. State critique theory predicts that posts critical of the state are those censored, regardless of their collective action potential. In contrast, the theory of collective action potential predicts that posts related to collective action events will be censored regardless of whether they criticize or praise

**Figure 6. Low Censorship on News and Policy Events (in 2011)****Figure 7. Two Topics with Continuous High Censorship Levels (in 2011)**

**Figure 8. Content of Censored Posts by Topic Area**

the state, with both critical and supportive posts uncensored when events have no collective action potential.

To conduct this test in a very large number of posts, we need a method of automated text analysis that can accurately estimate the percentage of posts in each category of any given categorization scheme. We thus adapt to the Chinese language the methodology introduced in the English language by Hopkins and King (2010). This method does not require (inevitably error prone) machine translation, individual classification algorithms, or identification of a list of keywords associated with each category; instead, it requires a small number of posts read and categorized in the original Chinese. We conducted a series of rigorous validation tests and obtain highly accurate results—as accurate as if it were possible to read and code all the posts by hand, which of course is not feasible. We describe these methods, and give a sample of our validation tests, in Appendix B.

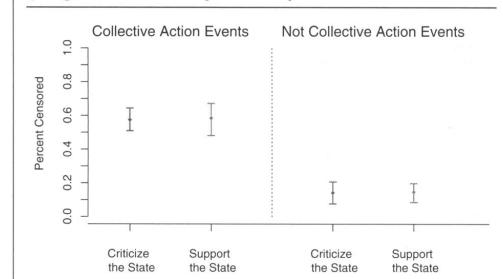
For our analyses, we use categories of posts that are (1) critical of the state, (2) supportive of the state, or (3) irrelevant or factual reports about the events. However, we are not interested in the percent of posts in each of these categories, which would be the usual output of the Hopkins and King procedure. We are also not interested in the percent of posts in each category among those posts which were censored and among those which were not censored, which would result from running the Hopkins-King procedure once on each set of data. Instead, we need to estimate and compare the percent of posts censored in each of the three categories. We thus develop a Bayesian procedure (described in Appendix B) to extend the Hopkins-King methodology to estimate our quantities of interest.

We first analyze specific events and then turn to a broader analysis of a random sample of posts from all of our events. For collective action events we choose those which unambiguously fit our definition—the arrest of the dissident Ai Weiwei, protests in Inner Mongolia, and bombings in reaction to the state's demolition of housing in Fuzhou city. Panel (a) of Figure 8 reports the percent of posts that are censored for each event, among those that criticize the state (right/red) and those which support the state (left/green); vertical bars are 95% confidence intervals. As is clear, regardless of

whether the posts support or criticize the state, they are all censored at a high level, about 80% on average. Despite the conventional wisdom that the censorship program is designed to prune the Internet of posts critical of the state, a hypothesis test indicates that the percent censorship for posts that criticize the state is not larger than the percent censorship of posts that support the state, for each event. This clearly shows support for the collective action potential theory and against the state critique theory of censorship.

We also conduct a parallel analysis for three topics, taken from the analysis in Figure 6, that cover highly visible and apparently sensitive policies associated with events that had no collective action potential—one child policy, corruption policy, and news of increasing food prices. In this situation, we again get the empirical result that is consistent with our theory, in both analyses: Categories critical and supportive of the state both fall at about the same, low level of censorship, about 10% on average.

To validate that these results hold across all events, we randomly draw posts from all volume bursts with and without collective action potential. Figure 9 presents the results in parallel to those in Figure 8. Here, we see that categories critical and supportive of the state again fall at the same, high level of censorship for collective action potential events, while categories

**Figure 9. Content of All Censored Posts (Regardless of Topic Area)**

critical and supportive of the state fall at the same, low level of censorship for news and policy events. Again, there is no significant difference between the percent censored among those which criticize and support the state, but a large and significant difference between the percent censored among collective action potential and noncollective action potential events.

The results are unambiguous: posts are censored if they are in a topic area with collective action potential and not otherwise. Whether or not the posts are in favor

*continue in next column*

“This is a city government [Yulin City, Shaanxi] that treats life with contempt, this is government officials run amuck, a city government without justice, a city government that delights in that which is vulgar, a place where officials all have mistresses, a city government that is shameless with greed, a government that trades dignity for power, a government without humanity, a government that has no limits on immorality, a government that goes back on its word, a government that treats kindness with ingratitude, a government that cares nothing for posterity....”

Another blogger wrote a scathing critique of China’s One Child Policy, also without being censored:

“The [government] could promote voluntary birth control, not coercive birth control that deprives people of descendants. People have already been made to suffer for 30 years. This cannot become path dependent, prolonging an ill-devised temporary, emergency measure.... Without any exaggeration, the one child policy is the brutal policy that farmers hated the most. This “necessary evil” is rare in human history, attracting widespread condemnation around the world. It is not something we should be proud of.”

Finally in a blog castigating the CCP (Chinese Communist Party) for its broken promise of democratic, constitutional government while, with reference to the Tiananmen square incident, another wrote, the following state critique, again without being censored:

“I have always thought China’s modern history to be full of progress and revolution. At the end of the Qing, advances were seen in all areas, but after the Wuchang uprising, everything was lost. The Chinese Communist Party made a promise of democratic, constitutional government at the beginning of the war of resistance against Japan. But after 60 years that promise is yet to be honored. China today lacks integrity, and accountability should be traced to Mao. In the 1980s, Deng introduced structural political reforms, but after Tiananmen, all plans were permanently put on hold...intra-party democracy espoused today is just an excuse to perpetuate one party rule.”

*continued*

of the government, its leaders, and its policies has no measurable effect on the probability of censorship.

Finally, we conclude this section with some examples of posts to give some of the flavor of exactly what is going on in Chinese social media. First we offer two examples, not associated with collective action potential events, of posts not censored even though they are unambiguously against the state and its leaders. For example, consider this highly personal attack, naming the relevant locality:

这是一个漠视生命的市政府[陕西省榆林市]——一个官员横行的市政府、一个没有公正的市政府，一个低级趣味的市政府，一个包二奶的市政府，一个为钱不要脸的市政府，一个为个权不要人格的市政府，一个没有血性的市政府，一个没有道德底线的市政府，一个出尔反尔的市政府，一个忘恩负义的市政府，一个不要子孙后代的市政府，一个什么怪事都出的市政府，一个什么的市政府，只要你想到的就有...

可以提倡人民自愿节育,但让人断子绝孙的强制节育,搞30年已是忍辱负重,不能形成路径依赖,将不得已的临时性恶政无限延长...可以毫不夸张地讲,计划生育是农民最痛苦的暴政。虽说是“必要的恶”,却是世界少有,遭到世界舆论的广泛谴责,实在不该以此为豪。

我一直将中国的近代史视为一场改良与革命的赛跑,在清末的大赛场上,最终革命跑到了头,改良的一切设计,在武昌起义枪声响起后成了废纸。中共的民主宪政承诺,是抗战结束前开出的远期支票,超过了一个甲子仍未兑现。当今中国社会缺乏诚信,要从毛泽东开始问责。邓小平在80年代提出的政治体制改革,在“8964”事件后被长期搁置...近年所谓“党主立宪”之说,也是主流学者为维系一党执政地位所做的政治设计。

These posts are neither exceptions nor unusual: We have thousands more. Negative posts, including those about “sensitive” topics such as Tiananmen square or reform of China’s one-party system, do not accidentally slip through a leaky or imperfect system. The evidence indicates that the censors have no intention of stopping them. Instead, they are focused on removing posts that have collective action potential, regardless of whether or not they cast the Chinese leadership and their policies in a favorable light.

*continue in next column*

“The bombing led not only to the tragedy of his death but the death of many government workers. Even if we can verify what Qian Mingqi said on Weibo that the building demolition caused a great deal of personal damage, we should still condemn his extreme act of retribution.... The government has continually put forth measures and laws to protect the interests of citizens in building demolition. And the media has called attention to the plight of those experiencing housing demolition. The rate at which compensation for housing demolition has increased exceeds inflation. In many places, this compensation can change the fate of an entire family.”

Another example is the following censored post supporting the state. It accuses the local leader Ran Jianxin, whose death in police custody triggered protests in Lichuan, of corruption:

“According to news from the Badong county propaganda department web site, when Ran Jianxin was party secretary in Lichuan, he exploited his position for personal gain in land requisition, building demolition, capital construction projects, etc. He accepted bribes, and is suspected of other criminal acts.”

## CONCLUDING REMARKS

The new data and methods we offer seem to reveal highly detailed information about the interests of the Chinese people, the Chinese censorship program, and the Chinese government over time and within different issue areas. These results also shed light on an enormously secretive government program designed to suppress information, as well as on the interests, intentions, and goals of the Chinese leadership.

The evidence suggests that when the leadership allowed social media to flourish in the country, they also allowed the full range of expression of negative and positive comments about the state, its policies, and its leaders. As a result, government policies sometimes look as bad, and leaders can be as embarrassed, as is often the case with elected politicians in democratic countries, but, as they seem to recognize, looking bad does not threaten their hold on power so long as they manage to eliminate discussions associated with events that have collective action potential—where a locus of

*continued*

To emphasize this point, we now highlight the obverse condition by giving examples of two posts related to events with collective action potential that support the state but which nevertheless were quickly censored. During the bombings in Fuzhou, the government censored this post, which unambiguously condemns the actions of Qian Mingqi, the bomber, and explicitly praises the government’s work on the issues of housing demolition, which precipitated the bombings:

爆炸案造成他本人和多名政府工作人员死亡的悲剧，即使钱明奇在微博里所称拆迁造成的个人损失是属实的，我们也应谴责他的极端报复行为……政府在连续出台保护被拆迁者利益的政策法规，媒体也在为公平对待被拆迁者大声疾呼，各地拆迁补偿款上升速度，大多高于商品房售价上升速度，在不少地方，补偿款已经足以改变一个家庭的命运。

湖北省巴东县委宣传部都在其官方网站发布新闻通稿称，再建新在担任利川市都亭办事处常委书记、主任期间，利用职务之便，在征地拆迁、工程发包等事项中为他人谋取利益，收受他人贿赂，涉嫌受贿犯罪。

power and control, other than the government, influences the behaviors of masses of Chinese people. With respect to this type of speech, the Chinese people are individually free but collectively in chains.

Much research could be conducted on the implications of this governmental strategy; as a spur to this research, we offer some initial speculations here. For one, so long as collective action is prevented, social media can be an excellent way to obtain effective measures of the views of the populace about specific public policies and experiences with the many parts of Chinese government and the performance of public officials. As such, this “loosening” up on the constraints on public expression may, at the same time, be an effective governmental tool in learning how to satisfy, and ultimately mollify, the masses. From this perspective, the surprising empirical patterns we discover may well be a theoretically optimal strategy for a regime to use social media to maintain a hold on power. For example, Dimitrov (2008) argues that regimes collapse when its people stop bringing grievances to the state, since it is an indicator that the state is no longer regarded

as legitimate. Similarly, Egorov, Guriev, and Sonin (2009) argue that dictators with low natural resource endowments allow freer media in order to improve bureaucratic performance. By extension, this suggests that allowing criticism, as we found the Chinese leadership does, may legitimize the state and help the regime maintain power. Indeed, Lorentzen (2012) develops a formal model in which an authoritarian regimes balance media openness with regime censorship in order to minimize local corruption while maintaining regime stability. Perhaps the formal theory community will find ways of improving their theories after conditioning on our empirical results.

More generally, beyond the findings of this article, the data collected represent a new way to study China and different dimensions of Chinese politics, as well as facets of comparative politics more broadly. For the study of China, our approach sheds light on authoritarian resilience, center-local relations, subnational politics, international relations, and Chinese foreign policy. By examining what events are censored at the national level versus a subnational level, our approach indicates some areas where local governments can act autonomously. Additionally, by clearly revealing government intent, our approach allows an examination of the differences between the priorities of various subnational units of government. Because we can analyze social media and censorship in the context of real-world events, this approach is able to reveal insights into China's international relations and foreign policy. For example, do displays of nationalism constrain the government's foreign policy options and activities? Finally, China's censorship apparatus can be thought of as one of the input institutions. Nathan (2003) identifies as an important source of authoritarian resilience, and the effectiveness and capabilities of the censorship apparatus may shed light on the CCP's regime institutionalization and longevity.

In the context of comparative politics, our work could directly reveal information about state capacity as well as shed light on the durability of authoritarian regimes and regime change. Recent work on the role of Internet and social media in the Arab spring (Ada et al. 2012; Bellin 2012) debate the exact role played by these technologies in organizing collective action and motivating regional diffusion, but consistently highlight the relevance of these technological innovations on the longevity of authoritarian regimes worldwide. Edmond (2012) models how the increase in information sources (e.g., Internet, social media) will be bad for a regime unless the regime has economies of scale in controlling information sources. While Internet and social media in general have smaller economies of scale, because of how China devolves the bulk of censorship responsibility to Internet content providers, the regime maintains large economies of scale in the face of new technologies. China, as a relatively rich and resilient authoritarian regime, with a sophisticated and effective censorship apparatus, is probably being watched closely by autocrats from around the world.

Beyond learning the broad aims of the Chinese censorship program, we seem to have unearthed a valuable source of continuous time information on the interests of the Chinese people and the intentions and goals of the Chinese government. Although we illustrated this with time series in 85 different topic areas, the effort could be expanded to many other areas chosen *ex ante* or even discovered as online communities form around new subjects over time. The censorship behavior we observe may be predictive of future actions outside the Internet (see Appendix C), is informative even when the traditional media is silent, and would likely serve a variety of other scholarly and practical uses in government policy and business relations.

Along the way, we also developed methods of computer-assisted text analysis that we demonstrate work well in the Chinese language and adapted it to this application. These methods would seem to be of use far beyond our specific application. We also conjecture that our data collection procedures, text analysis methods, engineering infrastructure, theories, and overall analytic and empirical strategies might be applicable in other parts of the world that suppress freedom of the press.

## APPENDIX A: TOPIC AREAS

Our stratified sampling design includes the following 85 topic areas chosen from three levels of hypothesized political sensitivity described in the section on data above. Although we allow overlap across topic areas, empirically we find almost none.

**High:** Ai Weiwei, Chen Guangcheng, Fang Binxing, Google and China, Jon Hunstman, Labor strike and Honda, Li Chengpeng, Lichuan protests over the death of Rao Jianxin, Liu Xiaobo, Mass incidents, Mergen, Pornographic Web sites, Princelings faction, Qian Mingqi, Qian Yunhui, Syria, Taiwan weapons, Unrest in Inner Mongolia, Uyghur protest, Wu Bangguo, Zengcheng protests

**Medium:** AIDS, Angry Youth, Appreciation and devaluation of CNY against the dollar, Bo Xilai, China's environmental protection agency, Death penalty, Drought in central-southern provinces, Environment and pollution, Fifty Cent Party, Food prices, Food safety, Google and hacking, Henry Kissinger, HIV, Huang Yibo, Immigration policy, Inflation, Japanese earthquake, Kim Jong Il, Kungfu Panda 2, Lawsuit against Baidu for copyright infringement, Lead Acid Batteries and pollution, Libya, Micro-blogs, National Development and Reform Commission, Nuclear Power and China, Nuclear weapons in Iran, Official corruption, One child policy, Osama Bin Laden, Pakistan Weapons, People's Liberation Army, Power prices, Property tax, Rare Earth metals, Second rich generation, Solar power, State Internet Information Office, Su Zizi, Three Gorges Dam, Tibet, U.S. policy of quantitative easing, Vietnam and South China Sea, Wen Jiabao and legal reform, Xi Jinping, Yao Jiaxin

**Low:** Chinese investment in Africa, Chinese versions of Groupon, Da Ren Xiu on Dragon TV (Chinese American Idol), DouPo CangQiong (serialized Internet novel), Education reform, Health care reform, Indoor smoking ban, Let the Bullets Fly (movie), Li Na (Chinese tennis star), MenRen XinJi (TV drama), New Disney theme park in Shanghai, Peking opera, Sai Er Hao (online game), Social security insurance, Space shuttle Endeavor, Traffic in Beijing, World Cup, Zimbabwe

## APPENDIX B: AUTOMATED CHINESE TEXT ANALYSIS

We begin with methods of automated text analysis developed in Hopkins and King (2010) and now widely used in academia and private industry. This approach enables one to define a set of mutually exclusive and exhaustive categories, to then code a small number of example posts within each category (known as the labeled “training set”), and to infer the proportion of posts within each category in a potentially much larger “test set” without hand coding their category labels. The methodology is colloquially known as “ReadMe,” which is the name of open source software program that implements it.

We adapt and extend this method for our purposes in four steps. First, we translate different binary representations of Chinese text to the same unicode representation. Second, we eliminate punctuation and drop characters that do not appear in fewer than 1% or more than 99% of our posts. Since words in Chinese are composed of one to five characters, but without any spacing or punctuation to demarcate them, we experimented with methods of automatically “chunking” the characters into estimates of words; however, we found that ReadMe was highly accurate without this complication.

And finally, whereas ReadMe returns the proportion of posts in each category, our quantity of interest here is the proportion of posts which are censored in each category. We therefore run ReadMe twice, once for the set of censored posts (which we denote  $C$ ) and once for the set of uncensored posts (which we denote  $U$ ). For any one of the mutually exclusive categories, which we denote  $A$ , we calculate the proportion censored,  $P(C|A)$  via an application of Bayes theorem:

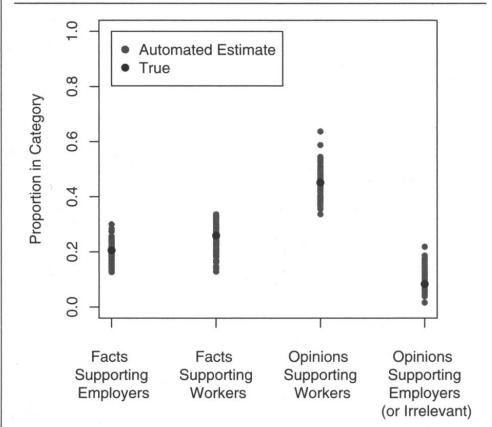
$$P(C|A) = \frac{P(A|C)P(C)}{P(A)} = \frac{P(A|C)P(C)}{P(A|C)P(C) + P(A|U)P(U)}.$$

Quantities  $P(A|C)$ ,  $P(A|U)$  are estimated by ReadMe whereas  $P(C)$  and  $P(U)$  are the observed proportions of censored and uncensored posts in the data. Therefore, we can back out  $P(C|A)$ . We produce confidence intervals for  $P(C|A)$  by simulation: we merely plug in simulations for each of the right side components from their respective posterior distributions.

This procedure requires no translation, machine or otherwise. It does not require methods of individual classification, which are not sufficiently accurate for estimating category proportions. The methodology is considered a “computer-assisted” approach because it amplifies the human intelligence used to create the training set rather than the highly error-prone process of requiring humans to assist the computer in deciding which words lead to which meaning.

Finally, we validate this procedure with many analyses like the following, each in a different subset of our data. First, we train native Chinese speakers to code Chinese language blog posts into a given set of categories. For this illustration, we use 1,000 posts about the labor strikes in 2010, and set aside 100 as the training set. The remaining 900 constituted the test set. The categories were (a) facts supporting employers, (b) facts supporting workers, (c) opinions supporting workers, and (d) opinions supporting employers (or irrelevant). The true proportion of posts censored (given vertically) in each of four categories (given horizontally) in the test set is indicated by four black dots in Figure 10. Using the text and categories from the training set and only the text from the test set, we estimate these proportions using our procedure above. The confidence intervals, represented as simulations from

**Figure 10. Validation of Automated Text Analysis**



the posterior distribution, are given in a set of red dots for each of the categories, in the same figure. Clearly the results are highly accurate, covering the black dot in all four cases.

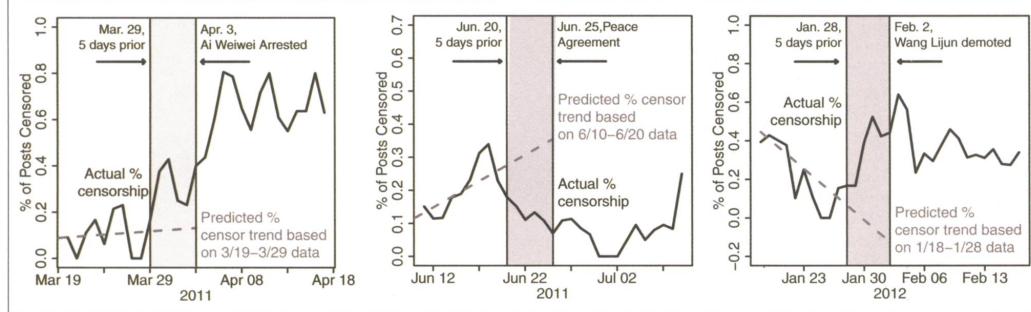
## APPENDIX C: THE PREDICTIVE CONTENT OF CENSORSHIP BEHAVIOR

If censorship is a measure of government intentions and desires, then it may offer some hints about future state action unavailable through other means. We test this hypothesis here. However, most actions of the Chinese state are easily predictable comments on or responses to exogenous events. The difficult cases are those which are not otherwise predictable; among those hard cases, we focus on the ones associated with events with collective action potential.

We did not design this study or our data collection for predictive purposes, but we can still use it as an indirect test of our hypothesis. We do this via well-known and widely used case-control methodology (King and Zeng 2001). First, we take all real world events with collective action potential and remove those easy to predict as responses to exogenous events. This left two events, neither of which could have been predicted with information in the traditional news media: the 4/3/11 arrest of Ai Weiwei and the 6/25/11 peace agreement with Vietnam regarding disputes in the South China Sea. We analyze these two cases here and provide evidence that we may have been able to predict them from censorship rates. In addition, as we were finalizing this article in early 2012, the Bo Xilai incident shook China—an event widely viewed as “the biggest scandal to rock China’s political class for decades” (Branigan 2012) and one which “will continue to haunt the next generation of Chinese leaders” (Economy 2012)—and we happened to still have our monitors running. This meant that we could use this third surprise event as another test of our hypothesis.

Next, we choose how long in advance censorship behavior could be used to predict these (otherwise surprise) events. The time interval must be long enough so that the censors can do their job and so we can detect systematic changes in the percent censored, but not so long as to make the prediction impossible. We choose five days as fitting these constraints, the exact value of which is of course arbitrary

**Figure 11. Censorship and Prediction (a) Ai Weiwei's Arrest (b) South China Sea Peace Agreement (c) Wang Lijun Demotion (Bo Xilai affair)**



but in our data not crucial. Thus we hypothesize that the Chinese leadership took an (otherwise unobserved) decision to act approximately five days in advance and prepared for it by making censorship patterns different from what they would have been otherwise.

In panel (a) of Figure 11, we apply the procedure to the surprise arrest of Ai Weiwei. The vertical axis in this time series plot is the percent of posts censored. The gray area is our five-day prediction interval between the unobserved hypothesized decision to arrest Ai Weiwei and the actual arrest. Nothing in the news media we have been able to find suggested that an arrest was imminent. The solid (blue) line is actual censorship levels and the dashed (red) line is a simple linear prediction based only on data greater than five days earlier than the arrest; extrapolating it linearly five days forward gives an estimate of what would have happened without this hypothesized decision. Then the vertical difference between the dashed (red) and solid (blue) lines on April 3rd is our causal estimate; in this case, the predicted level, if no decision had been made, is at about baseline levels at approximately 10%; in contrast, the actual levels of censorship is more than twice as high. To confirm that this result was not due to chance, we conducted a permutation test, using all other five-day intervals in the data as placebo tests, and found that the effect in the graph is larger than all the placebo tests.

We repeat the procedure for the South China Sea peace agreement in panel (b) of Figure 11. The discovery of oil in the South China Sea led to an ongoing conflict between Beijing and Hanoi, during which rates of censorship soared. According to the media, conflict continued right up until the surprise peace agreement was announced on June 25. Nothing in the media before that date hinted at a resolution of the conflict. However, rates of censorship unexpectedly plummeted well before that date, clearly presaging the agreement. We also conducted a permutation test here and again found that the effect in the graph is larger than all the placebo tests.

Finally, we turn to the Bo Xilai incident. Bo, the son one of the eight elders of the CCP, was thought to be a front runner for promotion to the Politburo Standing Committee in CPC 18th National Congress in Fall of 2012. However, his political rise met an abrupt end following his top lieutenant, Wang Lijun, seeking asylum at the American consulate in Chengdu on February 6, 2012, four days after Wang was demoted by Bo. After Wang revealed Bo's alleged involvement in homicide of a British national, Bo was removed as Chongqing party chief and suspended from the Politburo. Because of the extraordinary nature of this event in re-

vealing the behaviors and disagreements among the CCP's top leadership, we conducted a special analysis of the otherwise unpredictable event that precipitated this scandal—the demotion of Wang Lijun by Bo Xilai on February 2, 2012. It is thought that Bo demoted Wang when Wang confronted Bo with evidence of his involved in the death of Neil Heywood.

We thus apply the same methodology to the demotion of Wang Lijun in panel (c) of Figure 11, and again see a large difference in actual and predicted percent censorship before Wang's demotion. Prior to Wang's dismissal, nothing in the media hinted at the demotion that would lead to the spectacular downfall of one of China's rising leaders. And for the third of three cases, a permutation test reveals that the effect in the five days prior to Wang's demotion is larger than all the placebo tests.

The results in all three cases confirm our theory, but we conducted this analysis retrospectively, and with only three events, and so further research to validate the ability of censorship to predict events in real time prospectively would certainly be valuable.

## REFERENCES

- Ada, Sean, Henry Farrell, Marc Lynch, John Sides, and Deen Freelon. 2012. "Blogs and Bullets: New Media and Conflict after the Arab Spring." <http://j.mp/WviJPk>.
- Ash, Timothy Garton. 2002. *The Polish Revolution: Solidarity*. New Haven: Yale University Press.
- Bamman, D., B. O'Connor, and N. Smith. 2012. "Censorship and Deletion Practices in Chinese Social Media." *First Monday* 17: 3–5.
- Bellin, Eva. 2012. "Reconsidering the Robustness of Authoritarianism in the Middle East: Lessons from the Arab Spring." *Comparative Politics* 44 (2): 127–49.
- Blecher, Marc. 2002. "Hegemony and Workers' Politics in China." *The China Quarterly* 170: 283–303.
- Branigan, Tania. 2012. "Chinese politician Bo Xilai's wife suspected of murdering Neil Heywood." *The Guardian* April 10. <http://j.mp/K189ce>.
- Cai, Yongshun. 2002. "Resistance of Chinese Laid-off Workers in the Reform Period." *The China Quarterly* 170: 327–44.
- Chang, Parris. 1983. *Elite Conflict in the Post-Mao China*. New York: Occasional Papers Reprints.
- Charles, David. 1966. The Dismissal of Marshal Peng Teh-huai. In *China Under Mao: Politics Takes Command*, ed. Roderick Mac Farquhar. Cambridge: MIT University Press, 20–33.
- Chen, Feng. 2000. "Subsistence Crises, Managerial Corruption and Labour Protests in China." *The China Journal* 44: 41–63.

- Chen, Xi. 2012. *Social Protest and Contentious Authoritarianism in China*. New York: Cambridge University Press.
- Chen, Xiaoyan, and Peng Hwa Ang. 2011. Internet Police in China: Regulation, Scope and Myths. In *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, eds. David Herold and Peter Marolt. New York: Routledge, 40–52.
- Crosas, Merce, Justin Grimmer, Gary King, Brandon Stewart, and the Consilience Development Team. 2012. "Consilience: Software for Understanding Large Volumes of Unstructured Text."
- Dimitrov, Martin. 2008. "The Resilient Authoritarians." *Current History* 107 (705): 24–9.
- Duan, Qing. 2007. *China's IT Leadership*. Vdm Verlag Saarbrücken, Germany.
- Economy, Elizabeth. 2012. "The Bigger Issues Behind China's Bo Xilai Scandal." *The Atlantic* April 11. <http://j.mp/JQBBv>.
- Edin, Maria. 2003. "State Capacity and Local agent Control in China: CPP Cadre Management from a Township Perspective." *China Quarterly* 173 (March): 35–52.
- Edmond, Chris. 2012. "Information, Manipulation, Coordination, and Regime Change." <http://j.mp/WviWlz>.
- Egorov, Georgy, Sergei Guriev, and Konstantin Sonin. 2009. "Why Resource-poor Dictators Allow Freer Media: A Theory and Evidence from Panel Data." *American Political Science Review* 103 (4): 645–68.
- Esarey, Ashley, and Qiang Xiao. 2008. "Political Expression in the Chinese Blogosphere: Below the Radar." *Asian Survey* 48 (5): 752–72.
- Esarey, Ashley, and Qiang Xiao. 2011. "Digital Communication and Political Change in China." *International Journal of Communication* 5: 298–C319.
- Freedom House. 2012. "Freedom of the Press, 2012." [www.freedomhouse.org](http://www.freedomhouse.org).
- Grimmer, Justin, and Gary King. 2011. "General purpose computer-assisted clustering and conceptualization." *Proceedings of the National Academy of Sciences* 108 (7): 2643–50. <http://gking.harvard.edu/files/abs/discov-abs.shtml>.
- Guo, Gang. 2009. "China's Local Political Budget Cycles." *American Journal of Political Science* 53 (3): 621–32.
- Herold, David. 2011. Human Flesh Search Engine: Carnivalesque Riots as Components of a 'Chinese Democracy.' In *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, eds. David Herold and Peter Marolt. New York: Routledge, 127–45.
- Hinton, Harold. 1955. *The "Unprincipled Dispute" Within Chinese Communist Top Leadership*. Washington, DC: U.S. Information Agency.
- Hopkins, Daniel, and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 229–47.
- Huber, Peter J. 1964. "Robust Estimation of a Location Parameter." *Annals of Mathematical Statistics* 35: 73–101.
- King, Gary, and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9 (2, Spring): 137–63. <http://gking.harvard.edu/files/abs/0s-abs.shtml>.
- Kung, James, and Shuo Chen. 2011. "The Tragedy of the Nomenklatura: Career Incentives and Political Radicalism during China's Great Leap Famine." *American Political Science Review* 105: 27–45.
- Kuran, Timur. 1989. "Sparks and Prairie Fires: A Theory of Unanticipated Political Revolution." *Public Choice* 61(1): 41–74.
- Lee, Ching-Kwan. 2007. *Against the Law: Labor Protests in China's Rustbelt and Sunbelt*. Berkeley, CA: University of California Press.
- Lindtner, Silvia, and Marcella Szablewicz. 2011. China's Many Internets: Participation and Digital Game Play Across a Changing Technology Landscape. In *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, eds. David Herold and Peter Marolt. New York: Routledge, 89–105.
- Lohmann, Susanne. 1994. "The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989–1991." *World Politics* 47 (1): 42–101.
- Lohmann, Susanne. 2002. "Collective Action Cascades: An Informational Rationale for the Power in Numbers." *Journal of Economic Surveys* 14 (5): 654–84.
- Lorentzen, Peter. 2010. "Regularizing Rioting: Permitting Protest in an Authoritarian Regime." Working Paper.
- Lorentzen, Peter. 2012. "Strategic Censorship." SSRN. <http://j.mp/Wvj3xx>.
- MacFarquhar, Roderick. 1974. *The Origins of the Cultural Revolution Volume 1: Contradictions Among the People 1956–1957*. New York: Columbia University Press.
- MacFarquhar, Roderick. 1983. *The Origins of the Cultural Revolution Volume 2: The Great Leap Forward 1958–1960*. New York: Columbia University Press.
- MacKinnon, Rebecca. 2012. *Consent of the Networked: The Worldwide Struggle For Internet Freedom*. New York: Basic Books.
- Marolt, Peter. 2011. Grassroots Agency in a Civil Sphere? Rethinking Internet Control in China. In *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, eds. David Herold and Peter Marolt. New York: Routledge, 53–68.
- Nathan, Andrew. 2003. "Authoritarian Resilience." *Journal of Democracy* 14 (1): 6–17.
- O'Brien, Kevin, and Lianjiang Li. 2006. *Rightful Resistance in Rural China*. New York: Cambridge University Press.
- Perry, Elizabeth. 2002. *Challenging the Mandate of Heaven: Social Protest and State Power in China*. Armonk, NY: M. E. Sharpe.
- Perry, Elizabeth. 2008. Permanent Revolution? Continuities and Discontinuities in Chinese Protest. In *Popular Protest in China*, ed. Kevin O'Brien. Cambridge, MA: Harvard University Press, 205–16.
- Perry, Elizabeth. 2010. Popular Protest: Playing by the Rules. In *China Today, China Tomorrow: Domestic Politics, Economy, and Society*, ed. Joseph Fewsmith. Plymouth, UK: Rowman and Littlefield, 11–28.
- Przeworski, Adam, Michael E. Alvarez, Jose Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Well-being in the World, 1950–1990*. New York: Cambridge University Press.
- Ratkiewicz, J., F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. 2010. Traffic in Social Media II: Modeling Bursty Popularity. In *Social Computing, 2010 IEEE Second International Conference*. Minneapolis, MN IEEE, 393–400.
- Reilly, James. 2012. *Strong Society, Smart State: The Rise of Public Opinion in China's Japan Policy*. New York: Columbia University Press.
- Rousseeuw, Peter J., and Annick Leroy. 1987. *Robust Regression and Outlier Detection*. New York: Wiley.
- Schurmann, Franz. 1966. *Ideology and Organization in Communist China*. Berkeley, CA: University of California Press.
- Shih, Victor. 2008. *Factions and Finance in China: Elite Conflict and Inflation*. Cambridge: Cambridge University Press.
- Shirk, Susan. 2007. *China: Fragile Superpower: How China's Internal Politics Could Derail Its Peaceful Rise*. New York: Oxford University Press.
- Shirk, Susan L. 2011. *Changing Media, Changing China*. New York: Oxford University Press.
- Teiwes, Frederick. 1979. *Politics and Purges in China: Retification and the Decline of Party Norms*. Armonk, NY: M. E. Sharpe.
- Tsai, Kellee. 2007a. *Capitalism without Democracy: The Private Sector in Contemporary China*. Ithaca, NY: Cornell University Press.
- Tsai, Lily. 2007b. *Accountability without Democracy: Solidary Groups and Public Goods Provision in Rural China*. Cambridge: Cambridge University Press.
- Whyte, Martin. 2010. *Myth of the Social Volcano: Perceptions of Inequality and Distributive Injustice in Contemporary China*. Stanford, CA: Stanford University Press.
- Xiao, Qiang. 2011. The Rise of Online Public Opinion and Its Political Impact. In *Changing Media, Changing China*, ed. Susan Shirk. New York: Oxford University Press, 202–24.
- Yang, Guobin. 2009. *The Power of the Internet in China: Citizen Activism Online*. New York: Columbia University Press.
- Zhang, Liang, Andrew Nathan, Perry Link, and Orville Schell. 2002. *The Tiananmen Papers*. New York: Public Affairs.

# Anti-Americanism and Anti-Interventionism in Arabic Twitter Discourses

Amaney A. Jamal, Robert O. Keohane, David Romney, and Dustin Tingley

Systematic investigation of attitudes expressed in Arabic on Twitter towards the United States and Iran during 2012–13 shows how the analysis of social media can illuminate the politics of contemporary political discourses and generates an informative analysis of anti-Americanism in the Middle East. We not only analyze overall attitudes, but using a novel events-based analytical strategy, we examine reactions to specific events, including the removal of Mohamed Morsi in Egypt, the *Innocence of Muslims* video, and reactions to possible U.S. intervention in Syria. We also examine the Boston Marathon bombings of April 2013, in which the United States suffered damage from human beings, and Hurricane Sandy, in which it suffered damage from nature. Our findings reinforce evidence from polling that anti-Americanism is pervasive and intense, but they also suggest that this animus is directed less toward American society than toward the impingement of the United States on other countries. Arabic Twitter discourses about Iran are at least as negative as discourses about the United States, and less ambivalent. Anti-Americanism may be a specific manifestation of a more general phenomenon: resentment toward powerful countries perceived as interfering in national and regional affairs.

One aspect of globalism—a state of the world involving networks of interdependence at multi-continental distances—is what could be called “social globalism,” entailing long-distance transnational transmission of ideas, information, and images.<sup>1</sup> Social globalism implies discord, since it brings groups with different interests and values into contact with one another.<sup>2</sup> Contemporary social media enable individuals who identify with different groups to express their views

in public in relatively safe ways. The result is a discordant set of discourses—contentious and not always deeply reflective, but revealing about values, perspectives, and emotions of large numbers of people who have politically relevant views and are ready to express them. We use the plural, “discourses,” because it is not clear that participants in social media are occupying a common public sphere, speaking to one another. There may be distinct discourses, with distinct populations, speaking to

---

*A permanent link to supplementary materials provided by the authors precedes the references section.*

*Amaney A. Jamal is the Edward S. Sanford Professor of Politics at Princeton University (ajamal@princeton.edu). She specializes on the politics of development and democratization in the Middle East. Robert O. Keohane is Professor of International Affairs at the Woodrow Wilson School of Public and International Affairs, Princeton University (rkeohane@Princeton.edu). He is the author of After Hegemony: Cooperation and Discord in the World Political Economy (1984) and Power and Governance in a Partially Globalized World (2002). He has served as president of the International Studies Association and the American Political Science Association. David Romney is a doctoral candidate at Harvard University whose primary research interests are the psychology of intergroup relations, ethnic and religious conflict, and the Middle East. His research interests include experimental political science, social media and the internet, and Southeast Asia (dromney@fas.harvard.edu). Dustin Tingley is the Paul Sack Associate Professor of Political Economy in the Government Department at Harvard University. His research interests include international relations, international political economy, experimental approaches to political science, and statistical methodology (dtingley@gov.harvard.edu). The authors wish to thank, for written comments, the editor; five anonymous reviewers; and multiple colleagues: Carla Beth Abdo, Giacomo Chiozza, Henry Farrell, Peter Katzenstein, Gary King, Marc Lynch, Helen Milner, Rich Nielsen, and Brandon Stewart. We also thank seminar participants at Harvard University, Princeton University, and NYU. Access to Crimson Hexagon via the Social Research Grant Program is acknowledged. Harvard University and Princeton University provided support for this research.*

doi:10.1017/S1537592714003132

© American Political Science Association 2015

March 2015 | Vol. 13/No. 1 **55**

one another and only glancing at other communities, perhaps negatively.

Discourses on social media are typically focused on contemporary topics—often events that have just occurred—and therefore illuminate how people in this partially globalized world interpret phenomena that are inherently ambiguous. These discourses expand the public sphere by enabling ordinary people to comment, in real time and for a potentially global audience, on world events. They also provide opportunities for political scientists who are interested in new and interactive patterns of globalization to explore them directly, by monitoring them and seeking to analyze their content. Students of politics have been quick to observe these discourses, and to analyze them.<sup>3</sup>

One theme of many of these discourses, especially in the contemporary Middle East, is anti-Americanism. Since World War II the United States has engaged in extensive military intervention in the Middle East—a region in which anti-American views have become predominant. Monitoring the Arabic-language social media provides a fascinating window into the cross-currents of opinion with respect to the United States in the Middle East, and therefore indicates how American policy and the policies of other countries generate animosity, ambivalence, and contentious discussion. We can learn a great deal about anti-Americanism, and more broadly about politics in an era of global social media, by analyzing these discourses.

Social scientists have sought to analyze whether anti-Americanism is principally political—hating America because of what it does, particularly its interventions—or principally social: hating America for what it is and what it represents.<sup>4</sup> Some of the best work on this subject, analyzing the 2004 Pew Global Attitudes Survey in detail, shows both that the Middle East stands out for its negative views of the United States and that even in the Middle East, anti-Americanism is differentiated by issue. Attitudes are most negative toward the diffusion of U.S. customs abroad, toward the war on terror, and toward the effects of American foreign policy on inequality between rich and poor countries. Attitudes toward U.S. political institutions and popular culture are somewhat negative but more balanced; views of U.S. science are quite positive. In general, what Giacomo Chiozza calls “socio-cultural anti-Americanism” seems to be much less intense than political anti-Americanism.<sup>5</sup>

We take advantage here of new technology to analyze the social media site, Twitter, in order to gain information that is not based on polls but on autonomous expressions of opinion by individuals about anti-Americanism in the Middle East. This technology enables us to observe responses to events on Twitter and to analyze the resulting discourses. In doing so, we generate multiple empirical observations about Arab attitudes, creating new information not only about anti-Americanism but also about how Arabic-speaking

publics react to events and to other states in the region, notably Iran. Our analysis helps to resolve a major ambiguity in the existing literature on anti-Americanism: whether Middle Eastern animus toward the United States is directed toward the nature of its society—what the United States “is”—or toward the effects of its policies and social practices on other societies—what the United States “does.”<sup>6</sup> The Twitter discourse is highly political and focused on the impact of the United States abroad more than on criticisms of American society. The specter of U.S. intervention hangs over the Twitter discourse.

Our findings about the effects of the impingement of the United States, and American society, on Arab societies led us to ask whether Arab publics express similar resentments toward Iran. We find that Arabic Twitter participants strongly dislike Iranian policy and do not express positive views toward Iranian society—unlike their more ambivalent views of the United States. Our interpretation of this data is that what is often labeled “anti-Americanism” reflects, to a considerable extent, fear of alien intrusions and hegemonic influence, from whatever source, into one’s own society. It may well reflect a desire for political and social autonomy rather than dislike for America *per se*.

Participants on Twitter do not constitute a random sample of any identifiable population: we do not claim that our data reflect Arab attitudes in a perfectly representative sense. However, Twitter participants are not a tiny portion of the Arabic-speaking population; according to the Third Wave of the Arab Barometer, close to 40 percent of the Arab public is now online and of this population, 30 percent is on Twitter. According to a report by the Dubai School of Government, the number of active Twitter users in the Arab World reached 3.7 million in 2013, up from 2 million in 2012. Saudi Arabia records the highest number of active Twitter users at 1.9 million. Egypt (519,000) and the UAE (401,000) come in second and third.<sup>7</sup> In these countries Arabic is the preferred language of those who tweet, with 73.6 percent of all activity in Arabic. No systematic analysis has been done to establish the profile of Arabic Twitter users, but it seems clear that they are relatively young. Sometimes the Arab youth population has been portrayed as more cosmopolitan, more moderate, less religious, and more pro-democratic than the Arab population as a whole.<sup>8</sup> But careful analysis of the Arab Barometer Second Wave data (2011–2012) indicates that Arab youth are not necessarily more democratic than older cohorts and are actually slightly more supportive of political Islam.<sup>9</sup> Little evidence therefore exists to support the possible objection that our analyses of Twitter data are biased in a pro-democratic, moderate, or secular direction.

Twitter discourses are distinctive, and not necessarily representative of mass public attitudes, but they are increasingly important as a medium of expression and communication, particularly for citizens who are especially

interested in politics. Arabic-language social media discourses affect participants' expectations about how other Twitter participants will respond to their own posts. They are therefore likely to affect participants' own expressions of views both through persuasion and socialization and by shaping their incentives with respect to their own contributions.<sup>10</sup> Hence, these discourses are politically important in their own right.<sup>11</sup> Insofar as public politics is migrating to the Worldwide Web, political scientists need to study social media.

Another major advantage of our focus on social media is that we can examine reactions to specific events. As we will show, these specific reactions are often more informative than aggregate findings, whether from polls or our own Twitter-based analysis, because they reflect interpretations of events as they unfold, thereby providing evidence of pre-existing attitudes, and because reactions to events are interactive, generating a social and political discourse. In situations characterized by ambiguity, people who are active on Twitter create a social reality, which may correspond more or less closely to what analysts later decide actually occurred. The technology that we employ therefore has wide implications for studying social and political discourses around the world, on a variety of topics. Its relevance is not limited to the Middle East or to the analysis of anti-Americanism. We hope that our analysis of anti-Americanism will help to stimulate creative and rigorous social media analysis in political science.

We first explain how we use social media analysis to analyze discourses about America and intervention, in the Middle East and pose our critical questions. Next we analyze general levels of anti-Americanism in the Middle East. We then move to our core contribution where we examine Twitter data following specific events. We first look at Arabic Twitter responses to actions or inactions by the United States that affected Arabs including the July 2013 coup against Mohamed Morsi's government in Egypt and the Syrian civil war. We then turn to responses to the publicity, between September 1 and November 10, 2012, over *Innocence of Muslims*, a video that was widely viewed in the Arab and Muslim world and discussed on Twitter. Next, we shift to events where America was the target, either of human action (the Boston Marathon bombing of April 2013) or of nature (as in Hurricane Sandy, October 2012).

Reactions by the Arabic Twitter public to these events show that in terms of volume of traffic, political anti-Americanism is much stronger than social anti-Americanism. But a more important distinction is between attitudes toward American domestic politics and society, which reveal considerable ambivalence, and attitudes toward how the United States affects politics and society in Arab countries, which are intensely negative. Evidence of Arab hostility

to Iran, which we then examine, provides crucial evidence that anti-Americanism is one aspect of a more generic phenomenon: publics in a variety of countries dislike the impingement of powerful countries on their societies.

## Analyzing Social Media Discourses

Studying social media such as Twitter implies studying discourses—how people talk to one another electronically. Expressions of opinion on social media are not necessarily representative of the views of the publics of countries in which they originate, but they are views that individuals have decided—unprompted—to express. Furthermore, people who express these views have been able to see others' views on Twitter and can therefore react to the changing social media universe as they perceive it. Data derived from monitoring Twitter traffic therefore enable us not only to answer questions similar to those posed by analysts relying on public opinion data, but also to answer questions about interactions among those people expressing opinions.<sup>12</sup> In carrying out this analysis we are able to learn about anti-Americanism in the Middle East on the basis of an analysis of Twitter feeds, without identifying individuals.

The existing literature on the Middle East is divided about the depth and nature of Arab animosity to the United States. One school of thought holds that levels of anti-Americanism are inherent to the culture and identity of Arabs and Muslims.<sup>13</sup> High levels of anti-Americanism are viewed as reactions to western and liberal values, which some in the region view as antithetical to Islamic precepts. In this vein, some studies show that levels of anti-Americanism increase when secular-religious tensions grow.<sup>14</sup> A different school of thought holds that levels of anti-Americanism reflect negative views of U.S. policies in the region rather than of the United States as a society.<sup>15</sup> Many of these scholars have shown that there is much admiration for basic American values and culture.<sup>16</sup>

Looking at responses to events sheds light on these questions. We can investigate responses to American intervention or possible interventions in the Middle East; we can examine Arabic responses to social events where American society could be seen as critical of Arabic society; and we can look at responses to events where U.S. society itself is a target. We find that Arabic-language political discourses are permeated with anti-Americanism, particularly when issues of intervention arise; but views toward American society are more complex. The volume of social expression is much lower than that of political opinion. When Americans are the victims rather than the perpetrators of harmful actions there is some animus toward American society but this animus is not as intense as Arab hostility toward American policy abroad.

## **Research Design and Methodology**

We use Arabic Twitter posts gathered and stored by Crimson Hexagon, a social media analytics company, employing a supervised text analysis model developed by Daniel Hopkins and Gary King.<sup>17</sup> Text analysis methods can be either “supervised” or “unsupervised.”<sup>18</sup> The difference lies in who (or what) determines the topics by which to classify texts. Using an unsupervised method, topics are not determined *ex ante*; rather, the statistical method itself helps determine them.<sup>19</sup> In contrast, using a supervised method, the researcher hand-codes a “training-set” of documents into pre-determined categories, and on the basis of these training documents the rest are classified by an algorithm.<sup>20</sup> In our study we seek to know the proportion of the population of tweets that fit in specific categories. The supervised ReadMe algorithm is highly suited to this calculation. Rather than estimating topic proportions based on the categorization of individual documents, ReadMe estimates these proportions using words within each text. Importantly, this means that ReadMe does not individually classify tweets. But we do get population-based estimates, which are the quantity of interest here, and we present both proportions and estimated volume, which is the product of the proportions and overall volume.

We gain access to these techniques and data through Crimson Hexagon,<sup>21</sup> a social media analytics company founded in 2007. Crimson Hexagon (CH) combines the text analysis method developed by Hopkins and King with a vast collection of social media data in an easy-to-use online platform. In particular, we use here, among other things, the universe of Arabic language Twitter data from 2012–2013.

The analysis proceeds as follows. First, the user determines the date range of interest and sources to draw upon, imposing language and geographical restrictions as desired.<sup>22</sup> After these basic parameters are set up, she determines a set of keywords on which to draw texts satisfying the other restrictions. CH offers the use of Boolean operators to create a complex set of keywords (or phrases) used to include or exclude texts.<sup>23</sup>

After setting up the basic parameters and the keyword restrictions—forming what CH calls a “monitor”—the user is then ready to begin training. Because Crimson Hexagon uses essentially the Hopkins-King algorithm—a supervised text analysis method—a set of pre-defined categories and training texts is required. To aid with the supervised document classification process, CH feeds the user, one-by-one, tweets that match the setup restrictions. The user places each tweet into the appropriate category, continuing this process until she feels that the monitor has been sufficiently trained.<sup>24</sup>

Once training is complete, the user runs the algorithm and accesses the analysis results. The main results consist

of daily estimated frequency data for each of the categories. The user also has limited access to the raw tweets themselves, either through Crimson Hexagon’s “bulk export” function, which is limited to 10,000 tweets per day, or through example tweets, which it uses a proprietary classifier to select. They are the classifier’s “best guess” of good examples of each category. These are helpful for getting a rough sense at whether the training has picked up on what the analyst is interested in, and they were used in our monitor review process.

The greatest advantage of Crimson Hexagon is that, as a Twitter-certified product with access to the “Twitter fire hose,” it provides every public tweet ever posted on Twitter—in any language and from any location—that matches the user-determined restrictions; and its commercial customers demand a high level of filtering out of spam and bots that could otherwise contaminate our analysis. Two weaknesses of using Crimson Hexagon should be mentioned. CH limits access to the underlying source texts; we have compensated by reading thousands of tweets manually. And CH does not enable us to distinguish between original tweets and re-tweets, so our implicit assumption is that re-tweets usually reflect sympathy with the original tweet.

Thus, with a reliable set of Arabic keywords referring to the United States, we can analyze every genuine Arabic tweet referring to the United States that has ever been sent. Of course no method comes without limitations. Ferreting out sarcasm is well beyond the Read-Me algorithm. And any text method requires extensive ex post checking and transparency, which we attempt to do in our extensive online supplementary materials.<sup>25</sup> Only small percentages of Twitter users indicate their country, making cross-national analysis difficult.

## **Analyzing General Levels of Anti-Americanism in the Middle East**

As mentioned earlier, examination of public opinion data from the Arab world has long revealed high levels of anti-Americanism.<sup>26</sup> Because Twitter posts result from choices to express oneself (rather than a response to a question someone might never have thought about), data from Twitter enable us to discover whether self-expressed views—the Arabic discourses—are similarly hostile to the United States.

Table 1 reports on two different sets of public opinion polls—the 2012 Pew Global Attitudes Poll and the 2011/2012 Arab Barometer—in seven Arab countries: Egypt, Jordan, Lebanon, Tunisia, Iraq, Algeria and Yemen. First, we look at two sets of questions in the Pew Global Attitude Poll: favorability scores toward the United States and agreement with the statement that it is good that U.S. ideas and customs are spreading to the region. Second, we turn to two Arab Barometer questions: whether respondents believe that “armed

**Table 1**  
**Arab public opinion toward the United States**

	Egypt	Jordan	Lebanon	Tunisia	Iraq	Algeria	Yemen
Favorability opinion of the United States, 2012 (Pew Global Attitudes 2012, % favorable)	19	12	48	45	NA	NA	NA
It's good American ideas and customs are spreading here (Pew Global Attitudes 2012, % agree)	11	10	41	25	NA	NA	NA
The United States' interference in the region justifies armed operations against the United States everywhere (Arab Barometer 2011/2012, % agree)	63	58	44	64	70	53	47
American and western culture have positive aspects (Arab Barometer 2011/2012, % agree)	63	55	75	83	67	50	66

operations” are justified against the United States because of its interference in the region, and whether respondents believe there are positive features linked to American and Western culture.

Favorability scores are quite low—remarkably low in some countries such as Egypt and Jordan, while even in Lebanon and Tunisia these scores are not predominantly positive. There is widespread opposition to the spread of Western values. Yet, quite remarkably in terms of the strongly negative views about American foreign policy and the spread of western values, 50 percent to 83 percent of respondents say that American culture has positive aspects. That is, in the polling data we see indications of the same distinction we will document for the Twitter data between Arab attitudes toward the impingement of American policy and society on Arab countries and toward American society as such.

Systematic polling data in the region is quite difficult to secure. The Pew Global Attitudes poll, for example, does not survey in many Arab countries and in the 2012 wave it only gathered data in four countries: Egypt, Lebanon, Jordan and Tunisia. The Arab Barometer data has expanded its number of countries in recent waves but problems remain. For example, some countries have not allowed social scientists to ask questions about attitudes toward the United States. Clearly, countries are worried about the negative ramifications of “exposing” the degree of anti-American sentiment in their countries. Morocco (2006), Egypt (2006), and Saudi Arabia (2013) are examples of countries restricting public opinion polling.

### General Attitudes toward America in the Twitter Universe

Our first monitor, called the “General Anti-Americanism” (General AA) monitor, looks at general trends in the Arabic Twitter conversation about the United States. We sought to make this monitor as broad as possible, choosing a large date range: from the earliest date for which Twitter data is available in Crimson Hexagon to the current date.

However, we quickly found that data after January 1, 2012 are the most reliable, and all of our figures are based on tweets after this date. We also aimed for breadth in our keyword criteria, producing a large set of keywords. Between January 1, 2012, and December 31, 2013, an astounding number of 33,009,354 tweets matched these date and keyword criteria and were classified in one of our six substantive categories.

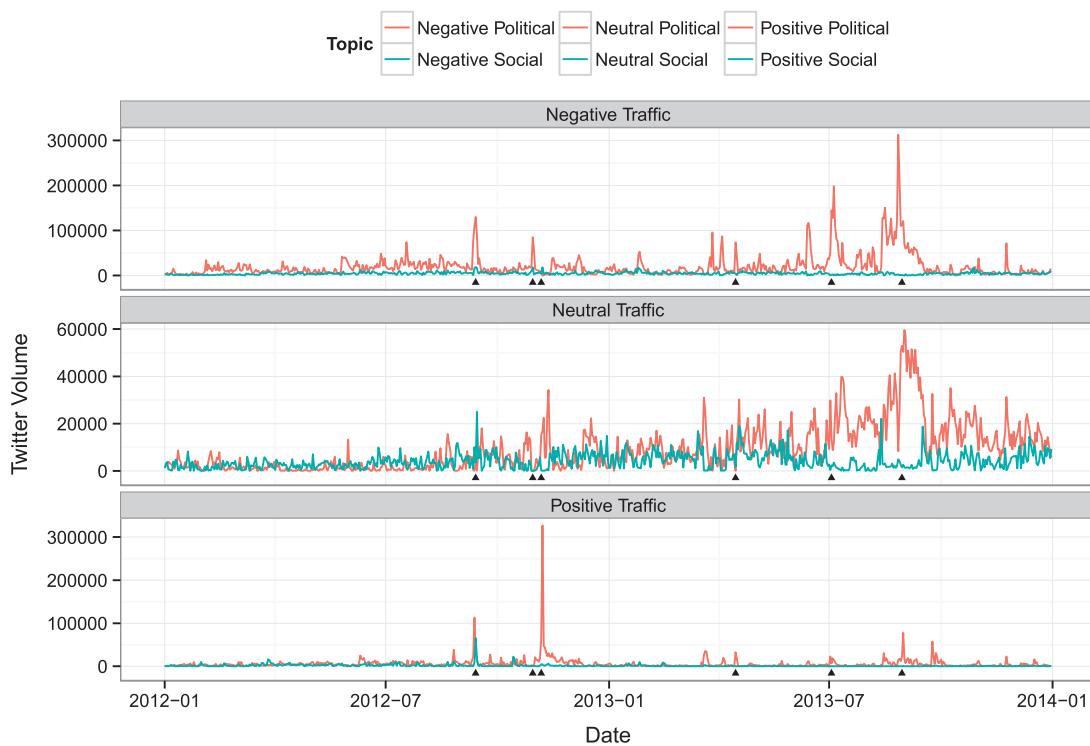
The goal of our general monitor is to get a general sense of Arabic-language tweets referring to the United States. Furthermore, this analysis helps give an indication of instances of when an “event” has happened that requires additional scrutiny. We developed the six training categories shown in Table 2 based on two distinctions that are important in the anti-Americanism literature: the topic of discussion (political vs. social) and its valence (negative, neutral, or positive).<sup>27</sup>

Which categories are most prominent? Is anti-Americanism rampant in the Arabic Twitter universe? Does it span both political and social categories, or is it circumscribed to political anti-Americanism? Table 2 plots the estimated number of tweets for each category. The results are striking. Although the ratio of negative to positive tweets is over 3:1 in both social and political categories, the volume of political traffic is nearly four times as great as social traffic.<sup>28</sup>

**Table 2**  
**Estimated number of tweets per category for general attitudes monitor**

Category	Frequency	Percentage
Positive social	957,922	3
Neutral social	2,986,740	9
Negative social	3,140,021	10
Positive political	4,389,598	13
Neutral political	6,759,360	20
Negative political	14,775,713	45

**Figure 1**  
**Total Arabic tweets by political and social category using all tweets in the world, plotted across time**



*Note:* The triangles indicate key events during this time period.

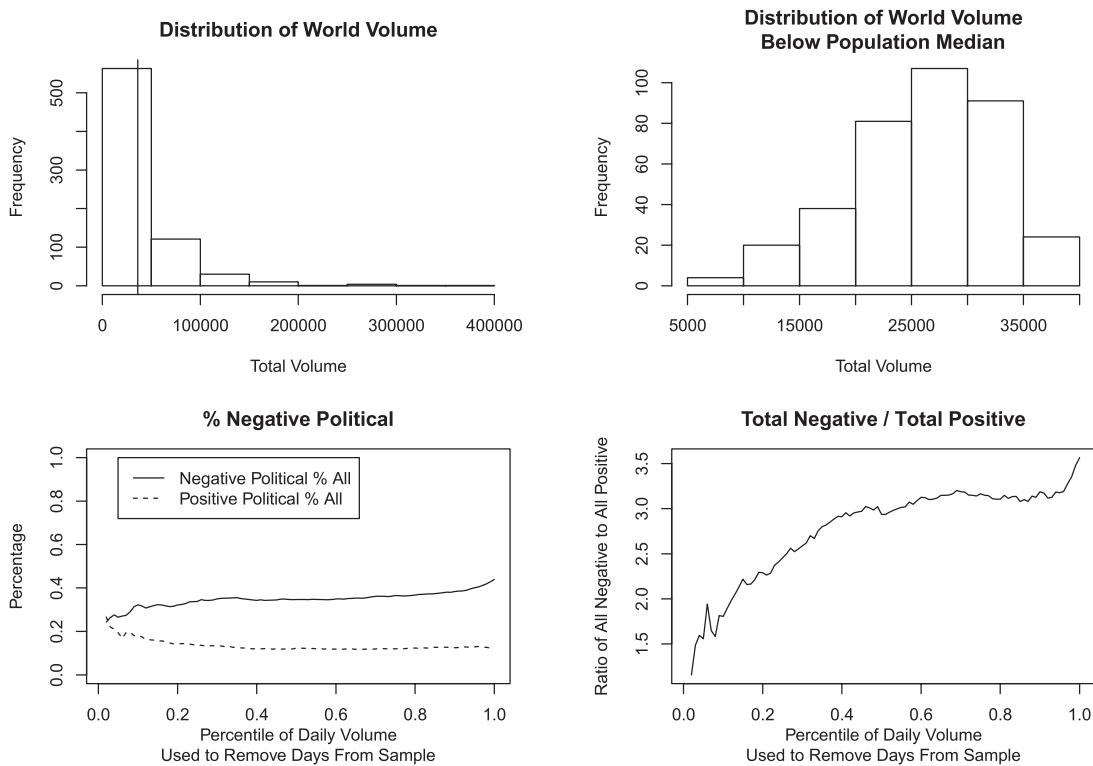
In figure 1 we plot over time each of the categories presented in table 2. The small black triangles along the horizontal axis represent a set of key events during our time period, many of which we analyze in detail in the next section. One thing that stands out is that for these events there are often dramatic spikes in Twitter volume. This is perhaps not surprising given the nature of Twitter. However, it raises an important question: when we remove high-category days, is the volume of negative political traffic still substantially higher than the other categories?

Figure 2 shows that even when we eliminate days where there are spikes in Twitter activity, negative political traffic remains dominant. The top left plot is a histogram of total on-topic volume at the daily level. The population median is given with a vertical line. This clearly shows the presence of spikes. The top right plot shows the same histogram, where we have removed all days above the fiftieth percentile. The bottom left plots the percent of total on-topic volume that is politically positive or negative (y-axis) against the percentile used to remove days that had a total volume above that percentile. For example, at an x-axis value of 1, all days are used, and at the x-axis of .5, we are only using days captured in the histogram in the top right

(those days with a volume below the median). We see that once we restrict our sample to days without events, we get a similar picture. Negative political volume stays close to 40 percent of the entire sample. We see a similar story if we take the ratio of negative (political and social) to positive tweets in the bottom right of figure 2. Looking at both the universe of tweets and only using days that did not have spikes in volume, we arrive at the same conclusion that negative political tweets dominate the Arabic language conversation about U.S. politics and society.

However, in spite of the predominance of negative political tweets over others, one political event stands out as prompting positive responses: the 2012 U.S. presidential election. The spike in positive traffic around this time (refer to figure 1) was the largest spike in traffic for our entire monitor. When we closely inspected this we found it was comprised principally of favorable reactions to President Obama's re-election and admiration for the institution of free and fair elections, coupled with the desire for similar institutions in one's own country. However, this positivity towards Obama's re-election disappeared very quickly, resulting in a very low percentage for this category overall.

**Figure 2**  
**Analysis excluding spikes in Twitter volume**



*Notes:* The top left figure plots the histogram of total on-topic daily volume with a vertical line at the median. The top right figure plots the same distribution but using observations below the population median. The bottom left figure plots the percentage of total on-topic volume that is negative or positive political as days are removed for being over a given percentile (x-axis) in total volume. The bottom right figure plots a similar line, but where the sum of negative tweets is divided by the sum of positive tweets.

In summary, the conversation on Twitter, in Arabic, about the United States is especially negative towards U.S. policy; the conversation about U.S. society is also mostly negative but with some positive elements and much smaller in volume than the conversation about U.S. policy. These results hold whether we look at all Twitter traffic or only those days that reflect more of a “baseline.”

### Analyzing Responses to Specific Events

We now analyze a set of events in more detail than was possible in the general monitor. We structure our analysis by looking both at events involving U.S. actions towards the Arabic speaking world, and at events in which the United States was the target. This distinction is important because we might expect different levels of traffic and different levels of positive and negative reactions to these two types of events. We might expect, because of generally negative attitudes towards the United States,

that actions the United States takes towards the Arabic speaking world would generate large amounts of negative traffic. On the other hand, we might expect different reactions when the United States is the target, perhaps even expressions of satisfaction. We first discuss political events in which the United States was seen as the impinging actor; we then analyze the responses to the *Innocence of Muslims* video emanating from American society; and next we turn to events in which the United States and the American people were affected by actions of people or natural forces such as Hurricane Sandy.<sup>29</sup>

Two of our monitors target political events: a monitor that tracks the U.S.-focused traffic surrounding Morsi's removal from office and the subsequent pro- and anti-military protests in Egypt; and a monitor that tracks U.S.-focused traffic surrounding the Syrian civil war. These two situations differ from each other in many ways, but their similarities invite interesting comparisons. In both situations, two camps dominated the conversation: on the

one hand, the “pro-regime” camps (those who supported Assad in the Syria monitor and those who supported the military in the Morsi monitor), and on the other, the “anti-regime” camps (those who supported the Syrian revolution and those who supported Morsi). Also, looking at these two events together allows us to understand situations in which the U.S. government is perceived to be able to influence the course of events due to its influence in the region. This is in contrast to situations where the United States is the target, which we will discuss.

### **Morsi and Egypt**

The Morsi monitor seeks to look at the Arabic conversation surrounding the removal of Mohammed Morsi from power on July 3, 2013 and the events that followed. For this monitor, we were interested in looking at how Arabic tweeters were talking about the United States in relation to the events that were going on in Egypt. How did the Arabic tweeter public respond to these events?

For this monitor tweets were only selected for analysis if they included both a reference to the United States and a reference to Egypt.<sup>30</sup> We ran the monitor from June 27, 2013 (approximately a week before Morsi was forced out of office) to September 8, 2013 (when it seemed that news coverage of the situation in Egypt was giving way to coverage of the crisis in Syria).

In designing the categories for our analysis we wanted to allow for both sides of the dispute to take on both pro- and anti-American positions, and to allow for both general pro and anti sentiments towards the United States that were not in the context of supporting one or the other domestic group. With these desires in mind, we came up the categories listed in table 3, where we report the results. To facilitate comparability among the event monitors, in all cases we exclude tweets that contained only neutral news reports, though our substantive conclusions do not change if we include these.

The first thing to notice is that positive views of the United States were a very small proportion of Twitter traffic: 4 percent overall. The top three categories encompassed tweets that expressed opposition to the

United States. The largest category was anti-American and anti-military, the next largest was general anti-Americanism, and third was anti-American and anti-Muslim Brotherhood. No matter which side of the domestic dispute an individual was on, he or she was likely to be opposed to the United States. Because of the small number of positive tweets, it is hard to determine how much of the positive traffic was coming from anti-military versus anti-Muslim Brotherhood tweeters. In summary, this analysis clearly shows sharp divisions within Egyptian society on the question of the Muslim Brotherhood, but a common opposition to the United States. Rather than an enemy of an enemy being a friend, the United States is consistently cast as an enemy.

### **Syria**

Our analysis of Syria is similar to the Morsi analysis in that it seeks to look at how Arabic tweeters are talking about the United States in relation to one particular event or topic—in this case the Syrian Civil War, which began on March 15, 2011. The explosion of discussion on Twitter about possible U.S. intervention in Syria following alleged chemical weapons attacks by the Assad regime provides a good opportunity to look into reactions towards U.S. foreign policy. Since this story in 2013 was one of U.S. inaction, we are asking here about levels of anti-American political opinion in an event where the United States had not used force but was widely perceived as being able to influence the course of events in Syria.<sup>31</sup>

We use tweets if they included a reference to the United States and a reference to Syria, following a similar procedure to that used in the Morsi monitor. We trained this analysis from January 1, 2012, to December 31, 2013. Modeling our categories on the Morsi monitor, but with slight adaptions based on what we found when looking at the Syria traffic, we obtained the five categories found in table 4, where we report the results.

Pro-U.S. traffic is between about 1 percent and 4 percent of total non-news traffic, depending on the time period. All of the pro-U.S. traffic comes from the anti-regime camp, and the amount of negative traffic coming from the pro-regime camp far outweighs the negative traffic from the revolutionaries. This is to be expected, given that the United States was seen as on the side of the Syrian revolution. However, even for the anti-regime tweeters there were 350 percent more anti-U.S. tweets than pro-U.S. tweets before the chemical weapons incident, and about 1,200 percent more after the event. In addition, anti-regime traffic that was ambiguous about the United States was large prior to the chemical weapons attack but declined sharply following the attacks, whereas general negative traffic against the United States increased after the attacks. The right side of table 4 presents figures for the post-August 20 period.

**Table 3**  
**Total estimated number of tweets in each category for Egypt, omitting 515,257 posts with News/Neutral content**

Category	Frequency	Percentage
Pro-U.S.	66,024	4
Anti-U.S. and anti-Muslim Brotherhood	220,906	13
General anti-American	500,311	29
Anti-American and anti-military	918,541	54

**Table 4**  
**Total estimated number of tweets in each category for Syria, split by date just prior to chemical weapons attacks**

Category	Jan. 1, 2012–Aug. 19, 2013		Aug. 20, 2013–Dec. 31, 2013	
	Frequency	Percentage	Frequency	Percentage
Pro U.S., anti-regime	67,704	4	7,856	1
Ambiguous about U.S., anti-regime	147,006	9	39,077	4
Negative U.S., anti-regime	252,958	15	94,898	10
Negative U.S., general	526,917	32	401,854	42
Negative U.S., pro-regime	663,838	40	416,112	43

Notes: News/Neutral posts are omitted.

Twitter responses to the Egypt and Syria situations dramatically demonstrate the breadth and depth of negative views toward American foreign policy in the Arabic Twitter universe. It is not surprising that supporters of the Muslim Brotherhood and of the Assad regime (not necessarily the same people) are negative toward the United States, since the United States was cool toward the Brotherhood and has been very negative toward the Assad regime. But what is dramatic is the extent to which Arabic Twitter users who favor the Egyptian military are not positive toward the United States, and the extent to which those who favor the Syrian revolutionaries are predominantly negative toward the United States. This is particularly striking for the Syria monitor, in which 87 percent of tweeters who expressed political views were antagonistic toward the United States, despite the fact that the United States opposed the Assad regime, which was also opposed by many Arabic tweeters. In the Egypt monitor, the proportion of political tweets that was anti-American reached 96 percent. We see a similar pattern if we only look at Twitter participants who were more or less on the same side as the United States: that is, they favored the Egyptian military or were opposed to the Assad regime. Once again, the fact that the United States was the enemy of their enemy did not make them regard it as their friend.<sup>32</sup>

#### **Arab Responses to Critical Views from American Society: *Innocence of Muslims***

Written and produced by an Egyptian-born Coptic Christian residing in the United States, the 14-minute long *Innocence of Muslims* film was widely considered an attack on Muslim society. After the first Arabic-dubbed version of the film appeared early September 2012, there were widespread demonstrations and riots in the Middle East, many with an anti-American tone, suggesting that the film's views were attributed by many people to the United States. Its appearance also led to an immense outpouring of commentary on Twitter. It was important to create a monitor for this event since it could be considered an attack on Muslim society by American

society, and thus responses to it could help indicate the extent to which perceptions of American society fuel anti-Americanism. If perceptions of American society do fuel anti-American sentiment, we would expect to see direct expressions of anger directed at American society in response to this event. We find, however, a much more complex pattern of responses.

We developed a short and specific list of base words to help us capture tweets related to this event.<sup>33</sup> We examined the two-month period right before and after the movie began making headlines in the Middle East, beginning on September 1, 2012 and continuing until November 10, 2012. As we trained the monitor, we found that responses were not easily mapped onto a simple pro- or anti-Americanism scale; such a scale did not capture the Twitter discussion about the film, which focused on what the best course of action for Muslims to address the film would be. For this reason, we modified our categories so that they captured different sentiments expressed about the film. This process led to the discovery of five categories that help us convey the substantive patterns we did find. These categories are listed in table 5.

Reviewing these results, we find that the biggest (non-news) category was the category encouraging people to ignore the film. The second largest category included posts supporting individual action against the film. The third largest category reflects posts arguing that Islam is stronger than the film. All of these reactions have negative implications but none of them represents a clear condemnation of American society in general. We did not find tweets with statements such as "This means that all Americans hate us" or "All of American society and people should be condemned." In this monitor, we see a clear negative reaction but we do not see a predominance of direct attacks on American society, unlike the direct attacks on American policy in the general, Morsi, and Syria monitors. But there is no doubt that both in this monitor, and the general anti-American monitor for

**Table 5**  
**Total estimated number of tweets in each category for the world**

Category	Frequency	Percentage
Demanding action from political leaders	10,735	2
Anger towards producers	26,376	4
Evidence Islam is stronger than film	80,383	14
Supporting individual action against film	214,337	36
Encouraging people to ignore film	260,124	44

Note: News/Neutral content posts (433,590) were omitted.

this time period, expressions of sentiment are preponderantly negative.<sup>34</sup>

#### **Arab Responses to Harms to America**

The previous analyses all dealt with events where the Arab world was the target. We now shift our attention to two events in which the United States was a target, providing an important symmetry to our paper. We study the Boston marathon bombings and Hurricane Sandy.

*Boston Marathon Bombing.* On April 15, 2013 a set of bombs exploded at the Boston Marathon, killing three people and injuring an estimated 264 others. The accused perpetrators purportedly carried out the attack for the sake of Islam. This monitor shows us how Arabic tweeters responded to this attack on the United States. How much sympathy was there for victims, and were responses principally focused on U.S. politics and policies or on U.S. society?

We developed a short and specific list of base words to help us capture tweets related to this event, as detailed in the on-line supplementary materials. As with the Innocence of Muslims monitor responses did not easily map onto a pro- or anti-Americanism scale, so we grouped responses by the topics that seemed prominent during our training. These categories are listed in table 6, where we report the results. This monitor was run from April 13 to June 10, 2013.

Figure 3 presents the Boston monitor results over time for each category, using the total number of tweets. The first small black triangle marks the date of the bombing and the second is the date of the final shoot-out with the alleged perpetrators. Table 6 presents the total number of tweets for each category, aggregated over time.

Several interesting patterns are present in the data. First, the ratio of tweets claiming that media attention to the event was undeserved to tweets expressing sympathy

**Table 6**  
**Estimated number of tweets per category for Boston Marathon bombing monitor**

Category	Frequency	Percentage
Sympathy for Arabs/ Muslims	7,999	3
Sympathy for victims	23,887	9
Conspiracy theories	39,742	15
Backlash	85,588	33
Not important	104,678	40

Note: News/Neutral content posts (262,761) were omitted.

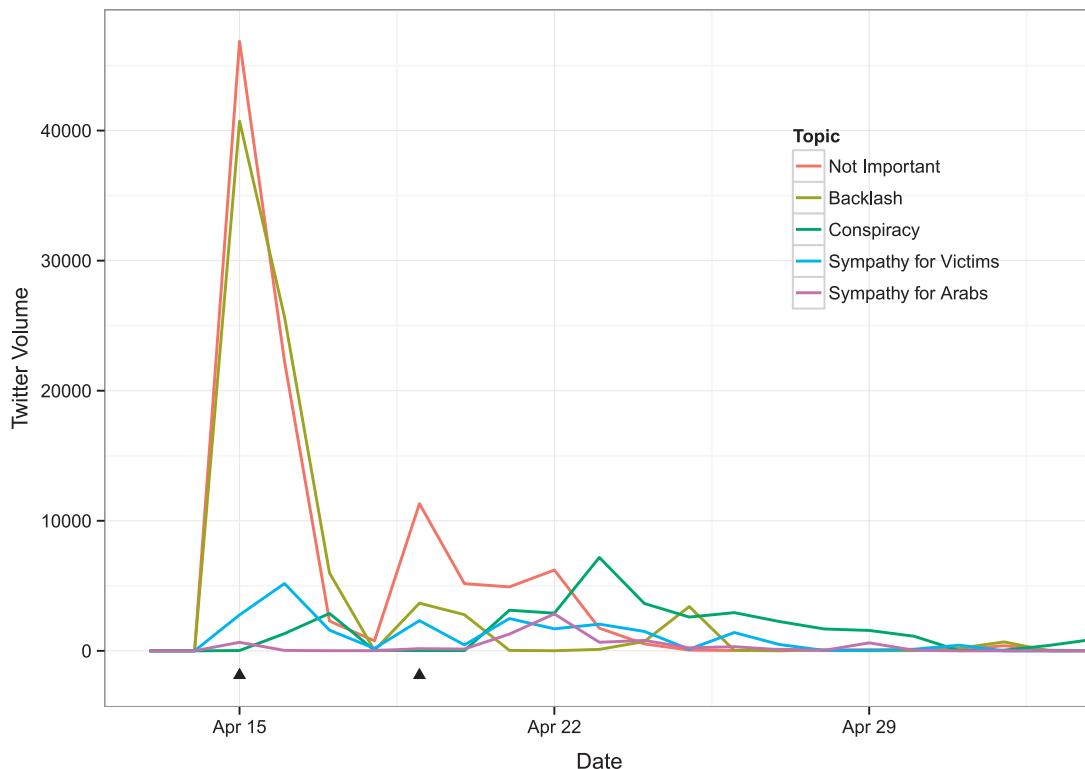
for the victims was almost 4:1. The low levels of sympathy towards innocent victims, and high levels of tweets dismissing the importance of their plight, suggest substantial negative sentiments toward the United States. Indeed, the amount of sympathy expressed for Arabs facing discrimination was about a third the size of sympathy for the victims, and the “Backlash” category, indicating concern about a backlash against people in the United States of Arab origin, had over 90,000 tweets. On the other hand, we did not observe tweets celebrating the attack on U.S. society. While many individuals argued that the attention was undeserved, they did so by pointing to deaths in the Arab world (especially Syria), often by arguing that behind these deaths are American policies. While overall sympathy for victims was a relatively small percentage of the conversation, it is American policies and interference in the Arab world that drew the ire of many tweeters.

We reflected a great deal about whether these negative views of the United States reflect “bias,” as distinguished from opinion.<sup>35</sup> Opinion reflects attitudes and reactions to events but no major cognitive distortions. People with negative opinions of other people, groups, societies, or institutions dislike what they actually do, or stand for symbolically. Bias, however, reflects distorted perceptions of reality: biased people misperceive the actions of the people, groups, societies, or institutions that they dislike.

Although in general we found it difficult to infer bias from the information we gathered, the clearest possible evidence of bias comes in the Boston Marathon monitor, specifically in the Conspiracy category. There were nearly twice as many tweets speculating about a conspiracy involving U.S. security agencies than tweets expressing sympathy for the victims.

We interpret the findings from the Boston monitor as suggesting a significant level of bias against the United States. Bias is based on a deeply negative view toward the United States that strongly colors respondent’s perceptions of observed U.S. actions and that may create negative beliefs about unobserved actions by the United States and

**Figure 3**  
**Estimated number of tweets per category over time for Boston Marathon bombing monitor**



the motivations of American decision-makers. Although there was never any credible evidence of a conspiracy beyond the plans of the two bombers, we see almost twice as many tweets that are coded as expressing belief in a conspiracy than tweets expressing sympathy for the victims of the bombings. However, our inability to discover clear evidence of bias, as opposed to negative opinion, in the other monitors makes us diffident about drawing strong general conclusions about bias.

To understand anti-Americanism it is important to ask whether the results would have been different were this event to have happened in another country. Our best effort to find a case as similar as possible comes from the Arabic Twitter conversation surrounding the death of British Army soldier Lee Rigby, who was attacked and killed by Islamist terrorists on May 22, 2013, in Woolwich, London. The results for London look very different than those for Boston. While we do not present the detailed results here,<sup>36</sup> relative to the total volume of activity there was much more sympathy for the victim; the view that this event is not important in light of attacks on Muslims in the Middle East is not expressed; and there was no discussion of a possible conspiracy. Apart from news reports, tweets fall

into three categories: expressions of sympathy with victims (60 percent), expressions of sympathy towards Arabs/ Muslims (20 percent), and expressions of concern about backlash against Muslims (20 percent). In London there were three times as many “Sympathy for Victims” tweets than “Backlash” tweets. The immediate response of Arabic tweeters was overwhelmingly one of sympathy towards Arabs and Muslims, with little to no explicit support for the attackers. There was so little discussion of a possible conspiracy that it was not included as one of our categories: the facts of the murder were accepted. This evidence suggests that hostility to the United States is greater than hostility to the United Kingdom.

*Hurricane Sandy.* In late October 2012, Hurricane Sandy caused extensive damage to the Northeastern United States. We explore this event—with results presented in table 7—to search for high levels of antipathy for American society that we have not heretofore detected.

In the Hurricane Sandy monitor most tweets were merely news items, leaving just over 310,000 tweets to analyze. Of these meaningful tweets, 10 percent (about 32,000) expressed the view that Hurricane Sandy is punishment for the *Innocence of Muslims* film or similar

**Table 7**  
**Estimated number of tweets per category for Hurricane Sandy monitor**

Category	Frequency	Percentage
Negative towards U.S. government	6,143	2
Negative general	6,594	2
Concern for Americans	7,869	3
Defending Americans	31,561	10
Negative mentioning <i>Innocence of Muslims</i>	32,125	10
Concern for Arabs/Muslims	49,276	16
Positive (towards U.S. government)	78,824	25
Not important	99,131	32

Note: News/Neutral content posts (451,728) were omitted.

prejudice against Muslims. In addition, about 2 percent (almost 13,000) were generally negative and 16 percent (almost 50,000) expressed particular concern about Arabs or Muslim Americans and residents. Over 30 percent (about 99,000) said that the Hurricane Sandy events were not important. In contrast, approximately 25 percent of tweeters (almost 80,000) commented favorably on the U.S. government's handling of the disaster, often as a contrast to the incompetence of Arab governments, and about 10 percent (over 30,000) condemned the view that Hurricane Sandy is punishment for the *Innocence of Muslims* film. In other words, of the Arabic Twitter participants who expressed an opinion, somewhat more than one-third expressed views that can be interpreted as generally favorable toward American society, somewhat less than one-third expressed explicitly negative views, and 30 percent regarded the events as not important. Even if we interpret "not important" as implying that harm to the United States is not bad, the difference between this monitor and those related to directly political events is striking.

The Boston Marathon and Hurricane Sandy monitors present something of a contrast. An overwhelming 88 percent of non-news responses to the Boston Marathon bombing are negative toward the United States: these events are seen as not important (in view of harm done to Arabs in the Middle East directly or indirectly by American policy); the greatest concern is for the welfare of Muslim Americans/residents; or a conspiracy of American intelligence agencies is viewed as responsible for the allegations. In contrast, in the Hurricane Sandy monitor over one-third of Twitter participants, apart from those merely re-tweeting news, expressed views quite favorable toward aspects of American society. One suggestion is that the Boston Marathon events were framed in an explicitly political way, evoking negative responses, while the Hurricane Sandy events were not.

### Summary

What conclusions can we draw about the nature of anti-Americanism from our findings? We find that two dimensions help us distinguish the volume and valence of Arabic-language responses: social vs. political content; and the U.S. impact abroad vs. U.S. domestic activities. Table 8 presents the most important results for previous surveys and our own analysis, placing negative responses in bold type. The table excludes neutral content and focuses on positive and negative reactions to events or topics. We find that events and activities in cell A—those with a focus on the political impact of the United States abroad—generate the largest volume and most negative responses. On the other hand, those in cell D—with a focus on U.S. domestic social activities and events—generate the smallest volume and most positive responses, although there are criticisms as well.

Table 8 shows that U.S. policy is the predominant target of anti-Americanism. Discussions of American society are more ambiguous. Arabic Twitter discourses on how American society affects the Arab world are particularly negative. Several pieces of evidence point toward this conclusion. First, public opinion polls (table 1) reveal overwhelming opposition to the proposition that it is good that American customs and ideas are spreading in the Arab world, but majority agreement that there are nevertheless positive aspects of U.S. society. Second, in our analysis of general anti-Americanism presented in table 2, the vast majority of tweets that were classified as "negative social" dealt with concerns about U.S. society impinging on Arabic society, especially dealing with women's issues. Finally, the dramatic response towards the *Innocence of Muslims* film is best understood as a response towards the impingement of U.S. society on Arabic society.

Lacking throughout our analyses is a discourse that is targeted towards the nature of U.S. norms and practices within the United States itself. People using social media in the Arabic-speaking world strongly dislike the impact of the United States, politically and socially, in their societies, but are less hostile to the United States as a society, with its very different customs and practices.

Further evidence from public opinion polling reinforces the significance of our distinction between views of American society and its impingement on the Arab world, although the argument is sometimes implicit. Recall from our introductory discussion that Chiozza found the most negative views of the United States to center on the diffusion of American customs abroad, the war on terror, and U.S. foreign economic policy—all issues in which the U.S. influence on the rest of the world is central. American political institutions, popular culture, and U.S. business all received a mix of evaluations, while evaluations of U.S. science were positive. In general, topics in this second set do not evoke thoughts about how United

**Table 8**  
**Summary of findings**

	<b>Impact of the United States Abroad</b>	<b>U.S. Domestic Activities</b>
<b>Political</b>	<p>A. Overwhelmingly negative and very large</p> <ul style="list-style-type: none"> <li>• U.S. alliance with Israel</li> <li>• Iraq War</li> <li>• Involvement in Syria/Egypt</li> <li>• Association of U.S. with disliked Arab leaders</li> <li>• U.S. foreign policy in general</li> </ul>	<p>B. Mixed and medium to small traffic</p> <ul style="list-style-type: none"> <li>• Admiration for U.S. political institutions, including elections</li> <li>• Favorable U.S.-M.E. leaders comparisons</li> <li>• Criticism of treatment of Muslims and Arabs in United States</li> <li>• Negative views of the United States in the Boston Marathon monitor, including conspiracy theory: FBI and CIA</li> </ul>
<b>Social</b>	<p>C. Negative and medium to small traffic</p> <ul style="list-style-type: none"> <li>• Salience of and negative responses to <i>Innocence of Muslims</i> video</li> <li>• Responses to Pew survey question about spread of American ideas and customs</li> <li>• Negative social comments outweigh positive social comments (3:1) in general AA monitor</li> </ul>	<p>D. Mostly positive and small traffic</p> <ul style="list-style-type: none"> <li>• Support in Pew poll for view that U.S. society has “positive aspects”</li> <li>• Sympathy by some tweeters for Sandy/Boston victims</li> <li>• Some positive interest in U.S. culture in General AA monitor<sup>a</sup></li> <li>• Criticism of treatment of Muslims and Arabs in United States</li> </ul>

Note: <sup>a</sup> Most of this interest is classified as neutral.

States power impinges on countries in the Middle East. Marc Lynch ascribes increases in anti-Americanism in the Middle East between 1999 and 2004 largely to the fact that “the American presence in the Arab world measurably increased” during this period.<sup>37</sup> Lars Berger finds, on the basis of a poll taken in 2008 in Egypt, Indonesia, and Pakistan, that fewer than ten percent of respondents approve of attacks on civilians in the United States, and that this view is strongly correlated with dislike for American culture but not with opposition to American foreign policy—opposition that is much more widely shared.<sup>38</sup> Berger therefore provides empirical support for a distinction that Amaney Jamal made between criticism of U.S. foreign policy and rejection of U.S. domestic practices. In her analysis, Jamal argues that fear is an important component of anti-Americanism.<sup>39</sup> Even Arab liberals, she claims, can become anti-American if they worry about U.S. actions, and the antipathy of sovereign nationalists to the United States is accentuated by fear.<sup>40</sup> Fear, of course, is only activated if the subject is thinking about the impact of the United States on him or her, and for most of our Twitter participants, that means foreign actions by the United States.<sup>41</sup>

These two points—that political anti-Americanism is more intense than social anti-Americanism but that the latter is still strong, and that objections are stronger to the impingement of American society on the Arab world than

to American society itself—are both important. They suggest that changes in United States policy alone are unlikely to transform Arab attitudes toward the United States, and that much will depend on how Arab attitudes themselves change.

### Is the United States Unique?

The literature on anti-Americanism has been written overwhelmingly by Americans with an exclusive focus on the United States, and this paper is no exception. The United States has actively intervened in world politics for the last 73 years, so it is not surprising that it generates strong feelings. But in what sense should we view these feelings as distinctively “anti-American” rather than being merely expressions of a more generic phenomenon: resentment of interference in one’s own affairs by a powerful state with a distinctly different set of cultural values as well as political interests? As suggested at the outset of this paper, we need to raise our sights a bit and look at countries other than the United States to see whether anti-Americanism is a highly specific or even a unique phenomenon, centered on the United States, or one aspect of a more general phenomenon: opposition to interference and influence by powerful states with different cultural values and political interests.

We approach this question first by asking how much Arabic tweeters talk about the United States, as opposed

**Table 9**  
**Volume of Arabic Twitter traffic for seven countries, Jan. 1, 2012–Dec. 31, 2013**

Country	Keywords	Total Traffic	Percentage
United States of America	America, American(s), the United States of America	40,845,963	25.88
Iran	Iran, Iranian(s)	27,634,417	17.51
Israel	Israel, Israeli(s), Zionism, Zionist(s), Judaism, Jew(s), Jewish <sup>31</sup>	27,266,590	17.28
Turkey	Turkey, Turkish, Turk(s)	18,062,640	11.45
India	India, Indian(s)	11,927,620	7.56
Russia	Russia, Russian(s), Soviet(s)	11,248,828	7.13
China	China, Chinese	11,123,497	7.05
United Kingdom	Britain, British, United Kingdom	9,707,759	6.15

to other countries that are important either globally or in the Middle East. To address this question, we compared traffic about the United States to traffic about six other countries: Iran, Israel, Turkey, India, Russia, China, and the United Kingdom. The results of this analysis are found in table 9. The United States is the most important focus of this Twitter traffic, an entire quarter of which was about the United States. However, other influential non-Arab countries that regularly get involved in Middle Eastern affairs commanded a large share of this traffic as well: Iran with 17.51 percent, Israel with 17.28 percent, and Turkey with 13.84 percent.

The large number of tweets about Iran enables us to ask whether negative attitudes toward the United States are unique or whether, on the contrary, they extend to other powerful states with different cultural values and political interests. To answer this question we conducted an analysis on Iran that is similar to the one we described earlier on general anti-Americanism. While the United States and Iran are both regarded as enemies, Arabic speaking tweeters appear to have distinct and cognitively consistent reasons to dislike both states, since both states are seen to interfere with their countries. So it does not require cognitive distortion to dislike both the United States and Iran.

As table 10 shows, non-neutral traffic referring to Iran was overwhelmingly negative. Indeed, we found so few tweets that were either positive, or related to Iranian society, that we were unable even to estimate the proportions of these categories. Hence there is indeed a contrast between feelings toward Iran and the United States, but it seems to favor the United States. Our findings reinforce the view taken by Peter J. Katzenstein and Robert O. Keohane, and Giacomo Chiozza, on the basis of public opinion polling, that views toward the United States elsewhere in the world are highly ambivalent. This ambivalence reflects the “polyvalence” of American symbols: “they embody a variety of values with different meanings to different people and indeed even to the same individual.”<sup>42</sup> In both the polling data and even more in our

Twitter data, one observes admiration for American popular culture, helping to create such ambivalence. There is no such Arab admiration for Iranian popular culture, and no discernible ambivalence.

### **Anti-Americanism, Anti-Interventionism, and the Politics of Social Media Discourses**

Anti-Americanism is an important political phenomenon and the subject of an extensive literature, with essentially all of the previous quantitative work relying exclusively on public opinion surveys and seeking to ascertain attitudes. In contrast, we focus on social media, specifically Twitter Arabic language feeds. We monitor millions of Arabic-language messages during 2012–13 on Twitter, enabling us to delineate the discourses that take place on social media in the Arabic-speaking world. Our aggregate monitor finds that discourses with respect to both political and social issues are overwhelmingly negative but that the volume of political traffic is four to five times as great as the volume of social traffic. Consistent with the findings of public opinion polls, distrust of the United States Government in the Middle East is deep.

Our analysis of Twitter data has allowed us to revisit many of the salient debates on anti-Americanism. First, with a high degree of confidence, we show that the primary focus of anti-Americanism in the Arabic Twitter

**Table 10**  
**Total estimated number of tweets in each category for Iran monitor**

Category	Frequency	Percentage
Negative political	12,141,844	28
Neutral political	30,510,853	72

Note: The number of tweets here is higher than in table 9 because references to “Shia” or “Shiite” were included in the Iran monitor.

universe is more about international politics than about domestic social norms. This finding has two important implications. First, much of the literature on anti-Americanism as driven by domestic social norms has posited that Arab societies, or even Muslim societies more generally, dislike the United States (and other Western countries) for the values that they express internally. If cultural distaste dictated high levels of anti-Americanism, then only fundamental cultural or normative change could provide a remedy to such intolerance. Direct policies would have to be aimed at inducing or supporting cultural shifts at the societal level. Such possible policies would include civil society initiatives to promote a democratic ethos, curricular development that emphasizes tolerance and western liberal values, the encouragement of “moderate” voices to assume the pulpit, and the strengthening of youth “liberal” voices on Arab streets. Their purpose would be to encourage the evolution of society in more liberal-democratic directions, generating also a more pro-American stance among Arab societies.

However, our results, especially as summarized in table 8, indicate that levels of anti-Americanism are primarily driven by the perceived impingement of America on the Middle East, and specifically by United States intervention in the region. We have further demonstrated that anti-American sentiment focuses around events in which the United States is seen as playing a major role. Overwhelmingly, citizens of the Arab world distrust the intentions of the United States: regardless of what the United States does, Arab publics will evaluate the United States through a deeply suspicious lens. As long as the United States continues to intervene militarily in the Middle East, its actions will generate resentment. Iranian intervention generates similar resentment.

Our analysis of attitudes toward Iran has led us to interrogate the very concept of anti-Americanism. The literature on anti-Americanism has been written overwhelmingly by Americans with an exclusive focus on the United States, and we began our analysis with the same mind-set. The United States has actively intervened in world politics for the last 73 years, so it is not surprising that it generates strong feelings. But in what sense should we view these feelings as distinctively “anti-American” rather than being merely expressions of resentment towards interference in one’s own affairs by a powerful state with a distinctly different set of cultural values and political interests? The concept of a distinct “anti-Americanism” has been reified by a polling technology that asked people around the world what they think of the United States, not their views of other powerful countries that are perceived to intervene in their affairs.

In light of our findings about Iran, political scientists should examine social media discourses and public opinion toward countries other than the United States to see whether anti-Americanism is a highly specific or even

a unique phenomenon, centered on the United States; or merely one aspect of a more general phenomenon: opposition to interference and influence by powerful states with different cultural values and political interests. Our analysis of Twitter data and public opinion analysis suggests that the latter interpretation is likely to be correct.

Arab anti-Americanism appears to be a specific version of a more general phenomenon: dislike and distrust of the impingement of other societies with different political and social values on the Arab Middle East. Seventy-five years ago these views may have been common toward Britain and France; but as the influence of these states has diminished, publics around the world no longer think of tweaking the tail of the British lion or calling out French republicans for their imperialist hypocrisy. Anti-Americanism arose in the wake of the rise of America to world power, and was strongest in areas such as the Middle East where intervention by the United States was particularly intrusive; regional antagonism to Iran is high in the wake of active Iranian intervention in Syria and neighboring countries. If China begins to intervene on a global basis, it may be next.

## Implications for Political Science and World Politics

Our analysis also raises more general questions, not limited to the study of anti-Americanism, which are relevant to the study of what we have called social globalism. Social media create a remarkable window not only into attitudes, as expressed by individuals, but into discourses. We can well imagine analyses that would try to identify different discourses on a set of topics, and their connections, and their disconnects. Twitter users with more moderate or extreme views may have different networks of whom they follow and who follows them.

These discourses are socially constructed and have their own dynamics. There may be considerable path-dependence, with initial themes helping to determine future themes: this is another set of questions worth analyzing. Social scientists should build on excellent existing work about how these discourses affect politics in the street or policy-making in governments.<sup>43</sup> To what extent do these discourses shape social movements and how closely are they responded to by policy-makers? Furthermore, how does social media engagement relate to other forms of expression, such as communications to leaders, commentary in newspapers, or informal discussions?<sup>44</sup> More broadly, political discussion includes people saying they like something, repeating what they heard, and replying to each other. But these actions on Twitter might engender substantively different dynamics from face-to-face interaction. The politics of social media discourses have created a subject that is ripe for innovative social scientific analysis.

Beyond the opportunities that social media discourses provide for social science, they raise major questions for the study of contemporary world politics. As we note in the introduction, the fact that voices of ordinary people can be heard, worldwide, is a new feature of contemporary globalism. It enables people with minority views within their own communities to find a broader community of sympathizers and to engage in debate. It also enables militant organizations, such as the Islamic State of Iraq and Syria (ISIS), to recruit, on a global basis, seeking to entice people who have never known anyone from ISIS to join these organizations. These organizations become focal points—both on social media and in the world of inter-personal politics and military action—for the expression of dissent and alienation. Prevailing high levels of anti-Americanism on social media may provide validation for decisions by some individuals to act against the United States, by joining ISIS or otherwise.

Military and organizational strength—whether on the part of Iran or the United States—does not enable modern states to control expressions on social media sites abroad, although some states such as China are quite effective in controlling domestic media sites.<sup>45</sup> On the contrary, the exertion of military power by these states to control events or influence political outcomes generates opposition in society that at least partially counterbalances their hard power. The United States or Iran may be able to influence the state-run media of sympathetic countries, but the extensive horizontal communication on social media reduces the dominance of state-run media and the efficacy of strategies that rely on them. Old-style intervention threatens to generate hostility that reverberates through social media, which amplifies reactions to every military action.

The dynamic and expansive character of social globalism, fostered by the internet and social media, does not imply that states have suddenly “lost control.” As Stephen Krasner has argued, states remain the most important actors in world politics and have historically been resilient, with state activity increasing along with economic and social interdependence.<sup>46</sup> But social media, and the discordant discourses that they engender, do generate new challenges for states, and particularly the United States, as they seek to develop strategies for effective action in world politics.

### **Online Supplemental Materials: “Anti-Americanism and Anti-Interventionism in Arabic Twitter Discourses”**

- Overview: Aggregate Twitter Volume
- The Hopkins-King “ReadMe” Method
- London Monitor Results
- Monitor Training Details

<http://dx.doi.org/10.7910/DVN/28171>

### **Notes**

- 1 Keohane and Nye 2000. The authors differentiate economic, military, environmental, and social globalism.
- 2 Keohane 2001.
- 3 Howard, 2011, Gainous and Wagner 2014, Gainous and Wagner 2013, Lynch 2011, Tarrow 2013, Bennett and Segerberg 2014, Norris 2001.
- 4 Katzenstein and Keohane 2007.
- 5 Chiozza 2009, table 3.1, 55; figures 4.2 and 4.3, 98 and 101.
- 6 Katzenstein and Keohane 2007.
- 7 Dubai School of Government 2013.
- 8 Esposito 2011.
- 9 Jamal 2012a.
- 10 See Mossberger et al. 2008.
- 11 Tarrow 2013; Gainous and Wagner 2014; O’Conner et al. 2010.
- 12 We note that since tweets are attributed to their authors, some tweeters may exhibit social desirability bias—expressing views that they want to be seen to have, rather than their own private views.
- 13 Joffe 2006, Ajami 2003, Lewis 1990, Huntington 1996.
- 14 Blaydes and Linzer 2012.
- 15 Gerges 1996, Jamal 2012b, Telhami 2002, Tessler 2003, Baxter and Akbarzadeh 2008, Makdisi 2002, Khalidi 2004, Dawisha 1985, Esposito 1998, Haddad 1996.
- 16 Chiozza 2007, Tessler 2003.
- 17 Hopkins and King 2010.
- 18 See Grimmer and Stewart 2013. For a more focused introduction to computer assisted methods of text analysis for comparative politics, see Lucas et al. forthcoming.
- 19 Blei 2012; Blei and Lafferty 2007; Blei; Ng, and Jordan, 2003; Robert et al. 2014.
- 20 Laver, Benoit, and Garry 2003; Hopkins and King 2010.
- 21 See <http://www.crimsonhexagon.com/>.
- 22 For details please refer to the online supplementary materials.
- 23 The user is limited to 5,000 characters for keyword restrictions, requiring the user to make judgments about which keywords are more or less important; these judgment calls are discussed in the online supplementary materials.
- 24 We placed tweets that do not fit into any category into an “irrelevant” category, so that the frequencies for the on-topic categories are not affected by this traffic. Refer to the online supplementary materials.
- 25 See Grimmer and Stewart 2013.
- 26 Jamal 2012b; Chiozza 2007; Lynch 2007.

- 27 We used verification methods as discussed in the online supplementary materials. As explained there, if our results are biased, they probably overestimate the volume of the Positive Political category—which is, relatively speaking, quite small.
- 28 This is evident over time as well. Throughout our time period the negative political conversation almost always dominates the other categories.
- 29 These monitors differ from the General AA monitor in three key ways. The time range for each of these monitors, except the Syria monitor, is significantly shorter. The category creation process for the event monitors was often inductive rather than defined *ex ante*, as was the case in the general monitor. We wanted to look at the conversation surrounding these events in detail rather than applying a specific pre-defined set of categories to them. Finally, we either used sets of keywords that were specific to the event (Boston, Sandy, and Innocence of Muslims monitors) or a combination of U.S. keywords from the General AA monitor and event-specific keywords (Egypt and Syria monitors) rather than just U.S. keywords.
- 30 For the United States, we used the general list of anti-American keywords with some minor alterations; for Egypt, we used four terms: “Egypt/Egyptian,” “Morsi,” “The Brotherhood,” and “Sisi” (the Commander-in-chief of Egypt’s armed forces). Refer to the online supplementary materials for details.
- 31 Our ability quickly to analyze Twitter reactions to the Syrian crisis contrasts with the fact that, as far as we know, there have been no polls conducted *in Arabic* anywhere on this issue.
- 32 We also ran a Libya monitor from February 1 to October 31, 2011, during the NATO campaign that overthrew the Qaddafi regime; details for this monitor can be found in the online supplementary materials. Caution is required since the number of tweets was dramatically lower then: apart from news and neutral items, fewer than 20,000 tweets mentioned Libya and the United States. The ratio of negative to positive comments was over 3:1, indicating high negativism toward U.S. intervention even though the United States was aligned with some Arab countries, such as Qatar in opposing the Qaddafi dictatorship. Yet this ratio is much lower than the 30:1 ratio in our Syria monitor. U.S. alignment with Arab regimes seems to have ameliorated anti-American sentiment but not to have reversed it.
- 33 For the list, refer to the online supplementary materials.
- 34 Although there is little directly negative traffic in the Innocence of Muslims monitor, the General AA monitor shows a spike in negative traffic during this time period. We checked this discrepancy by creating a monitor using the same keywords as the General AA monitor, around the spike between September 10–17, in which we found quite a bit of negative traffic, much of which is political.
- 35 Katzenstein and Keohane 2007.
- 36 Refer to the online supplementary materials for these results.
- 37 Lynch 2007, 223.
- 38 Berger 2014.
- 39 Jamal 2012b.
- 40 See also Katzenstein and Keohane 2007, 34–35.
- 41 Our inference that Arabic anti-Americanism principally reflects fear and distrust of the impact of the United States on the Arab world is admittedly a matter of interpretation, and subject to considerable uncertainty, but it seems to us a plausible interpretation and consistent with polling data.
- 42 Katzenstein and Keohane 2007, 317.
- 43 Tarrow 2013; Gainous and Wagner 2014; Howard 2011.
- 44 Cramer Walsh 2007, Lee 2002, Lynch 2011.
- 45 King, Pan, and Roberts 2013.
- 46 Krasner 1999, 223.

## References

- Ajami, Fouad. 2003. “The Falseness of Anti-Americanism.” *Foreign Policy* 138: 52–61.
- Almond, Gabriel A. 1950. *The American People and Foreign Policy*. New York: Harcourt, Brace and Company.
- Baxter, Kylie, and Shahram Akbarzadeh. 2008. *US Foreign Policy in the Middle East: The Roots of Anti-Americanism*. London: Routledge.
- Bennett, Lance, and Alexandra Segerberg. *The Logic of Connective Action: Digital Media and the Personalization of Contentious Politics*. New York: Cambridge University Press.
- Berger, Lars. 2014. “Foreign Policies or Culture: What Shapes Muslim Public Opinion on Political Violence against the United States?” 2014. *Journal of Peace Research* 51(6): 782–796.
- Blaydes, Lisa, and Drew A. Linzer. 2012. “Elite Competition, Religiosity, and Anti-Americanism in the Islamic World.” *American Political Science Review* 106(2): 225–43.
- Blei, David M. 2012. “Probabilistic Topic Models.” *Communications of the ACM* 55(April): 77–84.
- Blei, David M., and John D. Lafferty. 2007. “A Correlated Topic Model of Science.” *Annals of Applied Statistics* 1(June): 17–35.
- Blei, David M., Y. Ng Andrew, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning* 3(March): 993–1022.
- Chiozza, Giacomo. 2007. “Disaggregating Anti-Americanism: An Analysis of Individual Attitudes toward the United States.” In *Anti-Americanisms in World Politics*, ed. Peter J. Katzenstein and Robert O. Keohane. Ithaca, NY: Cornell University Press.

- \_\_\_\_\_. 2009. *Anti-Americanism and the American World Order*. Baltimore, MD: Johns Hopkins University Press.
- Cramer Walsh, Katherine 2007. *Talking about Politics: Informal Groups and Social Identity in American Life*. Chicago, IL: University of Chicago Press.
- Dawisha, Adeed. 1985. "Anti-Americanism in the Arab World: Memories of the Past and Attitudes of the Present." In *Anti-Americanism in the Third World: Implications for U.S. Foreign Policy*, ed. Alvin Z. Rubinstein and Donald Eugene Smith. New York: Praeger.
- Dubai School of Government. 2013. "Arab Social Media Report," [http://www.arabsocialmediareport.com/UserManagement/PDF/ASMR\\_5\\_Report\\_Final.pdf](http://www.arabsocialmediareport.com/UserManagement/PDF/ASMR_5_Report_Final.pdf)
- Esposito, John L. 1998. *Islam and Politics*. Syracuse: Syracuse University Press.
- \_\_\_\_\_. 2011. "Arab Youth Want Democracy, not Theocracy." CNN, February 28, <http://www.cnn.com/2011/OPINION/02/28/protests.democracy.islam/>.
- Gainous, Jason, and Kevin Wagner. 2013. "Digital Uprising: The Internet Revolution in the Middle East." *Journal of Information Technology and Politics* 10: 3.
- \_\_\_\_\_. 2014. *Tweeting to Power: The Social Media Revolution in American Politics*. London: Oxford University Press.
- Gerges, Fawaz A. 1996. "The 1967 Arab-Israeli War: U.S. Actions and Arab Perceptions." In *The Middle East and the United States: A Historical and Political Reassessment*, ed. David W. Lesch. Boulder, CO: Westview Press.
- Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (January): 1–31.
- Haddad, Yvonne Yazbeck. 1996. "Islamist Perceptions of U.S. Policy in the Middle East." In *The Middle East and the United States: A Historical and Political Reassessment*, ed. David W. Lesch. Boulder, CO: Westview Press.
- Hopkins, Daniel, and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1): 229–47.
- Howard, Philip. 2011. *The Digital Origins of Dictatorship and Democracy: Information Technology and Political Islam*. London: Oxford University Press.
- Huntington, Samuel P. 1996. *The Clash of Civilizations and the Remaking of World Order*. New York: Simon & Schuster.
- Jamal, Amaney. 2012a. "The Youth and the Arab Spring: Cohort Differences and Similarities." *Middle East Law and Governance* 4(1): 168–88.
- \_\_\_\_\_. 2012b. *Of Empires and Citizens: Pro-American Democracy or No Democracy at All?* Princeton, NJ: Princeton University Press.
- Joffe, Josef. 2006. *Überpower: The Imperial Temptation of America*. New York: W.W. Norton & Co.
- Katzenstein, Peter J., and Robert O. Keohane 2007. "Varieties of Anti-Americanism: A Framework for Analysis." In *Anti-Americanisms in World Politics*, ed. Peter J. Katzenstein and Robert O. Keohane. Ithaca, NY: Cornell University Press.
- Keohane, Robert O. 2001. "Governance in a Partially Globalized World." *American Political Science Review* 95(1): 1–13.
- Keohane, Robert O., and Joseph S. Nye Jr. 2000. "Governance in a Globalizing World." In *Governance in a Globalizing World*, ed. Joseph S. Nye Jr., and John D. Donahue. Washington, DC: Brookings Institution.
- Khalidi, Rashid. 2004. *Resurrecting Empire: Western Footprints and America's Perilous Path in the Middle East*. Boston: Beacon Press.
- King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(2): 326–43.
- Krasner, Stephen D. 1999. *Sovereignty: Organized Hypocrisy*. Princeton, NJ: Princeton University Press.
- Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97 (2): 311–31.
- Lee, Taeku 2002. *Mobilizing Public Opinion: Black Insurgency and Racial Attitudes in the Civil Rights Era*. Chicago, IL: University of Chicago.
- Lewis, Bernard. 1990. "The Roots of Muslim Rage." *Atlantic Monthly* 266(3): 47–60.
- Lucas, Christopher, Richard Nielsen, Margaret Roberts, Brandon Stewart, Alex Storer, and Dustin Tingley. forthcoming. "Computer Assisted Text Analysis for Comparative Politics." *Political Analysis*.
- Lynch, Marc. 2007. "Anti-Americanism in the Arab World." In *Anti-Americanisms in World Politics*, ed. Peter J. Katzenstein and Robert O. Keohane. Ithaca, NY: Cornell University Press.
- \_\_\_\_\_. 2011. "After Egypt: The Limits and Promise of Online Challenges to the Authoritarian Arab State." *Perspectives on Politics* 9(2): 301–310.
- Makdisi, Ussama. 2002. "'Anti-Americanism' in the Arab World: An Interpretation of a Brief History." *Journal of American History* 89(September): 538–57.
- Mossberger, Karen, Caroline J. Tolbert, and Ramona S. McNeal. 2008. *Digital Citizenship : The Internet, Society, and Participation*. Cambridge, Mass: MIT Press.

- 
- Norris, Pippa. 2001. *Digital Divide: Civic Engagement, Information Poverty and the Internet Worldwide*. New York: Cambridge University Press.
- O'Connor et al. 2010 – see n.11
- Robert, Margaret, Brandon Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Gadarian, Bethany Albertson, and David Rand. 2014. “Structural Topic Models for Open-Ended Survey Responses.” *American Journal of Political Science* 58(4): 1064–82.
- Tarrow, Sydney 2013. *The Language of Contention: Revolutions in Words, 1688–2012*. New York: Cambridge University Press.
- Telhami, Shibley. 2002. *The Stakes: America and the Middle East*. Boulder, CO: Westview Press.
- Tessler, Mark. 2003. “Arab and Muslim Political Attitudes: Stereotypes and Evidence from Survey Research.” *International Studies Perspectives* 4(May): 175–81.

# Méthodes non-supervisées



# Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It\*

Matthew J. Denny<sup>†</sup>      Arthur Spirling<sup>‡</sup>

## Abstract

Despite the popularity of unsupervised techniques for political science text-as-data research, the importance and implications of preprocessing decisions in this domain have received scant systematic attention. Yet, as we show, such decisions have profound effects on the results of real models for real data. We argue that substantive theory is typically too vague to be of use for feature selection, and that the supervised literature is not necessarily a helpful source of advice. To aid researchers working in unsupervised settings, we introduce a statistical procedure and software that examines the sensitivity of findings under alternate preprocessing regimes. This approach complements a researcher's substantive understanding of a problem by providing a characterization of the variability changes in preprocessing choices may induce when analyzing a particular dataset. In making scholars aware of the degree to which their results are likely to be sensitive to their preprocessing decisions, it aids replication efforts.

preText software available: [github.com/matthewjdenny/preText](https://github.com/matthewjdenny/preText)

Word count: 10461 (excluding online appendices)

---

\*First version: September, 2016. This version: September 27, 2017. We thank Will Lowe and Brandon Stewart for comments on an earlier draft, and Pablo Barbera for providing the Twitter data used in this paper. Replication data for this paper are available on the Political Analysis Dataverse here: [dx.doi.org/10.7910/DVN/XRR0HM](https://doi.org/10.7910/DVN/XRR0HM).

<sup>†</sup>mdenny@psu.edu; 203 Pond Lab, Pennsylvania State University , University Park, PA 16802

<sup>‡</sup>arthur.spirling@nyu.edu; Office 405, 19 West 4th St., New York University, New York, NY 10012

# 1 Introduction

Every quantitative study that uses text as data requires decisions about how words are to be converted into numbers. These decisions, known collectively as ‘preprocessing’, aim to make the inputs to a given analysis less complex in a way that does not adversely affect the interpretability or substantive conclusions of the subsequent model. In practice, perfecting this tradeoff—simpler data, but not too much information loss—is a non-trivial matter, and scholars have invested considerable energies in exploring the optimal way to proceed (see, e.g. Sebastiani, 2002, for a review). Unsurprisingly, such advice, which includes the merits of operations like decapitalization, pruning words back to their stems and removing very common words, can be found in textbooks for natural language processing and information retrieval (e.g. Jurafsky and Martin, 2008; Manning, Raghavan and Schütze, 2008). Subsequently, political scientists have suggested scholars in their field employ similar steps (e.g. Grimmer and Stewart, 2013).

On its face, this technology transfer from computer science to political science has much to recommend it. Clearly, political texts—be they treaties, manifestos, speeches or press releases—are not so different in substance or style to non-political texts—such as product or movie reviews—as to imply that such advice is *a priori* inappropriate. And, models that employ such preprocessing steps have been successful insofar as they are very widely cited, provide valid measures, and produce findings in keeping with qualitative understandings of fundamental political processes (e.g. Monroe, Colaresi and Quinn, 2008; Slapin and Proksch, 2008; Quinn et al., 2010). But as we explain in this paper, there are reasons for extreme caution when moving from one field to another, completely separate to matters of different substantive focus.

Ideally, as in any scientific measurement problem, feature selection decisions in political science ought to be based on ‘theory’, broadly construed. That is, researchers should match their preprocessing choices to their knowledge—in terms of what is likely to be important for understanding the data generating process—of the substantive matter at hand.<sup>1</sup> Our central contribution here recognizes, and is entirely compatible with, the primacy of this position. We note that in practice however, this is not how business is done. For one thing, faced with an area of study not yet widely examined with text analysis, scholars may have weak, possibly wrong, priors over what will be an important feature to include or discard, to weight up or to weight down. Thus, they simply follow and cite the extant literature, and the decisions taken there, with little understanding for how different preprocessing might affect their conclusions. Arguably worse, some may try a few specifications and report (only!) the one that returns results closest to their expectations, with the obvious consequences for reproducibility that such ‘cherry picking’ usually delivers.

This operating procedure may be innocuous if it were not for the fact that, ultimately, the providence of much preprocessing advice is the world of *supervised* techniques, yet it is applied in the world of *unsupervised* techniques. And there is worryingly little discussion about whether the shift from one form of learning, where effective classification is the goal, to another—where the goal is to reveal interesting latent structure—has consequences for the best approach to preprocessing. Because unsupervised learning typically requires careful and deep interpretation of results after a technique has been applied, scholars using such models have relatively little room—intellectually or physically in terms of time—to discuss what alternate specifications of the preprocessing steps would have suggested. The dimen-

---

<sup>1</sup>Consider a trivial case: one’s documents might contain no numbers, in which case it makes no practical difference whether one ‘removes’ them or not. More generally though, we may know from past experience of a document collection what is likely to reveal the structure we care about: an obvious example is the removal of certain words—like ‘Right Honourable’—which are essentially mandated and which we believe add nothing to our understanding of a speech’s content in the House of Commons.

sions of this problems are stark: notice that for just seven possible (binary) preprocessing steps, there would be  $2^7 = 128$  possible models to run and analyze (and that's before any model parameters, such as the number of clusters or topics, are adjusted).

Of course, in principle, it could be the case that what ‘works’ for supervised learning is fine for unsupervised problems, and that findings are anyway generally robust. Sadly, there is no *a priori* reason to believe this, and as we will show below, the idea that conclusions are not sensitive to perturbations of the preprocessing steps is wishful thinking. Curiously though, our paper is the first that we know of that explores exactly this question in the context of unsupervised approaches to social science text data.

By looking at two real datasets, we demonstrate that the inferences one draws can be extremely sensitive to the preprocessing choices the researcher makes. This ‘possibility result’ means that otherwise diligent researchers are in danger of drawing highly variable lessons from their documents, depending on the particular specification of preprocessing steps they adopt. Further, by transforming their text data in a given way and then substantively interpreting their post-model results without considering the patterns and differences that would have emerged, scholars can find themselves heading down “forking paths” of inference based on early data coding—i.e. preprocessing—decisions (see Gelman and Loken, 2014, S3 for discussion of this idea). More worryingly, researchers with malfeasant intent are able to try multiple different specifications until they find one that fits their preference or theory (known elsewhere as ‘fishing’).

With the above in mind, our second contribution is to provide a convenient method for assessing the sensitivity of inference for unsupervised models in the face of a large number of possible preprocessing steps. We describe, design and implement a measure based on

the way that pairwise distances between documents move as one tries alternative specifications. Helpfully, this allows us to make some general comments about how harmful—in the sense of how ‘unusual’ the resulting document term matrix (DTM) is relative to all other possibilities—a given choice might be. To be very clear, we envision that the typical use case of our technique is as a *complement*, not a substitute, for researchers’ domain-specific understandings about their data. That is, if a researcher has ‘theory’ about what should or should not be done in terms of pre-processing which still allows for doubt over exactly what is an optimal specification—which we maintain is the case for the vast majority of practitioners—our method will allow them to see how robust their findings are likely to be to reasonable perturbations of those choices. To reiterate, it is neither our aim, nor a well-defined objective, to provide the ‘right answer’ for preprocessing decisions: instead, we show how researchers can avoid getting a possibly ‘wrong answer’ with a method that provides the equivalent of a warning. And if researchers follow the procedure we lay out—made simple via our software—they can feel considerably more confident about the robustness of their findings under different transformations of their DTMs.

In the next section, we review some common text preprocessing choices the consequences of which we will investigate in more detail. We then review the differences between supervised and unsupervised methods, and why advice from the former need not apply to the latter. Moving to more practical matters, we then briefly describe the (representative) data sets we will operate on for the rest of the paper. This is followed by our troubling examples, in which we show inference is highly variable depending on small differences in preprocessing. We then move to our more general testing approach and present some advice for practitioners working with unsupervised models. The final section concludes.

## 2 Words to Numbers: Text Preprocessing Choices

Quantitative analysis requires that we transform our texts into numerical data. Accepting the wisdom that word order may be disposed of with minimal costs for inference (see Grimmer and Stewart, 2013, for discussion)—and a ‘bag of words’ representation employed—researchers typically apply (some subset of) several further binary preprocessing steps in constructing the relevant document term matrix. We now describe these in some detail, since they are the focus of our efforts below.

**P Punctuation:** The first choice a researcher must make when deciding how to preprocess a corpus is what classes of characters and markup to consider as valid text. The most inclusive approach is simply to choose to preprocess all text, including numbers, any markup (html) or tags, punctuation, special characters (\$, %, &, etc), and extra white-space characters. These non-letter characters and markup may be important in some analyses (e.g. hashtags that occur in Twitter data), but are considered uninformative in many applications. It is therefore standard practice to remove them. The most common of these character classes to remove is punctuation. The decision of whether to include or remove punctuation is the first preprocessing choice we consider.

**N Numbers:** While punctuation is often considered uninformative, there are certain domains where numbers may carry important information. For example, references to particular sections in the U.S. Code (“Section 423”, etc.) in a corpus of Congressional bills may be substantively meaningful regarding the content legislation. However, there are other applications where the inclusion of numbers may be less informative.

**L Lowercasing:** Another preprocessing step taken in most applications is the lowercasing of all letters in all words. The rationale for doing so is that whether or not the first letter of a word is uppercase (such as when that word starts a sentence)

most often does not affect its meaning. For example, “Elephant” and “elephant” both refer to the same creature, so it would seem odd to count them as two separate word types for the sake of corpus analysis. However, there are some instances where a word with the same spelling may have two different meanings that are distinguished via capitalization, such as “rose” (the flower), and “Rose” the proper name.

**S Stemming:** The next choice a researcher is faced with in a standard text preprocessing pipeline is whether or not to stem words. Stemming refers to the process of reducing a word to its most basic form (Porter, 1980). For example the words “party”, “partying”, and “parties” all share a common stem “parti”. Stemming is often employed as a vocabulary reduction technique, as it combines different forms of a word together. However, stemming can sometimes combine together words with substantively different meanings (“college students partying”, and “political parties”), which might be misleading in practice.

**W Stopword Removal:** After tokenizing the text, the researcher is left with a vector of mostly meaningful tokens representing each document. However, some words, often referred to as “stop words”, are unlikely to convey much information. These consist of function words such as “the”, “it”, “and”, and “she”, and may also include some domain-specific examples such as “congress” in a corpus of U.S. legislative texts. There is no single gold-standard list of English stopwords, but most lists range between 100 and 1,000 terms.<sup>2</sup> Most text analysis software packages make use of a default stopword list which the software authors have attempted to construct to provide “good performance” in most cases. There are an infinite number of potential stopword lists, so we restrict our attention to the choice of whether to remove words on the default list provided by the `quanteda` software package in R.

---

<sup>2</sup>See, for example: <http://www.ranks.nl/stopwords>

**3 n-gram Inclusion:** While it is most common to treat individual words as the unit of analysis, some words have a highly ambiguous meaning when taken out of context. For example the word “national” has substantially different interpretations when used in the multi-word expressions: “national defense”, and “national debt”. This has lead to a common practice of including *n*-grams from documents where an *n*-gram is a contiguous sequence of tokens of length *n* (Manning and Schütze, 1999). For example, the multi-word expression “a common practice” from the previous sentence would be referred to as a 3-gram or tri-gram (assuming stopwords were not removed). Extracting *n*-grams and adding them to the DTM can improve the interpretability of bag-of-terms statistical analyses of text, but also tends to lead to an explosion in the vocabulary size, due to the combinatorial nature of *n*-grams. Previous research has tended to use 1,2, and 3-grams combined, because this combination offers a reasonable compromise between catching longer multi-word expressions and keeping the vocabulary relatively smaller. After extracting all n-grams from a document, a number of approaches have been proposed to filter the resulting *n*-grams (Justeson and Katz, 1995), but here we choose to focus only on the most basic case of considering all 1,2, and 3-grams together without any filtering. So, the decision of whether include 2 and 3-grams (along with unigrams, which are always included) is the sixth preprocessing choice we consider.

**I Infrequently Used Terms:** In addition to removing common stopwords, researchers often remove terms that appear very infrequently as part of corpus preprocessing. The rationale for this choice is often two-fold; (1) theoretically, if the researcher is interested in patterns of term usage across documents, very infrequently used terms will not contribute much information about document similarity. And (2) practically, this choice to discard infrequently used terms may greatly reduce the size of the vocabulary, which can dramatically speed up many corpus analysis tasks. A commonly used rule of thumb is to discard terms that appear in less than 0.5-1% of documents (Grimmer,

2010; Yano, Smith and Wilkerson, 2012; Grimmer and Stewart, 2013), however, there has been no systematic study of the effects this preprocessing choice has on downstream analyses. The decision of whether include or remove terms that appear in less than 1% of documents is the seventh and final preprocessing choice we consider.

For reasons of notational sanity, we will refer to these steps via the characters we use as bullet markers. In particular, a string of such characters will describe what has been done to a given set of documents. Thus, N-L-S-3-I means that numbers were removed, then the document was lower-cased, then stemmed, then bigrams and trigrams were included and then infrequent terms were removed. In this case, the document did not have punctuation removed, nor stop words. We note that the order in which these steps are applied can have a substantial effect on the final DTM (such as stemming before or after removing infrequent terms). Therefore, in practice we will always apply (or not apply) the steps in the order of the bulleted items though, as we will see, other authors proceed in different orders. Together, these seven binary preprocessing decisions<sup>3</sup> lead to a total of  $2^7 = 128$  different possible combinations meaning a total of 128 DTMs, one of which is the original DTM with no preprocessing, and the other 127 involve at least one step.<sup>4</sup>

---

<sup>3</sup>Many of these decisions are, of course, not binary in the strictest sense. For example, there are an infinite number of ngram combinations one might consider (1 and 2 grams, 2 through 4 grams, etc.). We choose to treat these decisions as binary for practical reasons, and because most users of preprocessing software will select a binary argument to turn each of these steps on or off.

<sup>4</sup>Theoretically, if we were to permute the order in which the steps were applied, this would result in a much larger number of possible combinations. However, many of these permutations would be inconsequential (such as whether to remove punctuation or numbers first). We believe the correct approach in most cases (and the one we take) is to simply apply the steps in the default order of the preprocessing software one is using.

### 3 Text Preprocessing as Feature Selection: Supervised vs Unsupervised Approaches

To reiterate a point we made above, whether one uses a particular preprocessing step or not should be guided by substantive knowledge. For instance, if one is working with legal data, there may be a strong *a priori* justification for using bigrams and trigrams such that, e.g., “*Roe v. Wade*” is regarded as a unified token worth counting. In practice, we observe scholars following previous work, without much theoretical basis to form an independent justification for the case at hand. And to the extent that the much emulated (see Table 1) pioneering pieces in the discipline justify their own decisions, it is typically via experiences from one set of *supervised* techniques that do not necessarily apply to another set—*unsupervised*. In the supervised case, which is relatively rare in political science applications (though see e.g. Laver, Benoit and Garry, 2003; Hopkins and King, 2010; Diermeier et al., 2011; D’Orazio et al., 2014; King, Lam and Roberts, 2017), researchers have a set of hand-labeled training documents and they wish to learn the relationship between the features (e.g. terms) those texts contain and the labels they were given.<sup>5</sup> This is helpful because it allows scholars to automatically and rapidly classify new documents into the classes they care about. By contrast, unsupervised methods—for example topic models—do not require the researcher to provide pre-labeled documents. Instead, the model or technique reveals hidden structure within and between documents, and it is the job of the analyst to then interpret that information in a way that is substantively informative.

In the supervised context, preprocessing—more commonly referred as ‘feature selection’—is done for three primary reasons (Manning, Raghavan and Schütze, 2008; James et al., 2013):

---

<sup>5</sup>For a further discussion of the use of supervised and unsupervised methods for text analysis, see Online Appendix A.

first, because it reduces the size and complexity of the vocabulary which will act as input to the prediction process. This can substantially cut the computational time it takes to learn a relationship in the data for obvious reasons: for a given number of documents, learning how their class is predicted from the 10,000 different terms they contain is likely to be slower than understanding the relationship between their class and, say, 100 different terms. Second, preprocessing reduces the number of irrelevant ('noise') features, the presence of which can make classification actively worse. The particular threat here is the inclusion of 'unhelpful' (often rare) terms will mean that new documents will be misclassified as a product of the technique 'overfitting' to such words in the training set. Thirdly, preprocessing makes models easier to interpret substantively because not only is the number of predictors reduced (mechanically lessening the workload for the researcher), but those that do remain are the more important ones for the problem at hand.

There are numerous feature selection methods in this literature, including mutual information and various  $\chi^2$  procedures along with more elaborate regularization procedures. Regardless of the method, in the supervised context, notice that whether or not a given preprocessing step is merited can be evaluated in a well-defined way. For example, applying a given transformation to the document vectors might improve or reduce the accuracy of a classifier—literally, the proportion of cases that the learner places in the correct class. This applies similarly to other measures of classifier effectiveness, including precision, recall or  $F_1$  score. Indeed, scholarly accounts of various areas of supervised learning include discussions of optimal feature selection on such grounds (see, e.g., Sebastiani, 2002, for a comprehensive review). In given substantive domains, such work allows researchers to conclude that, for example, "bigram information does not improve performance beyond that of unigram presence", when classifying movie reviews in terms of their sentiment (Pang, Lee and Vaithyanathan, 2002, 6).

The principle that one should seek to preserve what is informative while jettisoning what is redundant or irrelevant is a sensible one. But the way that this principle ought to be applied in an unsupervised setting is far from obvious.<sup>6</sup> This is because, for a start, we do not typically evaluate our unsupervised models in sharp, statistical effectiveness terms. While there are ways to determine the best fitting model from a set of candidates—for example, via a perplexity measure for topic modeling (e.g. Wallach et al., 2009)—they are predicated on, and have nothing directly to say about the merits of, a given set of preprocessing decisions which determine the precise format of the data to be fed to the algorithm. Instead, whether an unsupervised model is useful is typically a matter of determining whether the latent patterns it uncovers in otherwise complex data are interesting or substantively informative. For example, the model results might help us understand how legislators’ attention to issues varies over time (Quinn et al., 2010), how Senators vary in their priorities and credit claiming (Grimmer, 2010), how Japanese politicians increasingly focus on foreign policy (Catalinac, 2016) or how citizens differ in terms of the themes they emphasize in open-ended survey questions (Roberts et al., 2014). And this is not simply a matter for approaches, like topic models, that treat documents as a mixture of discrete elements or that assign texts to clusters. Indeed, unsupervised *scaling* models attempt to reveal latent traits or positions on a continuum without training on a subset of documents: thus, for example, Proksch and Slapin (2010) examine the changing ideological nature of German politics over time. In all cases, some form of validation is required (see Quinn et al., 2010, for an overview of methods); that is, scholars should carefully examine the output of their models and be sure it makes sense substantively and intuitively. But to underline the point, when undertaking such a task, there is no simple analogy to the fit statistics or effectiveness measures we mentioned above. Furthermore, specifically in the case of topic models, there is some evidence that what is

---

<sup>6</sup>To see why preprocessing matters at all, Online Appendix B gives some basic intuition.

Citation	Steps	Cites
Slapin and Proksch (2008)	P-S-L-N-W	427
Grimmer (2010)	L-P-S-I-W	258
Quinn et al. (2010)	P-L-S-I	275
Grimmer and King (2011)	L-P-S-I	109
Roberts et al. (2014)	P-L-S-W	117

Table 1: Preprocessing steps taken/suggested in recent notable papers that deal with unsupervised learning methods. The cite total is taken from Google Scholar at the time of writing. In the case of Slapin and Proksch (2008), we consulted their Wordfish manual (version 1.3). In the case of Roberts et al. (2014), the authors suggest further steps might be appropriate for a given application.

deemed to be a model that fits well statistically need not be one that is easily or sensibly interpretable (Chang et al., 2009).

Given this ambiguity, it is not surprising that scholars use (and by implication suggest) different steps in practice. In Table 1, we report the steps taken in some notable recent papers that use unsupervised learning methods. Clearly authors do different things (and in different orders) meaning that, for example, it is hard to know *a priori* whether one should or should not remove stop words, or whether one should or should not remove infrequent terms. Theoretically at least, one could imagine trying multiple different specifications and verifying that the substantive inferences (their usefulness and sensibleness) one draws are similar. In practice, with 128 different transformations at a minimum, this is no easy task especially considering that various other options (such as numbers of topics) would increase the implied model space significantly and prohibitively. Of course, it might be that such choices are generally inconsequential for the conclusions we would draw from typical social science datasets. In that context, inference would be consistent across specifications, and there is neither a danger that a researcher stumbles on to a transformation that is unrepresentative nor could they deliberately manipulate their results via such choices. Sadly, as we show shortly, this is false. Before doing that, we briefly describe the data we will use in the

rest of the paper.

## 4 Description of Datasets Used in Analyses

We make use of eight corpora in the analyses and examples presented in this study: they are described in Table 2, and are representative of data commonly used in the discipline.<sup>7</sup> The ‘UK Manifestos’ are the 69 manifestos released by the Conservative, Labour and Liberal party for every general election in the United Kingdom, 1918–2001. The ‘State of the Union Speeches’ are by the President of the United States and generally given annually, 1790–2016. They can be found on numerous websites. The ‘Death Row Statements’ are transcriptions of the last words recorded by the Texas Department of Criminal Justice for 439 inmates executed by the state between 1982 and 2009. They are available on the official website of that bureau. The ‘Indian Treaties’ are historic treaties signed between the United States government and various Indian tribes between 1784 and 1911 which fall into the categories “Valid and Operable”, “Ratified Agreements”, “Rejected by Congress” and “Unratified Treaties” as described by Spirling (2012). We then make use of a sample of 1,000 Senate press releases originally compiled by Grimmer (2010). The full dataset contains 72,785 press releases<sup>8</sup> which makes it prohibitively computationally costly to preprocess 128 different ways, so we randomly selected 10 Senators, and their 100 most recent press releases as our corpus. The ‘Congressional Bills’ are a sample of 300 bills introduced in the U.S. House of Representatives during the 113th session of Congress and originally collected by Handler et al. (2016). The ‘New York Times’ corpus comprises 494 articles of varying length, published between 1987 and 2007, sampled from the digital NYT archive. The Trump Campaign Tweets dataset contains 2,000 tweets written by Donald Trump between April and June 2015.<sup>9</sup>

---

<sup>7</sup>Replication data for this paper are available on the Political Analysis Dataverse here: [dx.doi.org/10.7910/DVN/XRR0HM](https://doi.org/10.7910/DVN/XRR0HM).

<sup>8</sup>The data are available for download here: [github.com/lintool/GrimmerSenatePressReleases](https://github.com/lintool/GrimmerSenatePressReleases)

<sup>9</sup>This dataset was originally compiled by Pablo Barbera and is [available here]

Corpus	Num. Docs	Word Types	Total Tokens	Tokens/Doc.
UK Manifestos	69	17,136	570,658	8,270
State of The Union Speeches	230	32,321	1,960,304	8,523
Death Row Statements	331	3,296	40,332	122
Indian Treaties	596	18,181	987,659	1,657
Congressional Press Releases	1,000	178,044	432,686	433
Congressional Bills	300	17,829	653,366	2,178
New York Times	494	30,759	312,389	632
Trump Campaign Tweets	2,000	7,466	46,056	23

Table 2: Corpus descriptive statistics for the five corpora used in our analyses. The number of word types, total tokens, and tokens per document were calculated using the no preprocessing specification.

## 5 What Could Possibly Go Wrong?

Our first claim is that while theory should guide our preprocessing choices, in practice, there is little concrete guidance for those embarking on an unsupervised analysis of a fresh dataset. Second, those choices are consequential for the inferences that can be made in terms of both substance and model fit. Here we provide evidence of these claims. In particular, we consider two small datasets and two extremely well-known and well-used techniques. We show that depending on the steps a researcher undertakes, their model results may differ in ways that could lead an honest researcher down forking paths of inference or allow a malfeasant researcher to support a range of hypotheses that may not be reflective of the universe of results.

### 5.1 Unsupervised Scaling: An Application of Wordfish

The Wordfish model of Slapin and Proksch (2008) has proved both popular and valuable for assessing the positions of parties in terms of their manifesto output, their positions in parliament (Proksch and Slapin, 2010) and other related matters (see, e.g., Lauderdale and Herzog, 2016, for a recent extension). Derivation and explication can be found in the original article, but it suffices to note that the core of the approach is based on a Poisson

model of word use rates, with an unknown (latent) ideological position for the text (or text author) being the crucial estimand. In our current toy example, we apply the model to UK Conservative and Labour manifestos over a 15 year, four general election period: 1983, 1987, 1992, 1997. We thus have eight texts. Importantly, since this period of British history is well studied (see, e.g., Jones, 1996; Kavanagh, 1997; Pugh, 2011), we have strong priors over the relative ideological positions of at least some of the documents. Famously the “longest suicide note in history”<sup>10</sup>, the 1983 Labour manifesto put forward a heady and unpopular brew of unilateral nuclear disarmament, higher taxes, withdrawal from the European Economic Community and substantial (re)nationalization of industry. Meanwhile, the 1997 Labour manifesto extolled what came to be seen as the ‘third way’ position which combined social democracy with more right-wing economic policies. This was in keeping with a Labour party that had subsequently ditched ‘Clause IV’ of its constitution (formally committing it to nationalization) and clawed its way back to the center of British politics.

Taking into account reviews of Conservative positions over the same time, a reasonable observer might place the documents, left to right (with conservative documents being associated with higher latent positions), as follows:

$$\text{Lab 1983} < \text{Lab 1987} < \text{Lab 1992} < \text{Lab 1997} < \text{Con 1992} < \text{Con 1997} < \text{Con 1987} < \text{Con 1983}.$$

In practice, we do not require readers agree with this rank ordering to make our main point, though it will help to set expectations.

Recall that we have a total of 128 different preprocessing possibilities for the document term matrix, where 127 of them are something other than the original DTM for which we

---

<sup>10</sup>An epithet attributed to Gerald Kaufman, MP.

Specification	Most Left	Most Right
P-N-S-W-3-I	Lab 1983	Cons 1983
N-S-W-3	Lab 1987	Cons 1987
N-L-3	Lab 1992	Cons 1987
N-L-S	Lab 1983	Cons 1992

Table 3: Some example specifications which differ in terms of the manifestos they place on the (far) left and (far) right under the *Wordfish* model.

have undertaken no preprocessing steps. Thus, we have 128 DTMs to pass to the *Wordfish* software.<sup>11</sup> Our interest here is the different rank orders the model suggests (again, in terms of the latent positions of the parties). In Figure 1 we report the results in a single plot, with 128 rows. Each row of the plot represents a different specification. A white bar implies that the manifesto for that year is in the correct place as regards our priors. A black bar implies it was misplaced. Thus, at the top of the plot we have specifications which placed the parties in the ‘correct’ order in terms of our priors. At the bottom we have specifications which were almost completely ‘wrong’ (no parties in the correct slot). An immediate observation is that different specifications produce different orderings. Indeed, there were a total of twelve unique orderings of the manifestos from the 128 possible preprocessing steps. Furthermore, the substantive differences between at least some of the specifications are quite stark. Consider Table 3, where we consider the ‘most left’ and ‘most right’ manifestos under some different specifications of the preprocessing steps. Clearly, depending on the particular choices the researcher makes, the poles of British politics move substantially: under P-N-S-W-3-I the manifestos are as expected, with Labour’s 1983 effort on the far left, and the Tory document of the same year on the far right. Meanwhile, under N-L-3, the researcher would be able to conclude that in fact Labour’s manifesto of 1992 marked the high watermark for socialism, while the Conservatives’ 1987 manifesto was the most extreme right-wing document for this period. To underline the main point here, none of these specifications are

---

<sup>11</sup>We use the standard defaults in `quanteda::textmodel_wordfish`, with the Labour and Conservative manifestos from 1983 as the anchors.

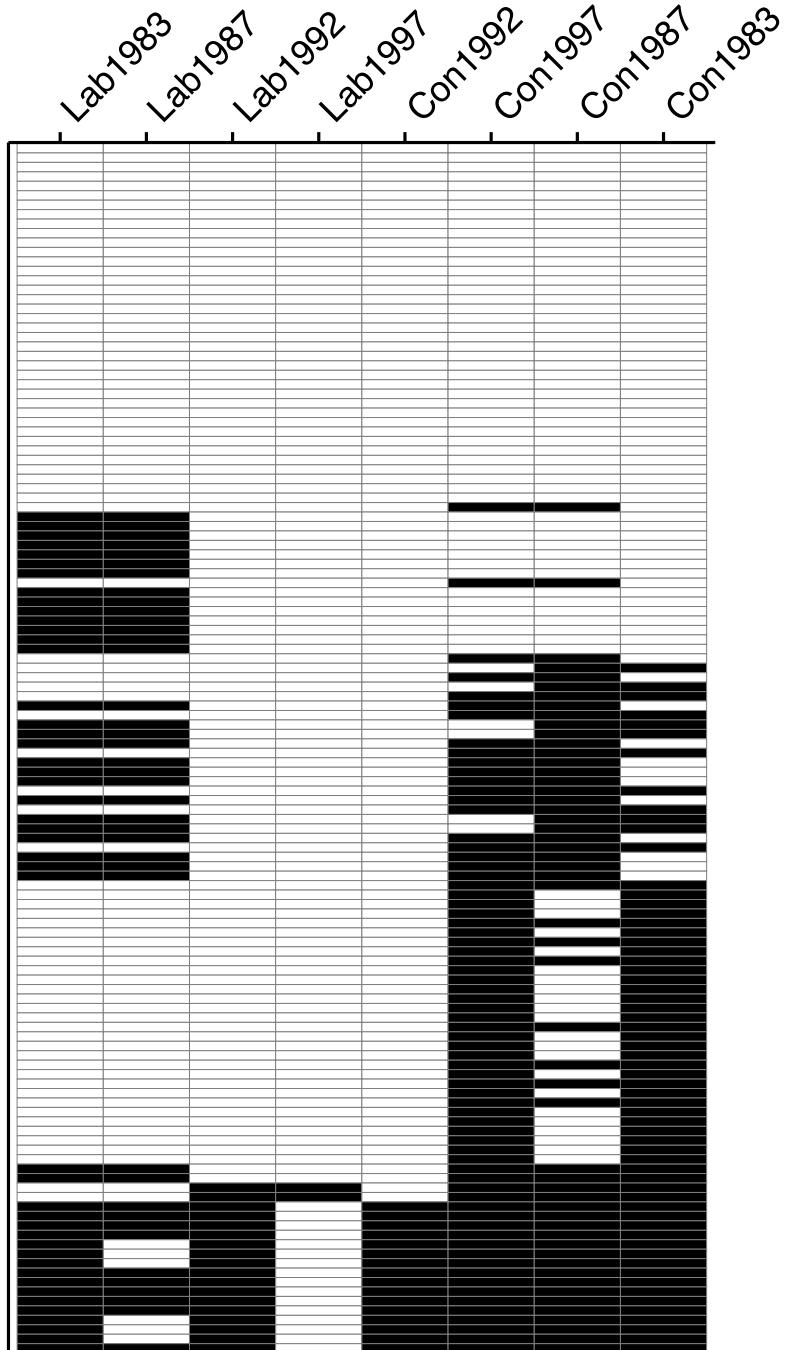


Figure 1: Wordfish results for the 128 different preprocessing possibilities. Each row of the plot represents a different specification. A white bar implies that the manifesto for that year is in the correct place as regards our priors. A black bar implies it was misplaced.

unreasonable *a priori*, but they yield very different conclusions. And this is true regardless of the strength of ones priors about the ‘correct’ rank ordering. In a more cynical light, our results suggest that a malfeasant researcher could, by fitting and refitting under different specifications, support an extremely diverse array of theories. They could conclude, for example, that Michael Foot as leader in 1983 was not especially left wing, and that it was Neil Kinnock (who lead the party to election defeat in 1992) that had to go before the party was electable.

## 5.2 Topic Modeling: An Application of LDA

In order to evaluate the effects of the preprocessing decisions we consider in topic modelling applications, we conducted a relatively simple experiment. We made use of the *Congressional Press Releases* corpus, which contains 1,000 documents written by ten different members of congress (100 each) for this experiment. For each preprocessing specification, we determined the optimal number of topics to characterize that specification using a perplexity criterion, which is discussed below. Note that we only consider the 64 specifications that do not include trigrams in this application, due to the very large computational costs associated with fitting topic models to corpora with large vocabularies. We then fit a topic model with the optimal number of topics to each DTM. Afterward, we looked at the top twenty terms associated with each topic and picked out a series of five “key terms” that strongly anchored our interpretation of the meaning of a topic in at least one of the preprocessing specifications. We then looked for the prevalence of these key terms in topics across all specifications. We found that the proportion of topics in which they appear varies dramatically across preprocessing specifications, likely leading a practitioner to draw different substantive conclusions.

A standard metric for evaluating a particular set of parameters for a probabilistic model is to measure the log-likelihood of a held-out test set under those parameters. In the topic model

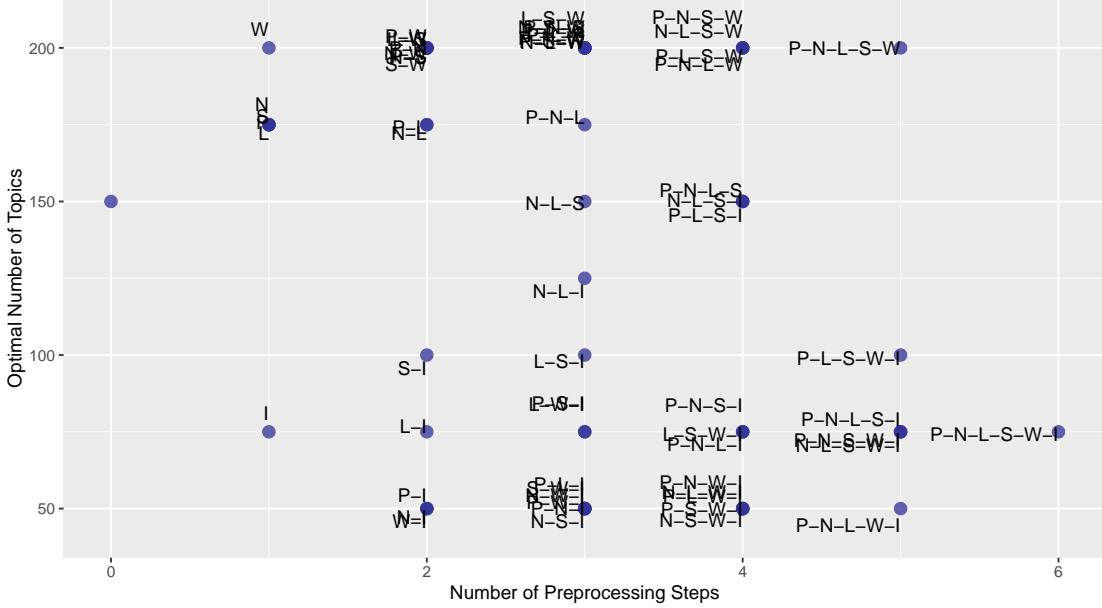


Figure 2: Plot depicting the optimal number of topics (as selected via perplexity) for each of 64 preprocessing specifications not including trigrams. On the x-axis is the number of preprocessing steps, and the y-axis is the number of topics. Each point is labeled according to its specification.

context, the most commonly used metric is a normalization of the held-out log-likelihood known as *perplexity*—and that is what we used to determine the optimal number of topics,  $k$  for a model fit to each preprocessing specification.<sup>12</sup>

In particular, to determine the optimal number of topics for a given combination of steps, we conducted a grid search over  $k = \{25, 50, 75, 100, 125, 150, 175, 200\}$  topics and calculated the perplexity for each choice of  $k$  using a 10-fold cross-validation procedure.<sup>13</sup> All topic

<sup>12</sup>See Online Appendix C for more details on held-out likelihood and perplexity.

<sup>13</sup>More specifically, for each preprocessing specification, we formed ten random splits of input documents into train and test sets. Each training set contained 800 documents (80%) and each test set contained 200 documents (20%). We used the same ten test-train splits for each specification. Then for each choice of  $k$  we fit a topic model to each of the ten training sets using that value of  $k$  as the number of topics. Each of the resulting ten fitted models was then used to calculate the perplexity of the corresponding test set, and these perplexities were averaged across splits. We note that the choice of ten splits is considered to be best practice in the machine learning literature (Jensen and Cohen, 2000). Thus, for each preprocessing

models were fit using the original variational inference algorithm for LDA, as proposed by Blei, Ng and Jordan (2003). The optimal number of topics associated with each specification is depicted in Figure 2: clearly  $k$  has a very large range, right across the different possibilities. Thus it as low as 50 topics in the case of P-N-L-W-I, but as high as 200 in the case of P-N-L-S-W. Similar results can be seen for N-S-I vs L-S-W. To underline the point here: the objectively ‘best’ topic model—by the industry standard measure<sup>14</sup>—varies widely and unpredictably, and depends very heavily on minor perturbations of preprocessing choices.

After determining the optimal  $k$  for each preprocessing specification, we then fit a single topic model to the entire corpus (1,000 documents) for each preprocessing specification, using the optimal number of topics.<sup>15</sup> A common substantive analysis step after fitting a topic model is to look at the top  $t$  terms associated with each topic. We chose to extract  $t = 20$  top terms associated with each topic, across all specifications. Examination of a sample of these topic-top words across several specifications revealed some “key terms” which in our evaluation anchor a particular topic. This means that when we looked at the top twenty words in the topic, those terms tended to provide substantial information as to what the topic is about. For example “stem” and “cell” anchor topics that relate to press releases about stem cell research, while “Iraq” anchors topics that relate to press releases about the war in Iraq. We do not claim that these are the only important terms in the results we looked at, but only that as well intentioned researchers, these terms tended to consistently

---

specification and each choice of  $k$ , we arrived at a perplexity score. Topic models were fit using the `LDA()` function from the `topicmodels` (v.0.2-4) **R** package. The held out document perplexities were calculated using the `topicmodels::perplexity()` function. More information on the method by which perplexity is calculated in the `topicmodels` package can be found in section 2.4 of the package vignette [cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf](http://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf)

<sup>14</sup>We recognize that there are other measures of topic model “quality” that seek to evaluate linguistic characteristics of a given set of topics, and might yield substantively different results. We choose to focus on perplexity in this application because it makes the least assumptions about the importance of including different kinds of features (numbers, punctuation, etc.) in determining the optimal number of topics.

<sup>15</sup>These results were generated using the default parameters (except for number of topics) for the `LDA()` function from the `topicmodels` (v.0.2-4) **R** package.

and strongly influence our evaluations as to what a topic was about.

To see if these prominent anchor terms appear in topic top-twenty terms across our pre-processing specifications, we performed a case-insensitive search for the stem of each of five key terms {“iraq”, “terror”(ism), (al) “qaeda”, “insur”(ance), “stem” (cell)} in all topics across all specifications. We then calculated the proportion of topics each key word appeared in, for each specification.<sup>16</sup> These results are presented in Figure 3.

As we can see, there is a great deal of variation in the proportion of topics each key

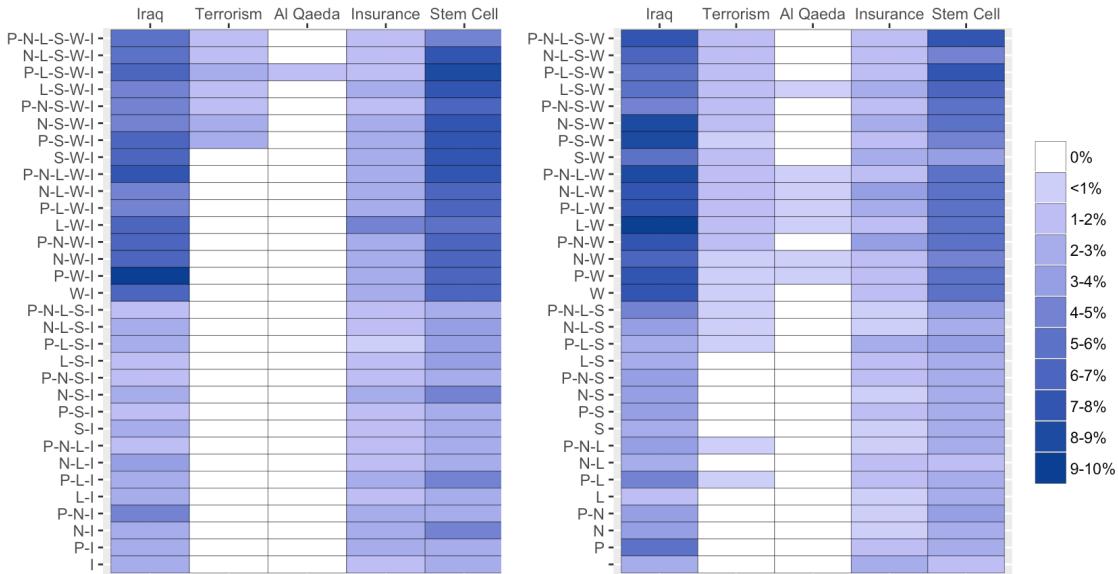


Figure 3: Plots depicting the percentage of topic top-20-terms which contain the stem of each of five keywords, for each of 64 preprocessing steps (thus excluding those which include trigrams). The number of topics for specifications fit to each of the 64 DFM were determined through ten-fold cross validation, minimizing the model perplexity.

<sup>16</sup>While we focus on variation in the *proportion* of topics each key word appeared in, our basic findings are not changed by simply comparing the raw *number* of topics each key term appears in. We chose to focus on the proportion of topics each key word appeared in because the denominator (total number of topics) changes for each specification, making the raw number more difficult to compare across specifications.

word appeared in. To verify that this variation is not simply being driven by the instability of LDA, we replicated our analysis with forty different initializations, and the average outcomes remain strikingly similar (see Online Appendix D). These results illustrate two related issues. First, some key terms do not appear at all in the top terms for some preprocessing specifications. This means that upon inspection of the topic model results, a researcher might be unaware that there were any press releases discussing some of these key issues such as “terrorism” or “Al Qaeda”. Thus the researcher could draw different substantive conclusions about what legislators have public views on, depending on which preprocessing specification they select. Second, for some key terms such as “Iraq” or “Stem Cell”, there is about an order of magnitude difference in the proportion of topics in which a term appears, depending on the specification. This could similarly lead a researcher to different conclusions about the partisan valence or salience of these issues to members of Congress. For example, a single topic related to “stem cells” might combine together different partisan terms related to the issue in the same topic, whereas these terms might separate into several topics in an alternate specification. However, there are also some key terms (such as “insurance”) which do not exhibit the same degree of variation in their prevalence across preprocessing specifications.

To reiterate a point we alluded to above, an experienced practitioner of topic modelling might not find the results presented here especially troubling. This is because the careful use of hyperparameter optimization, asymmetric priors, or an alternate estimation method might make the results look more similar across preprocessing specifications. Furthermore, they might deliberately employ a specific set of preprocessing steps to highlight certain types of terms in topic model results. However, the vast majority of social scientists engaged in exploratory analysis of topic top terms are unlikely to be aware of the current state of the art. Our main point then remains: it is possible that a well intentioned researcher could be

lead to radically different conclusions depending on how they preprocess their data.

## 6 preText: a new method for assessing sensitivity

In the previous section, we showed that even ‘reasonable’ preprocessing decisions can have large and unexpected consequences for both substantive inferences and the appropriate model specification for the data. In this section, we turn to ways that researchers might assess how their preprocessing decisions are likely to affect their results, and try to offer some general advice about how to proceed. As always, we would suggest that researchers should first consult theory, and our approach here is intended to complement that crucial step. Putting that point aside for now, we need to arrive at a basic metric by which to measure how different one DTM is from another.

As we noted above, researchers undertaking unsupervised analyses are typically looking to explore or describe somewhat complex datasets, and to throw (possibly hidden, latent) relationships between observations into starker relief than as they originally appear. With that in mind, we claim that what matters to researchers who are seeking to see these new patterns is how documents ‘move’ relative to one another when they apply some transformation to the DTM, be it a topic model, scaling routine or some decomposition. One of the simplest operations a researcher can undertake—and indeed, one that forms the basis for many more complicated approaches such as principal components analysis—is to generate pairwise distances between documents. It is this very basic step that we focus on here.

To fix ideas, consider the following toy example. A researcher has three documents,  $\text{doc}_1$ ,  $\text{doc}_2$  and  $\text{doc}_3$ . These might be single texts, or three sets of multiple texts where each set is written by a different author. The distance, say measured in Euclidean or cosine terms,

between any two of the documents indexed by  $i$  and  $j$  is  $d(i, j)$ . When using the original document term matrix, for which the researcher has undertaken no preprocessing at all, the distances are  $d(1, 2) = 1$  while  $d(1, 3) = 3$  and  $d(2, 3) = 2$ . So, relatively speaking,  $\text{doc}_1$  and  $\text{doc}_3$  are far apart. Suppose now we impose a particular preprocessing step, such as the removal of stop words and rerun our similarity analysis. On inspection we see that now,  $d(1, 2) = 2$  while  $d(1, 3) = 6$  and  $d(2, 3) = 4$ . While all the distances have been doubled, the ranking of pairwise distances has remained the same:  $d(1, 3) > d(2, 3) > d(1, 2)$ . In this context, we suspect that a researcher would think these specifications are equivalent (in substance terms): things are (up to a constant) as they were previously.

By contrast, suppose that a given preprocessing step altered the distances as follows:  $d(1, 2) = 4$  while  $d(1, 3) = 1$  and  $d(2, 3) = 6$ . Now, the distance between documents 2 and 3 has grown in relative terms, while  $\text{doc}_1$  and  $\text{doc}_3$  are more similar than previously thought. Most importantly, the *order* of the distances is different:  $d(2, 3) > d(1, 2) > d(1, 3)$ . This would imply a new substantive conclusion, or at least provide an opening for one.

Given that researchers in the social sciences typically do more than inspect similarities, why focus our concerns on pairwise distances? First, as we noted above, changing distances between documents have mechanical, ‘knock on’ consequences for the data fed to a more complex technique and thus the (substantive) conclusions that may be drawn from them. Second, specifically on the issue of the importance of *pairwise* comparisons, we would contend that as a behavioral regularity, researchers—either implicitly or explicitly—commonly use them to validate and interpret their findings. This is because scholars typically have strong priors about (only) one or two or a few particular units, be they manifestos (e.g. Labour 1983 v Labour 1997), or Senators (Elizabeth Warren vs Marco Rubio), or parties (Front National vs Parti socialiste in France) in terms of where they lie in some space. If

such ‘landmark’ distances change rapidly and unpredictably between specifications of preprocessing steps, we claim that researchers would (or should!) regard this fact as concerning. For example, if removing punctuation meant that the distance between the Labour 1983 manifesto and Conservative 1983 manifesto was the largest pairwise distance observed in the DTM (something which makes substantive sense), but removing punctuation *and numbers* meant that this pairwise distance was the fifth largest observed (which makes little sense), red flags should be raised. Put very crudely, our logic is as follows: pairwise distance changes are not all that matters, but if they don’t matter, it’s not clear what does.

To begin to formalize our intuition here, consider a researcher starting with the original DTM, and considering one specification, denoted  $M_1$  of the possible 127 preprocessing specifications we identified. They apply the specification in question and ask themselves “when I use this preprocessing step, which document pair changes the most in rank order terms?” Just assuming away ties for the moment, one pair of document must move up or down (i.e. in absolute terms) the rank order more than any other. In the running example above,  $d(1, 3)$  moved from third to first place in rank terms, relative to  $d(2, 3)$  and  $d(1, 2)$ . Thus  $d(1, 3)$  was the biggest mover. Suppose now that the researcher asks whether  $d(1, 3)$  is the biggest mover (again in rank order terms) when going from the original DTM to the next preprocessing option (of the 126 remaining), which we denote as  $M_2$ . She finds that it was not: now, it is  $d(2, 3)$  that changes the most. Similarly, for the third possible preprocessing step  $M_3$  (of 125 remaining) she finds that  $d(2, 3)$  is the biggest mover. And, in fact, for every other one of the 124 remaining preprocessing specifications  $M_3, \dots, M_{127}$ , the researcher finds that  $d(2, 3)$  moves most. What this implies is that the first preprocessing specification,  $M_1$ , was something of an ‘outlier’—it moved  $d(1, 3)$  the most, but every other specification did not. Those other specifications favored a different pair in terms of top mover.

We can make this idea more helpful by switching the focus from pairwise distances to the specifications themselves. Again, consider  $M_1$  and suppose there were now 6 documents in the corpus (meaning there are 15 pairwise distances). As we saw, a given specification, like  $M_1$ , will rank a particular distance at ‘number 1’ in absolute terms of its movement. Now consider seeing where specification  $M_2$  ranks that largest mover from  $M_1$ , and where  $M_3$  ranks that largest mover from  $M_1$  and so on through all the specifications. For 127 different specifications, we will have vector of length 126 for  $M_1$ , looking something like this

$$\mathbf{v}_{M_1} = (2_{M_2}, 14_{M_3}, 2_{M_4}, 3_{M_5}, 8_{M_6}, 7_{M_7}, \dots, 15_{M_{127}}).$$

In this particular example, the pairwise distance most affected by  $M_1$  was only the second most affected under  $M_2$  (thus  $2_{M_2}$ ) while it was the 14th most affected under  $M_3$  (thus  $14_{M_3}$ ), the second most affected under  $M_4$  and so on down to the 127 specification where, in fact, that pairwise distance was the smallest mover as one went from the original DTM to  $M_{127}$ . In principle, the researcher could undertake this exercise for every single one of the specifications. In the limit, a vector of ones, i.e.  $\mathbf{v}_{M_i} = (1, 1, \dots, 1)$  implies that the given specification  $M_i$  gives similar results—at least in terms of the largest pairwise distance mover—to all other specifications. By contrast, a vector of  $\frac{n(n-1)}{2}$  where  $n$  is the size of the corpus (in the example here,  $\frac{5 \cdot 6}{2} = 15$ ) implies that, in terms of what it suggests is the largest mover, this specification is completely dissimilar to every other one (they rank that pair last in terms of absolute distance movement).

## 6.1 preText Scores

Of course, it may be misleading to only look at the single pair that is induced to change most in rank order terms. A more general approach is to look at the top  $k$  pairs which change the most in rank order terms. Then for each of these pairs we can calculate  $\mathbf{v}_{M_i}^{(k)}$ —the rank

difference for pair  $k$  between specification  $i$  and all others—and take the average of these differences across the top  $k$  pairs. Doing so gives us a more general sense of the degree to which a particular preprocessing specification is unusual compared to others, as it is less likely to be affected by any one unusual document pair. One question is why not compare the rank orders of all document pairs, and the answer is primarily a practical one: this is incredibly computationally intensive, to the point where it is impractical for even moderately sized corpora. However, in practice, we have found that setting  $k = 50$  provides results which are stable to increasing the value of  $k$ . As an example, we calculated the average difference in pairwise rank orderings for  $k = 100$  for our Indian Treaties corpus. Cosine distance was used as the underlying distance metric in this example (and all others in this study), but the results were not particularly sensitive to using Euclidean distance as an alternative measure. The results are illustrated in Figure 4, with each row on the  $y$ -axis corresponding to one of the 128 choice combinations, and each point on the  $x$ -axis being the mean rank difference for every one of the top-100 pairs as we move across specifications. The plot suggests that these differences can vary about three-fold in magnitude.

In order to make these rank differences comparable across corpora (and to look for common trends), we normalize the rank differences for each preprocessing specification. This is accomplished by dividing the rank differences by the maximal possible rank difference for that corpus. For example, a preprocessing specification whose average rank difference was 17,000 in our Indian Treaties corpus would be normalized to:

$$\frac{17,000}{177,309} \approx 0.0959 \quad (1)$$

because the maximal difference in rank orderings for this corpus is 177,309. We call this normalized average rank order difference the **preText score**  $\in [0, 1]$  for that particular pre-

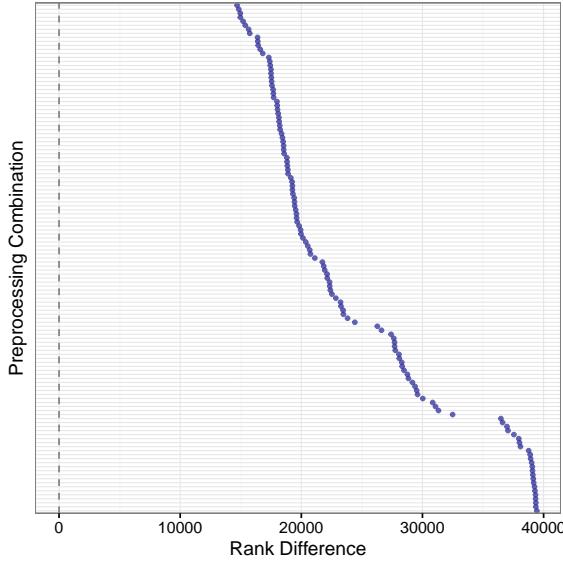


Figure 4: Rank test average difference results for  $k = 100$  for the Indian Treaties corpus ( $n = 596$ ). The maximum possible rank difference for a given document pair, for this corpus, is 177,309.

processing combination. The lower the score for a particular preprocessing specification, the more ‘usual’ it is, while higher scores denote an ‘unusual’ preprocessing specification.

While finding a preprocessing specification with minimal `preText` score for a particular corpus is a valuable diagnostic tool, we also want to understand the impact of each particular preprocessing decision conditional on all other decisions. We can do this by specifying a linear regression with the `preText` score for a particular preprocessing specification as the dependent variable, and dummy variables for each preprocessing decision as predictors. The parameter estimate associated with each preprocessing step will thus tell us that on average, performing that step (controlling for all other steps), has the following marginal effect on

the mean movement of the `preText` score.

$$\begin{aligned} \text{preText score}_i = & \beta_0 + \beta_1 \text{Punctuation}_i + \beta_2 \text{Numbers}_i + \beta_3 \text{Lowercase}_i + \beta_4 \text{Stem}_i + \\ & \beta_5 \text{Stop Words}_i + \beta_6 \text{N-Grams}_i + \beta_7 \text{Infrequent Terms}_i + \varepsilon_i \end{aligned} \quad (2)$$

We performed this regression analysis for each of these corpora, and results are presented in Figure 5. We replicated our analysis using the top 10, 50, and 100 maximally different pairs as a basis for `preText` scores, in order to assess the degree to which the number of top pairs we examine affects our analysis. The  $R^2$  for these regressions (for 100 maximally different pairs) range between 0.4 and 0.82.<sup>17</sup>

The interpretation of these regression results is as follows: a negative parameter estimate for a particular preprocessing step for a given corpus indicates that it tends to reduce the `preText` score for a given specification, thus reducing the risk of drawing unusual conclusions from an analysis with that preprocessing specification applied. A positive parameter implies the opposite: that performing the preprocessing step increases the risk of drawing unusual conclusions from an analysis with that preprocessing specification applied. Just as an example, consider the fourth column of results in Figure 5, which deals with the Death Row Statements. In the first, second and third subfigure, we see that the coefficient on using  $n$ -grams ('3') is negative, as is the coefficient on removing punctuation (P), while the coefficient on removing infrequent terms (I) is positive. This implies that, for this corpus, the choices of whether to add  $n$ -grams, remove punctuation, or remove infrequent terms may have a significant influence on the DTM.

More generally, for a given corpus, if all regression parameter estimates are not significantly

---

<sup>17</sup>The  $R^2$  statistics for each corpus are as follows. UK Manifestos: 0.5, State of The Union Speeches: 0.399, Death Row Statements : 0.764, Indian Treaties: 0.7, Congressional Press Releases: 0.828.

different from zero, then any given preprocessing choice is unlikely to be overly important for the substantive conclusions drawn. Therefore, even if the researcher's theory about which preprocessing specification is most appropriate is not particularly strong, the conclusions they draw from the analysis of their data (under their favored preprocessing specification) are not likely to be sensitive to their choice of preprocessing specification. If, however, a number of regression parameter estimates are significantly different from zero, then the conclusions they draw from the analysis of their data are likely to be particularly sensitive to their choice of preprocessing specification. In the first case, theoretical certainty about the correct preprocessing specification is less important, while in the second case, it is imperative.

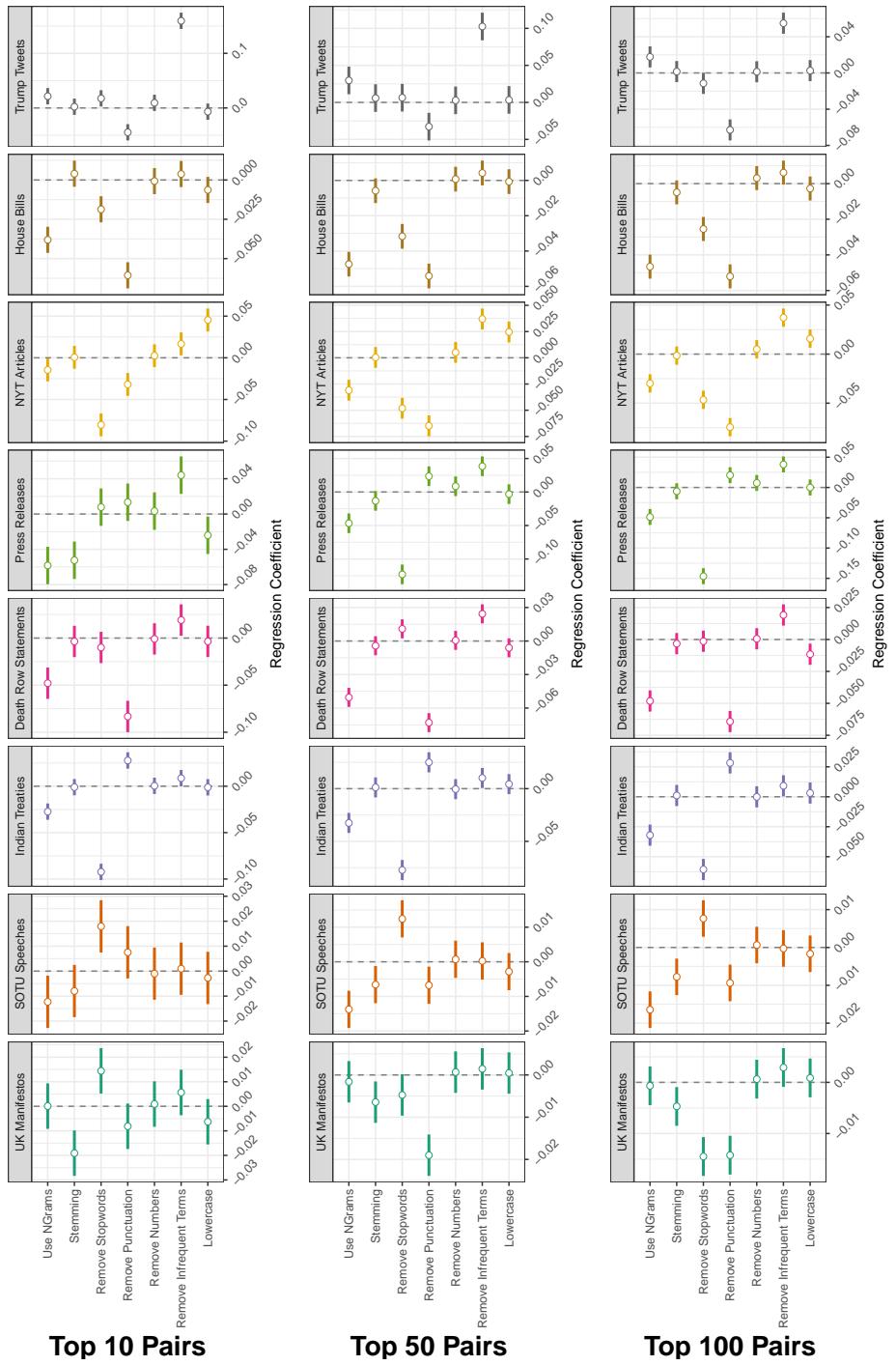


Figure 5: Regression results depicting the effects of each of the seven preprocessing steps on the preText score for that preprocessing combination.

## 6.2 Advice for Practitioners

We suggest the following workflow for researchers seeking to draw conclusions from the analysis of DTM that are robust to their choice of preprocessing specification.

1. Use theory, to the extent it exists for the problem, to choose potential preprocessing steps on the basis that the information this removes or preserves is reasonable for the application.
2. Having carefully selected a theoretically motivated preprocessing specification, generate `preText` score regression results similar to those in Figure 5 for a random sample of up to 500-1,000 documents from the corpus. Such a sample size balances the goals of accurately approximating the entire corpus with keeping runtime under 24 hours so as not to slow down the analysis process too much.
3. Examine the `preText` score regression results. Depending on the nature of these results and the strength of the researcher’s theory about their specification, we advocate for one of three courses of action:
  - (a) **All Parameter Estimates Are Not Significantly Different From Zero:** In this case, the researcher’s conclusions are unlikely to be highly sensitive to their choice of preprocessing specification, and it is therefore reasonable to proceed with the analysis.
  - (b) **Strong Theory, Some Parameter Estimates Are Significantly Different From Zero:** In this case, assess which parameter estimates are different from zero. If there is a strong theoretical reason to prefer a particular choice on each of those preprocessing steps, then (cautiously) proceed with the analysis. However, a more conservative approach would be to replicate the analysis across all combinations of preprocessing steps whose parameter estimates are significantly

different from zero, and include these results as a robustness check in an appendix.

(c) **Weak Theory, Some Parameter Estimates Are Significantly Different**

**From Zero:** Again, assess which parameter estimates are different from zero. If there is not a strong theoretical justification for the preferred choices of these preprocessing steps, the appropriate course of action is to replicate the analysis across all combinations of preprocessing steps whose parameter estimates are significantly different from zero, and then to average or otherwise aggregate over those results in their final analysis.

Taking the steps described above may increase analysis time, but if the researcher is guided by strong theoretical expectations about the appropriate preprocessing specification for their dataset, then replication across some steps may be reasonably avoided. However, in general, the most conservative approach is simply to replicate one's analysis across all steps with `preText` score regression parameters that are significantly different from zero, and include the results of those analyses as a robustness check for one's preferred preprocessing specification.

As an illustration, we apply the approach outlined above to the Wordfish example from Section 5.1. We selected a “theoretically motivated” preprocessing specification of P-N-L-S-W-I, following Grimmer and Stewart (2013) and based on our expectations about what will and will not matter for the application. The Wordfish scores (with 95% confidence intervals) are presented in the left panel in Figure 6. Next, we averaged the Wordfish scores of eight models using every combination of the three preprocessing steps with significant parameter estimates in Figure 5: stemming (or not), stopping (or not) and removing punctuation (or not). The average Wordfish scores for these specifications, along with the appropriately adjusted 95% confidence intervals (Buckland, Burnham and Augustin, 1997), are displayed in the right panel of Figure 6.

Documents are ordered from top to bottom based on our theoretical ranking, going from most conservative to most liberal. As we can see, while both our theoretically selected and the averaged results produce the “correct” order for Labour manifestos, in both cases, Wordfish places the Conservative manifestos in the theoretically incorrect order. Looking at our theoretically selected specification (left panel), we would conclude that there is a clear ordering of:

$$\boxed{\text{Con } 1992 < \text{Con } 1987 \sim \text{Con } 1983 < \text{Con } 1997}$$

where Con 1997 is significantly more conservative than Con 1987 or Con 1983. However, when we incorporate the additional uncertainty from averaging across the eight possible preprocessing specifications representing the three preprocessing steps with significant parameter estimates in Figure 5, we can no longer distinguish between Con 1997, Con 1987 and Con 1983 (right panel). We note that even without model averaging we could not statistically distinguish between Con 1987 and Con 1983. Such estimation uncertainty may mean that model averaging does not make a practical difference, in some cases. However, this can only be verified through comparison to the averaged results, so we recommend performing this procedure even when there is relatively substantial estimation uncertainty.

$$\boxed{\text{Con } 1992 < \text{Con } 1997 \sim \text{Con } 1987 \sim \text{Con } 1983}$$

Coming back to our general argument, the `preText` score regression results for the UK manifestos provided guidance regarding which preprocessing steps to focus on in order to assess the robustness of our results in this context. By focussing on three likely-consequential preprocessing decisions and averaging across those eight specifications, our results changed substantively—to be closer to the “ground truth” we identified in our example earlier.

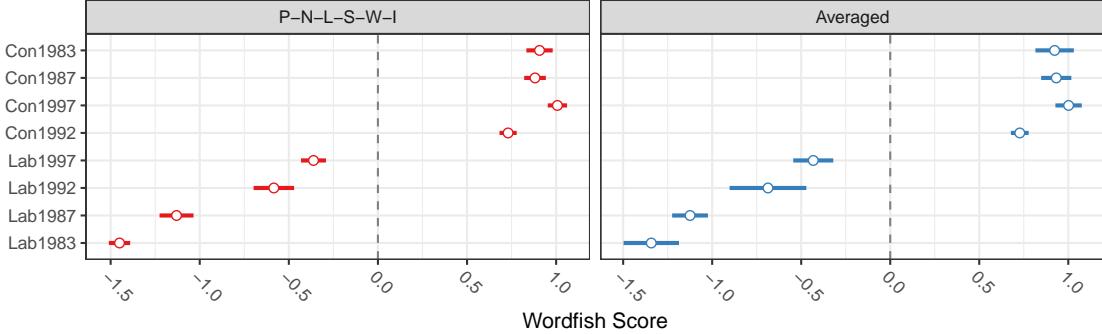


Figure 6: Wordfish scores for eight UK party manifestos generated using a theoretically selected preprocessing specification (P-N-L-S-W-I), and averaged across the eight possible DTM<sub>s</sub> generated using stemming (or not), stopping (or not) and removing punctuation (or not). These choices correspond to the choices with parameter estimates that were significantly different from zero in Figure 5

While we believe our working example using the UK Manifestos corpus is compelling, it is reasonable to wonder whether the issues we raise are a peculiarity of the example we selected, or whether they actually effect published findings. To examine this possibility, we replicated the Wordfish analysis in Lowe and Benoit (2013) using their theoretical preprocessing specification, and model averaging implied by preText regression results. Lowe and Benoit were primarily interested in comparing human coding and Wordfish scores, but we find that it is possible to arrive at a substantively different interpretation of the Wordfish results when we use model averaging. Our replication is detailed in Online Appendix E, and we feel that it highlights the value of model averaging when the researcher does not have strong theoretical reasons for selecting a particular preprocessing specification.

## 7 Discussion

It is hard to deny that the quantitative analysis of text is now a force to be reckoned with in political science: our leading journals devote special issues to its developments, scholars design easy-to-use software for its processing, and recent innovations in modeling documents

rack up thousands of citations. Unsupervised methods have played a key part in this growing interest, not least because scholars often find themselves in situations where they suspect a latent structure or continuum in their data, and need some exploratory technique to help them uncover it. Indeed, outside of some very specific applications, political scientists have made relatively little use of supervised techniques—especially ‘off the shelf’ machine learning tools. This may be because the output of those models does not easily lend itself to answering substantive questions (see, e.g. Monroe, Colaresi and Quinn, 2008) or perhaps because the assumptions underlying those techniques are both consequential, and non-trivial to fully understand (see, e.g. Lowe, 2008).

Despite this ambivalence about supervised approaches for inference, political scientists have been very happy to import advice about preprocessing steps from that literature. This is sometimes done knowingly, but more often in a way that substitutes ‘theory’ on a given problem with citation of current—though unexamined—practice in previous studies. To reiterate, we can find little discussion of, or evidence for, whether those preprocessing choices ‘work’ or are optimal for the question under consideration. With that in mind, our paper makes sobering reading. Above we took two real data sets and showed that under relatively small perturbations of preprocessing decisions—none of which were *a priori* unreasonable—very different substantive interpretations would emerge. Furthermore, we showed that other modeling choices, such as the optimal number of topics, were also startlingly dependent on one’s earlier preprocessing decisions. Our specific examples were of scaling and topic modeling, but we have no reason to believe it would be not be true for larger data sets where priors on what ‘should’ be seen are more diffuse.

But all is not doom and gloom. A further contribution of our paper was the proposal of a new procedure to analyze the sensitivity of results to preprocessing decisions. Our method

essentially compares the relative movement of pairwise document distances under different preprocessing specifications. Our approach is built on what we believe to be a reasonable theoretical base, and we outline a conservative approach to applying it which we believe is likely to minimize the risk of a researcher drawing conclusions which are sensitive to poorly motivated preprocessing choices, while balancing the additional analysis time needed to determine the robustness of their results. Our more general point stands, though: it is not generally appropriate to arbitrarily pick one particular preprocessing combination and just hope for the best.

To underline our philosophical point here, note that the issue is not simply that dishonest researchers might cynically pick a specification they like and run with it, to the detriment of scientific inquiry. The more subtle problem is that well-meaning scholars would have no idea of the truth value of their findings. A particular feature of unsupervised models of text is that there are typically many possible specifications, and many plausible ‘stories’ about politics that can be fit to them, and validated, after estimation. Fundamentally then, a lack of attention to preprocessing produces a potentially virulent set of “forking paths” (in the sense of Gelman and Loken, 2014) along which researchers interpret their results and then suggest further cuts, tests and validation checks without realizing that they would have updated had they preprocessed their documents differently.

Clearly, we believe that being systematic and transparent about how preprocessing choices affect inferences is important. We are certainly not alone in this broad concern: scholars in psychology, for example, have recently mooted the idea of running a given regression analysis on every possible data set that emerges from coding variables differently, and then comparing the resulting  $p$ -values (Steegen et al., 2016).<sup>18</sup> In line with that paper, we would hope

---

<sup>18</sup>See also Moore, Powell and Reeves (2013), Appendix 3, for a political science example

that, ideally, researchers would motivate their specification choices from theory and their substantive understanding of a given area. Typically in unsupervised work, however, they do not—or perhaps cannot—and it is to that scenario that we speak here. For those working specifically with texts, we hope this paper and its attendant software helps brings research using unsupervised models into line with efforts to further replication and the permanence of findings elsewhere in the discipline (see, e.g., Gelman, 2013, on preregistration). Nonetheless, we make no claims that our method is the last word: we have not been encyclopedic in checking all possible text datasets, or in deriving formal properties of our approach, or in exploring the multiple other steps scholars might take in preparing their data. We leave such efforts for future work.

## References

- Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research* 3:993–1022.
- Buckland, S. T., K. P. Burnham and N. H. Augustin. 1997. “Model Selection: An Integral Part of Inference.” *Biometrics* 53(2):603–618.  
**URL:** <http://www.jstor.org/stable/2533961>
- Catalinac, Amy. 2016. “Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections.” *Journal of Politics* 78(1):1–18.
- Chang, Jonathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Neural Information Processing Systems*.
- Diermeier, Daniel, Jean-François Godbout, Bei Yu and Stefan Kaufmann. 2011. “Language and Ideology in Congress.” *British Journal of Political Science* 42(01):31–55.

D’Orazio, Vito, Steven Landis, Glenn Palmer and Philip Schrodt. 2014. “Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines.” *Political Analysis* 22(2).

Gelman, Andrew. 2013. “Preregistration of Studies and Mock Reports.” *Political Analysis* 21(1):40–41.

Gelman, Andrew and Eric Loken. 2014. “The Statistical Crisis in Science.” *American Scientist* 102(6):460–465.

Grimmer, J. 2010. “A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases.” *Political Analysis* 18(1):1.

Grimmer, Justin and Brandon M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21(3):267–297.

Grimmer, Justin and Gary King. 2011. “General purpose computer-assisted clustering and conceptualization.” *Proceedings of the National Academy of Sciences of the United States of America* 108(7):2643–50.

Handler, Abram, Matthew J. Denny, Hanna Wallach and Brendan O’Connor. 2016. Bag of What? Simple Noun Phrase Extraction for Text Analysis. In *Proceedings of the Workshop on Natural Language Processing and Computational Social Science at the 2016 Conference on Empirical Methods in Natural Language Processing*.

**URL:** <https://brenocon.com/handler2016phrases.pdf>

Hopkins, Daniel and Gary King. 2010. “A Method of Automated Nonparametric Content Analysis for Social Science.” *American Journal of Political Science* 54(1):229–247.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. New York: Springer.

Jensen, David D. and Paul R. Cohen. 2000. “Multiple Comparisons in Induction Algorithms.” *Machine Learning* 38:309–338.

Jones, Tudor. 1996. *Remaking the Labour Party: From Gaitskell to Blair*. New York: Routledge.

Jurafsky, Daniel and James H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition*. Prentice Hall.

Justeson, John S. and Slava M. Katz. 1995. “Technical terminology: some linguistic properties and an algorithm for identification in text.” *Natural Language Engineering* 1(01).

Kavanagh, Dennis. 1997. *The Reordering of British Politics: Politics after Thatcher*. Oxford University Press.

King, Gary, Patrick Lam and Margaret E Roberts. 2017. “Computer-Assisted Keyword and Document Set Discovery from Unstructured Text.” *American Journal of Political Science* 00(00):1–18.

**URL:** <http://onlinelibrary.wiley.com/doi/10.1111/ajps.12291/abstract>

Lauderdale, Benjamin and Alexander Herzog. 2016. “Measuring Political Positions from Legislative Speech.” *Political Analysis* 24(2):1–21.

Laver, Michael, Kenneth Benoit and John Garry. 2003. “Extracting Policy Positions from Political Texts Using Words as Data.” *American Political Science Review* 97(2):311–331.

Lowe, Will. 2008. “Understanding wordscores.” *Political Analysis* 16(4 SPEC. ISS.):356–371.

- Lowe, Will and Kenneth Benoit. 2013. “Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark.” *Political Analysis* 21(3):298–313.
- Manning, Christopher D and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.
- Manning, Christopher D, Prabhakar Raghavan and Hinrich Schütze. 2008. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. “Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict.” *Political Analysis* 16:372–403.
- Moore, Ryan, Elinor Powell and Andrew Reeves. 2013. “Driving support: workers, PACs, and congressional support of the auto industry.” *Business and Politics* 15(2):137–162.
- Pang, Bo, Lillian Lee and Shivakumar Vaithyanathan. 2002. “Thumbs up? Sentiment classification using machine learning techniques.” *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 79–86.
- Porter, M.F. 1980. An algorithm for suffix stripping. In *Program: electronic library and information systems*. Vol. 14 pp. 130–137.
- Proksch, Sven-Oliver and Jonathan B. Slapin. 2010. “Position Taking in European Parliament Speeches.” *British Journal of Political Science* 40(03):587–611.
- Pugh, Martin. 2011. *Speak for Britain!: A New History of the Labour Party*. New York: Random House.
- Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin and Dragomir R. Radev. 2010. “How to analyze political attention with minimal assumptions and costs.” *American Journal of Political Science* 54(1):209–228.

- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. “Structural topic models for open-ended survey responses.” *American Journal of Political Science* 58(4):1064–1082.
- Sebastiani, Fabrizio. 2002. “Machine learning in automated text categorization.” *ACM Computing Surveys* 34(1):1–47.
- Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. “A Scaling Model for Estimating Time-Series Party Positions from Texts.” *American Journal of Political Science* 52.
- Spirling, Arthur. 2012. “U.S. treaty making with American Indians: Institutional change and relative power, 1784-1911.” *American Journal of Political Science* 56(1):84–97.
- Steegen, Sara, Francis Tuerlinckx, Andrew Gelman and Wolf Vanpaemel. 2016. “Increasing Transparency through a Multiverse Analysis.” *Perspectives on Psychological Science* 11(5):702–712.
- Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov and David Mimno. 2009. “Evaluation methods for topic models.” *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09* (4):1–8.
- Yano, Tae, Noah a Smith and John D Wilkerson. 2012. “Textual Predictors of Bill Survival in Congressional Committees.” *Conference of the North American Chapter of the Association for Computational Linguistics* pp. 793–802.

## Online Appendix A Google Scholar Results

To investigate the use of supervised and unsupervised methods for text analysis in Political Science over time, we collected data from Google Scholar. Google Scholar allows users to search for the number of results containing a key term in a particular year, thus giving us a sense of the use of a term in academic research over time. We collected data on five search terms over the past 9 years (since the first Wordfish results appeared on Google Scholar) to examine trends related to supervised and unsupervised learning. Figure 7 depicts the relative increase in the number of results returned by Google Scholar (with the number of results for each term in 2008 used as the baseline for that term) over time between 2008 and 2016.

We included three general terms in our search (“Supervised Learning”, “Unsupervised Learning”, and “Text Analysis”). As we can see from Figure 7, the growth in the use of these three terms tracked closely together over time. While these terms appear in papers published in a wide range of fields, they serve as a good baseline against which to compare changes in the political science literature. To examine that field specific part, we selected two unsupervised models prominent in the Political Science literature (“Topic Model”, and “Wordfish”). As we can see, the use of these key terms increased at a much higher rate over the time period than the baseline terms. These results are far from exhaustive, but they demonstrate the growth in importance of unsupervised methods in Political Science and in text analysis more broadly. We feel that they highlight the importance of taking preprocessing seriously.

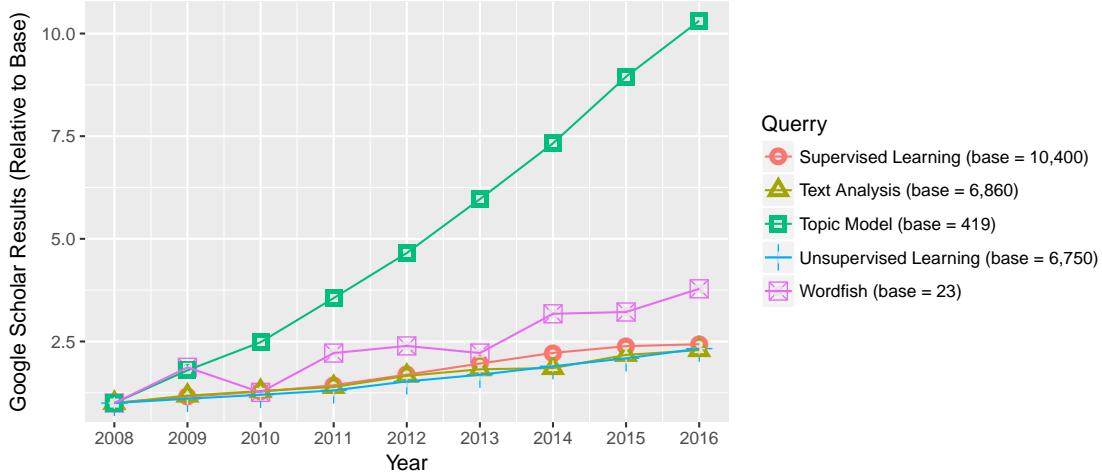


Figure 7: Google scholar results.

## Online Appendix B Why Preprocessing Matters: An example and Intuition

To see why preprocessing matters, consider the following sentences dealing with Britain’s nuclear defence system, Trident. The first is from the UK Labour manifesto in 1983:

The next Labour government will cancel the Trident programme.

The second is from the same party in 1997:

A new Labour government will retain Trident.

Clearly, these represent very different positions. The question though, is in what ways preprocessing might affect our sense of how different they are. We note, to begin, that the cosine similarity of these snippets is 0.51. The relevant document frequency matrix looks as that in Table 4 (assuming we only lowercase words)

	the	next	labour	government	will	cancel	trident	programme	.	a	new	retain
1983	2	1	1	1	1	1	1	1	1	0	0	0
1997	0	0	1	1	1	0	1	0	1	1	1	1

Table 4: Document frequency matrix with stop words retained, no stemming.

Consider two researchers, *A* and *B*. Researcher *A* decides to remove stop words from the documents—‘and’, ‘the’ and so on—while Researcher *B* keeps stop words in, but decides to stem the words back to their ‘roots’. In this particular case, Researcher *B*’s decision has zero effect on the distance between the documents: this is because, the words that were stemmed (‘government’, ‘programme’) were common to both documents. Table 5 shows the relevant document term matrix: with minor column name changes, it is identical to Table 4.

	the	next	labour	govern	will	cancel	trident	programm	.	a	new	retain
1983	2	1	1	1	1	1	1	1	1	0	0	0
1997	0	0	1	1	1	0	1	0	1	1	1	1

Table 5: Document frequency matrix with stop words retained, and stemming.

What about Researcher *A*? In practice, removing stop words changes the documents in different ways. In particular, the 1983 manifesto had more incidences of ‘the’. With those removed—as pictured in Table 5—the documents now look *more similar* than before. Indeed, the cosine distance between them rises from 0.51 to 0.62. Thus, when Researcher *A* and Researcher *B* are asked how similar the documents are, their conclusions differ. This matters because document similarity is not some abstruse property: in various forms, it is at the core of almost all unsupervised techniques—be they scaling or clustering or something else.

	next	labour	govern	cancel	trident	programm	.	new	retain
1983	1	1	1	1	1	1	1	0	0
1997	0	1	1	0	1	0	1	1	1

Table 6: Document frequency matrix with stop words removed, no stemming.

## Online Appendix C Held-out likelihood and Perplexity

Consider a split of documents in a corpus into a training set  $\mathbf{w}$ , and a test set  $\mathbf{w}'$ . In the case of LDA (Blei, Ng and Jordan, 2003), the predictive distribution for the model is characterized by the document-topic probability matrix  $\Phi$ , and hyperparameter  $\alpha$  (controlling document-topic distributions). The log-likelihood of the held out test set is thus:

$$\mathcal{L}(\mathbf{w}') = \log p(\mathbf{w}'|\Phi, \alpha) = \sum_d \log p(\mathbf{w}'_d|\Phi, \alpha). \quad (3)$$

This log-likelihood of unseen documents can thus be used to compare models, with a higher log-likelihood implying a “better” model. The *perplexity* of a test set is a closely related to its log-likelihood and is defined as:

$$\text{perplexity}(\mathbf{w}') = \exp \left\{ -\frac{\mathcal{L}(\mathbf{w}')}{\text{count of tokens in } \mathbf{w}'} \right\} \quad (4)$$

which is essentially a normalization of the held-out log-likelihood. Perplexity is the most commonly used metric for evaluating topic model fit. It is intractable because calculating  $\mathcal{L}(\mathbf{w}')$  is intractable, however approximation methods have been developed (Wallach et al., 2009) and implemented in numerous software packages.

## Online Appendix D Replication of Topic Model Results

Figure 8 displays the average percentage of topic top-20-terms which contain the stem of each of five keywords across 40 different initializations of LDA. Comparison to Figure 3 illustrates highly similar results, indicating that the potential instability of LDA is unlikely to be driving our results.

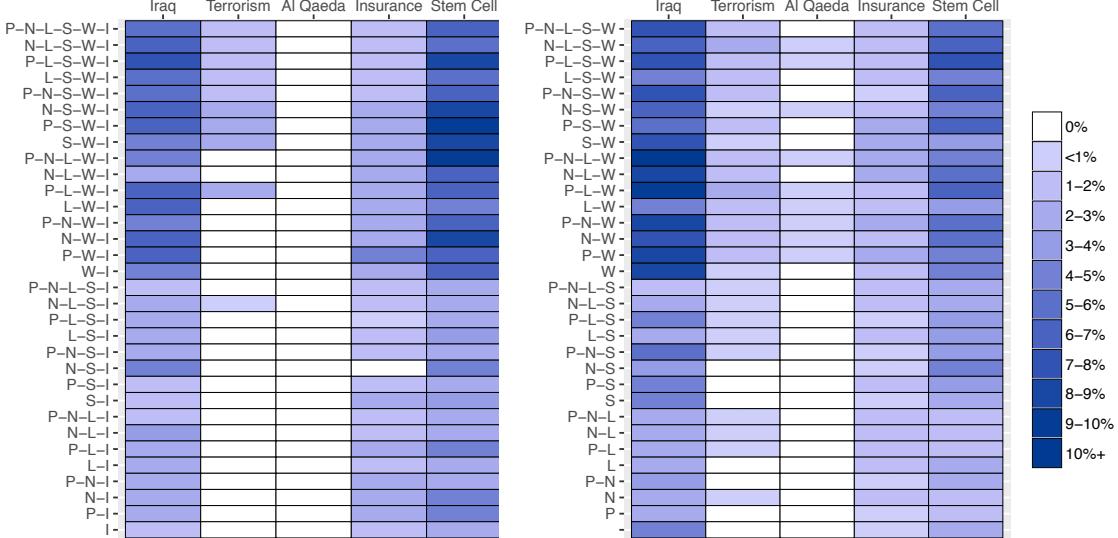


Figure 8: Plots depicting the average percentage of topic top-20-terms which contain the stem of each of five keywords, for each of 64 preprocessing steps (thus excluding those which include trigrams) across 40 different initializations of LDA. The number of topics for specifications fit to each of the 64 DFM were determined through ten-fold cross validation, minimizing the model perplexity.

## Online Appendix E Applying preText to Lowe and Benoit (2013)

In this Appendix, we replicate the Wordfish scaling results from Lowe and Benoit (2013) using the author’s preferred preprocessing specification, as well as model averaging suggested by `preText` regression results. Lowe and Benoit apply a Wordfish scaling model to 14 Irish parliamentary budget debate speeches from 2009, and then compare the results of their analysis to human expert coding results. The authors are very careful throughout the paper, and place a strong emphasis on validating their results.

The authors selected a relatively standard preprocessing specification of removing all punctuation, numbers, and lowercasing all text (P-N-L). The authors did not stem, or remove stopwords or infrequently occurring words, and did not include n-grams in their analysis. They also noted that they replicated their results with stemming, but this did not change their substantive conclusions at all (something that is backed up by our results). Furthermore, the authors note that they did not remove stopwords or infrequently occurring terms primarily because they did not have a-priori information about which terms might be important to their analysis. We feel that this study represents a case where conscientious and

experienced authors used their best judgement in preprocessing, but did not have the luxury of obvious theoretical guidance for all of their preprocessing decisions.

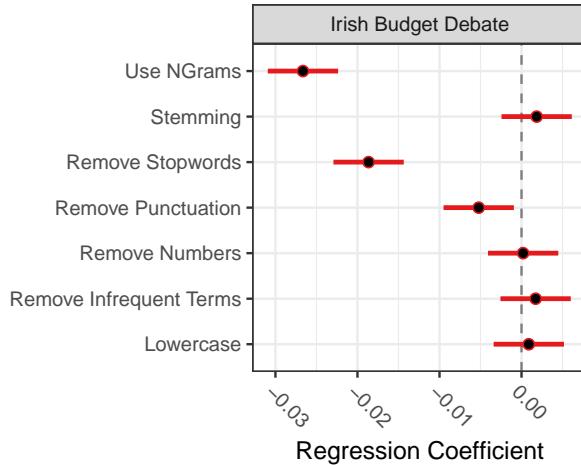


Figure 9: PreText results for 14 Irish parliamentary budget debate speeches from Lowe and Benoit (2013)

To assess the sensitivity of the findings of Lowe and Benoit (2013) to their preprocessing specification, we performed a `preText` regression analysis of the corpus. Regression results are displayed in Figure 9, and indicate that the choices of whether to use ngrams, remove stopwords, and remove punctuation all had significant effects on `preText` scores. While there was a significant effect of including ngrams (or not), we decided to focus our attention on stopwords and punctuation. The choice to include ngrams has not been standard in the literature using Wordfish, and should be further explored in terms of its consequences for the estimation procedure.

Following our own advice to practitioners (see Section 6.2), we averaged Wordfish estimation results over four possible combinations of preprocessing steps (P-N-L, P-N-L-W, N-L, N-L-W) implied by the `preText` regression analysis (excluding ngrams). The averaged parameter estimates are compared to those from the theoretically justified specification of Lowe and Benoit (2013) in Figure 10. Going by point estimates, we can see that the median legislator is somewhere between OCaolain and ODonnell for both the theoretical and averaged results. But, once we look at the confidence intervals, life is more interesting: for Lowe and Benoit, Gilmore is almost certainly to the ‘left’ of OCaolain, and Ryan is almost certainly to the ‘right’ of Morgan (the confidence intervals do not overlap). But using the averaged results, this need not be the case—because we can switch people’s point estimates around based on uncertainty bands: now Gilmore and OCaolain overlap, as do Ryan and Morgan. While Lowe and Benoit were primarily interested in comparing these Wordfish estimates to human

coding, our results suggest that a researcher could be led to different conclusions from the averaged Wordfish results.

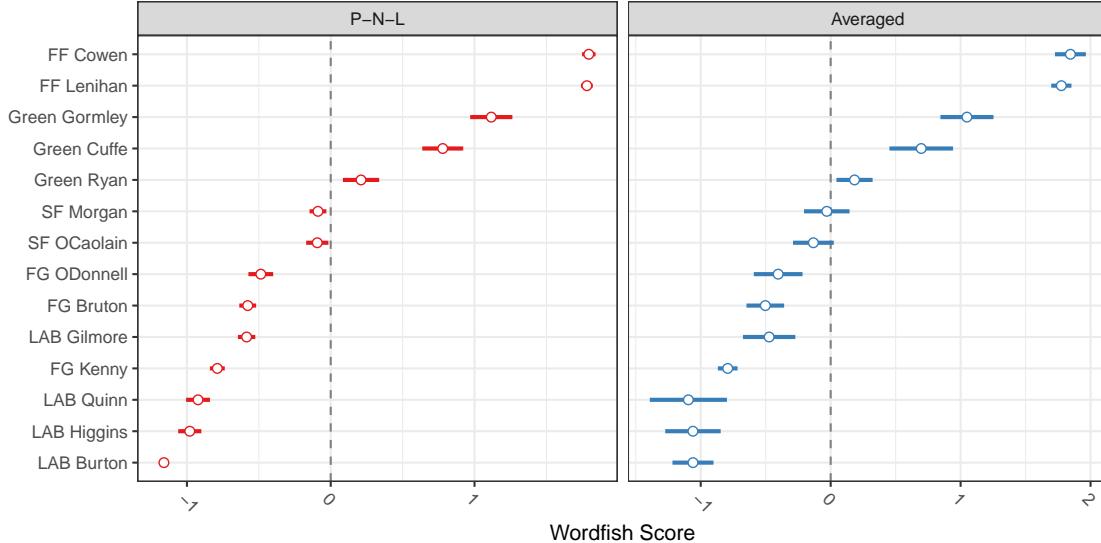


Figure 10: Wordfish scores for 14 Irish parliamentary budget debate speeches from Lowe and Benoit (2013), generated using the authors' selected preprocessing specification (P-N-L), and averaged across the four possible DTMAs generated using stopping (or not), and removing punctuation (or not). These choices correspond to the choices with parameter estimates that were significantly different from zero in Figure 9, but exclude n-grams.

# A Scaling Model for Estimating Time-Series Party Positions from Texts

**Jonathan B. Slapin** Trinity College, Dublin

**Sven-Oliver Proksch** University of California, Los Angeles

*Recent advances in computational content analysis have provided scholars promising new ways for estimating party positions. However, existing text-based methods face challenges in producing valid and reliable time-series data. This article proposes a scaling algorithm called WORDFISH to estimate policy positions based on word frequencies in texts. The technique allows researchers to locate parties in one or multiple elections. We demonstrate the algorithm by estimating the positions of German political parties from 1990 to 2005 using word frequencies in party manifestos. The extracted positions reflect changes in the party system more accurately than existing time-series estimates. In addition, the method allows researchers to examine which words are important for placing parties on the left and on the right. We find that words with strong political connotations are the best discriminators between parties. Finally, a series of robustness checks demonstrate that the estimated positions are insensitive to distributional assumptions and document selection.*

Many theories of comparative politics rely on the ability of researchers to locate political parties in a policy space. Theories of coalition formation and duration use party positions to predict which governments form and how long they survive (Baron 1991; Crombez 1996; de Swaan 1973; Druckman and Thies 2002; Druckman, Martin, and Thies 2005; Strom 1984; Warwick 1992). Likewise, theories of lawmaking use distances between parties to predict policy change (Bawn 1999; Hallerberg and Basinger 1998; Tsebelis 2002), as do analyses of budgetary politics (Franzese 2002), globalization and the social welfare state (Garrett 1998), and labor politics (Wallerstein 1999). In fact, all tests of spatial models in comparative politics rely on the ability to estimate party positions.

Despite the importance of party positions to the study of comparative politics, locating parties in a political space over time is a difficult task. Although one might have a good intuition about where parties stand relative to each other, the positions themselves are abstract concepts that cannot be observed directly (Benoit and Laver 2006b, chap. 3). To facilitate empirical work,

scholars have developed numerous methods for estimating party positions. The existing methodological arsenal includes expert surveys (Benoit and Laver 2006b; Castles and Mair 1984; Huber and Inglehart 1995; Laver and Hunt 1992), hand coding of party manifestos (Budge, Robertson, and Hearl 1987; Budge et al. 2001), and more recently computer coding of manifestos (Laver, Benoit, and Garry 2003). Despite the widespread use of these methods, we argue that they face several challenges in producing valid and reliable time-series position estimates. This leaves a gap in the literature on estimating party ideology.

This article presents a statistical model that adds to and improves upon the existing methodologies by estimating party positions, and their associated uncertainty, over time using word frequencies from manifestos. The remainder of the article reviews the existing methods for estimating party positions, then introduces a new model and compares it to other methods. Finally, we use this model to estimate party positions from manifestos in postreunification Germany. In addition, we describe the lexicon of German politics during this era. The new

---

Jonathan B. Slapin is lecturer in political science, Trinity College, University of Dublin, Dublin 2, Ireland ([jonslapin@gmail.com](mailto:jonslapin@gmail.com)). Sven-Oliver Proksch is PhD candidate, Department of Political Science, University of California, Los Angeles, CA 90095-1472 ([proksch@ucla.edu](mailto:proksch@ucla.edu)).

We would like to thank Kathleen Bawn, Ken Benoit, Jim DeNardo, Tim Groseclose, James Honaker, Thomas König, Jeff Lewis, Will Lowe, Burt Monroe, George Tsebelis, several anonymous reviewers, and participants at the UCLA Methods Workshop and the 2007 annual meeting of the Midwest Political Science Association for their comments and suggestions. In addition, we thank the Zentralarchiv für Empirische Sozialforschung at the University of Cologne, Germany, for providing us with the German party manifestos in electronic format. The order of authors' names reflects the principle of rotation. Both authors have contributed equally to all work.

*American Journal of Political Science*, Vol. 52, No. 3, July 2008, Pp. 705–722

©2008, Midwest Political Science Association

ISSN 0092-5853

705

estimates are robust to various model specifications, correlate highly with other estimates, but are indeed an improvement over previous party position estimates.

## Current Methods for Estimating Party Positions

Party positions are unobservable and must therefore be treated as a latent variable in empirical work. Scholars face the challenge of measuring these underlying party positions and policy dimensions. Parties reveal their positions indirectly through a variety of activities. They publish manifestos prior to elections in which they state policy goals, they make political statements and speeches, and their members cast votes in parliaments (Benoit and Laver 2006b). Currently, there are three primary methods for estimating latent party positions. Hand coding and computer-based analysis of manifestos assume that election manifestos contain precise information about party positions at a particular point in time. Expert surveys measure the positions not from primary sources, but indirectly through judgments of country specialists who rely on a variety of sources beyond manifestos to form an opinion.<sup>1</sup>

### Expert Surveys

In an ideal world, regularly conducted expert surveys may provide the best means for estimating party positions. Experts are able to synthesize large quantities of information from various sources, including manifestos, speeches, voting patterns, and media reports (Benoit and Laver 2006b). Moreover, surveys may be able to examine when new issues arise and determine their relative importance (Castles and Mair 1984; Huber and Inglehart 1995). Experts are able to tell researchers what, in their opinion, are the salient dimensions, rather than leaving the researcher to guess or assign arbitrary weights. From a pragmatic standpoint, however, expert surveys are difficult and expensive to repeat over time and across countries, requiring continuous sources of funding to conduct new surveys at regular intervals. Often, they require multilingual research teams. If a researcher realizes that a survey failed to include a question, it is impossible to go back in time to retrieve that information. Frequently, surveys phrase questions differently, making the comparisons across surveys question-

<sup>1</sup>A possible fourth method is to analyze the voting records of party members in legislatures. This is the most prominent approach used in presidential systems (e.g., roll-call analysis using NOMINATE [Poole and Rosenthal 1985]). However, in parliamentary systems voting patterns unsurprisingly reveal only a division between government parties and opposition parties due to high levels of party discipline and government agenda control (Laver 2006, 137).

able. Moreover, it is difficult to know whether different experts across countries and over time understand and answer the questions in a similar manner. While surveys often come up short as pooled cross-sectional time-series data, they do provide researchers with a method for checking the validity of position estimates from other methods in addition to providing a snapshot of party positions at one point in time (Gabel and Huber 2000).

### Hand Coding: Comparative Manifestos Project

Probably the most well-known and widely used method for generating party positions is hand coding of party manifestos. The Comparative Manifestos Project (CMP; Budge, Robertson, and Hearl 1987; Budge et al. 2001) has greatly advanced the ability of scholars to conduct comparative research by providing estimates of party positions across countries and over time. The CMP group has created 56 issues, which fall into seven major categories. To generate party positions, the CMP group codes the number of quasi-sentences which fall into each issue and then divide by the total number of quasi-sentences in the manifesto to control for manifesto length. Thus, the score for each party for each issue is simply the percentage of total sentences which fall into this issue.

To calculate party positions on a left-right dimension from these data, scholars have employed several methods. Laver and Budge (1992) provide one of the more commonly used approaches. They identify several important issues as left-wing issues and others as right-wing issues. Then they simply sum the left-wing scores and the right-wing scores and subtract the right totals from the left totals. The problem is that not all 56 categories can be attributed to the left or to the right. Thus, even though two parties may discuss the left-wing issues in an identical manner, if one party mentions neutral issues while the other does not, the positions of these parties will be coded differently.<sup>2</sup>

In addition, left and right issues may vary across countries and over time. This may create problems for constructing a valid left-right scale. For example, in

<sup>2</sup>For example, imagine two parties with very short manifestos. The first party's manifesto reads: "We support more social welfare spending." The second party's manifesto reads: "We support more social welfare spending. Decisions about this spending should be made at the local levels." Because 100% of the first party's manifesto deals with a left-wing issue, the party's score on the left-right dimension would be 1, or as far left as possible. The second party's score, on the other hand, would be 0.5 by this coding scheme. The first sentence, 50% of the manifesto, falls into a left-wing category. The second sentence, however, deals with decentralization, an issue which is coded neither left nor right. We would not want to conclude, though, that party 1 is actually located to the left of party 2 simply because party 1 remained silent on a neutral issue.

the United States, decentralization would probably be a right-wing issue while in other countries it may be a neutral issue, or even a left-wing issue. Moreover, it is not clear that all issues should be given the same weight in determining party positions, and weights may vary across countries and time. The fixed coding scheme of the CMP also means important new issues must be placed into existing categories (e.g., global terrorism after 9/11). Other categories may no longer be relevant (e.g., foreign special relations between West and East Germany after 1990).

There have been several attempts to fix the manifesto scheme. Gabel and Huber (2000), for example, suggest simply extracting the first principal component from the 56 issues, an approach they refer to as the *vanilla method*. Others have retained the seven main categories in the original dataset and then extracted principal components from each category (Klingemann 1995).

The hand-coding approach provides the only cross-sectional time-series database on party positions to date. It has the advantage that researchers know exactly what issues are included in the left-right dimension because categories are defined. However, the coding scheme of left-right positions itself is problematic and can lead to invalid positions. Moreover, because the manifestos have been coded only once, researchers do not know the uncertainty associated with this technique.<sup>3</sup> Finally, such a project is costly and difficult to replicate.

### Computer-Based Content Analysis

The most recent innovation in estimating party positions involves computer-based content analysis of party manifestos. This method attempts to reduce both the costs and likelihood of human error associated with hand coding texts. Laver, Benoit, and Garry (2003) make great advances in computer-based content analysis by suggesting the use of reference texts rather than hand-coded dictionaries.<sup>4</sup> Using this approach, researchers first identify reference texts known to represent the extremes of the political space (and possibly the center as well). This one-dimensional space is anchored by assigning reference values to the reference texts, ideally obtained from previous

<sup>3</sup>A recent paper attempts to fix the uncertainty problem and generates confidence intervals by bootstrapping quasi-sentences (Benoit, Laver, and Mikhaylov 2007).

<sup>4</sup>Earlier computer coding schemes relied on linking texts with computer-based dictionaries containing words or phrases associated with predetermined policy positions (Laver 2001). However, as Laver, Benoit, and Garry (2003, 312) note, this method does not actually cut down on the human effort as it requires teams of researchers to input large, hand-coded dictionaries, and therefore the likelihood of human error remains.

expert surveys. Laver, Benoit, and Garry's computer program *Wordscores* then counts the number of times each word occurs in the reference texts and compares these counts to word counts from the texts being analyzed. The manifestos are placed on a continuum between the reference texts depending on how similar the word counts are to each reference text. This method clearly constitutes a breakthrough for quantitative content analysis of manifestos. It is easy to implement, and researchers can apply it in almost any setting.

Nevertheless, there are several issues with the *Wordscores* technique, which our approach aims to address. First, the usefulness of the *Wordscores* approach hinges on the ability of the researcher to identify appropriate reference texts and reference values. Scholars or experts can reasonably disagree about the extremes of the political space. The choice of reference values becomes even more critical when positions are estimated for more than one dimension. To estimate multiple dimensions, Laver and his co-authors propose that researchers use different reference values on the exact same references texts. This is problematic for two reasons. First, they suggest that it is feasible to generate specific policy dimension estimates from the entire manifesto, even though only some parts of the text deal with the issue under investigation. Second, if analysts have the same two extreme reference texts for all policy dimensions, then party placements hinge on the reference values attributed to the center parties alone. Exogenous measures of a single reference party position could therefore determine the *Wordscores* results.<sup>5</sup>

Second, *Wordscores* assigns all words the same weight in the estimation process. Thus, words that occur frequently in all texts and provide little political information, such as conjunctions and articles, pull the document scores towards the center of the space, making these scores incomparable with the original reference values assigned to the reference texts. To make these scores comparable, Laver, Benoit, and Garry (2003) rescale the raw scores by stretching the variance of document scores to equal the variance of the reference text scores. Martin and Vanberg (2008) point out, however, that the particular rescaling algorithm used by Laver, Benoit, and Garry (2003) does not place the transformed scores on the same metric as the reference texts. They offer a new rescaling technique which

<sup>5</sup>This is the case for the U.K. example in their article (Laver, Benoit, and Garry 2003). If the researcher fails to use the Liberal Democratic party's manifesto as a reference text, only unidimensional estimation is possible. It is possible to get around this issue by using only sections of the manifesto which deal with the policy dimensions of interest. Proksch and Slapin (2006) parse the reference texts into economic and social sections and then estimate positions using the respective sections only.

leads to different results from those produced by the original rescaling procedure. We avoid this problem entirely by estimating the importance of words for discriminating between party positions rather than treating all words equally.

Finally, time-series estimation is problematic using *Wordscores*. The *Wordscores* authors argue that their technique should not be used for time-series analysis because the political lexicon is constantly in flux (Benoit and Laver 2006a, 133). Nevertheless, scholars seem willing to assume that political language is sufficiently stable to use this technique for time-series estimation (Budge and Pennings 2006; Hug and Schulz 2007; McGuire and Vanberg 2005). The bigger issue for time-series estimation using *Wordscores* is the proper identification of reference texts. This challenge has led researchers to adopt various approaches in order to apply *Wordscores* to time-series data, all of which come with their own problems. Some analysts concatenate all manifestos over the entire time period in order to produce long reference texts (Budge and Pennings 2006), others run the algorithm twice using two different sets of reference texts from different time periods (Hug and Schulz 2007), and, lastly, some pick two reference texts from different time periods assuming that these constitute the extremes during the entire period (McGuire and Vanberg 2005).<sup>6</sup> Time-series party positions can be estimated with *Wordscores* if one is ready to make three assumptions. First, the political lexicon remains sufficiently stable over time, second, chosen reference texts include *all* relevant words over time, and third, the reference texts represent the most extreme positions during the time period. We propose an approach which does not rely on reference texts and therefore does not make the latter two assumptions.

<sup>6</sup>Budge and Pennings (2006) apply *Wordscores* for a 20-year period by concatenating reference manifestos over this period and assigning averaged left-right scores from the CMP dataset as reference values. As Benoit and Laver point out, "such a procedure is guaranteed to produce flat times series, with the only difference between party estimates being associated with the average positions over the time period—not individual changes at different time periods" (Benoit and Laver 2006a, 134). Hug and Schulz (2007) address the time problem by estimating two different sets of Swiss party positions, using reference values from surveys in 1991 and 2003. The first reference values and texts are used to estimate positions between 1947 and 1995, the second for positions between 1995 and 2003. This creates two problems. First, the vocabulary in the 1991 reference texts might miss important words relevant in the previous elections (1947–91). Second, the authors present the two different sets of estimates as a single time series by concatenating the estimates, even though different texts were used to anchor the parties. Finally, McGuire and Vanberg (2005), estimating the positions of U.S. Supreme Court decisions on religion, chose a conservative decision from 1962 and a liberal decision from 2000 as reference texts, simply asserting that these cases mark the extremes over time.

## A Scaling Approach to Party Positions

This article presents an easy-to-implement statistical scaling model to estimate time-series policy positions from political texts. Like other manifesto-based position estimates, this approach assumes that relative word usage of parties provides information about their placement in a policy space. The advantage of this new approach is three-fold: its ability to produce time-series estimates, the fact that it does not require the use of reference texts because it instead assumes an underlying statistical distribution of word counts, and, lastly, the ability to use all words in every document and to estimate the importance of each of these words.

This approach draws on a long tradition of quantitative analysis of text. Authorship studies, for example, try to identify authors based on their literary styles. To do so, linguists attempt to uncover characteristics of a particular author by measuring and counting stylistic traits (Holmes 1985; Peng and Hengartner 2002). This technique has been prominently applied in political science to identify authorship of the unsigned *Federalist Papers* (Mosteller and Wallace 1964).

The process by which words are generated in a text is highly complex, but to facilitate analysis, linguists commonly use a *naïve Bayes* assumption in applied work (Eyheramendy, Lewis, and Madigan 2003; Lewis 1998). A text is represented as a vector of word counts or occurrences. Individual words are assumed to be distributed at random. Put differently, the probability that each word occurs in a text is independent of the position of other words in the text. It has been pointed out that "while this assumption is clearly false in most real-world tasks, naïve Bayes often performs classification very well" (McCallum and Nigam 1998, 1). Scholars then have tried to determine statistical distributions which most accurately approximate word usage. Commonly used distributions include the Poisson (Mosteller and Wallace 1964), the negative binomial (Mosteller and Wallace 1964) and other Poisson mixtures (Church and Gale 1995), as well as zero-inflated (binomial) distributions (Jansche 2003). All of these distributions are heavily skewed, as is the case of word usage.

Political scientists have started to make use of the *naïve Bayes* assumption and word frequency distributions to analyze political text. Monroe and Maeda (2004) use a Poisson word count distribution to extract multidimensional positions of U.S. legislators from their speeches. They find that the principal dimension of speech in the U.S. Congress is of a linguistic nature, with the second dimension yielding policy-relevant results.

We analyze word frequencies of party manifestos and assume the frequencies are generated by a Poisson process.

This particular distribution is chosen because of its estimation simplicity: it only has one parameter,  $\lambda$ , which is both the mean and the variance. This assumption means that the number of times party  $i$  mentions word  $j$  in election year  $t$  is drawn from a Poisson distribution. This model specification is essentially a Poisson *naïve Bayes* model and has also been used by Monroe and Maeda (2004). We later apply other distributions to test the robustness of our findings to the distributional assumption. The functional form of the model is as follows:

$$\begin{aligned} y_{ijt} &\sim \text{Poisson}(\lambda_{ijt}) \\ \lambda_{ijt} &= \exp(\alpha_{it} + \psi_j + \beta_j * \omega_{it}) \end{aligned}$$

where  $y_{ijt}$  is the count of word  $j$  in party  $i$ 's manifesto at time  $t$ ,  $\alpha$  is a set of party-election year fixed effects,  $\psi$  is a set of word fixed effects,  $\beta$  is an estimate of a word specific weight capturing the importance of word  $j$  in discriminating between party positions, and  $\omega$  is the estimate of party  $i$ 's position in election year  $t$  (therefore  $it$  is indexing one specific manifesto). We include word fixed effects to capture the fact that some words are used much more often than other words by all parties. The party-election year effects control for the possibility that some parties in some years may have written a much longer manifesto. The parameters of interest are the  $\omega$ 's, the position of the parties in each election year, and the  $\beta$ 's because they allow us to analyze which words differentiate between party positions.

This model treats each election manifesto as a separate party position and all positions are estimated simultaneously. In other words, the position of party  $i$ 's manifesto in election  $t-1$  does not constrain the position of party  $i$ 's manifesto in election  $t$ . If a party maintains a similar position from one election to the next, it means the party has used words in similar relative frequencies over time. On the other hand, if the model indicates that a party moves away from its former position and closer to the position of a rival, it implies that the party's new word choice more closely resembles that of the rival's than of its former self. An alternate specification might assume that a party's position at time  $t$  is both a function of its word choice at time  $t$  and its position in previous elections. Such a specification might ensure smooth party movement over time, but the movement would both be a function of the word usage and the assumptions about the model's functional form. The current specification has the advantage that observed party movement is, in fact, due to changes in word frequencies and is not an artifact of the model.

As specified, the model estimates positions on a single dimension. Using the entire manifesto text as data, we expect this dimension to correspond to a left-right politics dimension, which we confirm by comparing the results

to other estimates of left-right positions. This expectation is justified if manifestos (or other documents being analyzed) are encyclopedic statements of the parties' positions.<sup>7</sup> To obtain specific policy positions, we modify the text data to be analyzed. For example, we estimate economic positions by running the model on manifesto sections regarding economic policy only. This approach is in contrast to Monroe and Maeda (2004) and other factor analytic techniques, which interpret multidimensional scores ex post. It is also different from Laver, Benoit, and Garry (2003), who estimate different dimensions not by altering the text inputs but by changing the reference values assigned to reference texts.

## Estimation

Unlike a standard Poisson regression model, the entire right-hand side of the equation needs to be estimated. To do this, we use an expectation maximization (EM) algorithm. The EM algorithm is an iterative procedure to compute maximum likelihood estimates for latent variables (McLachlan and Krishnan 1997). The E step involves calculating the expectation of the latent variable as if it were observed. The M step then maximizes the log-likelihood conditional on the expectation. The implementation of this algorithm entails several steps:

### Step 1: Calculate starting values.

We obtain starting values for word fixed effects ( $\psi$ ) by calculating the logged mean count of each word. For the party fixed effects ( $\alpha$ ), we use the logged ratio of the mean word count of each party-election manifesto relative to the first party election in our dataset. We set the starting values relative to the first party-election because this party fixed effect is set to zero during the estimation in order to identify the model. To obtain starting values for word weights ( $\beta$ ) and party positions ( $\omega$ ) from the word frequencies, we first subtract the starting values for the word and party fixed effects from the logged word frequencies. We then use the left- and right-singular vectors from a singular value decomposition of this matrix as starting values for  $\omega$  and  $\beta$ .

### Step 2: Estimate party parameters.

We estimate party parameters ( $\omega$  and  $\alpha$ ) conditional on our expectation for the word parameters. In the first iteration, our expectation of those word parameters equals their starting

<sup>7</sup>We thank an anonymous reviewer for pointing this out to us.

values calculated in step 1. We maximize the following log-likelihood for each party-election  $i$ :

$$\sum_{j=1}^m (-\lambda_{ijt} + \ln(\lambda_{ijt}) * y_{ijt}),$$

where

$$\lambda_{ijt} = \exp(\alpha_{it} + \psi_j^{start} + \beta_j^{start} * \omega_{it}).$$

We use  $\omega_{it}^{start}$  and  $\alpha_{it}^{start}$  as starting values in the maximization stage. To identify the model, in addition to setting  $\alpha_1$  to 0, we set the mean of all party positions across all elections to 0 and the standard deviation to 1. This identification strategy allows party positions to change over time relative to the mean position because we fix the total variance of all positions over time. We do not hold the variance or the mean in each election constant, as this would not allow us to make interpretations about party movements over time.

#### Step 3: Estimate word parameters.

We estimate word parameters ( $\psi$  and  $\beta$ ) conditional on our expectation for the party parameters, which we obtain in step 2. For each word  $j$ , we maximize the log-likelihood:<sup>8</sup>

$$\sum_{it=1}^n (-\lambda_{ijt} + \ln(\lambda_{ijt}) * y_{ijt}),$$

where

$$\lambda_{ijt} = \exp(\alpha_{it}^{step2} + \psi_j + \beta_j * \omega_{it}^{step2}).$$

#### Step 4: Calculate log-likelihood.

The log-likelihood of our model is the sum of the individual word log-likelihoods from step 3, which are themselves calculated conditional upon the party log-likelihoods from step 2:

$$\sum_j^m \sum_{it=1}^n (-\lambda_{ijt} + \ln(\lambda_{ijt}) * y_{ijt}).$$

#### Step 5: Repeat steps 2–4 until convergence.

Using the new expectations for the word parameters, we reestimate party parameters (step 2). Then, using those expectations, we reestimate word parameters (step 3). This process is repeated until an acceptable level of convergence, measured as the difference in the log-likelihood from

<sup>8</sup>We include in this log-likelihood a relatively diffuse word-specific prior in order to prevent words from carrying infinite weight. The prior belief is that  $\beta$ s are distributed normally with mean of zero and standard deviation  $\sigma$ . This reduces the weight given to words that are mentioned very infrequently (e.g., by only one party in one election) which might otherwise discriminate perfectly. The prior solves a technical problem, but has no effect on our estimated party positions.

step 4 between the current and the previous iteration, is reached.

## 95% Confidence Intervals

We obtain confidence intervals for the estimates using a parametric bootstrap. We first estimate all parameters by running the EM algorithm described above. From these ML estimates, we calculate  $\lambda_{ijt}$  for each cell in the dataset. We then generate 500 new datasets, each time taking random draws from a Poisson distribution with parameter  $\lambda_{ijt}$  for each cell in the word count matrix. Finally, using the ML estimates as starting values, we rerun the algorithm on each of these datasets and estimate 500 new party positions. We use the 0.025 and the 0.975 quantiles of the simulated party positions as an approximate 95% confidence interval.<sup>9</sup> Our method for estimating party positions is one of few which allows researchers to measure the uncertainty associated with the estimation.<sup>10</sup>

The parametric bootstrap has the desirable property that the confidence intervals shrink as the number of words increases, something which should be true of confidence intervals of estimates from text analysis (Benoit, Laver, and Mikhaylov 2007; Laver, Benoit, and Garry 2003). We have tested this with a Monte Carlo simulation (Appendix B). First, true parameter values for the party positions were fixed, and the remaining parameter values were drawn from random distributions. Second, simulated word frequencies were generated by taking random draws from a Poisson distribution using the true parameter values to calculate  $\lambda_{ijt}$ . Finally, the simulation generated confidence intervals from 100 bootstraps. We repeated this procedure, each time increasing the number of unique words being used in the estimation, starting with 25 words and ending with 10,000 words. Because we only increase the number of unique words in this procedure while holding party positions fixed, only the error surrounding these estimates should vary. The simulation demonstrates that the average confidence

<sup>9</sup>The same is possible for the word weights.

<sup>10</sup>We are not alone in relying on the parametric bootstrap to produce standard errors for this type of analysis. Lewis and Poole (2004) suggest a parametric bootstrap to generate confidence intervals for ideal point estimates obtained from NOMINATE. As far as text-based approaches are concerned, *Wordscores* generates standard errors through the dispersion of individual word scores around the text's mean score, but these error estimates need to be transformed and rescaled in the same manner as the raw text score (Laver, Benoit, and Garry 2003, 317). Monroe and Maeda (2004) use Gibbs sampling embodied in Bayesian approaches to generate confidence intervals. A recent paper by Benoit, Laver, and Mikhaylov (2007) bootstraps quasi-sentences to generate error estimates for the CMP data. The different approaches to generate standard errors make their comparability across methods difficult.

interval for party positions decreases substantially as texts get longer. The average 95% confidence interval is almost six times larger for 25 unique words than for 500 unique words, and the interval is still 2.5 times larger for 500 words compared with 5,000 unique words. The reason for this decrease is that the model treats each unique word as an independent observation. More words mean more data for estimating party positions, and hence smaller confidence intervals.

We have tested several alternatives to this method for producing confidence intervals, but believe the parametric bootstrap provides a good compromise between all of these approaches. The first alternative to our method would involve a nonparametric bootstrap. This approach would sample words from each text with replacement to generate new manifestos. In simulations, we have found this problematic for text data. The simulated manifesto data do not correspond on average to actual manifesto word counts. Infrequent words in the manifesto rarely appear in the simulated data, leading to confidence intervals that do not encompass the ML position estimate. As a second alternative, after obtaining the ML estimates, one could numerically calculate a Hessian matrix, take the negative inverse of this matrix to obtain a variance/covariance matrix for the entire parameter space, and take draws from a multivariate normal distribution to obtain simulated parameter values. However, given the number of parameters typically being estimated in our model, computational obstacles make it impossible to calculate such a large variance/covariance matrix. Third, rather than using a Poisson model, one could revert to a negative binomial model with an overdispersion parameter. Because we use a parametric bootstrap, the confidence intervals we generate are sensitive to our distributional assumptions. Wrong distributional assumptions will generate poor simulated data and lead to invalid estimates of uncertainty. King notes, for example, that the Poisson model will produce biased standard errors in the presence of over- or underdispersion (King 1998, 128). Simulations reveal, however, that confidence intervals produced using the negative binomial model only increase slightly compared with the Poisson model, while the computational effort to generate them vastly increases. This leaves us with the Poisson model using a parametric bootstrap as the most feasible method to obtain confidence intervals.

## Implementation in R: WORDFISH

To implement the routine, we have written a computer program *Wordfish* for the *R* statistical language.<sup>11</sup> As

<sup>11</sup> *Wordfish* is available at [www.wordfish.org](http://www.wordfish.org).

input, the program requires a word frequencies matrix.<sup>12</sup> The code then takes the word frequency dataset, generates starting values, and runs the algorithm. It outputs the party positions along with the word weights and party and word fixed-effects. In addition, the program can generate confidence intervals from a parametric bootstrap.<sup>13</sup>

Like all statistical models, *Wordfish* makes several assumptions which researchers should keep in mind when using the method. To estimate positions over time, the model assumes—like users of *Wordscores* do—that word meanings remain stable. An alternative estimation strategy would hold only a subset of word weights fixed, while allowing the remaining words to have different weights in different time periods. Such an approach would naturally come at the cost of making the model more time consuming to estimate. In addition, it would require subjective judgments on the part of the researcher as to which word parameters to allow to vary and which ones to hold fixed. Researchers would have to state *a priori* which words' meanings have changed over time and which have not. Because of the inherent difficulty of this task, we opt to assume that all word parameters are fixed over time. Moreover, it is not possible to allow all word parameters to vary across time because the model would be unidentified. To identify the model, we would have to hold party positions fixed, and, given we are interested in party movement over time, this would make little sense. However, we do believe that our approach has an advantage in estimating time-series positions because it uses words from all documents. If the political lexicon changes through words entering and exiting the political dialogue, rather than through words changing meaning, our method does take these changes into account when estimating positions.

With regard to dimensionality, *Wordfish* assumes the principle dimension extracted from texts captures the political content of those texts. In other words, if researchers want estimates of party positions regarding foreign policy, they should run the program on documents containing information about foreign policy only. Such a decision is

<sup>12</sup> Easy-to-use programs are *Yoshikoder* and *jfreq*, the latter of which can be called from within *R* (Lowe 2007), available at <http://people.iq.harvard.edu/~wl Lowe/Software.html>.

<sup>13</sup> To demonstrate that our program produces valid parameter estimates, we run a simulation generating word counts using our Poisson model as the data-generating process. First, we set the true parameter values. With the exception of party positions, which we fix, these are drawn at random from a distribution so that the resulting word counts resemble real manifesto data. Second, we generate the word frequencies by taking random draws from a Poisson distribution using the true parameter values to calculate  $\lambda_{ijt}$ . Finally, we run the code which calculates the starting values and then performs the EM algorithm. The estimated parameters correlate highly with the true values. The correlation between estimated party positions and the truth is always greater than 0.99. The other parameter estimates correlate with the truth at .9 or greater.

nontrivial. It means that a researcher must carefully read the manifesto to be able to divide it into issue areas, or policy dimensions. Naturally, this requires the knowledge of the document language. Different researchers may make different decisions about which parts of the manifesto refer to economic policy. This leads to an additional source of error which we do not take into account here. If the researcher is not concerned about specific dimensions and is confident the texts under investigation represent the totality of the authors' policy positions, he or she can confidently extract a left-right dimension.

Therefore, when analyzing more than one dimension, we recommend that researchers first define the dimensions *ex ante* and, second, use only documents that contain information relevant to that dimension. Defining the dimension includes being transparent about what information is being used. For example, a researcher might define a foreign policy dimension as including texts on security, defense, and the United Nations. Others might disagree with this definition and develop a different one. However, only documents which deal with the dimension and issue of interest should be compared. In practice, parties divide manifestos into issue areas themselves to make them more readable and accessible to party members and the electorate. This facilitates the task of defining policy dimensions. In addition, *Wordfish* gives researchers the ability to analyze the degree to which the estimates capture the dimension under investigation by estimating the word-discrimination parameters. For example, words related to foreign policy should presumably receive a great deal of weight when examining foreign policy texts. If they do not, the researcher may want to consider reexamining the source documents.

## Estimates for German Parties, 1990–2005

We apply this new technique to estimate the positions of German parties in the postreunification era (1990–2005).<sup>14</sup> The estimation requires three steps: defining policy dimensions, generating the word frequency dataset, and running the algorithm. We perform two analyses: a

<sup>14</sup>German Manifestos in electronic format were made available from the Zentralarchiv für Empirische Sozialforschung, Universität zu Köln. The manifestos were transferred into electronic format by Paul Pennings and Hans Keman, Vrije Universiteit Amsterdam, Comparative Electronic Manifestos Project, in cooperation with the Social Science Research Centre Berlin (Andrea Volkens, Hans-Dieter Klingemann), the Zentralarchiv für empirische Sozialforschung, GESIS, Universität zu Köln, and the Manifesto Research Group (chairman: Ian Budge).

left-right dimensional analysis using the entire manifesto of each party in each election, and a multidimensional analysis using particular sections of each manifesto (economic, societal, and foreign policies).

Our first analysis uses the entire manifesto text, and we expect our results to capture a basic left-right dimension of German politics. In the second analysis, we calculate positions for individual dimensions of interest. Here, we concentrate our analysis on economic, societal, and foreign policies.<sup>15</sup> Each manifesto text is thus divided into three separate files. We then run our algorithm on each dimension separately and retrieve three positions for each party.<sup>16</sup>

We follow a scheme applied to German manifestos by König, Blume, and Luig (2003) to divide up the manifestos into policy-specific sections.<sup>17</sup> The economic dimension captures socioeconomic policies including taxes, revenues, and spending. The foreign dimension covers international political and economic affairs as well as relations with the European Union. Finally, the societal dimension includes diverse areas such as law and order, gender equality, higher education, immigration, housing, and sport. Once the dimensions are defined and the manifesto texts are compiled, we generate a word frequency dataset. The rows of this matrix correspond to a party manifesto from a particular election and the columns to all unique words mentioned in the texts. This means that we have 25 rows (five parties, five elections) and several thousand columns depending on the number of unique

<sup>15</sup>We use the term "societal" rather than "social" because we believe the term "societal" is broader. We include several issues in this dimension, such as environmental politics, which are not usually categorized as social politics, but they clearly have societal ramifications.

<sup>16</sup>These are three separate unidimensional positions. In the present context of our model, it is not possible to determine whether these dimensions are orthogonal to one another, nor do we know the relative weights of the dimensions.

<sup>17</sup>The scheme divides up the manifesto as follows. *Economic Policy*: agriculture, budget, revenue, taxes, consumer protection, deregulation, energy, future policies, general health policy, industrial policy, infrastructure, labor market, pensions, policies concerning Eastern Germany, research and development, trade, welfare state. *Societal Policy*: animal rights, culture, direct democracy and constitutional reform, anti-drug and HIV policies, children, education (including higher education), environmentalism (except energy policy), family, fight against extremism and terrorism (except on the international level), gender equality, housing, immigration, law and order, traditional morals, multiculturalism, seniors (except pensions), sport. *Foreign Policy*: defense and security, European Union, global affairs, international terrorism, world trade. *Left-Right*: economic sections + societal sections + foreign sections. We excluded the following manifesto sections from the analysis: general introduction of a manifesto/preamble, review of the previous parliamentary term, reference to other parties and their manifesto, conclusion of a manifesto.

words for each dimension. While it is possible to estimate positions using the entire party-word matrix, we remove words that parties use infrequently and thus contain little information about their placement. We include a word in the estimation if it was mentioned at least once on average by each party during the period between 1990 and 2005. This has three practical advantages. First, it speeds up the estimation process by eliminating the “long tail” in our dataset. Second, it ensures that our estimation results do not hinge on these infrequently mentioned words. Lastly, it eliminates the possibility that spelling mistakes or other minor and infrequent errors affect our estimates.<sup>18</sup>

## Position Estimates

Figure 1a plots the party position estimates ( $\omega$ ) for the main left-right dimension.<sup>19</sup> The estimates reflect several important changes in the party system over time. Since reunification, the former East German communist *Party of Democratic Socialism* (PDS) has occupied the left end of the political spectrum. The *Greens* start out on the left in 1990, but move slightly towards the political center up until the most recent election in 2005. This movement reflects the transformation of the *Greens* from an environmentalist fringe party in the 1980s to a mainstream governing party by 1998. Most importantly, our estimates pick up the significant right shift of the *Social Democratic Party* (SPD) throughout the 1990s. This matches conventional wisdom that Chancellor Gerhard Schröder moved the traditional left-wing socialist party to the political center to recapture government in same way that Tony Blair moved the British Labour Party to the center with his “Third Way.” In addition, we see a left shift by both the SPD and the PDS in 2005. This may be explained by a split in the SPD. The left wing of the SPD, led by former party

<sup>18</sup>We have run the analysis using all words, and the result correlates very highly with the results we report ( $r = .98$ ); however, the estimation does take substantially longer.

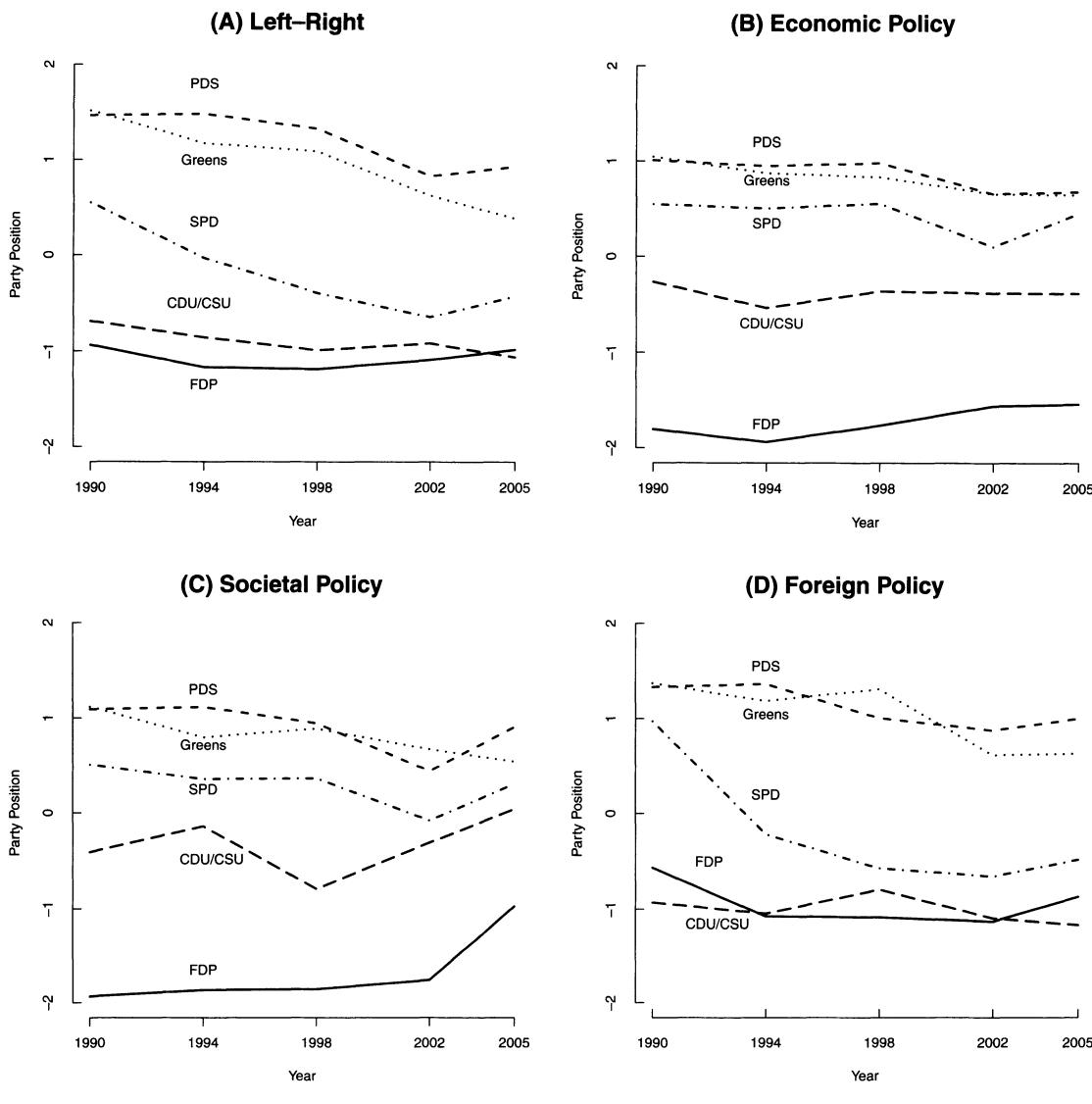
<sup>19</sup>Appendix A lists the estimated German party positions since 1990 on all dimensions with their respective confidence intervals. It also presents a summary of the estimation results, including the number of unique words, the number of party elections, the number of iterations, the log-likelihood, and the mean absolute difference in the estimated party positions between the last and the previous iteration. To give a rough indication of estimation time, it takes about 45 minutes for the code to converge estimating the main left-right positions (25 documents containing approximately 9,000 unique words). Estimation times will increase with both the number and length of texts and also depend upon computing speed. This analysis was performed on a PC with a 1.73 Ghz Intel processor and 760 MB RAM. The bootstrap procedure generating the 95% confidence intervals can take up to a few days, depending on the number of bootstraps specified.

leader Oskar Lafontaine, was upset by the party’s rightward movement under Schröder and split off to form a new party together with the PDS, *Die Linke*. The SPD needed to move left to placate their base and to avoid losing even more party members to *Die Linke*. Finally, the liberal *Free Democrats* (FDP) and the conservative *Christian Democrats* (CDU-CSU) are further to the right and remain relatively stable over time. The FDP tends to be slightly to the right of the CDU-CSU up until 2005, when it moves to the center. The confidence intervals, reported in the appendix, reveal that we can distinguish between parties in all elections except between the Greens and PDS in 1990 and between the CDU-CSU and FDP in 2005. We also find a statistically significant time trend for all parties. Nevertheless, there are several instances in which we cannot statistically distinguish between a party’s position and its position in the previous election.

Figures 1b through 1d plot our party estimates for the economic, societal, and foreign dimensions. On the economic dimension, our analysis confirms that the liberal FDP is clearly the most conservative party, demanding lower taxes and less public spending. This is reflected by the large gap between this party and the CDU-CSU. The two largest German parties (SPD and CDU-CSU) are closest to each other in 2002 and 2005. Following the 2005 election, the two parties formed a grand coalition government. In general, all party positions remain relatively stable over time on this dimension.

The societal dimension captures a wide range of policies, including immigration, education, and environment. The most significant finding for this dimension is that all parties except the *Greens* move to the left in 2005. In the context of German electoral politics, this was the year when the SPD chancellor decided to hold early elections because some of his own party members had switched over to the PDS. The FDP is still to the right of all parties. This party is often thought to be located between the SPD and the CDU-CSU on social policies. However, the dimension includes more than just social policies, making it difficult to compare this dimension to other estimates of social policy positions.

On foreign policy, a similar ranking of the parties emerges. The *Greens*, which emerged from an antiwar, pro-environmental social movement, and the PDS are located closely to each other during the first half of the 1990s. Once the *Greens* enter government in 1998, their policy positions shifts slightly towards the center. The SPD makes its most significant ideological shift throughout the 1990s, when it moves from a leftist position towards a centrist position on foreign policy. Again, this change is likely to be associated with the SPD taking over government responsibility in 1998. The CDU-CSU and the FDP

**FIGURE 1** Estimated Party Positions in Germany, 1990–2005

have similar positions. In 1990 and 2005, the FDP is more centrist and located between the two major parties.

A comparison of the size of the confidence intervals reveals that positions estimated from fewer words have larger intervals. For example, the average confidence interval for the economic policy dimension (4,714 words) is 54% larger than the average confidence interval for the left-right dimension (8,995 words). These results confirm the Monte Carlo simulation that more words reduce the uncertainty surrounding the estimates.

### Word Analysis: The Political Lexicon

To further confirm our findings, we check the validity of our results both internally and externally. For internal validation, we examine the word parameters. We expect to find a particular pattern in the results. Frequent words (e.g., conjunctions, articles, prepositions, etc.) should not discriminate between party manifestos because they do not contain any political meaning. Therefore, they should have large fixed effects associated with weights close to

**FIGURE 2 Word Weights vs. Word Fixed Effects, Left-Right Dimension, Germany 1990–2005 (Translations given in text)**

---

zero. In contrast, as words are mentioned more infrequently, they are more likely to be part of politically relevant language and discriminate between the parties. These words should therefore have smaller fixed effects associated with either positive or negative weights, depending on whether the words place parties on the left or on the right.

Figure 2 plots the estimated word fixed effects against the word weights. The scatterplot confirms our expectations and takes the shape of an “Eiffel Tower of words.” Words with a high fixed effect have zero weight, but words with low fixed effects have either negative or positive weight. The graph also highlights some words as examples. Most importantly, words with large weights have a politically relevant connotation. Manifestos on the left mention words like “fascism,” “professional ban,” “male

violence,” “emancipation,” and “pornography” more often than the ones placed on the right. The largest weight on the left is for the word “BRD,” the abbreviation for Federal Republic of Germany, a word that is used primarily by one party, the PDS. While this may appear rather trivial, in the German political context of reunification it is, in fact, an interesting result. It is well known that the official doctrine of the former communist party of East Germany (SED), the predecessor to the PDS, was to refer to West Germany in its abbreviated form in order to demonstrate its rejection of West Germany’s claim for sole right of representation. However, the official position of West German governments was to use the full constitutional name (Stevenson 2002, 50). This pattern seems to continue after reunification. On the right, parties use words such as “income taxation,” “nonwage labor costs,”

and “education vouchers” more often. The highest weight on this side is for the word “general welfare payments,” related to a long-standing proposal by the liberal democratic FDP to bundle up all welfare payments and pay them out to eligible citizens in one lump-sum payment.

Finally, words with large fixed effects do not have discriminating value. The plotted words “entry into force,” “protects,” “safe,” “they/she,” “the,” and finally the word “and” with the largest fixed effect do not contain much politically relevant information. Their associated weight is close to zero.

Table 1 completes the word analysis for all dimensions and reports the top 10 words placing parties on the left and the right. For instance, in addition to the words shown already in the figure, parties on the left use “womens’ movement” and “stratosphere” much more often, whereas parties on the right talk more about “business location” and “mobility.”

On the economic dimension, words such as “workers’ participation,” “quota,” “mobility,” and “negotiated wages” matter most. All of these are words associated with economic and labor policy. Likewise, on the societal dimension we find references to “process of reunification,” “university graduates,” “sexuality,” and “climate catastrophe.” With words as diverse as these, the results reinforce our belief that this is a category capturing societal politics broadly defined. Lastly, words such as “unilateral,” “NGOs,” “weapons production,” and “armies” all clearly refer to the foreign and defense policy domain. In addition, right parties often refer to the European defense and security policy (EDSP), the European police agency (Europol) and to the EU budget. In sum, the fact that the weights are largest for words carrying political meaning demonstrates that our model is capturing the policy space.

## Cross-Validation

Next, we cross-validate our results with existing methods (hand coding of manifestos, expert surveys, and *Wordscores*). First, we compare our results with the *Comparative Manifestos Project* left-right scale and three policy scales for Germany, 1990–98 (Budge et al. 2001). The CMP data constitute the only comparable time-series dataset. The three policy scales are market economy (MARKECO), welfare state (WELFARE), and international peace (INT-PEACE). We assume that these correspond to our economic, societal, and foreign dimensions. Second, we use expert survey estimates from Benoit and Laver (2006b) on a left-right dimension and on a taxes versus spending dimension for 2002–2003. Finally, we compare our estimates to *Wordscores* estimates on an economic and social

dimension from Laver, Benoit, and Garry (2003) for 1990 and 1994 and from Proksch and Slapin (2006) for 2005.

Table 2 presents the correlations between our and other position estimates. The correlations between our Poisson scaling model and the other three methods is high, suggesting that the techniques provide similar placement of parties in the political space. Unlike what Monroe and Maeda (2004) find in U.S. Congressional speeches, this indicates that the dimension we estimate is political and not solely linguistic. Almost all coefficients range between 0.8 and .98. Only our broad societal category corresponds less well to social and welfare categories of the other measures.

As an additional cross-validation, Figure 3 directly compares our left-right dimension with the *Comparative Manifestos Project* left-right scale for the years 1990–98. The CMP data suggest major changes in the party system that are inconsistent with standard accounts of German politics. First, it locates the conservative CDU-CSU closer to the *Greens* than to any other party in 1990, including its governing partner the FDP. Second, it suggests that the social-democratic SPD shoots from being next to the former communists to the position of the free-market Free Democrats, crossing the position of the Green party. It is inconceivable that a major centrist party in an established multiparty system would make such a jump. Moreover, expert survey data do not find that the SPD is to the left of the Greens in 1990 (Huber and Inglehart 1995). In contrast, our method provides less extreme party movements in the 1990s, eliminating the unlikely crossovers suggested by the CMP data. We find that the SPD makes a more modest move relative to the other parties, remaining in the center of the space throughout the period. Our estimates furthermore match the rankings of the parties from the Huber and Inglehart expert survey data. In general, our findings for the German party system correspond well with other methods for estimating party positions. When used as time-series data, our estimates substantially improve upon previous estimates by providing smoother party movements than those found in the CMP data.

## Robustness Checks

While the analysis of the word weights, together with the method’s high correlation to other estimates of party positions, indicate that we are capturing a primary left-right dimension in German politics, questions may remain about how robust this technique is to the texts we chose and the model specification we use. Here we demonstrate that our technique is robust to the selection of texts and our assumption about the underlying statistical distribution of word counts.

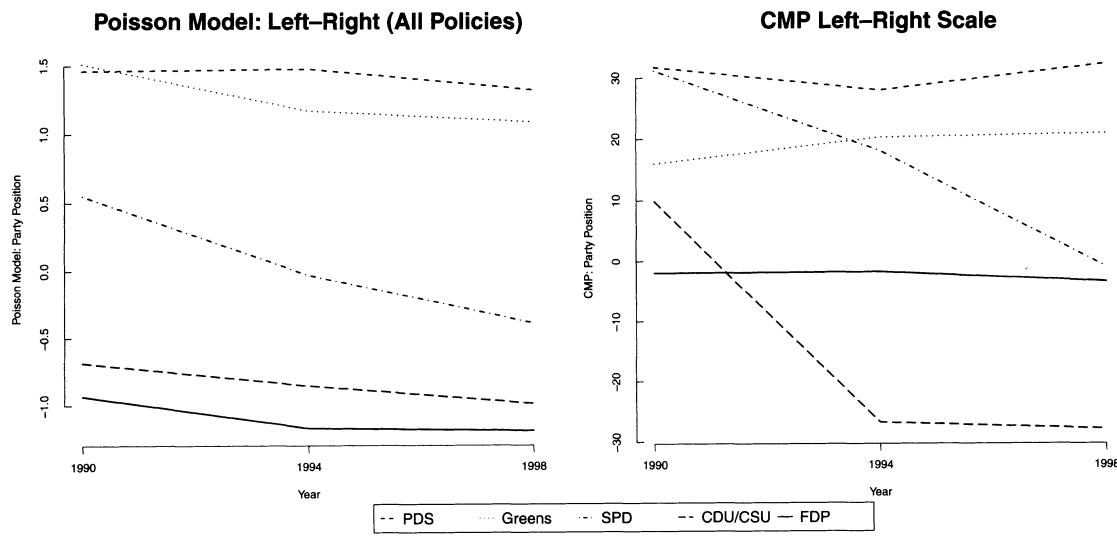
**TABLE 1 Top 10 Words Placing Parties on the Left and Right**

Dimension	Top 10 Words Placing Parties on the . . .	
	Left	Right
<b>Left-Right</b>	Federal Republic of Germany (BRD) immediate (sofortiger) pornography (Pornographie) sexuality (Sexualität) substitute materials (Ersatzstoffen) stratosphere (Stratosphäre) women's movement (Frauenbewegung) fascism (Faschismus) Two thirds world (Zweidrittewelt) established (etablierten)	general welfare payments (Bürgergeldsystem) introduction (Heranführung) income taxation (Einkommensbesteuerung) non-wage labor costs (Lohnzusatzkosten) business location (Wirtschaftsstandort) university of applied sciences (Fachhochschule) education vouchers (Bildungsgutscheine) mobility (Beweglichkeit) peace tasks (Friedensaufgaben) protection (Protektion)
<b>Economic</b>	Federal Republic of Germany (BRD) democratization (Demokratisierung) to prohibit (verbieten) destruction (Zerstörung) mothers (Mütter) debasing (entwürdigende) weeks (Wochen) quota (Quotierung) unprotected (ungeschützter) workers' participation (Mitbestimmungs-möglichkeiten)	to seek (anzustreben) general welfare payments (Bürgergeldsystem) inventors (Erfinder) mobility (Beweglichkeit) location (Standorts) negotiated wages (Tarif-Löhne) child-raising allowance (Erziehungsgeld) utilization (Verwertung) savings (Ersparnis) reliable (verlässlich)
<b>Societal</b>	Federal Republic of Germany (BRD) climate catastrophe (Klimakatastrophe) sexuality (Sexualität) pornography (Pornographie) fascism (Faschismus) irreplaceable (ersatzlos) process of reunification (Wende) women's movement (Frauenbewegung) substitute materials (Ersatzstoffen) nuclear facilities (Atomanlagen)	data processing (Datenverarbeitung) contraception counseling (Verhütungsberatung) requested (aufgefordert) questions regarding property (Eigentumsfragen) competitive sports (Leistungssport) leisure activities (Freizeitverhalten) in general (generell) animal protection law (Tierschutzgesetzes) social housing fee (Fehlbelegungsabgabe) university graduates (Hochschulabsolventen)
<b>Foreign</b>	Federal Republic of Germany (BRD) immediately (sofort) departure (Aufbruch) foreign political (aussenpolitischer) unilateral (einseitiger) Two thirds world (Zweidrittewelt) emancipation (Emanzipation) NGOs (NGOs) armies (Armeen) weapons production (Rüstungs-produktion)	cultural policy (Kulturpolitik) foreign (auswärtige) Europol (Europol) legal protection (Rechtsschutz) delimitation of competences (Kompetenz-abgrenzung) neglected (vernachlässigt) EDSP (EVSP) euro-atlantic (euro-atlantischen) introduction (Heranführung) EU budget (EU-Haushalt)

**TABLE 2 Cross-Validation: Correlations between German Party Position Estimates**

	Poisson Scaling Model			
	Left-Right	Economic	Societal	Foreign
<b>Hand-coding manifestos</b>				
CMP: Left-Right (n = 15, 1990–1998)	-0.82			
CMP: Markeco (n = 15, 1990–1998)		0.81		
CMP: Welfare (n = 15, 1990–1998)			0.58	
CMP: Intpeace (n = 15, 1990–1998)				0.81
<b>Expert Survey</b>				
Benoit/Laver 2006: Left-Right (n = 5, 2002)	-0.91			
Benoit/Laver 2006: Taxes-Spending (n = 5, 2002)		0.86		
<b>Wordscores</b>				
Laver et al. 2003: Economic (n = 10, 1990–1994)		0.93		
Laver et al. 2003: Social (n = 10, 1990–1994)			-0.47	
Proksch/Slapin 2006: Economic (n = 5, 2005)		0.98		
Proksch/Slapin 2006: Social (n = 5, 2005)			-0.47	

**FIGURE 3 Comparison of Left-Right Positions in Germany, 1990–98**



The model specification means that adding or subtracting elections or parties may affect the positions of all remaining parties. To test the extent to which our results hinge on the elections and parties we include in the analysis, we rerun the results dropping single manifestos (one party in one election year), an entire party, and an entire election year. In all cases we get results which correlate very highly with our original estimates of party positions. Our

lowest correlation with the original party positions estimates occurs when we drop an entire party, the FDP. When we do this, our results correlate with the remaining original estimates at 0.94. When we drop the entire election year 2005, the remaining party positions correlate with the original positions at 0.99. Likewise, we correlate very highly ( $r = 0.99$ ) with our original results when we drop individual manifestos (the CDU-CSU's 1990 manifesto

and the FDP's 2005 manifesto). This would suggest that even if researchers are unable to obtain all party manifestos, they can still use our method and have a high degree of confidence in their results.

In addition, we examine how well our results hold when we alter our assumption about the underlying statistical distribution of word counts. Although, from the standpoint of estimation, the Poisson distribution has the nice feature that its mean equals its variance, this assumption is unlikely to hold for word-count data (Jansche 2003; Mosteller and Wallace 1964). Therefore, we also estimate our model using a negative binomial distribution with a separate overdispersion parameter for each manifesto.<sup>20</sup> The additional parameters vastly increase the computation time, in particular when running the parametric bootstrap. The results again correlate very highly with our original party position estimates using the Poisson distribution ( $r = 0.97$ ). The only major difference between the negative binomial estimation and the Poisson estimation is that using the negative binomial estimation we find that the liberal FDP is located just to the left of the CDU, in between the CDU and the SPD, while the FDP was the most right-wing party all years except 2005 using the Poisson distribution.

Finally, we estimate our results using the simplest distribution possible, a log-normal distribution. Here, we simply regress party and word parameters on logged word counts. This also gives us virtually identical results, correlating with our Poisson estimates at 0.94. Moreover, using the log-normal, we get the same party ordering that we had in the Poisson model. In both the log-normal and negative binomial models, all the party trends remain the same, with SPD and the Greens moving to the center of the political space as they enter government.

## Conclusion

Comparative politics research requires accurate time-series estimates of party positions. Surprisingly, there is currently no easy-to-implement method that provides valid time-series positions along with measures of their uncertainty. We have presented a methodology which aims to fill this gap. We assume an underlying word frequency distribution in political text and use an EM algorithm to estimate party parameters (positions and fixed

<sup>20</sup>We use the NB2 parameterization of the negative binomial found in Cameron and Trivedi (1998). We again include our diffuse normal prior over the estimation of our word weights. An alternative implementation would be to estimate a word-specific overdispersion parameter.

effects) as well as word parameters (weights and fixed effects). Our approach adds to existing methods by providing a computer-based text analysis program, *Wordfish*, which does not require the use of reference texts. Like the *Comparative Manifestos Project*, our method can create rich time-series data, but does not require teams of potentially error-prone hand coders to do so. At the same time, like *Wordscores*, we provide easy-to-implement computer code which researchers can apply to virtually any set of political texts. Our method only requires party manifestos of those parties whose positions are to be estimated.

We have demonstrated that our approach produces estimates of party positions which correspond well with positions from other estimation techniques. We are able to accurately portray the German party system in the 1990–2005 postreunification era. Our estimated positions correlate highly with other methods. However, our approach is much less cost and time intensive, it is easily replicable, and it produces a more accurate time series with uncertainty estimates. In addition, the results for word parameters suggest that the technique captures a political, rather than linguistic, dimension.

Nevertheless, when deciding how to estimate party positions, researchers should carefully assess our set of assumptions compared with those of other computer content analysis programs. First, our method does require analysts to assume word meanings do not vary over time; however, if words enter and exit the political lexicon our approach will still capture their relative importance. Other computer content analysis methods, such as the Bayesian approach taken by Monroe and Maeda (2004), make the same assumptions about word meanings as we do, but are significantly more complicated to implement. *Wordscores* requires the additional assumptions that all words of interest are contained in the reference texts specified, and these texts represent the extremes over time. Second, researchers must decide whether they prefer to set dimensions ex ante or interpret them ex post. If researchers prefer the latter, only the Bayesian approach of Monroe and Maeda (2004) is currently able to extract more than one dimension from texts. *Wordscores* requires researchers to identify new reference values and then to assume that both their reference texts and documents of interest contain sufficient information about their dimension of interest to produce meaningful results. We suggest analysts use only documents specifically pertaining to their dimension of interest. This requires that analysts carefully select the texts they are using as data and be familiar with their content. However, we hope that all researchers using computer content analysis do this regardless of the methodology they employ. Lastly, researchers may want to consider speed and ease of use when selecting a methodology.

When estimating positions at a single point in time for which it is fairly easy to assess the political extremes and identify appropriate reference texts, *Wordscores* provides the fastest and easiest method for obtaining valid and replicable positions. In situations such as estimating a time series, for which identifying appropriate references texts is difficult or perhaps impossible, our technique provides researchers a straightforward and relatively fast technique for avoiding many of the assumptions necessary when using *Wordscores*. Finally, researchers may prefer our technique even when estimating positions at a single point in time because we are able to estimate the importance of words for discriminating between texts, thus avoiding the rescaling problem inherent in *Wordscores*.

There are certain limitations associated with our current model specification and the algorithm which open up a research agenda and which should be addressed in future work. First, although our results appear relatively robust to dropping texts (e.g., removing a party, an election, and an individual manifesto), our algorithm is sensitive to the overall number of texts used. Because each word parameter is estimated using all manifestos, the data must include a sufficient number of manifestos to avoid a small-N problem in the estimates. Second, we have shown that our results seem robust to different distributional assumptions. Nevertheless, future work should examine in

more detail the consequences of choosing one distribution over another, paying specific attention to the consequences for uncertainty estimates. Third, our current model assumes that the political lexicon remains similar over time. This is because word parameters are estimated in a time-insensitive manner in order to identify the statistical model allowing all party positions to move. Future versions of the model could relax this assumption for longer time periods and allow weights for a subset of words to vary over time, although researchers would have to make judgments about which words to allow to vary. Finally, we have opted to extract a single dimension over time and our results suggest that it is policy relevant. It would also be possible to rewrite the model to extract more than one dimension. However, we believe that in comparative politics research scholars may prefer policy dimensions whose meaning is set *ex ante* rather than interpreted *ex post*.

This set of questions opens up exciting new avenues for research on party positions and ideology estimated from political texts, which is reflected by an increasing number of studies that combine quantitative linguistic analysis with the study of political ideology. Our method takes this approach to examine party ideology over time. The results provide new insights into the German postreunification party system and its political lexicon.

## Appendix A

**TABLE A1 Party-Position Estimates (95% confidence intervals in parentheses)**

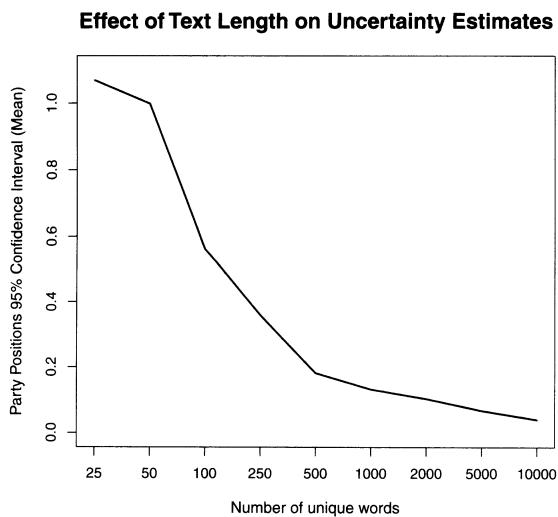
Election	Party	Left-Right	Economic	Societal	Foreign
2005	PDS	0.93 (0.87,1.01)	-0.68 (-0.75,-0.59)	0.91 (0.81,1.00)	1.00 (0.83,1.13)
	Greens	0.39 (0.36,0.46)	-0.65 (-0.70,-0.59)	0.54 (0.47,0.60)	0.64 (0.53,0.76)
	SPD	-0.42 (-0.49,-0.35)	-0.45 (-0.52,-0.36)	0.31 (0.14,0.44)	-0.48 (-0.69,-0.26)
	CDU	-1.06 (-1.12,-1.00)	0.38 (0.29,0.53)	0.04 (-0.20,0.22)	-1.16 (-1.30,-1.02)
	FDP	-0.98 (-1.02,-0.94)	1.54 (1.46,1.61)	-0.97 (-1.09,-0.86)	-0.87 (-0.95,-0.74)
2002	PDS	0.83 (0.78,0.90)	-0.66 (-0.73,-0.58)	0.44 (0.31,0.54)	0.89 (0.79,0.96)
	Greens	0.63 (0.60,0.70)	-0.66 (-0.70,-0.60)	0.67 (0.60,0.74)	0.62 (0.50,0.74)
	SPD	-0.64 (-0.69,-0.59)	-0.11 (-0.17,0.00)	-0.08 (-0.20,0.02)	-0.66 (-0.78,-0.53)
	CDU	-0.92 (-0.96,-0.87)	0.37 (0.31,0.50)	-0.31 (-0.46,-0.20)	-1.10 (-1.20,-0.99)
	FDP	-1.09 (-1.14,-1.06)	1.56 (1.49,1.62)	-1.76 (-1.82,-1.67)	-1.13 (-1.22,-1.04)
1998	PDS	1.32 (1.27,1.35)	-0.98 (-1.07,-0.95)	0.95 (0.89,1.00)	1.01 (0.89,1.09)
	Greens	1.09 (1.06,1.12)	-0.83 (-0.89,-0.80)	0.89 (0.84,0.94)	1.31 (1.24,1.37)
	SPD	-0.39 (-0.46,-0.32)	-0.56 (-0.61,-0.48)	0.36 (0.25,0.47)	-0.57 (-0.75,-0.39)
	CDU	-0.99 (-1.04,-0.94)	0.36 (0.27,0.50)	-0.79 (-0.97,-0.66)	-0.79 (-0.90,-0.68)
	FDP	-1.19 (-1.24,-1.17)	1.77 (1.67,1.81)	-1.86 (-1.91,-1.76)	-1.09 (-1.18,-1.00)

**TABLE A1 (CONTINUED)**

Election	Party	Left-Right	Economic	Societal	Foreign
1994	PDS	1.48 (1.42,1.50)	-0.95 (-1.05,-0.89)	1.12 (1.07,1.19)	1.36 (1.21,1.49)
	Greens	1.17 (1.13,1.20)	-0.87 (-0.93,-0.84)	0.80 (0.75,0.84)	1.19 (1.10,1.25)
	SPD	-0.03 (-0.08,0.05)	-0.51 (-0.56,-0.42)	0.36 (0.25,0.45)	-0.22 (-0.41,-0.03)
	CDU	-0.86 (-0.91,-0.80)	0.53 (0.44,0.68)	-0.14 (-0.29,-0.02)	-1.05 (-1.17,-0.92)
	FDP	-1.17 (-1.21,-1.14)	1.94 (1.85,1.95)	-1.87 (-1.90,-1.78)	-1.08 (-1.15,-1.01)
1990	PDS	1.46 (1.40,1.48)	-1.01 (-1.11,-0.98)	1.09 (1.04,1.16)	1.33 (1.20,1.44)
	Greens	1.51 (1.46,1.52)	-1.05 (-1.15,-1.03)	1.12 (1.10,1.18)	1.37 (1.31,1.44)
	SPD	0.55 (0.47,0.65)	-0.55 (-0.63,-0.44)	0.51 (0.36,0.64)	0.97 (0.79,1.13)
	CDU	-0.69 (-0.78,-0.60)	0.26 (0.11,0.50)	-0.41 (-0.60,-0.23)	-0.93 (-1.10,0.74)
	FDP	-0.94 (-0.98,-0.89)	1.81 (1.71,1.84)	-1.93 (-1.99,-1.84)	-0.57 (-0.67,-0.45)
Unique Words		8995	4714	4817	2200
Iterations		111	178	26	18
Log-Likelihood		841237.4	233023.5	247732.2	62880.7
Difference in $\omega$		$8.22 * 10^{-4}$	$2.48 * 10^{-3}$	$2.19 * 10^{-3}$	$2.10 * 10^{-3}$

## Appendix B

**FIGURE B1 Simulation: Text Length and Uncertainty Estimates**



*Note:* Simulations based on 25 parties and 100 bootstraps. Party positions are identified with mean = 0 and stand. dev. = 1.

## References

- Baron, David P. 1991. "A Spatial Bargaining Theory of Government Formation in Parliamentary Systems." *American Political Science Review* 85(1): 137–64.
- Bawn, Kathleen. 1999. "Money and Majorities in the Federal Republic of Germany: Evidence for a Veto Players Model of Government Spending." *American Journal of Political Science* 43(3): 707–36.
- Benoit, Kenneth, and Michael Laver. 2006a. "Benchmarks for Text Analysis: A Response to Budge and Pennings." *Electoral Studies* 26(1): 130–35.
- Benoit, Kenneth, and Michael Laver. 2006b. *Party Policy in Modern Democracies*. London: Routledge.
- Benoit, Kenneth, Michael Laver, and Slava Mikhaylov. 2007. "Estimating Party Policy Positions with Uncertainty Based on Manifesto Codings." Prepared for the 2007 annual meeting of the American Political Science Association.
- Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tanenbaum. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998*. Oxford: Oxford University Press.
- Budge, Ian, and Paul Pennings. 2006. "Do They Work? Validating Computerised Word Frequency Estimates Against Policy Series." *Electoral Studies* 26(1): 121–29.
- Budge, Ian, David Robertson, and Derek Hearl. 1987. *Ideology, Strategy, and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press.
- Cameron, A. Colin, and Pravin K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge University Press.

- Castles, Francis G., and Peter Mair. 1984. "Left-Right Political Scales: Some Expert Judgements." *European Journal of Political Research* 12(1): 73–88.
- Church, Kenneth W., and William A. Gale. 1995. "Poisson Mixtures." *Natural Language Engineering* 1(2): 163–90.
- Crombez, Christophe. 1996. "Minority Governments, Minimal Winning Coalitions and Surplus Majorities in Parliamentary Systems." *European Journal of Political Research* 29(1): 1–29.
- de Swaan, Abram. 1973. *Coalition Theories and Cabinet Formation*. Amsterdam: Elsevier.
- Druckman, James N., Lanny W. Martin, and Michael F. Thies. 2005. "Influence without Confidence: Upper Chambers and Government Formation." *Legislative Studies Quarterly* 30(4): 529–48.
- Druckman, James N., and Michael F. Thies. 2002. "The Importance of Concurrence: The Impact of Bicameralism on Government Formation and Duration." *American Journal of Political Science* 46(4): 760–71.
- Eyheramendy, Susana, David Lewis, and David Madigan. 2003. "On the Naive Bayes Model for Text Categorization." Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics.
- Franzese, Robert J. 2002. *Macroeconomic Policies of Developed Democracies*. Cambridge: Cambridge University Press.
- Gabel, Matthew J., and John D. Huber. 2000. "Putting Parties in Their Place: Inferring Left-Right Ideological Positions from Party Manifestos Data." *American Journal of Political Science* 44(1): 94–103.
- Garrett, Geoffrey. 1998. *Partisan Politics in the Global Economy*. Cambridge: Cambridge University Press.
- Hallerberg, Mark, and Scott Basinger. 1998. "Internationalization and Changes in Tax Policy in OECD Countries: The Importance of Domestic Veto Players." *Comparative Political Studies* 31(3): 321–52.
- Holmes, D. I. 1985. "The Analysis of Literary Style—A Review." *Journal of the Royal Statistical Society* 148(4): 328–41.
- Huber, John D., and Ronald Inglehart. 1995. "Expert Interpretations of Party Space and Party Locations in 42 Societies." *Party Politics* 1(1): 73–111.
- Hug, Simon, and Tobias Schulz. 2007. "Left-Right Positions of Political Parties in Switzerland." *Party Politics* 13(3): 305–30.
- Jansche, Martin. 2003. "Parametric Models of Linguistic Count Data." 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 288–95.
- King, Gary. 1998. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor: University of Michigan Press.
- Klingemann, Hans-Dieter. 1995. "Party Positions and Voter Orientations." In *Citizens and the State*, ed. Hans-Dieter Klingemann and Dieter Fuchs. Oxford: Oxford University Press, 183–205.
- König, Thomas, Till Blume, and Bernd Luig. 2003. "Policy Change without Government Change? German Gridlock After the 2002 Elections." *German Politics* 12(2): 86–146.
- Laver, Michael, ed. 2001. *Estimating the Policy Positions of Political Actors*. London: Routledge.
- Laver, Michael. 2006. "Legislatures and Parliaments in Comparative Context." In *Oxford Handbook of Political Economy*, ed. Barry Weingast and Donald Wittman. Oxford: Oxford University Press, 121–40.
- Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(2): 311–32.
- Laver, Michael, and Ian Budge, eds. 1992. *Party, Policy, and Government Coalitions*. London: St. Martin's Press.
- Laver, Michael, and W. Ben Hunt. 1992. *Policy and Party Competition*. New York: Routledge, Chapman and Hall.
- Lewis, David D. 1998. "Naïve (Bayes) at Forty: The Independence Assumption in Information Retrieval." *Proceedings of the 10th European Conference on Machine Learning*, 4–15.
- Lewis, Jeffrey B., and Keith T. Poole. 2004. "Measuring Bias and Uncertainty in Ideal Point Estimates via the Parametric Bootstrap." *Political Analysis* 12(2): 105–27.
- Lowe, Will. 2007. "Yoshikoder: Multilingual Content Analysis Software in Java." <http://www.yoshikoder.org>.
- Martin, Lanny W., and Georg Vanberg. 2008. "A Robust Transformation Procedure for Interpreting Political Text." *Political Analysis* 16(1): 93–100.
- McCallum, Andrew, and Kamal Nigam. 1998. "A Comparison of Event Models for Naïve Bayes Text Classification." AAAI-98 Workshop on Learning for Text Categorization.
- McGuire, Kevin T., and George Vanberg. 2005. "Mapping the Policies of the U.S. Supreme Court: Data, Opinions, and Constitutional Law." Prepared for delivery at the annual meeting of the American Political Science Association, Washington, DC.
- McLachlan, Geoffrey J., and Thriyambakam Krishnan. 1997. *The EM Algorithm and Extensions*. New York: Wiley.
- Monroe, Burt L., and Ko Maeda. 2004. "Talk's Cheap: Text-Based Estimation of Rhetorical Ideal-Points." 21st Annual Summer Meeting, Society for Political Methodology, Stanford University.
- Mosteller, Frederick, and David L. Wallace. 1964. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York: Springer Verlag.
- Peng, Roger D., and Nicolas W. Hengartner. 2002. "Quantitative Analysis of Literary Style." *The American Statistician* 56(3): 175–85.
- Poole, Keith, and Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29(2): 357–84.
- Proksch, Sven-Oliver, and Jonathan B. Slapin. 2006. "Institutions and Coalition Formation: The German Election of 2005." *West European Politics* 29(3): 540–59.
- Stevenson, Patrick. 2002. *Language and German Disunity: A Sociolinguistic History of East and West in Germany, 1945–2000*. Oxford: Oxford University Press.
- Strom, Kaare. 1984. "Minority Governments in Parliamentary Democracies." *Comparative Political Studies* 17(2): 199–227.
- Tsebelis, George. 2002. *Veto Players: How Political Institutions Work*. Princeton, NJ: Russell Sage/Princeton University Press.
- Wallerstein, Michael. 1999. "Wage-Setting Institutions and Pay Inequality in Advanced Industrialized Societies." *American Journal of Political Science* 43(3): 649–80.
- Warwick, Paul. 1992. "Ideological Diversity and Government Survival in Western European Parliamentary Democracies." *Comparative Political Studies* 25(3): 332–61.

## **Position Taking in European Parliament Speeches**

SVEN-OLIVER PROKSCH AND JONATHAN B. SLAPIN\*

This article examines how national parties and their members position themselves in European Parliament (EP) debates, estimating the principal latent dimension of spoken conflict using word counts from legislative speeches. We then examine whether the estimated ideal points reflect partisan conflict on a left-right, European integration or national politics dimension. Using independent measures of national party positions on these three dimensions, we find that the corpus of EP speeches reflects partisan divisions over EU integration and national divisions rather than left-right politics. These results are robust to both the choice of language used to scale the speeches and to a range of statistical models that account for measurement error of the independent variables and the hierarchical structure of the data.

How do legislators, and the parties they belong to, position themselves in legislative speeches? And how can political scientists systematically analyse the content of legislative speeches to gain insight into party positions? Until recently, legislative speeches have remained a largely untapped resource when examining position taking in parliamentary arenas. Instead, researchers have focused on voting behaviour to study ideology in legislatures. New advances in computer-based content analysis, however, have opened up the possibility of treating written or spoken text as data to study ideology.<sup>1</sup> In many ways, the ability to examine legislative speeches represents an improvement upon existing methodologies. In parliamentary systems, it is well known that voting behaviour does not reflect ideology due to high party discipline and government agenda setting. Therefore, roll-call votes provide very little information about the placement of parties in an ideological space and show instead the division between government and opposition parties. Even in other political systems, roll-call

\* Mannheim Centre for European Social Research, University of Mannheim; and Trinity College Dublin, respectively (email: proksch@uni-mannheim.de). The authors thank Ken Benoit, James Honaker, Thomas König, Jeff Lewis, Michael Peress, George Tsebelis, Albert Weale and several anonymous reviewers for their helpful comments and suggestions. A previous version of this article was presented at the Annual Meeting of the Midwest Political Science Association in Chicago and at the Workshop on Estimating Policy Preferences at the Mannheim Centre for European Social Research in 2008. Both authors have contributed equally to all work.

<sup>1</sup> Michael Laver, Kenneth Benoit and John Garry, ‘Extracting Policy Positions from Political Texts Using Words as Data’, *American Political Science Review*, 97 (2003), 311–32; Burt L. Monroe and Ko Maeda, ‘Talk’s Cheap: Text-Based Estimation of Rhetorical Ideal-Points’ (presented at the 21st Annual Summer Meeting, Society for Political Methodology, Stanford University, 2004); Daniel Hopkins and Gary King, ‘A Method of Automated Nonparametric Content Analysis for Social Science’, *American Journal of Political Science*, forthcoming; Daniel Diermeier *et al.*, ‘Language and Ideology in Congress’ (presented at the annual meeting of the Midwest Political Science Association, 2007); Jonathan B. Slapin and Sven-Oliver Proksch, ‘A Scaling Model for Estimating Time-Series Party Positions from Texts’, *American Journal of Political Science*, 52 (2008), 705–22.

votes may only account for a small and potentially biased sample of all votes. Members of parliaments, on the other hand, deliver speeches on a wide variety of topics on an almost daily basis. We argue that the content of these speeches provides a great deal of information about partisan ideology and position taking.

To examine parliamentary speech in the European Union, we have constructed a new dataset of all speeches given during the 5th session (1999–2004) of the European Parliament (EP). The EP provides an excellent but particularly hard case for the study of position taking in parliamentary speech. First, the EP has many more parties than other parliaments. Voters elect members of the European Parliament (MEPs) from national party lists. Although these national party MEPs do form political groups within the parliament, there were almost 130 national parties represented in the EP during the time period we investigated, and this was prior to the enlargement of the EU in 2004. With so many parties and political views, it will be difficult to find ideological structure in the speeches. Secondly, because there is no government–opposition divide as in parliamentary systems, there is less structure to EP debates than there might be in a national parliament. Lastly, the EP being a multilingual political body, all legislative speech occurs in translation. This may add an additional layer of error in the data and raises the question of whether some languages are more suitable for computer-based content analysis than others. If we are able to extract meaningful party positions from these speech data, and if we can do this regardless of the language we choose, then the approach we employ here should be able to estimate party positions from speeches in other political systems as well.

The remainder of the article examines the structure of parliamentary debate in the EP. We apply a novel method called Wordfish to extract policy positions from the speeches.<sup>2</sup> We then test whether these estimated positions correspond to (1) left–right ideology, (2) positions on European integration, or (3) a national dimension in the European Union. We find that the primary dimension of speech in the EP is best explained by national divisions and parties' positions towards deeper EU integration. In contrast, national parties do not appear to position themselves primarily according to their left–right ideology. These results are in contrast to findings of voting behaviour studies in the EP. Moreover, we show that our findings are robust to the choice of language and translation, to various independent measures of left–right and EU positions, and to the type of statistical model used.

#### REVEALING PARTY POSITIONS IN THE EUROPEAN PARLIAMENT

National parties reveal their positions in the EP through the actions taken by their members. There are two primary ways for MEPs to reveal both their positions and those of their parties: they give speeches and they subsequently vote on legislative proposals. Votes on legislative proposals and resolutions have been the primary source of data to study MEPs' revealed preferences. If such votes are recorded as a roll call, then this information can be used to estimate positions of MEPs as well as of national parties by aggregating individual MEP positions. Roll calls have therefore been used by numerous scholars to study national party positions in the European Parliament by applying various scaling techniques. The studies either use roll-call samples from

<sup>2</sup> Slapin and Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts'.

specific periods<sup>3</sup> or most recently cover all available roll calls from the beginning of the EP.<sup>4</sup>

Estimating ideal points from EP roll-call data is not unproblematic and scholars must regard such estimates with caution. Carruba *et al.* have pointed out a well-known selection problem associated with roll-call votes in the EP.<sup>5</sup> Not only are less than a third of all votes in the EP by roll call, but party groups use roll-call votes 'in a fashion that would introduce selection bias into the roll-call vote sample'.<sup>6</sup> Contrary to the common belief that roll calls represent votes on significant issues, the authors actually find that such rolls are taken disproportionately on (inconsequential) resolutions rather than on (consequential) legislative proposals under the co-decision procedure.<sup>7</sup> They conclude that roll-call votes are biased towards overestimating inter-party group cohesion, because MEP attendance on over-sampled resolutions is significantly different from attendance on co-decision votes, representing those who tend to vote the party line. Together, these results would suggest that ideological estimates from roll calls in the EP 'are most likely incorrectly characterizing the policy space'.<sup>8</sup> Even though scholars applying scaling techniques to roll-call data state explicitly that they are 'less interested in the estimation of the ideal points of individual MEPs than in the number of dimensions of politics',<sup>9</sup> researchers might nevertheless be tempted to use their scores to test models requiring ideal-point estimates.<sup>10</sup>

Speeches may offer a useful alternative to recorded votes. On the one hand, speaking in parliament and voting share a common feature in the sense that they are public. As a consequence, depending on the context, MEPs may make statements that are either symbolic, and include cheap talk, or strategic. On the other hand, there are fewer constraints in the EP on speeches than on votes. Speeches about European policies contain more nuanced arguments than simple 'Yes' or 'No' votes. Moreover, selective roll-call

<sup>3</sup> Fulvio Attina, 'The Voting Behaviour of the European Parliament Members and the Problem of the Europarties', *European Journal of Political Research*, 18 (1990), 557–79; Joanne Bay Brzinski, 'Political Group Cohesion in the European Parliament, 1989–1994', in Carolyn Rhodes and Sonia Mazey, eds, *The State of the European Union* (London: Longman, 1995), pp. 64–83; Tapio Raunio, *The European Perspective: Transnational Party Groups in the 1989–1994 European Parliament* (Sudbury, Mass.: Dartmouth/Ashgate, 1997); Amie Kreppel and George Tsebelis, 'Coalition Formation in the European Parliament', *Comparative Political Studies*, 32 (1999), 933–66; Simon Hix, 'Legislative Behaviour and Party Competition in the European Parliament: An Application of Nominate to the EU', *Journal of Common Market Studies*, 39 (2001), 663–88; Abdul Noury, 'Ideology, Nationality and Euro-Parliamentarians', *European Union Politics*, 3 (2002), 33–58; Abdul Noury and Gerard Roland, 'More Power to the European Parliament?', *Economic Policy*, 17 (2002):35, 281–319; Gail McElroy, 'Committee Representation in the European Parliament', *European Union Politics*, 7 (2006), 5–29; Jeong-Hun Han, 'Analysing Roll Calls of the European Parliament: A Bayesian Application', *European Union Politics*, 8 (2007), 479–507.

<sup>4</sup> Simon Hix, Abdul Noury and Gerard Roland, 'Dimensions of Politics in the European Parliament', *American Journal of Political Science*, 50 (2006), 494–511; Simon Hix, Abdul Noury and Gerard Roland, *Democratic Politics in the European Parliament* (Cambridge: Cambridge University Press, 2007).

<sup>5</sup> Clifford J. Carruba *et al.*, 'Off the Record: Unrecorded Legislative Votes, Selection Bias and Roll-Call Vote Analysis', *British Journal of Political Science*, 36 (2006), 691–704.

<sup>6</sup> Carruba *et al.*, 'Off the Record', p. 692.

<sup>7</sup> In their sample (5th EP, 1999–2000) co-decision votes were significantly under-sampled: only 0.77 per cent of co-decision votes were by roll call, see Carruba *et al.*, 'Off the Record'.

<sup>8</sup> Carruba *et al.*, 'Off the Record', p. 702.

<sup>9</sup> Hix, Noury and Roland, *Democratic Politics in the European Parliament*, p. 166.

<sup>10</sup> Gail McElroy, 'Legislative Politics as Normal? Voting Behaviour and Beyond in the European Parliament', *European Union Politics*, 8 (2007), 433–48, p. 437.

data are likely to be endogenous to the true, unobserved preferences of delegates, and are affected by partisan and institutional constraints in the Parliament (such as the strategic decision to demand a roll call). In contrast, speeches are more likely to yield preference data that are relatively free from such constraints for two reasons. First, MEPs give speeches on issues which never make it to a roll-call vote, and, secondly, all speeches are recorded and, therefore, do not have the same potentially problematic sample bias as roll-call votes. Legislative speeches are therefore an obvious, yet unexplored, source of data for research into partisan position taking inside the EP.<sup>11</sup>

### *The Structure of Debates in the EP*

Before examining position taking in EP speeches, it is helpful to understand when and how MEPs participate in legislative debate. The plenary sessions of the European Parliament take place every month for a week in Strasbourg, France, with additional meetings held in Brussels, Belgium. Debates in the plenary are primarily held on legislative and non-legislative reports. In addition, the EP exercises supervision of the other institutions through written and oral questions by MEPs to the Council and the Commission with subsequent debate. Furthermore, the EP may debate statements made by the President of the European Council, the Commission or the Council.<sup>12</sup> Finally, the EP has time set aside for debates on 'breaches of human rights, democracy and the rule of law' as well as for short-notice reactions to major events.

Independent of the agenda item being under discussion, structuring the debates always involves the allocation of speaking time.<sup>13</sup> Specific speaking time is reserved for the Commission and the Council (which we do not analyse here), but several MEPs also have reserved speaking time. These include rapporteurs and draftsmen of opinions and authors of motions for resolutions. The largest proportion of speaking time is allocated to the political groups of the EP. These political groups are made up of individual national party

<sup>11</sup> There are other methodological approaches for studying positions of national parties in the European Union, but they do not focus specifically on parliamentary behaviour. These approaches include expert surveys (Liesbet Hooghe and Gary Marks, 'Chapel Hill 2002 Expert Survey on Party Positioning on European Integration', <http://www.unc.edu/> (2002); Gary Marks, Liesbet Hooghe, Moira Nelson and Erica Edwards, 'Party Competition and European Integration in the East and West – Different Structure, Same Causality', *Comparative Political Studies*, 39 (2006), 155–75; Kenneth Benoit and Michael Laver, *Party Policy in Modern Democracies* (London: Routledge, 2006); Kenneth Benoit and Gail McElroy, 'Party Groups and Policy Positions in the European Parliament', *Party Politics*, 13 (2007), 5–28; Marco R. Steenbergen and Gary Marks, 'Evaluating Expert Judgments', *European Journal of Political Research*, 46 (2007), 347–66); and there are also MEP surveys (David Farrell *et al.*, 'EPRG 2000 and 2006 MEP Surveys Dataset', <http://www.lse.ac.uk/collections/EPRG/> (2006)), mass survey research (Simon Hix and Christopher Lord, *Political Parties in the European Union* (Basingstoke, Hants.: Macmillan, 1997)), interest group ratings, and European election manifestos (Matthew J. Gabel and Simon Hix, 'Defining the EU Political Space: An Empirical Study of the European Elections Manifestos, 1979–1999', *Comparative Political Studies*, 35 (2002), 934–64). However, none of these approaches actually studies the revealed preferences of the MEPs themselves. In addition, the alternative approaches have some methodological problems. For instance, McElroy points out that elite surveys suffer from sample response issues, preference measures on the basis of constituency characteristics are difficult given the weak electoral connection in the European Parliament, and interest group ratings tend to have selective samples, thus potentially exaggerating extreme positions (McElroy, 'Legislative Politics as Normal?', p. 437).

<sup>12</sup> Richard Corbett, Francis Jacobs and Michael Shackleton, *The European Parliament*, 5th edn (London: John Harper, 2003), p. 145.

<sup>13</sup> Corbett, Jacobs and Shackleton, *The European Parliament*, 5th edn.

delegations and loosely correspond to traditional party families. Each political group receives speaking time roughly in proportion to its seat share.<sup>14</sup>

A typical debate on legislation starts with an opening statement from the European Commission. This is followed by the rapporteur presenting the response of the relevant EP committee. If applicable, draftsmen of opinions from other committees may speak after the rapporteur. Then, the general debate follows with each political group speaking on the issue under debate, starting with the largest group. Party groups decide internally how to divide time among their MEPs, with the time for individual speeches being strictly limited, usually not more than three minutes.<sup>15</sup> At the end of the debate, the Commission replies to the speeches and indicates its position on proposed amendments to the legislative proposal.<sup>16</sup>

Legislative speeches in the EP cover a wide range of topics. To understand the structure of debates better, we identified all agenda items under debate during the 5th European Parliament (1999–2004), as well as the number of speeches given for each item.<sup>17</sup> In total, we found 2,000 different agenda items in the debates. We then put these items into pre-defined categories which follow standard categories of EU policies. Figure 1 presents the results. The largest number of speeches were delivered in the form of explanations of votes during voting time (around 20 per cent). The EP agenda does not break down the type of legislation being debated or the length of debate, so we must assume that this category includes all sorts of policies in which the EP has co-decision power. Three categories are not about specific policies (question time, procedural issues, and other speeches), making up another 20 per cent of the total speeches. Debates on specific policies constitute the largest category (60 per cent). They include speeches on internal (EU) policies (around 45 per cent of all debates) and on foreign policies (about 15 per cent). In short, speeches cover all policy areas of the EU. But how do parties position themselves in these speeches?

### *Hypotheses*

Existing empirical research has highlighted the presence of two major dimensions in European Union politics: a traditional left-right dimension and a European integration dimension.<sup>18</sup> Studies of voting behaviour inside the European Parliament find that left-right politics is the best predictor of MEP voting patterns,<sup>19</sup> even though replication of these analyses with more sophisticated statistical techniques finds that both left-right

<sup>14</sup> See Rule 149 of the EP Rules of Procedure.

<sup>15</sup> David Judge and David Earnshaw, *The European Parliament* (Basingstoke, Hants.: Palgrave Macmillan, 2003), p. 239.

<sup>16</sup> Corbett, Jacobs and Shackleton, *The European Parliament*.

<sup>17</sup> To automate this task, we wrote a computer script which automatically extracted the agenda item and the number of speeches from the information available on the EP website.

<sup>18</sup> Kreppel and Tsebelis, 'Coalition Formation in the European Parliament'; Gary Marks, Carole Wilson and Leonard Ray, 'National Political Parties and European Integration', *American Journal of Political Science*, 46 (2001), 585–94; Gabel and Hix, 'Defining the EU Political Space'; Mark Aspinwall, 'Preferring Europe: Ideology and National Preferences on European Integration', *European Union Politics*, 3 (2002), 81–111; Hix, Noury and Roland, 'Dimensions of Politics in the European Parliament'; Hix, Noury and Roland, *Democratic Politics in the European Parliament*.

<sup>19</sup> Hix, Noury and Roland, 'Dimensions of Politics in the European Parliament'; Hix, Noury and Roland, *Democratic Politics in the European Parliament*.



Fig. 1. Speeches by Agenda Item, 1999–2004

and pro/anti-Europe positions are contained in the first dimension extracted from roll-call data.<sup>20</sup>

Analogous to revealed behaviour in roll-call data, we might expect positions extracted from MEP speeches to line up along either an ideological left-right dimension or a European integration dimension. Speeches could reveal similar ideological positions to those uncovered through votes since parties might try to limit access to the floor and only allow those MEPs to deliver speeches who represent the official party line. We might, therefore, find left-right ideology in a speech dimension. But as debates cover a much larger range of topics, including issues not subject to roll-call votes, it is also possible that legislative speeches reflect positions on European integration rather than left-right politics. Conceptually, the EU integration dimension (or pro-/anti-Europe) is a rather narrow, well-defined dimension, while the left-right dimension is very broad.<sup>21</sup> Left-right commonly refers to a socio-economic dimension, but it may also include aspects of social

<sup>20</sup> Kenneth Benoit, Michael Laver and Slava Mikhaylov, 'Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions', *American Journal of Political Science*, 53 (2009), 495–513.

<sup>21</sup> We thank an anonymous referee for pointing this out to us.

conservatism or liberalism, and potentially even nationalism and militarism. Studies of roll-call votes have compared positions extracted from votes to both the left-right scale created by the Comparative Manifestos Project and expert surveys.<sup>22</sup> The CMP left-right scale includes issues related to economic ideology as well as features of culture, society and militarism.<sup>23</sup> Asked to assess the left-right positions of parties, experts must construct their own notion of what left-right ideology means. For instance, in their expert survey of EP group positions, Benoit and McElroy asked experts to locate the groups on a left-right dimension 'taking all aspects of group policy into account'.<sup>24</sup> Despite its scope, the left-right dimension is often viewed as orthogonal to the dimension of EU integration. There are both leftist and rightist parties opposed to integration. In some countries, the left may be more willing to support integration, while in other countries it is the right that prefers deeper integration. For this reason, it is possible to examine position taking in the EP both in terms of left-right ideology and integration, even if one category encompasses a great deal more than the other.

Besides such ideological factors, national party delegations may express national differences in speeches. The issue categories for the speeches (Figure 1) suggest that national factors might in fact play a role. This is especially true given the amount of time spent debating the annual budget, agricultural subsidies, institutional issues and foreign policy. These areas are likely to separate MEPs from different countries. Financial issues might cause MEPs from net paying and receiving countries to use different arguments in speeches, institutional issues can reveal a divide between small and large countries, and foreign policy might add a similar national dimension to the debates. Using the whole set of speeches from each party, we test the following hypotheses of national party position taking:

HYPOTHESIS 1 – Left-right position taking: national parties, through speeches given by their MEPs, position themselves in the EP according to their national left-right ideology.

HYPOTHESIS 2 – European position taking: national parties, through speeches given by their MEPs, position themselves in the EP according to their position on European integration.

HYPOTHESIS 3 – National position taking: national parties, through speeches given by their MEPs, position themselves in EP speeches according to national factors such as country size, wealth and net payer/receiver status.

#### EXTRACTING THE PRINCIPAL DIMENSION OF SPEECH: THE WORDFISH TECHNIQUE

To examine these hypotheses, we extract the principal dimension of speech using a new computer-based technique called Wordfish.<sup>25</sup> Computer-based content analysis aiming to extract political positions from texts has been applied to multiple sources of political text,

<sup>22</sup> Hix, Noury and Roland, 'Dimensions of Politics in the European Parliament'.

<sup>23</sup> Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara and Eric Tanenbaum, *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998* (Oxford: Oxford University Press, 2001).

<sup>24</sup> Benoit and McElroy, 'Party Groups and Policy Positions in the European Parliament', p. 22.

<sup>25</sup> Slapin and Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts'. Wordfish is implemented in R and available at [www.wordfish.org](http://www.wordfish.org).

including party manifestos,<sup>26</sup> legislative speeches,<sup>27</sup> campaign speeches,<sup>28</sup> constitutional negotiations,<sup>29</sup> and judicial decisions.<sup>30</sup> The Wordfish method uses unique words as the unit of analysis and compares political texts (e.g. manifestos, speeches, etc.) on the basis of relative word usage in each. As a scaling technique, Wordfish does not require an *a priori* definition of the dimension being estimated (for instance, by anchoring specific reference speeches). The technique uses an explicit parametric model of word counts and simply scales the word counts to reduce the data to a single dimension. Wordfish assumes that word frequencies are generated by a Poisson distribution.<sup>31</sup> This distribution is simple and has only one parameter that needs to be estimated,  $\lambda$ , which is both the mean and the variance. The functional form of the Wordfish model is as follows:

$$\begin{aligned}Wordcount_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\alpha_i + \psi_j + \beta_j * \omega_i)\end{aligned}$$

where  $\alpha$  is a set of national-party fixed effects,  $\psi$  is a set of word-fixed effects,  $\beta$  is an estimate of a word-specific weight capturing the importance of the word  $j$  in discriminating between positions, and  $\omega$  is the estimate of party  $i$ 's position. Word-fixed effects capture the fact that some words are used much more often than other words by all parties. National party effects control for the possibility that some parties speak more than other parties.

To estimate the parameters of this item-response model, Wordfish uses an expectation maximization (EM) algorithm, alternating between estimating word-specific parameters holding the party-specific parameters fixed and estimating party-specific parameters holding the word-specific parameters fixed. The process is repeated until a convergence criterion is met (i.e. log-likelihoods do not change any more from one iteration to the next). The resulting positions are located on a dimension which is (arbitrarily) scaled to a mean of 0 and a standard deviation of 1 to identify the likelihood function.<sup>32</sup>

<sup>26</sup> Laver *et al.*, 'Extracting Policy Positions from Political Texts Using Words as Data'; Sven-Oliver Proksch and Jonathan B. Slapin, 'Institutions and Coalition Formation: The German Election of 2005', *West European Politics*, 29 (2006), 540–59; Slapin and Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts'; Simon Hug and Tobias Schulz, 'Left–Right Positions of Political Parties in Switzerland', *Party Politics*, 13 (2007), 305–30.

<sup>27</sup> Michael Laver and Kenneth Benoit, 'Locating TDs in Policy Spaces: Wordscoring Dail Speeches', *Irish Political Studies*, 17 (2002), 59–73; Laver *et al.*, 'Extracting Policy Positions from Political Texts Using Words as Data'; Monroe and Maeda, 'Talk's Cheap: Text-Based Estimation of Rhetorical Ideal-Points'; Daniela Giannetti and Michael Laver, 'Policy Positions and Jobs in the Government', *European Journal of Political Research*, 44 (2005), 91–120; Diermeier *et al.*, 'Language and Ideology in Congress'.

<sup>28</sup> Michael Laver, Kenneth Benoit and Nicolas Sauger, 'Policy Competition in the 2002 French Legislative and Presidential Elections', *European Journal of Political Research*, 45 (2006), 667–97.

<sup>29</sup> Kenneth Benoit *et al.*, 'Measuring National Delegate Positions at the Convention on the Future of Europe Using Computerized Word Scoring', *European Union Politics*, 6 (2005), 291–313.

<sup>30</sup> Kevin T. McGuire and Georg Vanberg, 'Mapping the Policies of the U.S. Supreme Court: Data, Opinions, and Constitutional Law' (prepared for delivery at the Annual Meeting of the American Political Science Association, Washington, D.C., 2005).

<sup>31</sup> Slapin and Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts'.

<sup>32</sup> We have applied this model to compare election manifestos from German parties between 1990 and 2005. We found that the technique is able to recover party positions estimated by other techniques (e.g. expert surveys and hand-coding of manifestos). Furthermore, the positions reflect important changes in the party system, in particular a rightward movement of the major social-democratic party, the SPD, in the 1990s. We could produce estimates over time by making the assumption that word weights are

The Wordfish algorithm is not the only computer based content analysis technique that can be applied to study ideology in political text. The Wordscores technique also uses relative word frequencies in text documents to place actors on a single dimension.<sup>33</sup> The choice of content analysis technique depends on the research question. For the purpose of our study, we are interested in examining the speech dimension in the EP and thus prefer to use Wordfish as it scales the word data to extract a single dimension. If our aim were to place parties on a pre-defined dimension, we could use Wordscores as it allows definition of the dimension *ex ante* via reference texts.<sup>34</sup>

#### THE DATA: SPEECHES IN THE EUROPEAN PARLIAMENT

We test the hypotheses using a newly collected dataset of legislative speeches in the 5th European Parliament (1999–2004). The number of speeches delivered during this time is impressive. Between 1999 and 2004, MEPs gave over 50,000 speeches in the plenary (Table 1).<sup>35</sup> This set of political statements constitutes a rich dataset for multilingual content analysis. We want to estimate and examine the principal dimension of speech in the EP and compare it to other measures of ideology. But even though legislative speeches do provide a rich source of information, they might be harder to compare to each other than to written texts such as party manifestos. Laver *et al.* describe the potential problems:

While the analysis of speeches holds considerable promise, it also raises new challenges for content analysis – whether computerized or traditional – because such speeches differ substantially from party manifestos in several key respects. First, manifestos are typically comprehensive documents addressing a wide range of policy issues, while speeches tend to be much more restricted in focus. Secondly, manifestos are published in a political context that is fairly well defined. Greater care must be taken in establishing the political context of speeches if we are to justify the comparison of different speeches in the same analysis.<sup>36</sup>

(*F*note continued)

time-invariant (see Slapin and Proksch, ‘A Scaling Model for Estimating Time-Series Party Positions from Texts’).

<sup>33</sup> Laver *et al.*, ‘Extracting Policy Positions from Political Texts Using Words as Data’. While the technique has mostly been used to study political manifestos, it has been applied to legislative speeches as well (Laver and Benoit, ‘Locating TDs in Policy Spaces: Wordscoring Dail Speeches’; Laver *et al.*, ‘Extracting Policy Positions from Political Texts Using Words as Data’; Giannetti and Laver, ‘Policy Positions and Jobs in the Government’). Laver and Benoit use speeches from a confidence debate in the Irish Dáil in October 1991 over the future of the incumbent coalition government. They postulate a ‘pro-versus anti-government’ dimension and use the speech of the prime minister and of the opposition leaders as reference texts. The resulting placement of political parties on a scale of government versus opposition ‘is readily recognisable by any observer of Irish politics’ (Laver *et al.*, ‘Extracting Policy Positions from Political Texts Using Words as Data’, p. 327).

<sup>34</sup> We did validate the Wordfish algorithm presented here with the Wordscores technique. To do so, we anchored the Wordfish dimension in Wordscores by using the speeches from the most extreme parties identified by Wordfish as reference texts. We estimated the Wordscores positions using a slightly updated version of the algorithm (Lanny W. Martin and Georg Vanberg, ‘A Robust Transformation Procedure for Interpreting Political Texts’, *Political Analysis*, 16 (2008), 93–100). As expected, the results correlate very highly across all languages between the two techniques (correlation of 0.91 or higher).

<sup>35</sup> This number excludes new member state MEPs joining in 2004 for only a few weeks before the next election, but includes the presidents and vice-presidents of the EP who deliver mostly procedural speeches.

<sup>36</sup> Laver *et al.*, ‘Extracting Policy Positions from Political Texts Using Words as Data’, p. 327.

To address these two potential problems, we first use all legislative speeches given during the 5th European Parliament, not limiting ourselves to only a few important ones. This way we ensure our data are not issue specific.<sup>37</sup> Speeches cover all categories listed in Figure 1. Secondly, in order to control for speaker-specific context, we chose national parties as the unit of analysis, and not individual MEPs. We decide to focus on national party positions rather than individual positions for both substantive and methodological reasons. Substantively, findings in the existing literature on the importance of national parties in the European Parliament justify this choice.<sup>38</sup> For example, national parties choose the candidates who run in European Parliament elections, organize the campaigns, choose which European political group to align with once in Parliament, and control to a large extent the allocation of political offices in the EP. Moreover, scholars are often more concerned with analysing the positions of party groups and the national parties that compose them, rather than the positions of individual MEPs.<sup>39</sup> Even scholars examining the dimensionality of positions extracted from roll-call votes usually aggregate up to the level of the national party rather than examine individual (MEP) ideology.<sup>40</sup>

Methodologically, by aggregating speeches from the individual to the national party level, we ensure that the positions are estimated from more comprehensive data. The aim is to eliminate situations in which short or trivial speeches heavily influence the estimation and the results. In addition, an individual level analysis requires throwing away a great deal of data that may be preserved in the analysis at the level of the national party. A substantial number of MEPs gave very few speeches. We would not be able to estimate positions for these individuals. If, for example, there were two individuals from the same national party and both made relatively few speeches, thus preventing the estimation of individual positions, we may still be able to estimate a position for their national party if their combined speeches are sufficiently long.

Nevertheless, aggregation potentially leads to a few problems. First, we necessarily overlook intra-party variation. We certainly do not claim that there is no intra-party variation. In fact, we expect MEPs to agree to various degrees with their party and this ought to be reflected not only in votes but also in legislative speeches. A possible objection to the choice of the national party as the unit of analysis, then, is that the findings will be valid for that particular level of analysis only and possibly cloud true differences between legislators.<sup>41</sup> A second objection to the use of national parties as the unit of analysis is that, if we truly wish to make comparisons between roll-call positions and speech positions, they should both be measured at the same level of aggregation. To address both of

<sup>37</sup> The inferences will only be valid for this total set of speeches and do not necessarily apply for subsets of speeches (e.g. specific policy areas).

<sup>38</sup> Hix and Lord, *Political Parties in the European Union*; Raunio, *The European Perspective*; Kreppel and Tsebelis, 'Coalition Formation in the European Parliament'; Amie Kreppel, *The European Parliament and Supranational Party System* (Cambridge: Cambridge University Press, 2002); Simon Hix, 'Parliamentary Behavior with Two Principals: Preferences, Parties, and Voting in the European Parliament', *American Journal of Political Science*, 46 (2002), 688–98; Hix, Noury and Roland, 'Dimensions of Politics in the European Parliament'; Hix, Noury and Roland, *Democratic Politics in the European Parliament*.

<sup>39</sup> Benoit and McElroy, 'Party Groups and Policy Positions in the European Parliament'.

<sup>40</sup> Hix, Noury and Roland, 'Dimensions of Politics in the European Parliament'. There are no independent measures of ideology available at the individual level with the exception of the EPRG survey of MEPs themselves, which suffers from low response rates (Farrell *et al.*, 'EPRG 2000 and 2006 MEP Surveys Dataset'). If the researchers wish to compare roll-call positions with expert survey positions or CMP data, they must aggregate up to the level of national party.

<sup>41</sup> We thank one of the anonymous referees for pointing this out.

these concerns, we therefore conduct an analysis of speeches at the individual level as well, in order to validate the results from our national party level analysis.

Our data collection involved the following steps. First, we identified all MEPs in the 5th European Parliament, restricting our sample to MEPs from the fifteen member states prior to enlargement in 2004.<sup>42</sup> Secondly, we downloaded all speeches given by these MEPs in the English, French and German translations from the EP website.<sup>43</sup> Thirdly, we combined the speeches of MEPs from the same national party using party labels contained in the EP roll-call dataset.<sup>44</sup>

Table 1 presents the summary statistics of the speeches in the 5th European Parliament. Each MEP gave on average seventy-six speeches. Some MEPs did not give speeches at all (usually those who held national offices, such as Italian Prime Minister Silvio Berlusconi), and the most active MEPs were from the EP leadership (president and vice-presidents). On average, members from each national party gave more than 400 speeches. The more robust measure of central tendency, the median, yields close to 200 speeches per national party.

TABLE 1 *Summary Statistics: Speeches in the 5th European Parliament (1999–2004)*

	Mean	Median	Min	Max
MEPs per national party	5.4	3	1	43
Speeches per national party	410.8	196	0	2,486
Speeches per MEP	76.4	45	0	2,030
Total number of speeches		52,988		
Number of national parties		129		

We construct a word-count dataset with unique words in rows and national parties in columns and use a word-count program to stem words in all languages.<sup>45</sup> To make the estimation more efficient, we reduce the data according to the following criteria. First, we drop national parties whose MEPs do not say anything or give only very few speeches. As the cut-off criterion, a national party's members must give speeches that total 10,000 words or longer. We also eliminate speeches given by members of the EP's Bureau. These members preside over the plenary sessions and their speeches are mostly procedural. This way, we eliminate twenty-three parties from the dataset, leaving us with 106 parties. In a second step, we drop words that are used very infrequently. As the criterion, we specify that a word should be kept in the dataset if members from at least thirty national parties (around one-third) use it in their speeches. This reduces the number of unique words significantly, makes estimation faster and more feasible, and ensures that the speeches have a minimum level of comparability. In order to determine whether the cut-off potentially biases our results, we also estimate positions using a less strict criterion (words mentioned by at least ten national parties).

<sup>42</sup> We exclude new member state MEPs as they were only represented in the 5th European Parliament by nominated members for a few weeks between the date of enlargement (1 May 2004) and the elections to the 6th European Parliament (June 2004).

<sup>43</sup> We used Perl scripts to automate this task. The speech archive of the European Parliament is available at <http://www.europarl.europa.eu/activities/archives/cre/search.do?language=EN>, last consulted in April 2008.

<sup>44</sup> Hix, Noury and Roland, 'Dimensions of Politics in the European Parliament'.

<sup>45</sup> We use Will Lowe's *jfreq* program, available at <http://www.williamlowe.net/software/>.

The results correlate very highly and we are not worried that the choice of cut-off criterion affects our results.<sup>46</sup>

Speeches in the EP pose an additional challenge to content analysis because all of the EP's business occurs in multiple languages and therefore in translation. Even though so much of international politics occurs in translation, scholars have not paid significant attention to the effects of translation when using computer-based content analysis. The European Union is perhaps the most prominent example of a multilingual political system. With twenty-seven member states, the EU now has twenty-three official languages.<sup>47</sup> Unlike other multilingual political bodies, such as the United Nations, where career diplomats are competent in multiple languages, the elected members of the European Parliament have the right to communicate in their national language(s) as an expression of national identities and cultures in Europe.<sup>48</sup> Every speech made in the EP must, therefore, be interpreted and translated into each of these twenty-three languages so that all MEPs are able to understand it. Moreover, all official laws and regulations must be translated as well.<sup>49</sup> Rather than treating the presence of multilingualism as an obstacle to the analysis, we consider the EP as the perfect political arena for testing how translation affects computer-based content analysis. Translated EP speeches provide a unique source of data to estimate the positions of members of the EP because we know *a priori* that the content of all speeches is the same across languages.<sup>50</sup>

#### ESTIMATED POSITIONS FROM EP SPEECHES

We run the Wordfish algorithm for 106 parties using English, German and French translations. The estimated positions, including their 95 per cent confidence intervals, are

<sup>46</sup> The English Wordfish results using words mentioned by at least ten parties correlate with results using words mentioned by at least thirty parties at 0.99.

<sup>47</sup> The EU has fewer official languages (twenty-three) than member states (twenty-seven). German is spoken in Germany and Austria, English in the United Kingdom and Ireland, Greek in Greece and Cyprus, and Belgium and the Netherlands share common languages with their neighbouring countries.

<sup>48</sup> Corbett, Jacobs and Shackleton, *The European Parliament*, p. 34; Judge and Earnshaw, *The European Parliament*, p. 163.

<sup>49</sup> These obligatory tasks result in considerable costs in the EU. In 2003, prior to the enlargement, EU institutions spent a combined 549 million euros on translation, and following enlargement to twenty-five members in 2004, the expense rose to an estimated 807 million euros per year, or approximately 1.78 euros per EU citizen (see European Commission Memo 05/10, January 2005, <http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/05/10>). In 2005, after enlargement by ten new member states, the EP had over one million pages of parliamentary documents translated. In addition, the EP provided interpretation services totalling 85,340 work days (see European Parliament Budget 2005, [http://www.europarl.europa.eu/pdf/budget/rapportpublic2005\\_en.pdf](http://www.europarl.europa.eu/pdf/budget/rapportpublic2005_en.pdf)).

<sup>50</sup> There are several reasons to believe that translation may affect the output of computer-based content analysis. The German language has a particular feature that allows the compounding of words to create new ones. For example, the phrase 'workers' rights' is described by two words in English, three in French ('*droits des travailleurs*'), but only one in German ('*Beschäftigtenrechte*'). Moreover, translation itself possibly adds error to the data, which could lead to different results across language. Translation theorists have suggested that one can view translation as a series of choices that can be modelled as a decision tree (Jiří Levý, 'Translation as a Decision Process', in *To Honor Roman Jakobson II* (The Hague: Mouton, 1967), 1171–82). Each language presents the translator with a set of possible choices about which particular translation to choose. A stylistic choice a translator makes at one node may affect how he or she translates the rest of the text. This means that additional error may enter into the data both because different languages offer different choice sets and translators will make different decisions within those choice sets. Thus, we might get different results because some languages use different words and grammatical structures to express exactly the same content and because translators might follow different strategies in translation.

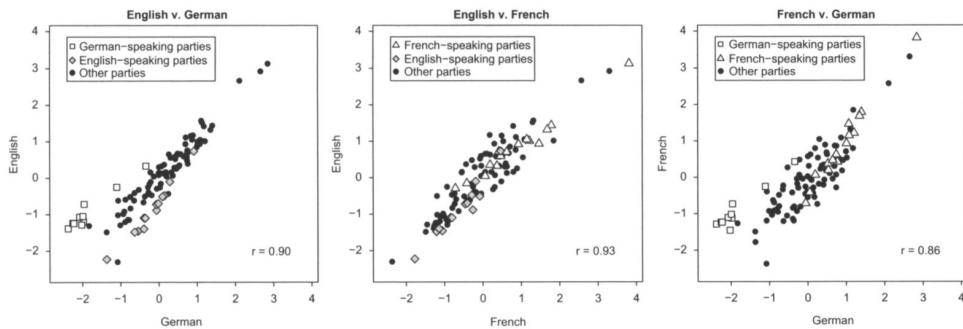


Fig. 2. Wordfish position estimates: comparison between languages ( $N = 106$ )

Note: German-speaking parties include parties from Germany, Austria and the Italian SVP. English-speaking parties include parties from the UK and Ireland. French-speaking parties include parties from France and French-speaking parties from Belgium.

presented in Appendix A.<sup>51</sup> First, we compare the correlation between the extracted positions from different translations to test the robustness of the technique across languages. Figure 2 shows position estimates for all three language combinations.

The comparison of the results across languages suggests that the position estimation technique is in fact highly robust to the choice of language (the correlation coefficient is 0.86 or higher). The highest correlation is between positions estimated from the English and French translations. These two languages are so similar to each other with regard to the information contained in words that they produce virtually identical position estimates. The relationship is slightly weaker between German and the other two languages. As we pointed out earlier, German words can be compounded and therefore contain more politically meaningful information than words in English and French.

The lower fit of German is visible in Figure 2, showing a modest heteroscedastic pattern for the German–English and German–French plots. To identify which parties form the clusters, we highlighted native German-speaking, English-speaking and French-speaking national parties in all plots. Speeches given by members of these parties are obviously only translated into the other two languages, whereas the original language version simply corresponds to the EP verbatim reports.<sup>52</sup> The labels show that German-speaking parties form a cluster and are at the extreme end of the speech dimension for the German translation. In contrast, the very same parties are more dispersed and not located at the extreme ends using the English and French translations. This suggests that the clustering may be due to a bias in the translation. EP speeches from native German speakers contain a set of words that does not appear in the German translation of speeches from non-German speaking parties. In other words, translators do not use the same words as native speakers when translating into German. This is likely to be due to the possibility of creating new words in German which are not used by translators. Because we do not find the cluster of German parties for English and French translations, the bias is most likely the result of translation rather than actual position taking. We do not find this clustering

<sup>51</sup> Appendix A shows the national party estimates using the English translations. The estimation is based on 4,859 unique words in English, 6,248 unique words in French and 7,369 unique words in German.

<sup>52</sup> In contrast, speeches from a party whose native language is not English, German or French are translated into all three languages.

effect for English-speaking or French-speaking parties, whose positions correlate highly across all language combinations. It is important to emphasize that our analysis does not determine a 'best language' to be used for automated content analysis. Even though the positions using English and French translations of the speeches correlate higher with each other than with German translations, all estimates are still quite robust to language choice. Our results should be encouraging to those who wish to use automated content analysis to extract political positions from texts. The results suggest that the techniques yield similar results in English, German and French.

Next, we estimate positions for individual MEPs rather than national party delegations to examine whether the results are comparable across levels of analysis. This means that the estimation is now based on fewer speeches per unit of analysis than before. After again excluding MEPs who did not deliver any speeches, we also remove MEPs who gave speeches that were shorter than 10,000 words (or approximately fifteen speeches). This leaves 427 MEPs in the sample. Furthermore, we exclude from the analysis infrequently used words (by less than 10 per cent of MEPs). We then extract positions for these 427 MEPs, who belong to 103 national parties, using the English translations of the speeches.<sup>53</sup> The results validate the findings from the national party level. The mean national party position from the individual level analysis correlates with the national-party level positions (English) at 0.79.<sup>54</sup> We later explore more systematically whether the dimension on the individual level actually resembles the one on the national party level.

#### TAKING NATIONAL PARTY POSITIONS IN EP SPEECHES

To test the *Left–Right*, *EU Integration*, and *National Politics* hypotheses, we run a multivariate regression using the estimated party positions from all three languages as the dependent variable. We relied on three data sources to measure left–right and European integration positions for national parties. First, we use an expert survey conducted in 2002–03, during the middle of the 5th European Parliament, by a research team from the University of North Carolina, Chapel Hill (UNC).<sup>55</sup> This survey polled national experts about European party positions regarding various aspects of EU integration. To capture a party's overall position with regard to EU integration, the survey asked experts to 'describe the general position on European integration that the party's leadership has taken over the course of 2002'. The survey also asked experts to place parties on a left–right spectrum in terms of their broad ideological stance. Of the 106 parties in our dataset, eighty-two are represented in the UNC data.<sup>56</sup>

To validate our findings from the Chapel Hill data, we use a second expert survey conducted at the same time by Benoit and Laver.<sup>57</sup> They asked numerous experts in

<sup>53</sup> The estimation is based on 4,765 unique words.

<sup>54</sup> We can also calculate the average standard deviation of national parties based on the results from the individual level analysis. For those national parties with more than one MEP ( $n = 71$ ), the average standard deviation of positions is 0.68, which is about two-thirds of the overall standard deviation of the positions (fixed at 1). If we include national parties with one MEP ( $n = 103$ ), the mean standard deviation of the positions across national parties drops to 0.47. It would be interesting to explore the reasons for the variation of individual-level positions in future research.

<sup>55</sup> Hooghe and Marks, 'Chapel Hill 2002 Expert Survey on Party Positioning on European Integration'; Marks *et al.*, 'Party Competition and European Integration in the East and West'; Steenbergen and Marks, 'Evaluating Expert Judgments'.

<sup>56</sup> In addition to missing several small parties, the UNC data do not include parties from Luxembourg.

<sup>57</sup> Benoit and Laver, *Party Policy in Modern Democracies*.

forty-seven European countries to place national parties on various policy dimensions. A party's position towards EU integration was captured by a question that asked whether the party 'favours increasing the range of areas in which the EU can set policy'.<sup>58</sup> For the left-right position we use the survey question asking experts to identify the parties' general left-right stance 'taking all aspects of party policy into account'.<sup>59</sup> Of the 106 parties in our dataset, the expert survey includes left-right and EU authority positions for sixty-four parties.<sup>60</sup>

To test the hypotheses using the full sample, we used roll-call votes as a third source of party position data. Specifically, we calculated the average first and second dimension Nominate scores for all 106 national parties during the 5th European Parliament.<sup>61</sup> It may seem objectionable to use positions based on voting behaviour to explain positions based on speech for two reasons. First, because MEPs speak about an issue before they vote on it, we would probably expect speech to explain votes and not vice versa. Secondly, we are not certain of the exact nature of the dimension that Nominate extracts. However, as previous research has demonstrated, first dimension Nominate scores correlate more highly with a traditional left-right dimension, while the second dimension correlates with positions on delegation of authority to the EU.<sup>62</sup>

Regardless of the direction of the causal arrow, the Nominate scores provide a good proxy for the survey data. Moreover, they allow us to examine our entire sample and uncover the relationship between positions estimated from voting and those estimated from speeches. Table 2 presents correlations between the average Nominate scores and the two expert survey variables. The first dimension scores correlate highly with left-right but not with the EU authority positions. The second dimension scores, by contrast, correlate with the EU authority variables but not with the left-right variables. In the following analysis, we use both the Nominate positions on the full sample and the survey data on the reduced sample and demonstrate that our results are robust regardless of which measure we use. Finally, to control for any national level effects, we include country dummies in our regression models.

We explicitly take into account that the position variables may contain measurement error. It is well known that the presence of measurement error in independent variables can lead to ordinary least squares (OLS) coefficients that are biased towards zero, thus underestimating the true effect of the independent variables. In order to correct for this bias, we use a technique called simulation-extrapolation or SIMEX.<sup>63</sup> This method adds measurement error to the model via simulations in order to establish a trend in the bias, and then reduces the effect of this measurement error.<sup>64</sup> Benoit *et al.* propose that

<sup>58</sup> Benoit and Laver, *Party Policy in Modern Democracies*, p. 229.

<sup>59</sup> Benoit and Laver, *Party Policy in Modern Democracies*, p. 131. The scales used for these questions range between 1 and 20. The Benoit/Laver survey includes other measures of EU support; however, they all correlate highly and produce the same result.

<sup>60</sup> Although most of the missing estimates are for smaller parties, positions for parties from Ireland and France are missing entirely from the survey on these questions.

<sup>61</sup> Hix, Noury and Roland, 'Dimensions of Politics in the European Parliament'.

<sup>62</sup> Hix, Noury and Roland, 'Dimensions of Politics in the European Parliament'; Hix, Noury and Roland, *Democratic Politics in the European Parliament*.

<sup>63</sup> J. R. Cook and L. A. Stefanski, 'Simulation-Extrapolation Estimation in Parametric Measurement Error Models', *Journal of the American Statistical Association*, 89 (1994), 1314-28.

<sup>64</sup> This method corrects for measurement error of the independent variables only. The dependent variable, the positions estimated from word counts in speeches, is also measured with error. Wordfish

TABLE 2 Correlation of Independent Variables: National Party Positions Estimated from Expert Surveys and Roll-Call Votes

	Expert surveys			
	Voting behaviour: Nominate (Hix <i>et al.</i> 2006)		(UNC 2002) (Benoit/Laver 2006)	
	1st Dimension	2nd Dimension	Left-right	EU integration
Nominate 1st dim	1.00 (n = 106)			
Nominate 2nd dim	0.22 (n = 106)	1.00 (n = 106)		
Left-right <sup>a</sup>	0.84 (n = 82)	-0.02 (n = 82)	1.00 (n = 82)	
EU integration <sup>a</sup>	0.19 (n = 82)	0.70 (n = 82)	-0.01 (n = 82)	1.00 (n = 82)
Left-right <sup>b</sup>	0.86 (n = 68)	-0.01 (n = 68)	0.97 (n = 62)	1.00 (n = 68)
EU authority <sup>b</sup>	0.08 (n = 64)	-0.57 (n = 64)	0.22 (n = 58)	0.22 (n = 58)
			1.00 (n = 64)	1.00 (n = 64)

<sup>a</sup>UNC (2002); <sup>b</sup>Benoit and Laver (2006).

researchers should use such a correction for models that involve ideological estimates as independent variables whenever possible.<sup>65</sup> Since we know the measurement error for two of our three position estimates, we follow their advice. Both surveys allow us to assess measurement error through the standard deviations of expert responses.<sup>66</sup> Although it is, in theory, possible to estimate the uncertainty surrounding the Nominate scores as well, the EP Nominate dataset does not contain such uncertainty measures, and we therefore run simple OLS models for these variables.<sup>67</sup>

Table 3 examines in detail which variables best explain the positions estimated by Wordfish for all three languages. All models include country dummies to capture any member state specific effects. Such dummies allow the speech positions to shift for parties from a particular country. Because we use three different ideology estimates, our sample size varies. The full sample (models 1, 4 and 7) uses Nominate scores to capture left-right ideology (first dimension scores) and positions towards EU integration (second dimension scores). The use of survey estimates restricts our sample size to eighty-two parties when using the UNC expert survey (models 2, 5 and 8), and to sixty-four parties when using the Benoit/Laver expert survey data (models 3, 6 and 9).

Regardless of the specific ideology measure and of the language, we find that the variables capturing party position towards EU integration are highly statistically significant. In contrast, left-right ideology variables were either not statistically significant or only marginally statistically significant (UNC survey data). Thus, positions extracted from MEP speeches appear to reflect party positions better towards deeper EU integration than left-right ideology. In addition to the importance of the EU integration variables, *F*-tests reveal that the country dummies explain the party positions as well. The country dummies are jointly significant at the 0.01 level in each of the models presented in Table 3, suggesting that there are national-specific effects reflected in the speech dimension.

As we pointed out previously, aggregating speeches into party units can pose a problem because the results may only hold for the aggregate level, not the individual level. Therefore, we run Model 1 from Table 3 on individual speech positions using the individual Nominate scores for each MEP. If the results from the national party aggregate data are accurate, we

(*F*'note continued)

allows researchers to estimate the fundamental uncertainty surrounding the positions via a parametric bootstrap. We have shown elsewhere through simulations that the confidence intervals of the estimated positions in Wordfish significantly decrease as the number of unique words used in the analysis increases (Slapin and Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts'). Because we use several thousand unique words to estimate the positions, the confidence intervals of those estimates are rather small (see Appendix B). Moreover, measurement error in the dependent variable will not cause the kind of attenuation bias in the regression coefficients that we worry about. (Keith T. Poole, 'Measuring Bias and Uncertainty in Ideal Point Estimates via the Parametric Bootstrap', *Political Analysis*, 12 (2004), 105–27). Alternatively, one could apply Bayesian statistical analysis to estimate positions and their uncertainty (Han, 'Analysing Roll Calls of the European Parliament').

<sup>65</sup> Benoit, Laver and Mikhaylov, 'Treating Words as Data with Error: Uncertainty in Text Statements of Policy Position', *American Journal of Political Science*, 53 (2009), 495–513.

<sup>66</sup> To estimate the SIMEX model as implemented in *R*, we use as the measurement error the mean standard deviation of responses across all parties.

<sup>67</sup> It is possible to generate uncertainty estimates for Nominate using a parametric bootstrap (Jeffrey B. Lewis and Keith T. Poole, 'Measuring Bias and Uncertainty in Ideal Point Estimates via the Parametric Bootstrap', *Political Analysis*, 12 (2004), 105–27). Alternatively, one could apply Bayesian statistical analysis to estimate positions and their uncertainty (Han, 'Analysing Roll Calls of the European Parliament').

TABLE 3 *Explaining Speech Positions: Regression Results with Country-Fixed Effects*

Variable	English			French			German		
	(OLS)	(SIMEX)	(SIMEX)	(OLS)	(SIMEX)	(SIMEX)	(OLS)	(SIMEX)	(SIMEX)
Nominate (1st dim)	-0.109 (0.135)			-0.1587 (0.138)			-0.1448 (0.120)		
Nominate (2nd dim)	-0.720*** (0.225)			-0.987*** (0.229)			-0.722*** (0.199)		
Left-right <sup>a</sup>		-0.065* (0.034)			-0.073** (0.034)			-0.063** (0.03)	
EU position <sup>a</sup>		-0.186*** (0.042)		-0.275*** (0.053)			-0.201*** (0.045)		
Left-right <sup>b</sup>		-0.014 (0.019)			-0.014 (0.021)			-0.015 (0.018)	
EU authority <sup>b</sup>		0.070** (0.029)			0.088*** (0.030)			0.078*** (0.027)	
Constant	-0.876** (0.363)	0.483 (0.451)	-1.471*** (0.501)	-0.829** (0.371)	1.055** (0.490)	-1.666*** (0.541)	-1.840*** (0.322)	-0.401 (0.408)	-2.527*** (0.47)
Observations	106	82	64	106	82	64	106	82	64

Note: Standard errors in parentheses. Dependent variables are estimated Wordfish positions for each national party. Nominate ideology scores for national parties are for the 5th EP from Hix *et al.*, 'Dimensions of Politics in the European Parliament'. The other two ideology estimates are from expert surveys ('Hooghe and Marks, 'Chapel Hill 2002 Expert Survey'; and 'Benoit and Laver, *Party Policy in Modern Democracies*'). Country-fixed effects are omitted from the table. \* $p \leq 0.1$ , \*\* $p \leq 0.05$ , \*\*\* $p \leq 0.01$ .

TABLE 4 Explaining Individual-Level Speech Positions (OLS)

	English Wordfish positions
Nominate (1st dimension)	-0.015 (0.082)
Nominate (2nd dimension)	-0.449 (0.124)***
Constant	-0.298 (0.193)
Country-fixed effects	Yes
Observations	427

*Note:* Standard errors in parentheses. Dependent variables are estimated Wordfish positions for each MEP. Nominate ideology scores for each MEP are for the 5th EP from Hix *et al.*, 'Dimensions of Politics in the European Parliament'). Country fixed-effects are omitted from table, but are jointly significant. \* $p \leq 0.1$ , \*\* $p \leq 0.05$ , \*\*\* $p \leq 0.01$ .

would expect the same strong effects of the second Nominate dimension and the country dummies, but not of the first Nominate dimension. This is exactly what we find (Table 4). The effect of the second Nominate dimension is statistically significant, whereas the first dimension does not explain the positions well. The *F*-test also reveals that the country-fixed effects are jointly significant ( $p < 0.001$ ). These individual level results reveal that positions extracted from speech are similar to the second dimension extracted by Nominate from votes, but they contain national-specific positions as well. This discrepancy may be due to constraints placed on MEPs by the party groups when voting that are not present when speaking.

The results from these regressions with country-fixed effects indicate that national level factors are important, but they do not tell us which national level factors matter. Therefore, we introduce variables that reflect national politics in speeches. Specifically, we add variables that measure the wealth, size and net contribution status of each country. We measure a member state's wealth by its gross domestic product per capita (*GDPpercap*), its size by the log of its population (*Logpop*), and the net contribution status by the average net contribution per capita to the EU budget between 1999 and 2003 (*Netconpercap*).<sup>68</sup> Both net contributions per capita and GDP per capita are included to capture a redistributive dimension. Parties from poor states, or states that are net receivers of EU money, may take positions different from parties coming from rich, or net payer, countries. Redistributive issues have, for example, been a major source of conflict when negotiating EU budgets and the Common Agricultural Policy (CAP). In addition, many EU institutional issues, such as distribution of votes in the Council of Ministers and the size of the Commission, create a large state vs. small state divide. We capture these potential divisions with the logged population variable.

Again, we use a SIMEX error-corrected OLS model when using the survey data to capture EU integration and left-right positions, and standard OLS for models using the Nominate data. By combining country-specific predictors with measures of party ideology,

<sup>68</sup> Average net contributions per capita for 1999–2003 are operating budgetary balances taken from the 2005 EU Commission report on the allocation of EU expenditures per member state divided by population, p. 138 ([http://ec.europa.eu/budget/documents/revenue\\_expenditure\\_en.htm](http://ec.europa.eu/budget/documents/revenue_expenditure_en.htm)). We include those years of the 5th European Parliament for which the budget lists the balances for EU-15 member states only.

our data acquire a two-level hierarchical structure (parties nested within countries). If we do not account for this hierarchical structure, we would ignore clustering in the data and violate the assumption that the observations are independent.<sup>69</sup> Errors associated with parties from the same country are likely to be positively correlated. This would lead to the attenuation of the standard errors of country-level coefficients, and may lead to a false rejection of the null hypothesis for country-level variables. To account for the multilevel data structure, we also estimate a mixed-effects multilevel model.<sup>70</sup> The results are very similar to error correction and OLS models, and we present the simpler model here and the hierarchical model in Appendix B.

Table 5 presents the results of the three national-level effects in addition to the party-level survey data variables. These models demonstrate that the best national-level explanations for the extracted positions are member states' net contributions to the EU per capita and GDP per capita. The net contributions per capita variable is significant in all but three models. The GDP per capita variable is significant in two of three models where the net contributions per capita variable is not significant.<sup>71</sup> Both these variables capture a redistributive dimension in EU politics, suggesting that EP speech, in part, reflects a national divide over the allocation of resources. The size of a member state measured by its population, however, seems to have little impact on the MEP speech positions. The population variable is only statistically significant in two models. This statistical significance disappears when controlling for the hierarchical nature of the data, while the statistical significance of the other two country-level variables does not (see Appendix B).

To capture the substantive importance of the variables in determining party positions, we examine how changing one independent variable would affect the movement of a hypothetical party on the speech dimension with regard to all the other parties in the dataset. We create our hypothetical party by setting all variables to their mean except for the party's position regarding EU integration. We set this variable to its minimum value – the position of the party least favourable towards integration. We then examine the percentage of parties that are estimated to be to the left of this hypothetical party. Next, we reset the hypothetical party's position regarding integration to the maximum value and again examine how many parties lie to the left of our hypothetical party on our dimension. We do this for all variables of interest and report the results in Table 6.

Table 6 demonstrates that across all measures and languages, the largest movements in party positions on the speech dimension occur when changing the country-level variables capturing redistribution and the variables capturing a party's position with regard to EU integration. It appears that the country-level effects are slightly larger than the effects of a party's position regarding integration. Nevertheless, both variables are very important when assessing the parties' positions extracted from speech. However, moving the left-right party positions from their minimum to their maximum does not affect the party positions estimated from speech very much at all.<sup>72</sup>

<sup>69</sup> Marco Steenbergen and Bradford S. Jones, 'Modeling Multilevel Data Structures', *American Journal of Political Science*, 46 (2002), 218–37, p. 233.

<sup>70</sup> Andrew Gelman and Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge: Cambridge University Press, 2007).

<sup>71</sup> GDP per capita is significant in the models using the UNC survey data, which excludes Luxembourg. Luxembourg is an outlier on GDP per capita, so excluding it from the analysis alters the results.

<sup>72</sup> To preserve space, we only report the predicted values for the country-level variables that attain statistical significance in the hierarchical model found in Appendix B.

TABLE 5 Explaining Speech Positions: Ideology and Country Effects

Variable	English			French			German		
	(1) (OLS)	(2) (SIMEX)	(3) (SIMEX)	(4) (OLS)	(5) (SIMEX)	(6) (SIMEX)	(7) (OLS)	(8) (SIMEX)	(9) (SIMEX)
Nominate (1st dim)	-0.117 (0.167)			-0.135 (0.162)			-0.089 (0.160)		
Nominate (2nd dim)	-0.675** (0.276)			-1.04*** (0.269)			-0.854*** (0.264)		
Left-right <sup>a</sup>		-0.054 (0.038)			-0.069* (0.04)			-0.063 (0.047)	
EU position <sup>a</sup>		-0.218*** (0.056)			-0.299*** (0.055)			-0.256*** (0.066)	
Left-right <sup>b</sup>			-0.007 (0.023)			-0.009 (0.023)			-0.005 (0.025)
EU authority <sup>b</sup>			0.086** (0.035)			0.099*** (0.035)			0.078** (0.038)
Population (Log)	-0.037 (0.106)	-0.028 (0.082)		-0.108 (0.118)	-0.137 (0.103)	-0.168* (0.0843)	-0.229** (0.113)	0.106 (0.101)	0.088 (0.095)
GDP (per capita)	-0.00003 (0.00002)	-0.0002*** (0.00003)	0.00001 (0.00002)	-0.00003 (0.00002)	-0.0002*** (0.00003)	0.000001 (0.00002)	-0.000002 (0.00002)	-0.0001*** (0.00003)	0.00003 (0.00003)
Net contribution (per capita)	0.001* (0.0007)	0.0003 (0.0006)	0.004*** (0.001)	0.0003 (0.0007)	0.0007 (0.0007)	-0.000039 (0.0006)	0.003*** (0.001)	0.002*** (0.0007)	0.004*** (0.001)
Constant	0.703 (0.845)	6.08*** (0.886)	-1.01 (1.10)	1.04 (0.823)	6.35*** (0.925)	-0.447 (1.07)	-0.432 (0.809)	4.33*** (1.05)	-1.70 (1.18)
Observations	106	82	64	106	82	64	106	82	64
Number of countries	15	14	13	15	14	13	15	14	13

Note: Simex standard errors based on jackknife estimation in parentheses. \* $p \leq 0.1$ , \*\* $p \leq 0.05$ , \*\*\* $p \leq 0.01$ . Dependent variables are estimated Wordfish positions for each national party. Nominate ideology scores for national parties are for the 5th EP from Hix *et al.*, 'Dimensions of Politics in the European Parliament'. The other two ideology estimates are from expert surveys (<sup>a</sup>Hooghe and Marks, 'Chapel Hill 2002 Expert Survey'; and <sup>b</sup>Benoit and Laver, *Party Policies in Modern Democracies*) and are estimated with measurement error (using the mean standard deviation of responses across all parties). We use the SIMEX package in R.

TABLE 6 *Change in Party Position (Percentiles) by Varying Independent Variables of Interest*

Ideology measure	Language	EU position variable Min → Max	GDP Min → Max	Net contribution Min → Max	Left-right Min → Max
Expert Survey <sup>a</sup>	English	0.84, 0.35	0.98, 0.15		0.59, 0.40
	German	0.89, 0.28	0.89, 0.21		0.56, 0.33
	French	0.93, 0.29	0.95, 0.16		0.65, 0.37
Expert Survey <sup>b</sup>	English	0.34, 0.83		0.25, 0.94	0.53, 0.52
	German	0.33, 0.80		0.17, 0.97	0.52, 0.47
	French	0.34, 0.86		0.31, 0.91	0.59, 0.53
Nominate	English	0.67, 0.33		0.41, 0.68	0.51, 0.45
	German	0.72, 0.25		0.27, 0.84	0.50, 0.41
	French	0.82, 0.25		0.47, 0.62	0.58, 0.47

Note: GDP and net contribution are per capita. The changes reported here reflect hypothetical party movements in terms of percentile when varying the independent variable of interest from its minimum value to its maximum value. All other independent variables are set to their means. <sup>a</sup>UNC (2002); <sup>b</sup>Benoit and Laver (2006).

#### CONCLUSION

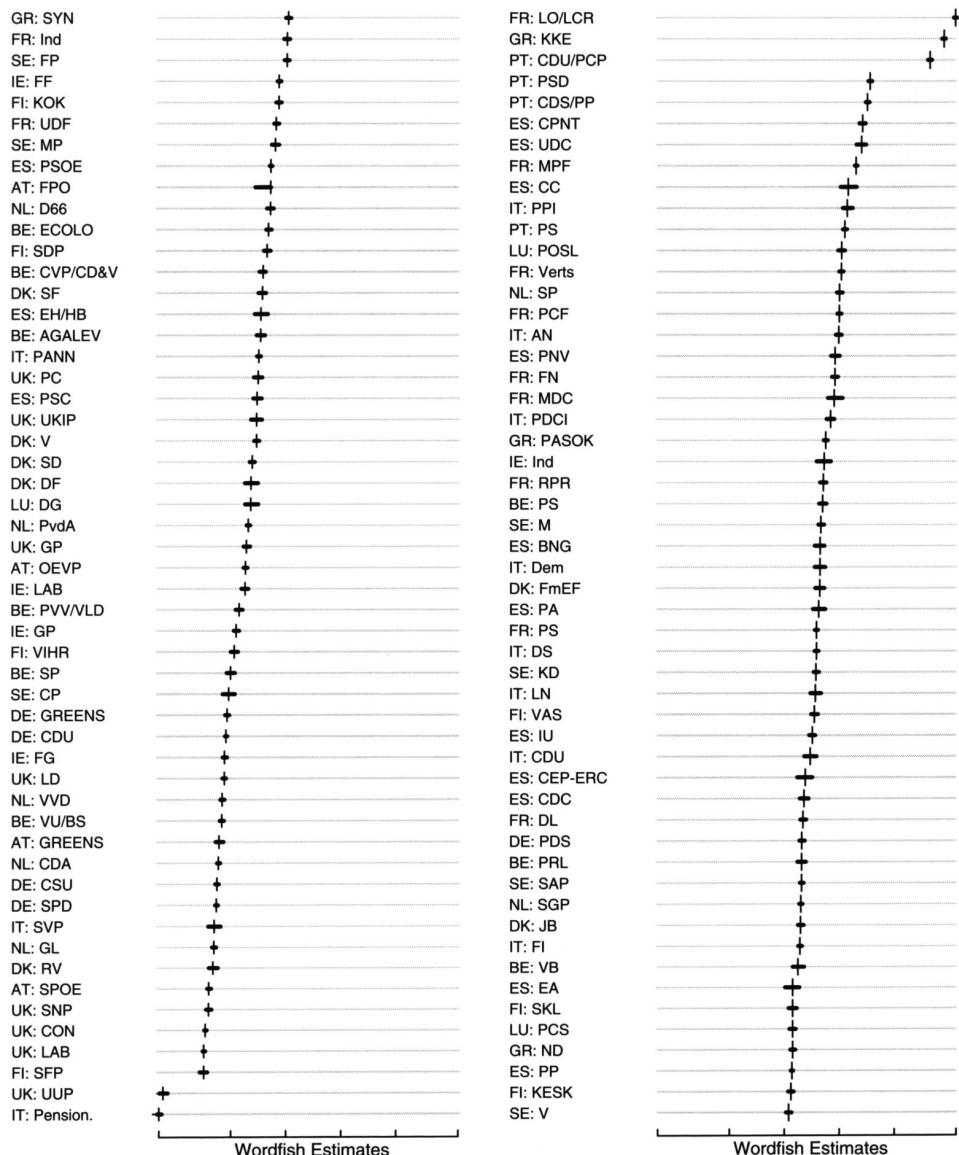
Previously, speeches have been an untapped source of information in the study of the European Parliament. We have estimated positions of national parties and MEPs using word counts from speeches delivered during the 5th European Parliament (1999–2004). The estimated positions reflect parties' stances with regard to EU integration as well as a strong national dimension. We could not find strong evidence that the estimates reflect parties' overall left-right positions. We have demonstrated the robustness of these findings using several datasets and statistical models. The basic findings hold regardless of (1) the language of translation used to estimate the positions, (2) the different methods used to estimate left-right and pro-/anti-European positions, (3) the type of statistical model used (OLS, error-corrected OLS, or multilevel regression analysis), and finally (4) the level of analysis (national party or MEP). The individual level analysis confirms that the positions derived from speech most closely reflect the second dimension of Nominate, but include national level factors not found in voting. These robust results are surprising because MEP behaviour has so far been best explained by partisan left-right ideology using roll-call votes as the primary source of data. They suggest that the party ideology reflected in speeches may not be identical to the ideology expressed through voting. Because voting and speaking are subject to different institutional constraints, these different data provide a different picture of ideology in the EP.

Our findings also have implications for users of computer-based text analysis more generally. Such methods enjoy more and more popularity in political science. Combined with electronically available political texts such as party manifestos, legislative speeches, newspaper reports and political blogs, scholars today have immense sources of data to study party systems, political campaigns, legislatures, media and international conflicts. In particular, in comparative politics and international relations, such analysis involves multiple languages. Our study analysed to what extent computer-based content analysis is sensitive to language choice. We examined the robustness of the Wordfish technique using translations of the exact same speeches in the three most common working languages of

the EU (English, French and German). Our results suggest that the Wordfish technique is highly robust to the choice of language, as estimated positions correlate highly across these languages.

Our findings open up an exciting avenue of new research on quantitative analysis of political speeches and on democratic politics in the European Union. Legislative speeches offer a valuable data source to study ideology, but the choice of words in speeches is likely to be different from the choice of words in written political texts such as election manifestos. Future studies could, therefore, address which speeches are particularly suited for an analysis of ideology. Studying the incentives and the institutional constraints MEPs face when delivering speeches in the European Parliament will provide deeper insights into when the corpus of speeches accurately represents opinions in the EP and when it reflects biased opinions. Our results, at a minimum, suggest that constraints on speeches are different from the constraints on voting. Even though we did not find strong evidence for left-right ideology in speeches, it may be the case that this matters more in some specific policy areas, and researchers could disaggregate speeches into areas of interest (for example: foreign policy, social policy, economic policy) prior to analysing ideology. In the future, scholars may wish to address these questions in more detail. This may also help us to understand the differences in ideology expressed by legislators through voting and through speaking.

## APPENDIX A: ESTIMATED SPEECH POSITIONS IN THE EP, 1999–2004

**Estimated Speech Positions in the EP (1999-2004)**

Note: Labels include country and party abbreviations (see Hix roll call data). English translation estimates shown, 95% CI generated from 200 bootstraps.

APPENDIX B: EXPLAINING SPEECH POSITIONS: MULTILEVEL REGRESSION

Variable	English			French			German		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
<i>Party level</i>									
Nominate	-0.113 (0.133)			-0.159 (0.135)			-0.14 (0.118)		
(1st dim)					-0.986*** (0.224)			-0.741*** (0.196)	
Nominate	-0.708*** (0.220)								
(2nd dim)									
Left-right <sup>a</sup>		-0.034 (0.029)			-0.038 (0.031)			-0.035 (0.026)	
EU position <sup>a</sup>		-0.152*** (0.041)			-0.227*** (0.044)			-0.166*** (0.037)	
Left-right <sup>b</sup>			-0.019 (0.018)			-0.02 (0.019)			
EU authority <sup>b</sup>			0.053** (0.021)			0.067*** (0.023)			0.054*** (0.020)
<i>Country level</i>									
Population	-0.154 (0.196)	-0.056 (0.120)		-0.133 (0.172)	-0.217 (0.175)	-0.17 (0.104)	-0.247* (0.138)	-0.008 (0.211)	0.068 (0.187)
(Log)						-0.00003 (0.00003)	-0.00015*** (0.00003)	-0.00000 (0.00003)	-0.00012** (0.00004)
GDP	-0.00003 (0.00003)	-0.00018*** (0.00003)						-0.00000 (0.00005)	0.00003 (0.00004)
(per capita)						0.004*** (0.0003)	0.003*** (0.0003)	0.002* (0.0003)	0.004** (0.0004)
Net contrib.	0.001 (0.001)	0 (0.001)		0.004*** (0.001)	0 (0.001)	0 (0.001)	0.003*** (0.001)	0.002* (0.001)	0.004** (0.002)
(per capita)									
Constant	1.032 (1.418)	5.627*** (1.100)		-0.476 (1.325)	1.288 (1.271)	5.740*** (0.991)	0.15 (1.099)	-0.16 (1.518)	3.644** (1.631)
SD (Intercept)	0.59	0.32		0.44	0.5	0.23	0.29	0.65	0.59
SD (Residual)	0.72	0.56		0.63	0.73	0.6	0.67	0.64	0.58
Observations	106	82		64	106	82	64	106	64
No. of countries	15	14		13	15	14	13	15	13
Log likelihood	-128.04	-75.9		-68.85	-128.13	-78.92	-69.57	-118.26	-74.74

Note: Standard errors in parentheses. Dependent variables are estimated Wordfish positions for each national party. Nominate ideology scores for national parties are for the 5th EP from Hix *et al.*, 'Dimensions of Politics in the European Parliament'. The other two ideology estimates are from expert surveys (Hooghe and Marks, 'Chapel Hill 2002 Expert Survey'; and Benoit and Laver, *Party Policies in Modern Democracies*). \*  $p \leq 0.1$ , \*\*  $p \leq 0.05$ , \*\*\*  $p \leq 0.01$ .

# General purpose computer-assisted clustering and conceptualization

Justin Grimmer<sup>a</sup> and Gary King<sup>b,1</sup>

<sup>a</sup>Department of Political Science, Stanford University, Encina Hall West, 616 Serra Street, Palo Alto, CA 94305; and <sup>b</sup>Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2010.

Contributed by Gary King, December 22, 2010 (sent for review September 23, 2010)

We develop a computer-assisted method for the discovery of insightful conceptualizations, in the form of clusterings (i.e., partitions) of input objects. Each of the numerous fully automated methods of cluster analysis proposed in statistics, computer science, and biology optimize a different objective function. Almost all are well defined, but how to determine before the fact which one, if any, will partition a given set of objects in an “insightful” or “useful” way for a given user is unknown and difficult, if not logically impossible. We develop a metric space of partitions from all existing cluster analysis methods applied to a given dataset (along with millions of other solutions we add based on combinations of existing clusterings) and enable a user to explore and interact with it and quickly reveal or prompt useful or insightful conceptualizations. In addition, although it is uncommon to do so in unsupervised learning problems, we offer and implement evaluation designs that make our computer-assisted approach vulnerable to being proven suboptimal in specific data types. We demonstrate that our approach facilitates more efficient and insightful discovery of useful information than expert human coders or many existing fully automated methods.

Creating categories and classifying objects in the categories “is arguably one of the most central and generic of all our conceptual exercises. It is the foundation not only for conceptualization, language, and speech, but also for mathematics, statistics, and data analysis in general. Without classification, there could be no advanced conceptualization, reasoning, language, data analysis or, for that matter, social science research” (1). An important step in the development of new hypotheses is the adoption of new ways of partitioning objects into categories. In this paper, we develop a method intended to assist in the creation of unique and insightful conceptualizations from a wide array of possible datasets and substantive problems. We focus on creating “clusterings” (i.e., partitions) of a given set of input objects in an “unsupervised” framework (i.e., with no training set).

Illustrations of useful clusterings in particular applications have been found for some of the existing individual cluster analysis methods. However, for a given application, no method exists for choosing before the fact which of these unsupervised approaches will lead to the most useful clusterings or the most insightful discoveries.

Although our approach builds on almost all prior methods, our goal diverges from the existing literature in one crucial respect: Whereas current cluster analysis methods are designed to produce fully automated clustering (FAC), we attempt to create a computer-assisted clustering (CAC) approach. The problem with FAC is that it requires a single, precisely defined objective function that works across applications. This is infeasible given that human beings are typically optimizing a (mathematically ill-defined) goal of “insightful” or “useful” conceptualizations; the definition of “insightful” differs to some degree by user; and codifying human creativity in a mathematical function is either logically impossible or well beyond current technology. (Existing methods, which we describe as FAC, do come with tuning para-

meters that enable a user to adjust the optimization function, but in our experience most adjustments turn out to have very small empirical effects, typically much smaller than the differences between methods.)

We develop a CAC approach that uses and encompasses all existing automated cluster analysis methods, numerous novel ones we create (based on combinations of existing solutions), and any others a researcher may create by hand or other technique. By using the collective wisdom of the statistical literature on cluster analysis, we generate a single approach applicable across many substantive problems, without having to know ahead of time which method to apply. We are able to do this by requiring interaction between our methodology and a human user.

In part because of the unsupervised learning nature of cluster analysis, the literature offers few satisfactory procedures for evaluating categorization schemes or the methods that produce them. Unlike in supervised learning methods or classical statistical estimation, straightforward concepts like unbiasedness or consistency do not immediately apply. We respond to this challenge by developing a design for evaluation experiments that reveal the quality of the results and the degree of useful information discovered. We implement these experimental designs in a variety of datasets and show that our CAC methods lead to more insightful conceptualizations than either subject matter experts or individual FAC methods can do alone.

In practice, before applying our algorithm and evaluation techniques, researchers may wish to set aside a randomly selected test set of observations. This holdout set could then be used as a way of making the researcher vulnerable to being wrong about the applicability or generality of a new conceptualization. This may also help prevent researchers from choosing clusterings that merely conform to preexisting conceptualizations, although of course researchers may also choose to let these preexisting views help guide their search for new conceptualizations. Below, we demonstrate that the clusterings and conceptualizations we discover in our subset of documents provide a useful way of analyzing the entire collection of documents.

Although our methods apply to categories of any type of object, we apply them here to clustering documents containing unstructured text. The spectacular growth in the production and availability of text makes this application of crucial importance in many fields.

## 2 Methodology

One way to think about CAC is to imagine presenting an extremely long list of clusterings (ideally, all of them) and letting the researcher choose the best one for his or her substantive pur-

Author contributions: J.G. and G.K. designed research, performed research, contributed new tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

<sup>1</sup>To whom correspondence should be addressed. E-mail: king@harvard.edu.

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1018067108/DCSupplemental](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1018067108/DCSupplemental).

poses. However, human beings do not have the patience, attention span, memory, or cognitive capacity to evaluate so many clusterings in haphazard order. Moreover, from the point of view of a human being, many clusterings are essentially the same. (Imagine 10,000 documents sorted into five categories and moving one document from category 3 to 4; these clusterings are essentially the same because few would even be able to perceive the difference.) Thus, we seek to organize these clusterings so researchers can quickly select the one that best satisfies their particular objectives.

Our procedure represents each clustering as a point in a two-dimensional visual space, such that clusterings (points) close together in the space are almost the same (and so can be disregarded except for fine tuning), and those farther apart may warrant a closer look because they differ in some important way. In effect, this visualization translates the uninterpretable chaos of huge numbers of possible clusterings into a simple framework that (we show) human researchers are able to comprehend and use to efficiently select one or a small number of clusterings that conveys the most useful information.

To create our space of clusterings, we follow six steps, outlined here and detailed below. First, we translate textual documents to a numerical dataset (Section 2.1). (This step is necessary only when the items to be clustered are text documents or in general not already numerical; all our methods would apply without this step to objects with preexisting numerical data.) Second, we apply (essentially) all clustering methods proposed in the literature, one at a time, to the numerical dataset (Section 2.2). Each approach represents different substantive assumptions that are difficult to express before their application, but the effects of each set of assumptions are easily seen in the resulting clusters, and it is the resulting clustering that is of most interest to applied researchers. (A new R package we have written makes this relatively fast.) Third, we develop a metric to measure the similarity between any pair of clusterings (Section 2.3). Fourth, we use this metric to create a metric space of clusterings, along with a lower dimensional Euclidean representation useful for visualization (Section 2.4).

Fifth, we introduce a “local cluster ensemble” method (Section 2.5) as a way to summarize any point in the space, including points for which there exist no prior clustering methods—in which case they are formed as local weighted combinations of existing methods, with weights based on how far each existing clustering is from the chosen point. This allows for the fast exploration of the space, ensuring that users of the software are able to quickly identify partitions useful for their particular research question. Sixth and finally, we develop a new type of animated visualization that uses the local cluster ensemble approach to explore the metric space of clusterings by moving around it while one clustering slowly morphs into others (Section 2.6), again to rapidly allow users to easily identify the partition (or partitions) useful for a particular research question. We also introduce an optional addition to our method that creates new clusterings (Section 2.7).

**2.1 Standard Preprocessing: Text to Numbers.** We begin with a set of text documents of variable length. For each, we adopt common procedures for representing them quantitatively: We transform to lower case, remove punctuation, replace words with their stems, and drop words appearing in fewer than 1% or more than 99% of documents. For English documents, about 3,500 unique word stems usually remain in the entire corpus. We then code each document with a set of (about 3,500) variables, each coding the number of times a word stem is used in that document.

Despite all the information discarded, these procedures are very common (2). The reason is that most human language is highly repetitive, and so this representation is usually more than adequate. For example, we need not read many sentences of a

vitriolic blog post about a political candidate before getting the point. Our general procedure also accommodates multiple representations of the same documents. These might include tf-idf or other term weighting representations, part of speech tagging, tokenization rules such as replacing “do” and “not” with “do\_not”, etc. (3). Likewise, the many variants of kernel methods—procedures to produce a similarity metric between documents without explicitly representing the words in a matrix—could also be included (4).

**2.2 The Collective Wisdom of the Statistical Community.** Second, we apply a large number of clustering methods, one at a time, to the numerical representation of our documents. To do this, we have written an R package that runs (with a common syntax) every published clustering method we could find that has been applied to text and used in at least one article by an author other than its developer; we have also included many clustering methods that have not been applied to text before. We developed computationally efficient implementations for the methods included in our program (including variational approximations for the Bayesian statistical methods) (5) so that we can run all the methods on a moderate sized dataset relatively fast; new methods can easily be added to the package as well. Although inferences from our method are typically not affected much, and almost never discontinuously, by including any additional individual method, there is no disadvantage in including as many methods as are available.

A complete list of the methods that we include in our application is available in the *SI Appendix*, but the method is extremely flexible. The only requirement is that each “method” form a proper clustering, with each document assigned either to a single cluster or to different clusters with weights that sum to 1.

**2.3 Distance Between Clusterings.** We next derive a metric for measuring how similar one clustering is to another. We do this stating three axioms that narrow the range of possible choices to only one. First, the distance is a function of the number of pairs of documents not placed together (i.e., in the same cluster) in both clusterings. (We also prove in the *SI Appendix* that focusing on pairwise disagreements between clusterings is sufficient to encompass differences based on all possible larger subsets of documents, such as triples, quadruples, etc.) Second, we require that the distance be invariant to the number of documents, given any fixed number of clusters in each clustering. Third, we set the scale of the measure by fixing the minimum distance to zero and the maximum distance to  $\log(k)$ . A key point is that none of these axioms requires that one artificially “align” clusterings before judging their distance, as some others have attempted; in fact, we do not even restrict the clusterings to have the same number of clusters.

As we prove in the *SI Appendix*, only one measure of distance satisfies all three axioms, the variation of information. This measure has also been derived for different purposes from a larger number of different first principles by Meila (6).

**2.4 The Space of Clusterings.** The matrix of distances between each pair in the set of  $J$  clusterings can be represented in a  $J$ -dimensional metric space. (The clusterings can each have the same number of clusters, if chosen by the user, or differing numbers.) We project this space down to two Euclidean dimensions for visualization. Because projection entails the loss of information, the key is to choose a multidimensional scaling method that retains the most crucial information. For our purposes, we need to preserve small distances most accurately, because they reflect clusterings to be combined (in the next section) into local cluster ensembles. As the distance between two clusterings increases, a higher level of distortion will affect our results less. This leads naturally to the Sammon multidimensional scaling algorithm

(7); in the *SI Appendix*, we define this algorithm and explain how it satisfies our criteria.

An illustration of this space is given in Fig. 1, *Middle*, with individual clusterings labeled (we discuss this figure in more detail below). Nearby points in this space represent similar clusterings, but by definition it eliminates the underlying diversity of individual clusterings and so does not work for our purposes. A related technique that is sometimes described by the same term organizes results by performing a “meta-clustering” of the individual clusterings. This alternative procedure has the advantage of preserving some of the diversity of the clustering solutions and letting the user choose, but because no method is offered to summarize the many clusterings within each “meta-cluster,” it does not solve the problem. Moreover, for our purposes, the technique suffers from a problem of infinite regress: Because any individual clustering method can be used to cluster the clusterings, a researcher would have to use them all and their combinations to avoid eliminating meaningful diversity in the set of clusterings to be explored. So whether the diversity of clusterings is eliminated by arbitrary choice of meta-clustering method rather than a substantive choice, or we are left with more solutions than we started with, these techniques, although useful for some other purposes, do not solve our particular problem.

Thus, to preserve local diversity and avoid the infinite regress resulting from clustering a set of clusterings, we develop here a method of generating local cluster ensembles, which we define as a new clustering created at a point in the space of clusterings from a weighted average of nearby existing clusterings. The procedure requires three steps. First, we define the weights around a user selected point in the space. Consider point  $\mathbf{x}^* = (x_1^*, x_2^*)$  in our space of clusterings. The new clustering defined at this point is a weighted average of nearby clusterings with one weight for each existing clustering in the space, so that the closer the existing clustering, the higher the weight. We define the weight for each existing clustering  $j$  on a normalized kernel as  $w_j = p(\mathbf{x}^*, \sigma^2) / \sum_{m=1}^J p(\mathbf{x}_m, \sigma^2)$ , where  $p(\mathbf{x}, \sigma^2)$  is the height of the kernel (such as a normal or Epanechnikov density) with mean  $\mathbf{x}^*$  and smooth-

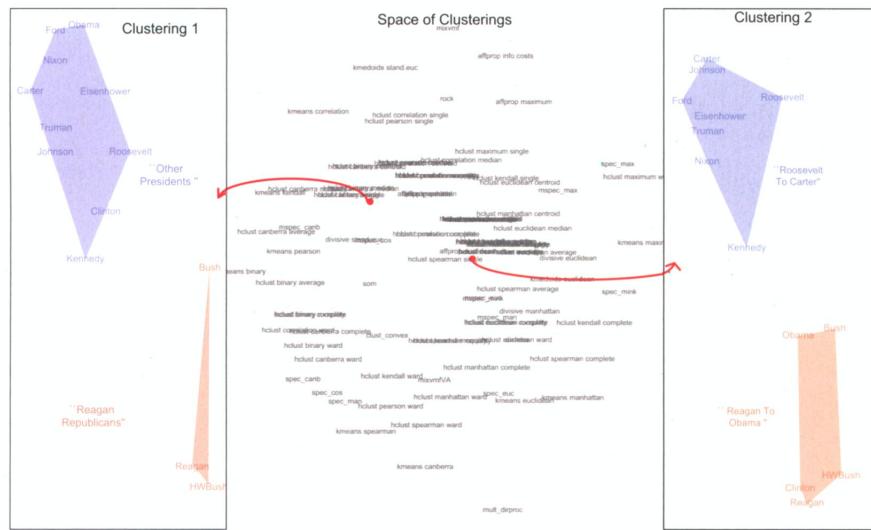
ing parameter  $\sigma^2$ . The collection of weights for all  $J$  clusterings is then  $\mathbf{w} = (w_1, \dots, w_J)$ . Note that although we are using a density to define the kernel, the approach requires no statistical or probabilistic reasoning.

Second, given the weights, we create a similarity matrix for the local cluster ensemble, where each clustering casts a weighted vote for whether each pair of documents appears together in a cluster in the new clustering. First, for a corpus with  $N$  documents clustered by method  $j$  into  $K_j$  clusters, we define an  $N \times K_j$  matrix  $\mathbf{c}_j$  that records how each document is allocated into (or among) the clusters (i.e., so that each row sums to 1). We then horizontally concatenate the clusterings created from all  $J$  methods into an  $N \times K$  weighted “voting matrix” of methods by document pairs,  $\mathbf{V}(\mathbf{w}) = \{\mathbf{w}_1 \mathbf{c}_1, \dots, \mathbf{w}_J \mathbf{c}_J\}$  (where  $K = \sum_{j=1}^J K_j$ ). The result of the election is a new similarity matrix, which we create as  $\mathbf{S}(\mathbf{w}) = \mathbf{V}(\mathbf{w}) \mathbf{V}(\mathbf{w})'$ . This calculation places priority on those cluster analysis methods closest in the space of clusters.

Finally, we create a new clustering for point  $\mathbf{x}^*$  in the space by applying any coherent clustering algorithm to this new averaged similarity matrix (with the number of clusters fixed to a weighted average of the number of clusters from nearby clusterings, using the same weights). As we demonstrate in the *SI Appendix*, our definition of the local cluster ensemble approach becomes invariant to the particular choice of clustering method applied to the new averaged similarity matrix as the number of clusterings increase. This invariance eliminates the infinite regress problem by turning a meta-cluster method selection problem into a weight selection problem (with weights that are variable in the method). The *SI Appendix* also shows how our local cluster ensemble approach is closely related to our underlying distance metric defined in Section 2.3. The key point is that the local cluster ensemble approach will approximate more possible clusterings as additional methods are included and of course will never be worse, and usually considerably better, in approximating a new clustering than the closest existing observed point.

**2.6 Cluster Space Visualization.** Fig. 1 illustrates our visualization of the space of clusterings, when applied to one simple corpora of documents. This simple and small example, which we choose for expository purposes, includes only the biographies of each US president from Roosevelt to Obama (see <http://whitehouse.gov>).

The two-dimensional projection of the space of clusterings is illustrated in the figure’s middle panel, with individual methods



**Fig. 1.** A clustering visualization. The center panel gives the space of clusterings, with each name printed representing a clustering generated by that method, and all other points in the space defined by our local cluster ensemble approach that averages near-by clusterings. Two specific clusterings (see red dots with connected arrows), each corresponding to one point in the central space, appear to the left and right; labels in the different color-coded clusters are added by hand for clarification, as is the spacing in each.

labeled. Each method corresponds to one point in this space and one set of clusters of the given documents. Points corresponding to a labeled method correspond to results from prior research; other points in this space correspond to new clusterings, each constructed as a local cluster ensemble.

A key point is that once the space is constructed, the labeled points corresponding to previous methods deserve no special priority in choosing a final clustering. For example, a researcher should not necessarily prefer a clustering from a region of the space with many prior methods as compared to one with few or none. In the end, the choice is the researcher's and should be based on what he or she finds to convey useful information. Because the space itself is crucial, but knowledge of where any prior method exists in the space is not, visualization software can easily toggle off these labels so that researchers can focus on clusterings they identify.

The space is formally discrete, because the smallest difference between two clusterings occurs when (for nonfuzzy partitions) exactly one document moves from one cluster to another, but an enormous range of possible clusterings still exists: Even this tiny dataset of only 13 documents can be partitioned in 27,644,437 possible ways, each representing a different point in this space. A subset of these possible clusterings appears in the figure corresponding to all those clusterings the statistics community has come up with, as well as all possible local cluster ensembles that can be created as weighted averages from them. (The arching shapes in the figure occur regularly in dimension reduction when using methods that emphasize local distances between the points in higher dimensional space; see ref. 14.)

Fig. 1 also illustrates two points (as red dots) in the middle panel, each representing one clustering and portrayed on one side of the central graph, with individual clusters color coded (and substantive labels added by hand for clarity). Clustering 1, in the left panel, creates clusters of "Reagan Republicans" (Ronald Reagan, George H. W. Bush, and George W. Bush) and all others. Clustering 2, in the right panel, groups the presidents into two clusters organized chronologically.

This figure summarizes snapshots of an animated software program at two points. In general, the software can be set up so a researcher can put a single cursor somewhere in the space of clusterings and see the corresponding set of clusters for that point appear in a separate window. The researcher can then move this point and watch the clusters in the separate window morph smoothly from one clustering to another. Our experience in using this visualization often leads us first to check about 4–6 well-separated points, which seems to characterize the main aspects of the diversity of all the clusterings. Then, we narrow the grid further by examining about the same number of clusterings in the local region. Although the visualization offers an enormous number of clusterings, the fact that they are highly ordered in this simple geography makes it possible to understand with greatly reduced time and effort.

**2.7 Optional New Clustering Methods to Add.** For most applications, beginning with the collective wisdom of the statistics community, and clusterings constructed from them, helps to narrow down the enormous space of all possible clusterings to a large (indeed larger than has ever before been explored) but yet still manageable set of solutions. However, there may well be useful insights to be found outside of the large space that we have already identified. Thus, we offer two methods to explore some of the remaining uncharted space. First, we randomly sample new clusterings from the entire space. Second, we define a Markov chain to move beyond the space of existing clusterings to the area around those clusterings. Details about both algorithms are available in the *SI Appendix*.

### 3 Evaluation Designs

The most important approach to evaluating a purely unsupervised learning approach to clustering is whether the user, or the user's intended audience, finds the chosen clustering useful or insightful. Thus, a perfectly reasonable approach is to use our method, choose a clustering and gather insight, and be done. However, one may also wish to go further in some circumstances and formally evaluate the clustering solutions.

Common approaches to evaluating the performance of cluster analysis methods, which include comparison to internal or supervised learning standards, have known difficulties. Internal standards of comparison define a quantitative measure indicating high similarity of documents within, and low similarity of documents across, clusters. But if this were the goal, we could define a cluster analysis method with an objective function that optimizes it directly; this may lead to a good answer but not one that is vulnerable to being proven wrong. Indeed, because any one quantitative measure is unlikely to reflect the actual substance a researcher happens to be seeking, "good scores on an internal criterion do not necessarily translate into good effectiveness in an application" (ref. 2, pp. 328–329).

An alternative evaluation approach is based on supervised learning standards, which involve comparing the results of a cluster analysis to some "gold standard" set of clusters, prechosen by human coders. Although human coders may be capable of assigning documents to a small number of given categories, they are incapable of choosing an optimal clustering or one in any sense better than what a CAC method could enable them to create. As such, using a supervised learning "gold standard" to evaluate an unsupervised learning approach is also of questionable value.

Success at facilitating discovery is difficult to formalize mathematically and easy to lead to unfalsifiable approaches. Indeed, some in the statistical literature have even gone so far as to chide those who attempt to use unsupervised learning methods to make systematic discoveries as unscientific (15).

To respond to these problems, we introduce and implement three direct evaluation approaches using insights from survey research and social psychology to compare to elicited human judgment in ways that people are capable of providing. We first evaluate cluster quality, the extent to which intraclass similarities outdistance intercluster similarities (Section 3.1). Cluster quality demonstrates that users of our approach are able to efficiently search through the space of clusterings to identify clusterings that are coherent and useful to others. Second is discovery quality, a direct evaluation by substance matter experts of insights produced by different clusterings in their own data (Section 3.2). This ensures that the clusterings identified are insightful for experts working in a field of study. Third and finally, we offer a substantive application of our method and show how it assists in discovering a specific useful conceptualization and generates new verifiable hypotheses that advance the political science literature (Section 3.3). For this third approach, the judge of the quality of the knowledge learned is the reader of this paper.

**3.1 Cluster Quality.** We judge cluster quality with respect to a particular corpus by randomly drawing pairs of documents from the same cluster and from different clusters and asking human coders unaware how each document was chosen to rate the similarity of the documents within each pair on a simple three point scale: (i) unrelated, (ii) loosely related, (iii) closely related. (Our extensive pretesting indicated that intercoder reliability suffers with more categories, but coders are able to understand and use effectively this coding scheme. We also found that the average code from 10 graduate students correlated with the average code from the Amazon Mechanical Turk system at 0.99.) The idea is to keep our human judges focused on well-defined tasks they are able to perform well, in this case comparing only two documents at a time. Then the numerical measure of cluster quality is the aver-

age rating of pair similarity within clusters minus the average of pairs in different clusters. (The *SI Appendix* also introduces a way to save on evaluation costs in measuring cluster quality.)

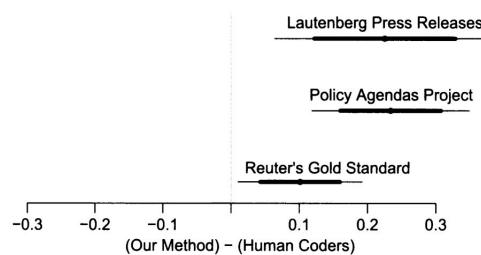
We apply this measure in each of three different corpora by choosing 25 pairs of documents (13 from the same clusters and 12 from different clusters), computing cluster quality, and averaging over the judgments about the similarity of each pair made separately by many different human coders. We then compare the cluster quality generated by our approach to the cluster quality from a preexisting hand-coded clustering. This comparison demonstrates that users of our method are able to identify clusterings that are coherent and are able to efficiently search through the millions of clusterings we present users.

What we describe as “our approach” here is a single clustering from the visualization we chose ourselves without participating in evaluating document similarity. This procedure is biased against our method because if we had let the evaluators use our visualization, our approach would almost by definition have performed much better. Although the number of clusters does not necessarily affect the measure of cluster quality, we constrained our method further by requiring it to choose a clustering with approximately the same number of clusters as the preexisting hand-coded clustering.

**Press releases.** We begin with 200 press releases we randomly selected from those issued by Senator Frank Lautenberg’s Senate office and categorized by him and his staff in 24 categories (<http://lautenberg.senate.gov>). These include appropriations, economy, gun safety, education, tax, social security, veterans, etc. These represent a difficult test for our approach because the documents, the categorization scheme, and the individual classifications were all created by the same people at great time and expense.

The top line in Fig. 2 gives the results for the difference in our method’s cluster quality minus the cluster quality from Lautenberg’s hand-coded categories. The point estimate appears as a dot, with a thick line for the 80% confidence interval and a thin line for the 95% interval. The results, appearing to the right of the vertical dashed line that marks zero, indicate that the clustering our method identified had unambiguously higher quality than the author of the documents produced by hand. This provides evidence that the clusterings are organized in a way that allows for the efficient search over many millions of different (but similar) conceptualizations. (We give an example of the substantive importance of our selected clustering in Section 3.3.)

**State of the Union messages.** Our second example comes from an analysis of all 213 quasi-sentences in President George W. Bush’s 2002 State of the Union address, hand coded by the Policy Agendas Project (<http://www.policyagendas.org>). Each quasi-sentence



**Fig. 2.** Cluster quality experiments. Each line gives a point estimate (dot), 80% confidence interval (dark line), and 95% confidence interval (thin line) for a comparison between our automated cluster analysis method and clusters created by hand. Cluster quality is defined as the average similarity of pairs of documents from the same cluster minus the average similarity of pairs of documents from different clusters, as judged by human coders one pair at a time.

(defined in the original text by periods or semicolon separators) takes the role of a document in our discussion. The authors use 19 policy topic-related categories, including agriculture, banking & commerce, civil rights/liberties, defense, education, etc. Quasi-sentences are difficult tests because they are very short and may have meaning obscured by the context, which most automated methods ignore.

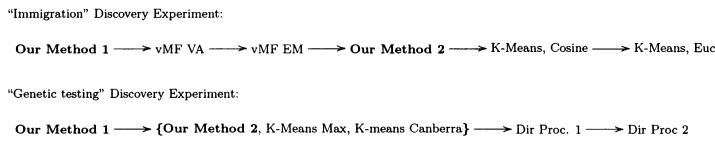
The results of our cluster quality evaluation appear as the second line in Fig. 2. Again, using our CAC methods we selected a clustering that turned out to have higher quality than the Policy Agendas Project coding scheme; this can be seen by the whole 95% confidence interval appearing to the right of the vertical dashed line. These results do not imply that anything is wrong with the Policy Agendas Project classification scheme, only that there seems to be more information in than the project’s chosen categories may indicate.

Substantively, our CAC approach led us to notice that the largest cluster of statements in Bush’s address were those that addressed the 9/11 tragedy, including many devoid of immediate policy implications, and so are lumped into a large “other” category by the project’s coding scheme, despite considerable political meaning. For example, “And many have discovered again that even in tragedy, especially in tragedy, God is near.” or “We want to be a Nation that serves goals larger than self.” This cluster thus conveys how the Bush administration’s response to 9/11 was sold rhetorically to resonate with his religious supporters and others, all with considerable policy and political content. For certain research purposes, this discovery may reflect highly valuable additional information.

**Reuters news stories.** For a final example of cluster quality, we use 250 documents randomly drawn from the Reuters-21578 news story categorization. This corpus has often been used as a gold standard baseline for evaluating clustering (and supervised learning classification) methods in the computer science literature (<http://www.daviddlewis.com/resources/testcollections/reuters21578/>). In this collection, each Reuters financial news story from 1987 has been classified by the Reuters news organization (with help from a consulting firm) into one of 22 categories, including trade, earnings, copper, gold, coffee, etc. We again apply the same evaluation methodology; the results, which appear as the bottom line in Fig. 2, indicate again that the clustering we identified turned out to have unambiguously higher cluster quality than Reuters’s own gold standard classification.

**3.2 Discovery Quality.** We show here that using our approach leads to more informative discoveries for researchers engaged in real scholarly projects. This is an unusually hard test for a statistical method and one rarely performed; it would be akin to requiring not merely that a standard statistical method possesses certain properties like being unbiased, but also, when given to researchers and used in practice, that they actually use it appropriately and estimate their quantities of interest correctly.

The question we ask is whether the computer assistance we provide helps. To perform this evaluation, we recruited two scholars in the process of evaluating large quantities of text in their own (independent) works in progress, intended for publication (one faculty member, one senior graduate student). In each case, we offered an analysis of their text in exchange for their participation in our experiment. One had a collection of documents about immigration in America in 2006; the other was studying a longer period about how genetic testing was covered in the media. Both had spent many months reading their documents. (To ensure the right of first publication goes to the authors, we do not describe the specific insights we found here and instead only report how they were judged in comparison to those produced by other methods.) Using a large collection of texts from each researcher, we spent about an hour using our method to



**Fig. 3.** Results of discovery experiments, where  $A \rightarrow B$  means that clustering  $A$  is judged to be "more informative" than  $B$  in a pairwise comparison, with braces grouping results in the second experiment tied due to an evaluator's cyclic preferences. In both experiments, a clustering from our method is judged to beat all others in pairwise comparisons.

identify two distinct clusterings from our space that we thought provided useful and distinct insights into the data. For comparison, we also applied the popular  $k$ -means clustering methodology (with variable distance metrics) and one of two more recently proposed clustering methodologies—the Dirichlet process prior and the mixture of von Mises Fisher distributions, estimated using a variational approximation (16). We used two different clusterings from each of the three cluster analysis methods applied in each case. For our method, we again biased the results against our method and this time chose the two clusterings ourselves instead of letting them use our visualization.

We then created an information packet on each of the six clusterings. This included the proportion of documents in each cluster, an exemplar document, and a brief automated summary of the substance of each cluster, using a technique that we developed. To create the summary, we first identified the 10 most informative words stems for each cluster, in each clustering (i.e., those with the highest "mutual information"). The summary then included the full length word most commonly associated with each chosen word stem. We found through much experimentation that words selected in this way usually provide an excellent summary of the topic of the documents in a cluster.

We then asked the researchers to familiarize themselves with the six clusterings. After about 30 min, we asked each to perform all  $\binom{6}{2} = 15$  pairwise comparisons, presented in random order, between the clusterings and in each case to judge which clustering within a pair they thought was "more informative." In the end, we want a cluster analysis methodology that produces at least one method that does well. Because the user ultimately will be able to judge and choose among results, having a method that does poorly is not material; the only issue is how good the best one is.

We are evaluating two clusterings from each cluster analysis method, and so we label them 1 and 2, although the numbers are not intended to convey order. Fig. 3 gives a summary of our results, with arrows indicating dominance in pairwise comparisons. In the first (immigration) example, illustrated at the top of the figure, the 15 pairwise comparisons formed a perfect Guttman scale (17) with "our method 1" being the Condorcet winner (i.e., it beat each of the five other clusterings in separate pairwise comparisons). (This was followed by the two mixtures of von Mises Fisher distribution clusterings, then "our method 2," and then the two  $k$ -means clusterings.) In the genetics example, our researcher's evaluation produced one cycle, and so it was close to but not a perfect Guttman scale; yet "our method 1" was again the Condorcet winner. (Ranked according to the number of pairwise wins, after "our method 1" was one of the  $k$ -means clusterings, then "our method 2," then other  $k$ -means clustering, and then the two Dirichlet process cluster analysis methods. The deviation from a Guttman scale occurred among the last three items.)

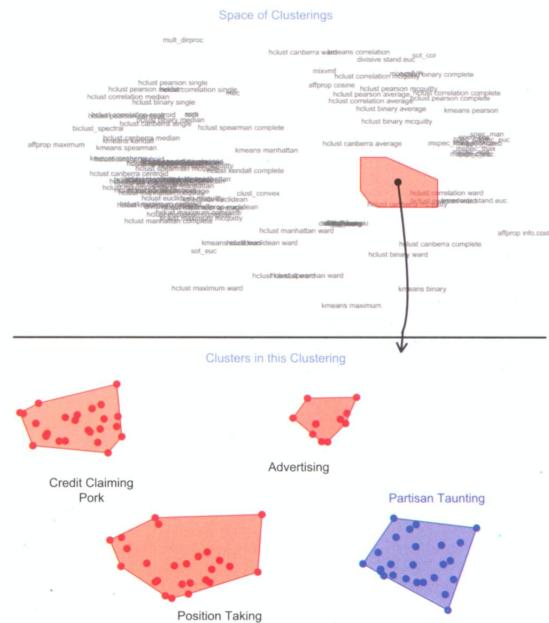
### 3.3 Partisan Taunting: An Illustration of Computer-Assisted Discovery

We now give a brief report of an example of the whole process of analysis and discovery using our approach applied to a real example. We develop a categorization scheme that advances one in the literature, measure the prevalence of each of its

categories in a new out-of-sample set of data to show that the category we discovered is common, develop a new hypothesis that occurred to us because of the new lens provided by our new categorization scheme, and then test it in a way that could be proven wrong. The degree of insight discovered can be judged by the reader.

In a famous and monumentally important passage in the study of American politics, (ref. 18, p. 49ff) Mayhew argues that "congressmen find it electorally useful to engage in...three basic kinds of activities"—credit claiming, advertising, and position taking. This typology has been widely used over the last 35 years, remains a staple in the classroom, and accounts for much of the core of several other subsequently developed categorization schemes (19–21). In the course of preparing our cluster analysis experiments in Section 3.1, we found much evidence for all three of Mayhew's categories in Senator Lautenberg's press releases, but we also made what we view as an interesting discovery.

We illustrate this discovery process in Fig. 4, where the top panel gives the space of clusterings we obtain when applying



**Fig. 4.** Discovering partisan taunting. The top portion of this figure presents the space of clustering solutions of Frank Lautenberg's (D-NY) press releases. Partisan taunting could be easily discovered in any of the clustering solutions in the red region in the top plot. The bottom plot presents the clusters from a representative clustering within the red region at the top (represented by the black dot). Three of the clusters (in red) align with Mayhew's categories, but we also found substantial partisan taunting cluster (in blue), with Lautenberg denigrating Republicans in order to claim credit, position-take, and advertise. Other points in the red polygon at the top represent different clusterings, but all clearly reveal the partisan taunting category.

**Table 1. Examples of partisan taunting in Senator Lautenberg's press releases**

Date	Lautenberg Category	Quote
2/19/2004	civil rights	"The Intolerance and discrimination from the Bush administration against gay and lesbian Americans is astounding."
2/24/2004	government oversight	"Senator Lautenberg Blasts Republicans as 'Chicken Hawks'"
8/12/2004	government oversight	"John Kerry had enough conviction to sign up for the military during wartime, unlike the Vice President [Dick Cheney], who had a deep conviction to avoid military service."
12/7/2004	homeland security	"Every day the House Republicans dragged this out was a day that made our communities less safe."
7/19/2006	health care	"The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then."

our methodology to Lautenberg's press releases (i.e., like Fig. 1). Recall that each name in the space of clusterings in the top panel corresponds to one clustering obtained by applying the named clustering method to the collection of press releases; any point in the space between labeled points defines a new clustering using our local cluster ensemble approach; and nearby points have clusterings that are more similar than those farther apart.

The clusters within the single clustering represented by the black point in the top panel is illustrated in the bottom panel, with individual clusters comprising Mayhew's categories of claiming credit, advertising, and position taking (all in red), as well as an activity that his typology obscures and he does not discuss. We call this new category *partisan taunting* (see blue region in Fig. 4) and describe it below. Each of the other points in the red region in the top panel represent clusterings that also clearly suggest partisan taunting as an important cluster although with somewhat different arrangements of the other clusters. That is, the user would only need to examine one point anywhere within this (red) region to have a good chance at discovering partisan taunting as a potentially interesting category.

Examples of partisan taunting appear in Table 1. Unlike any of Mayhew's categories, each of the colorful examples in the table explicitly reference the opposition party or one of its members, using exaggerated language to put them down or devalue their ideas. Most partisan taunting examples also overlap two or three of Mayhew's existing theoretical category definitions, which is good evidence of the need for this separate, and heretofore unrecognized, category. We did find that the documents were relatively easy to distinguish from Mayhew's existing categories.

Partisan taunting provides a new category of Congressional speech that emphasizes the interactions inherent between members of a legislature. Mayhew's (1974) original theory supposed that members of Congress were atomistic rational actors, concerned only with optimizing their own chance of reelection. Yet legislators interact with each other regularly, criticizing and supporting ideas, statements, and actions. This interaction is captured with partisan taunting but is absent from the original typology. In the *SI Appendix*, we detail how analyzing partisan taunting provides additional insights in addition to Mayhew's (1974) original typology.

Our technique has thus produced a new and potentially useful conceptualization for understanding Senator Lautenberg's 200 press releases. Although asking whether the categorization is "true" makes no sense, this modification to Mayhew's categorization scheme would seem to pass the tests for usefulness given in Section 3.1. We now show that it is also useful for out-of-sample descriptive purposes and separately for generating and rigorously testing other hypotheses suggested by this categorization.

We begin with a large out-of-sample test of the descriptive merit of the new category, for which we analyze all 64,033 press releases from all 301 senator-years during 2005–2007. To do this, we developed a coding scheme that includes partisan taunting, other types of taunting (to make sure our first category is well defined), and other types of press releases, including Mayhew's three categories. We then randomly selected 500 press releases

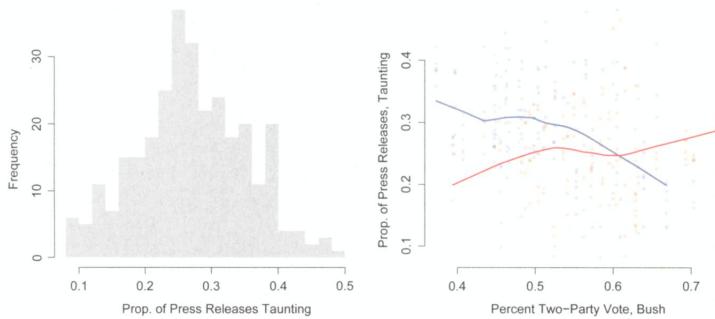
and had three research assistants assign each press release to a category (we had approximately 83% agreement and resolved disagreements by reading the press releases ourselves). Finally, we applied the supervised learning approach to text analysis given by ref. 22 to the entire set of 64,033 press releases to estimate the percent of press releases that were partisan taunts for each senator in each year. (By setting aside a portion of this training set, we verified that the Hopkins–King methodology produced highly accurate estimates in these data.)

Overall, we find that 27% of press releases among these 301 senator-years were partisan taunts, thus confirming that this category was not merely an idiosyncrasy of Senator Lautenberg. Instead partisan taunting seems to play a central role in the behavior of many senators. Indeed, it may even define part of what it means to be a member of the party in government. The histogram in the left panel of Fig. 5 gives the distribution of taunting behavior in our data; it conveys the large amount of taunting across numerous press releases, as well as a fairly large dispersion across senators and years in taunting behavior.\*

Finally, analyzing Senator Lautenberg's press releases led us to consider the role of taunting behavior in theories of democratic representation. Almost by definition, partisan taunting is antithetical to open deliberation and compromise for the public good (23). Thus, an important question is who taunts and when—which led us to the hypothesis that taunting would be less likely to occur in competitive senate seats. The idea is that taunting is most effective when a senator has the luxury of preaching to the choir and warning his or her partisans of the opposition (which has few votes); if instead a politician's electoral constituency is composed of large numbers of opposition party members, we would expect partisan taunting to be less effective and thus less used. If true, this result poses a crucial tension in democratic representation. Deliberation is seen as a normative good, but the degree to which a representative is a reflection of his or her constituency is also often seen to be an important component of democracy (24, 25). However, if our hypothesis is empirically correct, then democracies may have a zero sum choice between deliberation, which occurs more often in the absence of partisan taunting and thus in the most competitive states, and reflection, which by definition occurs in the least competitive states.

By using our large dataset of press releases, we construct an out-of-sample test of our hypothesis. The right panel of Fig. 5 gives the results. Each dot in this figure represents one senator-year, with red for Republicans and blue for Democrats. The horizontal axis is the proportion of the 2004 two-party vote for George W. Bush—a measure of the size of the underlying Republican coalition in each state, separate from all the idiosyncratic features of individual senatorial campaigns. We also portray the dominant patterns with a smoothed (LOESS) line for the Republicans (in red) and Democrats (in blue). The results overall clearly support the hypothesis: As states become more Republi-

\*The top 10 senator-year tauntings include Baucus (D-MT), 2005; Byrd (D-WV), 2007; Thune (R-SD), 2006; Ensign (R-NV), 2005; McConnell (R-KY), 2006; Biden (D-DE), 2005; Reid (D-NV), 2005; Coburn (R-OK), 2007; Sarbanes (D-MD), 2006; Kennedy (D-MA), 2007.



**Fig. 5.** Partisan taunting hypothesis verification. The left panel shows the distribution in partisan taunting in senators' press releases, and the right panel demonstrates that taunting is more likely when senators are in less competitive states. Each of the 301 points in the right panel represents the results of an analysis of one year's worth of a single senator's press releases, with blue for Democrats and red for Republicans.

can (moving from left to right), partisan taunting by Republicans increases, whereas partisan taunting by Democrats declines.

Of course, much more can be done with this particular empirical example, which is in fact the point: Our clustering methodology helped us choose a new categorization scheme to understand an aspect of the world in a new way, a new concept represented as a new category, a new hypothesis capable of being proven wrong, and a rigorous out-of-sample validation test for both describing and explaining the variation in the prevalence of this category among all senators.

#### 4 Concluding Remarks

We introduce in this paper a computer-assisted approach to unsupervised learning through cluster analysis. We also develop empirically based procedures for evaluating this and other cluster analytic methods and their resulting clusterings that use human judgment in a manner consistent with their cognitive strengths. Through a variety of examples, we demonstrate how this approach can relatively easily unearth new discoveries of useful information from large quantities of unstructured text.

Given the ongoing spectacular increase in the production and availability of unstructured text about subjects of interest to social scientists, and the impossibility of assimilating, summarizing, or even characterizing much of it by reading or hand coding, the most important consequence of this research may be its potential

for scholars to help efficiently unlock the secrets this information holds.

For methodologists and statisticians working on developing new methods of cluster analysis, this research also offers techniques for evaluating their efforts. Research that follows up on our strategy by creating new ways of encompassing existing methods might be designed to make the process easier, visualized in other ways, or computationally faster. Most of the research currently being done is focused on developing individual (i.e., nonencompassing) methods; we know that, by definition, any one individual method cannot outperform the approach proposed here, but new individual methods may be able to improve our approach if included in the cluster methods we encompass. For that purpose, we note that the most useful new individual methods would be those that fill empty areas in the space of clusterings, especially those outside the convex hull of existing methods in this space. Methods that produce clusterings for many datasets close to others would not be as valuable.

**ACKNOWLEDGMENTS.** For helpful advice, coding, comments, or data we thank John Ahlquist, Jennifer Bachner, Jon Bischof, Matt Blackwell, Heidi Brockman, Jack Buckley, Jacqueline Chattopadhyay, Patrick Egan, Adam Glynn, Emily Hickey, Chase Harrison, Dan Hopkins, Grace Kim, Elena Llaudet, Katie Levine, Elena Llaudet, Scott Moser, Jim Pitman, Matthew Platt, Ellie Powell, Maya Sen, Arthur Spirling, Brandon Stewart, and Miya Woolfolk.

1. Bailey KD (1994) *Typologies and Taxonomies: An Introduction to Classification Techniques* (Sage, Beverly Hills, CA).
2. Manning CD, Raghavan P, Schütze H (2008) *Introduction to Information Retrieval* (Cambridge Univ Press, New York).
3. Monroe Burt, Colaresi M, Quinn K (2008) Fighting words: Lexical feature selection and evaluation for identifying the content of political conflict. *Polit Anal* 16:372–403.
4. Shawe-Taylor J, Cristianini N (2004) *Kernel Methods for Pattern Analysis* (Cambridge Univ Press, Cambridge).
5. Jordan Michael, Ghahramani Z, Jaakkola T, Saul L (1999) An introduction to variational methods for graphical models. *J Mach Learn Res* 3:183–233.
6. Meila M (2007) Comparing clusterings: An information based distance. *J Multivariate Anal* 98:873–895.
7. Sammon J (1969) A nonlinear mapping for data structure analysis. *IEEE T Comput C-18:401–409.*
8. Strehl Alexander, Grosh J (2003) Cluster ensembles: A knowledge reuse framework for combining multiple partitions. *J Mach Learn Res* 3:583–617.
9. Fern X, Brodley C (2003) Random project for high dimensional data clustering: A cluster ensemble approach. *Proceedings of the Twentieth International Conference on Machine Learning* (International Machine Learning Society, Washington).
10. Law M, Topchy A, Jain A (2004) Multi-objective data clustering. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, Washington).
11. Caruana R, Elhawary M, Nguyen N, Smith C (2006) Meta clustering. *ICDM'06. Sixth International Conference on Data Mining* (SIAM, Bethesda, MD), pp 107–118.
12. Gionis A, Mannila H, Tsaparas P (2005) Clustering aggregation. *Proceedings of the 21st International Conference on Data Engineering* (IEEE Computer Society, Tokyo).
13. Topchy A, Jain AK, Punch W (2003) Combining multiple weak clusterings. *Proceedings IEEE International Conference on Data Mining* (IEEE Computer Society, Melbourne).
14. Diaconis P, Goel S, Holmes S (2008) Horseshoes in multidimensional scaling and local kernel methods. *Ann Appl Stat* 2:777–807.
15. Armstrong JS (1967) Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. *Am Stat* 21:17–21.
16. Blei D, Jordan M (2006) Variational inference for dirichlet process mixtures. *Bayesian Analysis* 1:121–144.
17. Guttman L (1950) The problem of attitude and opinion measurement. *Measurement and Prediction* 4:46–59.
18. Mayhew D (1974) *The Electoral Connection* (Yale Univ Press, New Haven, CT).
19. Fiorina M (1989) *Congress, Keystone of the Washington Establishment* (Yale Univ Press, New Haven, CT).
20. Eulau H, Karps P (1977) The puzzle of representation: Specifying components of responsiveness. *Legis Stud Quart* 2:233–254.
21. Yiannakis DE (1982) House members communication styles: Newsletters and press releases. *J Polit* 44:1049–1071.
22. Hopkins D, King G (2010) A method of automated nonparametric content analysis for social science. *Am J Polit Sci* 54:229–247 <http://gking.harvard.edu/files/abs/words-abs.shtml>.
23. Gutmann A, Thompson D (1996) *Democracy and Disagreement* (Harvard Univ Press, Cambridge, MA).
24. Miller WE, Stokes DE (1963) Constituency influence in Congress. *Am Polit Sci Rev* 57:45–56.
25. Pitkin HF (1972) *The Concept of Representation* (Univ of California Press, Berkeley, CA).

# Computer-Assisted Keyword and Document Set Discovery from Unstructured Text



**Gary King** Harvard University

**Patrick Lam** Thresher

**Margaret E. Roberts** University of California, San Diego

**Abstract:** The (unheralded) first step in many applications of automated text analysis involves selecting keywords to choose documents from a large text corpus for further study. Although all substantive results depend on this choice, researchers usually pick keywords in ad hoc ways that are far from optimal and usually biased. Most seem to think that keyword selection is easy, since they do Google searches every day, but we demonstrate that humans perform exceedingly poorly at this basic task. We offer a better approach, one that also can help with following conversations where participants rapidly innovate language to evade authorities, seek political advantage, or express creativity; generic web searching; eDiscovery; look-alike modeling; industry and intelligence analysis; and sentiment and topic analysis. We develop a computer-assisted (as opposed to fully automated or human-only) statistical approach that suggests keywords from available text without needing structured data as inputs. This framing poses the statistical problem in a new way, which leads to a widely applicable algorithm. Our specific approach is based on training classifiers, extracting information from (rather than correcting) their mistakes, and summarizing results with easy-to-understand Boolean search strings. We illustrate how the technique works with analyses of English texts about the Boston Marathon bombings, Chinese social media posts designed to evade censorship, and others.

**Replication Materials:** The data, code, and any additional materials required to replicate all analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse, at: <http://doi:10.7910/DVN/FMJDCD>.

Boolean keyword search of textual documents is a generic task used in numerous methods and application areas. Sometimes researchers seek one or a small number of the most relevant documents, a use case we call *fact finding* and for which Google, Bing, and other search engines were designed. For example, to find the capital of Montana, a weather forecast, or the latest news about the president, the user only wants one site (or a small number of sites) returned. In the second *collecting* use case, which we focus on, researchers do not try to find the needle in the haystack, at least at first; instead, they seek all documents that describe a particular literature, topic, person, sentiment, event, or concept.

Collecting is typically performed by attempting to think of all keywords that represent a specific concept,

and selecting documents that mention one or more of these keywords. Yet, this keywords selection process is known to be a “near-impossible task” for a human being (Hayes and Weinstein 1990), which we demonstrate can greatly bias inferences. Although no researchers should be selecting keywords for this purpose on their own, many applications require keywords. For example, applications of sophisticated methods of automated text analysis, designed to get around simplistic keyword matching and counting methods, are often preceded by selecting keywords to narrow all available documents to a manageable set for further analysis. Similarly, search engines are optimized for fact finding, but regularly used for collecting, even though they are suboptimal for this alternative purpose. Indeed, as we discuss in the third section

---

Gary King is Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge, MA 02138 ([King@Harvard.edu](mailto:King@Harvard.edu)). Patrick Lam is Lead Data Scientist at Thresher and Visiting Fellow at the Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge, MA 02138 ([patrick@thresher.io](mailto:patrick@thresher.io)). Margaret E. Roberts is Assistant Professor, Department of Political Science, University of California, San Diego, Social Sciences Building 301, 9500 Gilman Drive, #0521, La Jolla, CA 92093-0521 ([meroberts@ucsd.edu](mailto:meroberts@ucsd.edu)).

Our thanks go to Dan Gilbert, Burt Monroe, Brandon Stewart, Dustin Tingley, and the participants at the Society for Political Methodology conference for helpful suggestions. Data and replication information is available at King, Lam and Roberts (2016).

*American Journal of Political Science*, Vol. 61, No. 4, October 2017, Pp. 971–988

©2017, Midwest Political Science Association

DOI: 10.1111/ajps.12291

971

(“The Unreliability of Human Keyword Selection”), human brains have well-studied inhibitory processes that, although adaptive for other reasons, explicitly prevent us from recalling many keywords when needed for the task of collecting.<sup>1</sup>

The problem of keyword discovery is easier when structured data are available to supplement the raw text, such as search query logs (e.g., Google’s AdWords Keyword Tool, or Overture’s Keyword Selection Tool), databases of meta-tags, or web logs (Chen, Xue, and Yu 2008), and a large literature of methods of “keyword expansion or suggestion” has arisen to exploit such information. In this article, we develop methods for the wide array of problems for which raw text is the sole, or most important, source of information. To avoid requiring a human user having to think of all relevant keywords, we introduce methods of computer-assisted keyword discovery. Our key motivating principle is that although humans perform very poorly in the task of *recalling* large numbers of words from memory, they excel at *recognizing* whether any given word is an appropriate representation of a given concept.

We begin by describing some of the application areas to which our methodology may provide some assistance. We then conduct an experiment that illustrates the remarkable unreliability of human users in selecting appropriate keywords. Next, we define the statistical problem we seek to solve, along with our notation. We then present our algorithm, several ways of evaluating it, and an illustration of how it works in practice. Lastly, we discuss related prior literature and conclude. The appendices give details on algorithm robustness and how to build queries for much larger data sets. Replication information is available at King, Lam and Roberts (2016).

## Application Areas

Algorithms that meet the requirements of the statistical problem as framed in the fourth section suggest many new areas of application. We list some here, all of which the algorithm we introduce below may help advance. Some of these areas overlap to a degree, but we present them separately to highlight the different areas from which the use of this algorithm may arise.

<sup>1</sup>Some algorithms have been proposed and implemented on search engines to provide assistance for collecting, but the approaches are based on methods of fully automated cluster analysis that perform poorly on most general problems (Grimmer and King 2011).

## Conversational Drift

Political scientists, lobby groups, newspapers, interested citizens, and others often follow social media discussions on a chosen topic but risk losing the thread of the conversation, and the bulk of the discussion, when changes occur in how others refer to the topic. Some of these wording changes are playful or creative flourishes; others represent political moves to influence the debate or frame the issues. For example, what was once called “gay marriage” is now frequently referred to by supporters as “marriage equality.” Progressive groups try to change the discussion of abortion policy from “pro-choice” and “pro-life,” where the division is approximately balanced, to “reproductive rights,” where they have a large majority. Conservatives try to influence the debate by relabeling “late-term abortion” as “partial-birth abortion,” which is much less popular. As these examples show, selecting an incomplete set of keywords can result in severe selection bias because of their correlation with the opinions of interest.

## Evading the Censors

Internet censorship exists in almost all countries to some degree. Governments and social media firms that operate within their jurisdictions use techniques, such as keyword-based blocking, content filtering, and search filtering, to monitor and selectively prune certain types of online content (Yang 2009). Even in developed countries, commercial firms routinely “moderate” product review forums, and governments require the removal of “illegal” material such as child pornography. In response to these information controls, netizens continually try to evade censorship with alternative phrasings. For example, immediately after the Chinese government arrested artist-dissident Ai Weiwei, many social media websites began censoring the Chinese word for Ai Weiwei (King, Pan, and Roberts 2013); soon after, netizens responded by referring to the same person as “AWW” and the Chinese word for “love,” which in Chinese sounds like the “ai” in “Ai Weiwei.” Other creative censorship avoidance techniques involve using homographs and homophones.

## Starting Point for Statistical Analyses of Text

Most methods of automated text analysis assume the existence of a set of documents in a well-defined corpus in order to begin their analysis. They then spend most of their effort on applying sophisticated statistical, machine learning, linguistic, or data-analytic methods to this given corpus. In practice, this corpus is defined in one of a variety

of ways, but keyword searching is a common approach (e.g., Eshbaugh-Soha 2010; Gentzkow and Shapiro 2010; Ho and Quinn 2008; Hopkins and King 2010; King, Pan, and Roberts 2013; Puglisi and Snyder 2011). In this common situation, our algorithm should help improve the inputs to, and thus the results from, any one of these sophisticated approaches. The same issue applies for simple analysis methods, such as keyword counting.

### Intuitive and Infinitely Improvable Classification

Because statistical classifiers are typically far from perfect (Hand 2006), ordinary users who find individual documents misclassified may question the veracity of the whole approach. Moreover, since most classifiers optimize a global function of the data set, even sophisticated users may find of value hybrid approaches for adding human effort and knowledge to improve classification at the level of smaller numbers of documents. In this situation, keyword-based classifiers are sometimes more useful because the reasons for mistakes, even if there are more of them, are readily understandable and easily fixable (by adding or removing keywords from the selection list) for a human user (Letham et al. 2015). Keyword classifiers are also much faster than statistical classifiers and can be improved to any higher level of accuracy, with sufficient effort, by continual refinement of the Boolean query.

### Online Advertising

Academics recruiting study participants often bid for ad space next to searches for chosen keywords (Antoun et al. 2015), just as firms do in advertising campaigns. This is common with Google Adwords, Bing Ads, Facebook, and so on. These systems, and other existing approaches, suggest new keywords to those spending advertising dollars by mining information from structured data such as web searches, weblogs from specific websites, or other ad purchases. Our approach can supplement these existing approaches by mining keywords relevant to the population of interest from raw unstructured text found in research documents, literature reviews, or information in private companies such as customer call logs, product reviews, websites, or a diverse array of other sources. Whereas keywords (or more general Boolean searches) for advertising on search engines can be mined from search engine query logs, or website logs, keywords that identify rarely visited pages, or for advertising on social media sites, can only be mined from the unstructured text.

### Long Tail Search

Modern search engines work best when prior searches and the resulting structured metadata on user behavior (e.g., clicking on one of the websites offered or not) are available to continuously improve search results. However, in some areas, such metadata are inadequate or unavailable, and keywords must be discovered from the text alone. These include (1) traditional search with unique or unusual search terms (the “long tail”); (2) searching on social media, where most searches are for posts that just appeared or are just about to appear, and so have few previous visits; and (3) enterprise search for (confidential or proprietary) documents that have rarely if ever been searched for before. In these situations, it may be useful to switch from the present fully automated searching to computer-assisted searching using our technology.

Consider social search. During the Boston Marathon bombings, many followed the conversation on Twitter by searching for *#BostonBombings*, but at some point the social media Boston authors expressed community spirit by switching to *#BostonStrong* and out-of-towners used *#PrayForBoston*. Since guessing these new keywords is nearly impossible, those who did not notice the switch lost the thread of the conversation.

### The Unreliability of Human Keyword Selection

Human beings, unaided by computers, seem to have no problem coming up with *some* keywords to enter into search engines (even if not the optimal ones). Everyone is accustomed to doing Google searches, after all. However, as we demonstrate in this section, for the more complicated task of choosing a *set* of keywords for the task of collection, even expert human users perform extremely poorly and are highly *unreliable* at this task. That is, two human users familiar with the subject area, given the same task, usually select keyword lists that overlap very little, and the list from each is a very small subset of those they would each recognize as useful after the fact. The unreliability is exacerbated by the fact that users may not even be aware of many of the keywords that could be used to select a set of documents. And attempting to find keywords by reading large numbers of documents is likely to be logically infeasible in a reasonable amount of time.

Here, we first demonstrate this surprising result with a simple experiment. Second, because this result is counterintuitive *ex ante*, we briefly summarize the

well-developed psychological literature that can be used to explain results like this. And finally, we show the severe statistical bias (or extra *ex ante* variance) that can result from selecting documents with inadequate keyword lists.

## Experiment

For our experiment, we asked 43 relatively sophisticated individuals (mostly undergraduate political science majors at a highly selective college) to recall keywords with this prompt:

We have 10,000 twitter posts, each containing the word “healthcare,” from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obamacare.

We also gave our subjects access to a sample of the posts and asked them not to consult other sources. We repeated the experiment with an example about the Boston Marathon bombings.

The median number of words selected by our respondents was 8 for the Obamacare example and 7 for the experiment about the Boston Marathon bombings. In Figure 1, we summarize our results with word clouds of the specific keywords selected. Keywords selected by one respondent and not by anyone else are colored red (or gray if reading black and white). The position of any one word within the cloud is arbitrary.

The results clearly demonstrate the remarkably high level of unreliability of our human keyword selectors. In the Obamacare example, 149 unique words were recalled by at least one of our 43 respondents. Yet, for 66% of those words, every single one of the remaining 42 respondents, when given the chance, failed to recall the same word (Figure 1, red or gray words in the left panel). In the Boston Marathon bombing example, the percentage of words recalled by a single respondent was 59% (right panel). The level of unreliability was so high that no two users recalled the same entire keyword list.

This extreme level of unreliability is not due to our research subjects’ being unaware of some of the words. Indeed, after the fact, it is easy to see from Figure 1 that almost all the words recalled are recognizably related to Obamacare or the Boston bombings, respectively. In other words, although humans perform extremely poorly at recall, they are excellent at remembering.

## Psychological Foundation

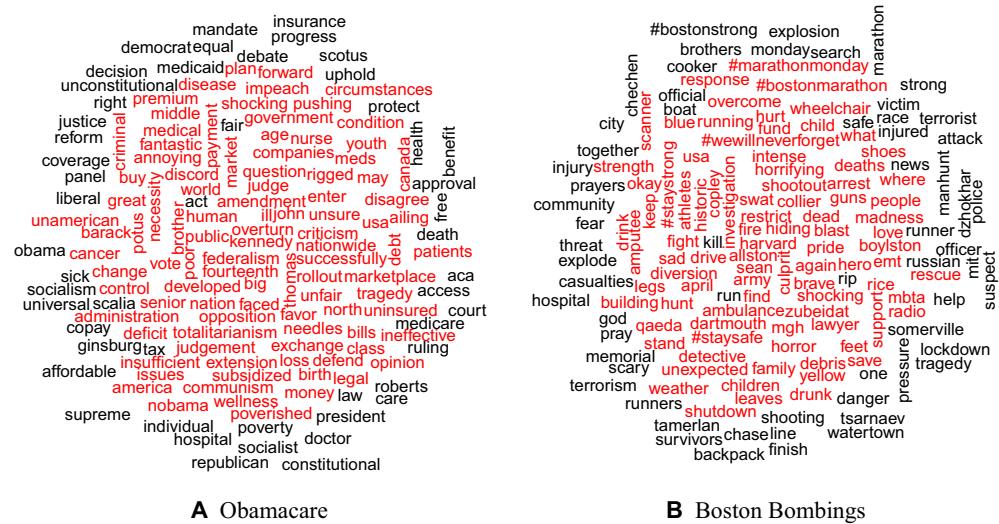
The counterintuitive result from our experiment is related to, and can be explained by, psychological research on “inhibitory processes” (and in particular, “part-list cuing”). The well-supported finding, from many experiments, is that revealing one word to the research subject facilitates remembering others, but the cue provided by revealing more than a few words strongly *inhibits* recall of the rest of the set, even though you would recognize them if revealed (Bauml 2008; Roediger and Neely 1982).

Why our brains would be constructed to stop us from remembering needed information deserves at least some speculation. One way to think about this is imagining memory as a network diagram with concepts represented as nodes, and connections between concepts represented as edges. Without inhibitory processes, activating any one concept by recall would activate all concepts connected to it, and all those connected to those, and so on (e.g., orange activates apple, apple activates banana, banana activates slip, slip activates...). Millions of concepts would come flowing into your comparatively tiny, short-term working memory and, unable to handle it all, you would likely be overwhelmed and perhaps unable to think at all. So either working memory would need to be much bigger, which does not seem to be on offer, or inhibitory processes are necessary.<sup>2</sup>

## Consequences for Statistical Bias

As is well known, the choice of a data selection rule, such as that defined by the choice of keywords, is only guaranteed to avoid bias if it is independent of the variables used to analyze the chosen document set. Obviously, this is a strong assumption, unlikely to hold in many applications, especially when using unreliable (i.e., human-only) methods of keyword selection. In other words, different keyword lists generate different document sets, which, in turn, can lead to dramatically different inferences, substantive conclusions, and biases.

<sup>2</sup>We can make this strange result somewhat more plausible by turning on an inhibitory process in your brain right now: Think of your bank password. Now think of your previous bank password. Assuming you listen to your bank and do not rotate them, now think of your bank password before that. Likely you cannot remember that one, but if someone showed it to you, we think you would agree that it would be easy for you to recognize it as correct. If so, then we have shown that the memory of that third password exists in your brain, even though something is causing you to not be able to access it. An example of inhibitory processes at work may even be the feeling that a thought you are having trouble remembering is “on the tip of your tongue”: It is stored in your brain, but you cannot access it.

**FIGURE 1** The Unreliability of Human Keyword Selection

Note: Word clouds of keywords were selected by human users; those selected by one and only one respondent are in red (or gray if printed in black and white). The position of each word within the cloud is arbitrary.

We now demonstrate these biases in an analysis of the data from our Boston Marathon bombings experiment. We study the well-known tendency for communities suffering a tragedy to turn public discourse from the obvious negative events into positive expressions based on solidarity, community spirit, and individual heroics. To do this, we use a simple, but still very common, analysis measure (Nielsen 2011). The idea is to code each word in a social media post as having negative ( $-1$ ), neutral ( $0$ ), or positive ( $+1$ ) sentiment (based on a fixed dictionary designed for Twitter) and to sum all the words in a post to give the final sentiment for that tweet. We use this method to compute the average sentiment of all tweets retrieved by each of the 43 keyword lists from our 43 subjects. The point estimates (dots) along with 95% confidence intervals (horizontal lines) for each appear in Figure 2, sorted from negative to positive sentiment.

The results vividly demonstrate the substantial effect the choice of a keyword list has on the sentiment of the document sets chosen by different research subjects given the identical prompt. Choosing some of the lists (on the bottom left) would lead a researcher to the conclusion that social media discourse was extremely negative during the month following the Boston Marathon bombing. If, instead, one were to choose other keyword sets (which appear in the middle of the graph), a researcher could report “evidence” that sentiment was only slightly negative. Alternatively, a researcher who used one of the keyword

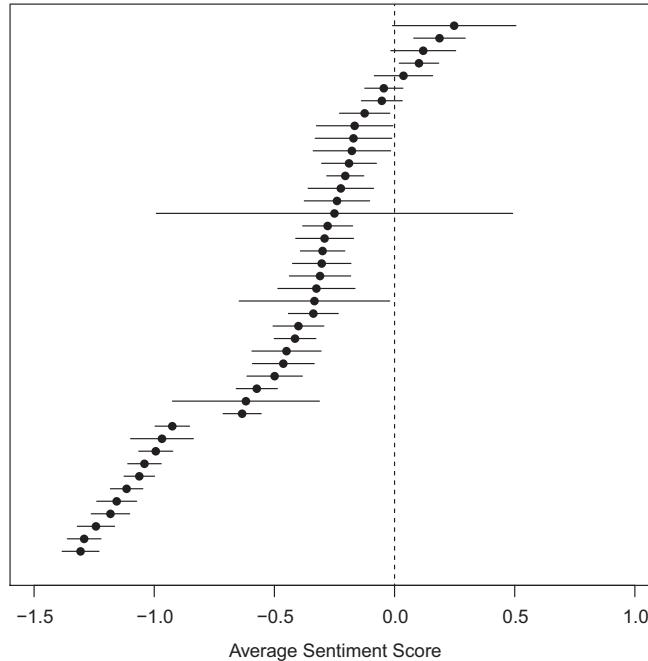
lists from the top right would be led to the conclusion that sentiment was relatively positive (by selecting documents that reflected expressions of community spirit). As is evident, almost *any* substantive conclusion can be drawn from these data by changing choice of the keyword list. This example clearly demonstrates the value of paying far more attention to how keyword lists are selected than has been the case in the literature.

## Defining the Statistical Problem

### Notation

We define the *reference set*,  $R$ , to be a set of textual documents, all of which are examples of a single chosen concept of interest (e.g., topic, sentiment, idea, person, organization, event). This set is defined narrowly so that the probability of documents being included that do not represent this concept is negligible. The reference set need not be a random or representative sample of all documents about the concept of interest (if such a process could even be defined), and may even reflect a subset of emphases or aspects of the concept (as was common for individual humans in the previous section).

Also define the *search set*,  $S$ , as a set of documents selected because it likely has additional documents of interest, as well as many others not of interest. The search set does not overlap the reference set,  $R \cap S = \emptyset$ . Our

**FIGURE 2 Average Sentiment of 43 Document Sets**

*Note:* Each document set was selected by a different keyword list, with point estimates (as dots) and 95% confidence intervals (horizontal lines) shown.

goal is to identify a *target set*,  $T$ , which is the subset of the search set ( $T \subset S$ ) containing documents with new examples of the concept defining documents in the reference set. Ultimately, we are interested in  $T \cup R$ , but, since we have  $R$ , the statistical task is to find  $T$  in  $S$ .

In practice, the reference set may be defined by choosing individual documents by hand, selecting an existing corpus, or using all available documents that contain text matching a specific Boolean query,  $Q_R$  (defined as a string containing user-defined keywords and Boolean operators, AND, OR, NOT, such that  $R = \{d : Q_R\}$ , for any document  $d$  under consideration). The search set can be defined as all websites on the Internet (after removing documents in  $R$ ), all available documents, a different selected existing corpus, or documents that match a Boolean query,  $Q_S$  (such that  $S = \{d : Q_S\}$ ). The elements of a Boolean query are “keywords.”<sup>3</sup>

<sup>3</sup>The simplest versions of keywords are unigrams, but they could also include higher-order  $n$ -grams, phrases, or any type of Boolean query. Common steps in automated text analysis, such as making all letters lowercase or stemming, can broaden the words that a single keyword will match (e.g., “consist” would then match “consist” as well as “Consist,” “consistency,” “Consisted,” “CONSIST-ING,” etc.). Other standard text-analytic preprocessing steps would

### An Unsupervised Statistical Problem

The statistical task of finding  $T$  is “unsupervised” in that the concept defining the reference and target sets may be broadened by the human user on the fly as part of the process of discovery (rather than, as in “supervised” analyses,  $T$  being a fixed quantity to be estimated). We thus seek to identify the target set  $T$  by first finding  $K_T$ , the set of all keywords in  $T$  ranked by likely relationships with the concept. We then use human input in specific ways to craft query  $Q_T$ , intended to retrieve  $T$  from  $S$ . Depending on the application, users may also be interested in the set of all keywords in the reference set  $K_R$ , the target and reference sets together  $T \cup R$ , a query that returns both the reference and target sets together  $Q_{RT}$ , or all of the above.

Our algorithm is human-led and computer-assisted rather than fully automated; it is related to semi-supervised learning (Zhu and Goldberg 2009). The more common fully automated approaches to document retrieval (e.g., spam filters) use statistical or machine learning classifiers that are viewed as a black box to the user.

remove words from the possible list of keywords, such as by removing stopwords or other very common words or very short words.

By restricting ourselves to a simple Boolean search, defined by a set of interpretable keywords, we empower users to control, understand, and continually improve the retrieval process.

Another reason for the choice of a human-powered approach is that the concept that the documents in the reference set share, and for which we seek a target set, is not a well-defined mathematical entity. Human language and conceptual definitions are rarely so unambiguous. For example, any two nonidentical documents could be regarded as the same (they are both documents), completely unrelated (since whatever difference they have may be crucial), or anything in between. Only additional information about the context (available to the person but not available solely from the data set) can informatively resolve this indeterminacy. To take a simple example, suppose one element of  $K_R$  is the keyword “sandy.” Should the target set include documents related to a hurricane that devastated New Jersey, a congresswoman from Florida, a congressman from Michigan, a cookie made with chopped pecans, a type of beach, a hair color, a five-letter word, or something else? To make matters worse, it could easily be the case that documents in the reference set represent two of seven of these examples, but two others in the search set are of interest to the human user. Of course, a user can always define the reference set more precisely to avoid this problem, but the nature of language means that some ambiguity will always remain. Thus, we use human input, with information from the text presented to the human user in a manner that is easily and quickly understood, to break this indeterminacy and grow the reference set in the desired direction.

## Algorithm

The algorithm first partitions  $S$  into two groups by classifying whether a document belongs in set  $T$  or its complement,  $S \setminus T$ . It mines  $S$  for all keywords  $K_S$  and then ranks keywords by how well they discriminate between  $T$  and  $S \setminus T$ . This results in two lists of keywords  $K_T$  and  $K_{S \setminus T}$  ranked in order of how well they discriminate each set from the other. The keyword lists themselves are often of interest to users who would like keyword recommendations for various uses. For document retrieval, the user would iterate through the two lists to produce a query  $Q_T$  that, when combined with the reference query  $Q_R$  to form  $Q_{RT}$ , best retrieves his or her desired document set of interest.

Table 1 gives a brief overview of the specific steps in our proposed algorithm.

**TABLE 1 The Keyword Algorithm**

- 
1. Define a reference set  $R$  and search set  $S$ .
  2. Using a diverse set of classifiers, partition all documents in  $S$  into two groups:  $T$  and  $S \setminus T$ , as follows:
    - (a) Define a training set by drawing a random sample from  $R$  and  $S$ .
    - (b) Fit one or more classifiers to the training set using as the outcome whether each document is in  $R$  or  $S$ .
    - (c) Use parameters from classifiers fit to the training set to estimate the predicted probability of  $R$  membership for each document in  $S$ . (Of course, every document is in  $S$ , and so the prediction mistakes can be highly informative.)
    - (d) Aggregate predicted probabilities or classifications into a single score (indicating probability of membership in  $T$ ) for each document in  $S$ .
    - (e) Partition  $S$  into  $T$  and  $S \setminus T$  based on the score for each document and a user-chosen threshold.
  3. Find keywords that best classify documents into either  $T$  or  $S \setminus T$ , as follows:
    - (a) Generate a set of potential keywords by mining  $S$  for all words that occur above a chosen frequency threshold,  $K_S$ .
    - (b) Decide whether each keyword  $k \in K_S$  characterizes  $T$  or  $S \setminus T$  better, by comparing the proportion of documents containing  $k$  in  $T$  with the proportion of documents containing  $k$  in  $S \setminus T$ .
    - (c) Rank keywords characterizing  $T$  by a statistical likelihood score that measures how well the keyword discriminates  $T$  from  $S \setminus T$ . Do the analogous ranking for keywords characterizing  $S \setminus T$ .
  4. Present keywords in two lists to the user, to iterate and choose words of interest or for use in building a document retrieval query.
  5. If sufficient computational power is available, rerun Steps 1–4 every time the user makes a measurable decision, such as adding a keyword to  $Q_T$  to improve the lists of keywords to consider.
- 

*Note:* The table displays a simple version of our algorithm, used in illustrations below. The algorithm also has numerous possible extensions, such as generating phrases or higher-order  $n$ -grams, clustering the documents in various different ways, redefining the reference set after the user chooses a keyword, and iterating between user input and the algorithm.

## Incrementally Defining $R$ and $S$

The simplest application of our algorithm has  $R$  and  $S$  defined at the outset, but alternatives are often easier in practice. For example, one may begin with a large document set and without any immediately obvious distinction between the two sets. This situation is common with large, continuously streaming, or even ill-defined data, such as being based on the entire Internet, all social media posts, or all documents narrowed by a set of very broad keywords. In this situation, we can define  $S$  and  $R$  adaptively, as part of the algorithm (e.g., D’Orazio et al. 2014).

Consider the following alternative adaptive strategy. The user begins by defining  $R$  narrowly based on one simple keyword search, as a subset of the existing corpus. We then add an intermediate step to the algorithm, which involves mining and displaying a list of keywords found in  $R$ ,  $K_R$ , ranked by a simple statistic such as document frequency or term frequency-inverse document frequency. The user then examines elements of  $K_R$  (aside from those used to define the set) and chooses some keywords to define  $Q_S$ , which in turn generates a definition for  $S$ , so that we can run the rest of the algorithm. The user can then continue to add keywords from  $K_R$  into the final desired query  $Q_{RT}$ . In this workflow,  $S$  can be neither predefined nor retrieved ex ante. This step also mitigates the issue of how to define a search set in large data sets that do not fit into memory all at once or may not even be able to be retrieved all at once. It also leverages additional information from  $R$  in the form of keywords likely to identify additional aspects of the concept and keywords the user may not have thought of for defining both  $R$  and  $S$ .

## Partitioning $S$ into $T$ and $S \setminus T$

To partition  $S$  into  $T$  and  $S \setminus T$ , we first we define a “training” set by sampling from  $S$  and  $R$ . We can repeat this step with different random subsettings to increase the diversity of keyword candidates that are surfaced. (Exemplars can substitute for random sampling as well.) Since  $R$  is typically much smaller than  $S$  and our test set for our classifiers is all of  $S$ , we often use the entire  $R$  set and a sample of  $S$  as our training set.

Next, we fit classifiers to the training set, using each document’s actual membership in  $R$  or  $S$  as the outcome variable. As predictors, we use any element of the text of the documents, as well as any available metadata. Any set of statistical, machine learning, or data-analytic classifiers can be used, but we recommend using as large and diverse a set of methods as is convenient and computationally feasible (e.g., Bishop 1995; Hastie, Tibshirani, and Friedman

TABLE 2 Classification Sets

		Classified	
		Search	Reference
Truth	Search	{S S}	{R S}
	Reference	{S R}	{R R}

Note: Classification sets are shown, where  $\{a|b\}$  is the set of documents in set  $b$  classified into set  $a$ ;  $S$  is the search set, and  $R$  is the reference set.

2009; Kulkarni, Lugosi, and Venkatesh 1998; Schapire and Freund 2012).

After fitting the classifiers, we use the estimated parameters to generate predicted probabilities of  $R$  membership for all documents in  $S$ . Of course, all the search set documents in fact fall within  $S$ , but our interest is in learning from the *mistakes* these classifiers make.

Although we do not need to transform the probabilities into discrete classification decisions for subsequent steps in the algorithm, we provide intuition into these mistakes by doing this now. Table 2 portrays the results for one example classifier, with the originally defined truth in rows and potential classifier decisions in columns. We will typically be interested in documents from the search set, (mis)classified into the reference set,  $\{R|S\}$ . The idea is to exploit these mistakes since documents in this set will reveal similarities to the reference set, and so they likely contain new keywords we can harvest to better represent the concept of interest.<sup>4</sup>

Once we have predicted probabilities of  $R$  membership for each document in  $S$  from the classifiers, we need to turn these into a single  $T$  membership “score” for the purpose of grouping documents. For a single classifier, the predicted probability of  $R$  membership from  $S$  is the predicted probability of  $T$  membership. In most situations, we recommend the use of multiple classifiers, so that we can extract their different “opinions” about in which set individual documents belong. The different classifiers will typically pick up on different aspects of the concept and thus highlight different keywords for the user to choose from. To ensure that this diversity of opinion is reflected in our keyword lists, we aggregate the probabilities across classifiers for a single document by taking the *maximum*

<sup>4</sup>Other groups defined by the classifier in Table 2 may also be useful. For example, the documents  $\{S|S\}$  contain keywords in the search set, classified into the search set, and so could be useful for identifying keywords to avoid when defining a topic of interest; in a Boolean query, these could be used with NOT. Similarly, the documents  $\{R|R\}$  can reveal keywords that select documents in the reference group. These can be used to refine the definition of the reference or search data sets. We also use these documents for model checking and for tuning in our classifiers.

probability across the classifiers as the membership score (i.e., rather than the usual approach of using the average or plurality vote). We then use this score to group documents into  $T$  and  $S \setminus T$ . Our simple aggregation rule thus boils down to placing all documents with at least one classifier “vote.”

## Discovering Keywords to Classify Documents

After partitioning  $S$  into our estimated target set  $T$  and nontarget set  $S \setminus T$ , we must find and rank keywords that best discriminate  $T$  and  $S \setminus T$ . We do this in three steps: (a) mine all keywords from  $S$  (perhaps limiting our list to those that meet thresholds such as a minimum document frequency of five documents), (b) sort them into those that predict each of the two sets, and (c) rank them by degree of discriminatory power.

Step (a) is accomplished by merely identifying all unique keywords in  $S$ . This is a simple step for our computer algorithm, but it is important in practice since a human who thinks of a word not in any documents in  $S$  will be useless, no matter how compelling the word seems to be.

For Step (b), we use the proportion of documents in which each keyword appears at least once. For example, if a keyword appears in 5 out of 10  $T$  documents and 15 out of 50  $S \setminus T$  documents, we put that keyword into the  $T$  list since it appears in 50% of  $T$  documents and 30% of  $S \setminus T$  documents, despite the fact that it appears in 10 more  $S \setminus T$  documents on an absolute scale. Keywords that appear in both sets with equal document proportions can be placed in either list or both lists.

In Step (c), we rank the keywords within lists, according to how well they discriminate the two sets. Although different scoring metrics could be used to accomplish this task, we find that a metric based on the following likelihood approach is quite effective (see Letham et al. 2013). For document  $d \in S$  at any point in using the algorithm, let  $y_d$  equal 1 if  $d \in T$  and 0 if  $d \in S \setminus T$ . For each keyword  $k$  in either list, denote  $n_{k,T}$  and  $n_{-k,T}$  as the number of documents in  $T$  that do and do not match  $k$ , respectively, and  $n_{k,S \setminus T}$  and  $n_{-k,S \setminus T}$  as the number of documents in set  $S \setminus T$  that do and do not match  $k$ , respectively. Also define the marginal totals so that  $n_{k,S} = n_{k,T} + n_{k,S \setminus T}$  and  $n_{-k,S} = n_{-k,T} + n_{-k,S \setminus T}$  denote the total number of documents in  $S$  that do and do not contain  $k$ , respectively, and  $N_T = n_{k,T} + n_{-k,T}$  and  $N_{S \setminus T} = n_{k,S \setminus T} + n_{-k,S \setminus T}$  denote the number of documents in  $T$  and  $S \setminus T$ , respectively.

This then leads to a convenient likelihood function for the model we use to distinguish  $T$  from  $S \setminus T$ :

$$\begin{aligned} p(y_1, \dots, y_n | \theta_k, \theta_{-k}, k) &= \text{Bin}(n_{k,T}, n_{k,S \setminus T} | n_{k,S}, \theta_k) \\ &\times \text{Bin}(n_{-k,T}, n_{-k,S \setminus T} | n_{-k,S}, \theta_{-k}), \end{aligned}$$

where  $\theta_k$  and  $\theta_{-k}$  are probability parameters with priors

$$\theta_k \sim \text{Beta}(\alpha_T, \alpha_{S \setminus T})$$

$$\theta_{-k} \sim \text{Beta}(\alpha_T, \alpha_{S \setminus T})$$

with  $\alpha_T = \alpha_{S \setminus T} = 1$  in our implementation. We want to then rank the keywords by how best they “fit” the actual distribution of documents into  $T$  and  $S \setminus T$  by calculating their scores from the likelihood function. Since the probability parameters  $\theta_k$  and  $\theta_{-k}$  are not of interest, we marginalize over them to get

$$\begin{aligned} p(y_1, \dots, y_n | \alpha_T, \alpha_{S \setminus T}, k) &\propto \\ &\frac{\Gamma(n_{k,T} + \alpha_T)\Gamma(n_{k,S \setminus T} + \alpha_{S \setminus T})}{\Gamma(n_{k,T} + n_{k,S \setminus T} + \alpha_T + \alpha_{S \setminus T})} \\ &\times \frac{\Gamma(N_T - n_{k,T} + \alpha_T)\Gamma(N_{S \setminus T} - n_{k,S \setminus T} + \alpha_{S \setminus T})}{\Gamma(N_T - n_{k,T} + N_{S \setminus T} - n_{k,S \setminus T} + \alpha_T + \alpha_{S \setminus T})}. \end{aligned}$$

We then calculate the value of the likelihood function for each keyword in each list and rank them all from highest to lowest likelihood.

## Human Input and Human-Computer Iteration

Our final step, prior to iterating, involves using human input to choose items from the two keyword lists and to build queries  $Q_T$  and  $Q_{RT}$ . Following the third section, we optimize so humans do what they are good at and computerize what they are not. We present all the keywords, so the humans do not need to recall anything, along with computerized rankings to organize best guesses about what may be of interest to them. Then the humans can use their detailed contextual knowledge, unavailable to our algorithm, to find different eddies of conversation and meanings of concepts of interest not previously recalled. This process of evaluating a list of words is of course considerably faster and much more reliable than asking humans to pull keywords out of thin air or thinner memories.

The algorithm is unsupervised so that human users can easily refine, improve, or totally redefine the concept of interest, as the keyword lists inspire them to think of new perspectives on the same material. Users may also discover new directions that cause them to begin again with a completely new reference set, or to add to the existing reference set or reference query  $Q_R$ .

At this point, the user can iterate with the algorithm in various ways to continue to adjust the partition of  $S$  into  $T$  and  $S \setminus T$  and to refine or redefine the concepts of interest. One way to iterate can be to simply update the reference query with the new selected words and rerun the algorithm. Another is for the user to designate specific keywords or documents of interest or not of interest, which gives the algorithm more information to update the definitions of  $T$  and  $S \setminus T$ .

## Evaluations

For our evaluations, we require a ground truth and a data set with documents properly coded to the concept of interest. Of course, the version of keyword selection we are studying is an unsupervised task, and so the concept initially chosen in real applications is not necessarily well defined, may differ from user to user or application to application, and can be refined or changed altogether while using the algorithm; indeed, the ability of the user to make these changes is an important strength of the algorithm in practice.

Thus, to make ourselves vulnerable to being proven wrong, we evaluate distinct parts of the algorithm in specifically designed experiments. For example, we consider a limited case with a specific and fixed concept of interest. To do this, we leverage the usage of Twitter hashtags as an explicit way users code their own concepts. The 4/15/2013 Boston Marathon bombings example used earlier was defined this way, with the hashtag `#bostonbombings`. We then construct a data set composed of three different sets of tweets. As the reference set, we use 5,909 English-language tweets that contain the hashtag `#bostonbombings` but not the word `boston` posted April 15–18, 2013. The target set  $T$  we hope the algorithm will identify contains 4,291 tweets during the same time period that contains both `#bostonbombings` and `boston`. We created the  $S \setminus T$  portion of the search set with the 9,892 tweets that were posted April 12–13, 2013, before the bombings, that contain the word `boston` but not `#bostonbombings`. The especially useful feature of these data is that the bombings were a surprise event that no one on social media was aware of ahead of time, which makes the demarcation between  $T$  and  $S \setminus T$  much clearer than it would normally be.

The task of identifying the target set is, of course, straightforward with the keywords `#bostonbombings` and `boston`, and so solely for this experiment we remove them from the text of each of the tweets before our analysis. We also do not use metadata indicating the date or time of the tweet. This is therefore an artificial example, but one

**TABLE 3** Top 25 Keywords in the Boston Bombings Validation Example

Target Keywords	Nontarget Keywords
peopl, thought,	marathon, celtic, game,
prayforboston, prayer, fbi,	miami, weekend heat,
affect, arrest, cnn, pray,	tsarnaev, new, play, red
video, obama, made,	watertown, open, back,
bomb, bostonmarathon,	sox, job mom, tonight,
heart, injur, attack, releas,	win, fan, monday
victim, terrorist, sad, news,	bruin, reaction, liam,
sick, rip, investig	tomorrow, payn

*Note:* The validation example is from the target  $T$  and nontarget  $S \setminus T$  search set lists produced by a single noniterative run of the algorithm, without human input.

constructed to make it possible to evaluate. The goal is for human users selecting keywords with our algorithm to be more accurate, more reliable, faster, and more creative than working on their own without it. Although this is the relevant goal for a single human user, it is a trivially easy standard for our algorithm to meet. To see this, consider a limited special case of our algorithm with keyword lists ordered *randomly*. Since we showed above that humans are usually incapable of recalling more than a small fraction of relevant keywords, but are very good at recognizing important keywords put before them, even randomly ordered keyword lists would still provide a great deal of help.

We thus seek to evaluate only the quantitative features of our algorithm here, and so we run the algorithm once without iteration, and also without any human input or interaction. To simplify the analysis, and to make replication of our results easier with fewer computational resources, we degrade our approach further by using only two fast classifiers (Naive Bayes and Logit). The estimated target set is designated as any document that receives at least one classifier vote, with probability above 0.5. We also preprocess the documents in standard ways, by stemming, and removing punctuation, stop words, numbers, and words with fewer than three characters.

## Qualitative Summary

We evaluate this analysis in three ways, beginning in this section with the qualitative summary in Table 3. This table lists the top 25 (stemmed) keywords from the target  $T$  and nontarget  $S \setminus T$  keyword lists produced by a single run of the algorithm, without human input. We can evaluate the algorithm informally by merely

looking at the words and seeing what readers recognize. It appears that most of the target keywords are closely related to the bombing incident (e.g., *#prayforboston, thought[s], prayer, fbi, arrest, bomb, injure, attack, victim, terrorist*). A few words are clearly related but may be too imprecise to be useful as keywords to select documents (e.g., *cnn, sad*). Most nontarget keywords do a good job of finding events related to Boston that are unrelated to the bombings, largely related to sports teams (e.g., *celtic game, miami, heat, red sox, bruin, win, fan*). They also include a few words that were apparently misclassified and so should be in the target set (e.g., *tsarnaev*). The word *bostonmarathon* in the target set and *marathon* in the nontarget set do not clearly discriminate posts related or unrelated to the bombings on their own to necessarily be useful—although interestingly, the algorithm discovered a pattern difficult for humans: that social media posts happened to use the former word to describe the bombings and the latter to describe the sporting event.<sup>5</sup>

### Grouping and Ranking Keywords

Second, we more formally evaluate the likelihood model used in our algorithm to group and rank keywords. Ideally, the target set list should have keywords that perform well on both recall and precision at the top, and the non-target set list should have keywords that perform poorly on both recall and precision.<sup>6</sup> Figure 3 reports the cumulative recall and precision for the first 100 keywords in each list (introduced one at a time from left to right in both graphs). The cumulative recall (left graph) and precision (right graph) are running estimates, as we add more and more terms into an “OR” Boolean query.

The key result in Figure 3 is that the target set line (in teal) is usually well above the nontarget set line (in red) for both recall and precision. In other words, our algorithm is doing a good job separating the two lists, which provides quantitative confirmation of the qualitative impression from the words in Table 3.

By definition, cumulative recall increases as we add more keywords. The fact that recall is not consistently zero for the nontarget set list speaks to both the need for human input as well as the nature of human language,

<sup>5</sup>Liam Payne was a 19-year-old singer inappropriately stopped by authorities in an underage establishment, and the subject of many social media posts. The word *rip* was, before removing punctuation and stemming, *R.I.P.*, which means “rest in peace.”

<sup>6</sup>The two common metrics we use are from the information retrieval literature. They include *precision*, the proportion of retrieved documents from each keyword that contains documents of interest, and *recall*, the proportion of all documents of interest that are retrieved by the keyword.

where the same keywords can often be used in social media posts with the opposite meanings—describing concepts of interest and not of interest. The general downward trend of the cumulative precision for the target set list shows that the general ordering of the keywords is also valid, with more precise words near the top of the list.

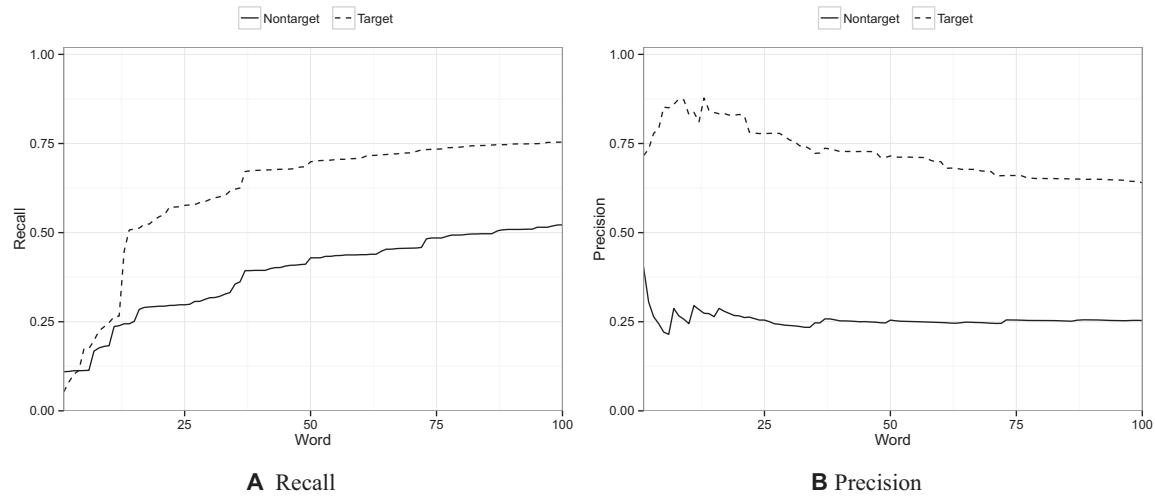
### Comparison to Human Users

For our final evaluation, we compare this single non-iterative run of our algorithm (with no human in the loop) with a purely human approach. We do this in two ways. First, we compare the top 145 words from our target set keyword list with the 145 unique keywords that the 43 undergraduates in our Boston Bombings experiment came up with in the experiment described in the third section. For this evaluation, we are therefore comparing the effort of 43 minds versus one single run of the computer algorithm without any human input. This is not a real comparison, of course, since in practice, researchers are unlikely to be able to hire 43 research assistants and would be able to use some human input to improve the algorithm, but it gives a useful baseline comparison.

Panels (a) and (b) in Figure 4 give density estimates for the overall precision and recall of the 145 words chosen by humans compared to the top 145 words from the target set list from our algorithm. The results show that recall of the algorithm is approximately the same as the collective work of 43 humans. Put differently, both the one-step algorithm and the humans come up with keywords of about the same quality. Of course, we constrained the algorithm to the same number of words as the 43 humans when, of course, our algorithm would produce *many* more than the 145 words shown in the graph.

To get a sense of the quality of the individual words in this comparison, we see from Panel (b) that the precision of the algorithm’s words is generally much higher than the precision of words from the humans. When restricted to 145 words, the algorithm produces the same level of recall as the effort of 43 different humans combined, but the words chosen by the algorithm contain much less noise and are therefore of substantially higher quality than human-only approaches.

Finally, we consider a more realistic comparison of a (still limited) one-step special case version of our algorithm without human input to one human research assistant at a time. Individual humans choose only about 7–8 words, with no one of our 43 individuals choosing more than 20. Panels (c) and (d) of Figure 4 give cumulative recall and precision for our algorithm out to 50

**FIGURE 3 Cumulative Recall and Precision**

words (although it could of course keep going) compared to each of our 43 human users. Individual undergraduate cumulative recall appears as separate black lines in Panel (c). The algorithm's cumulative recall is better than most of the human users until about 12 words are recalled, at which point the algorithm's performance soars well beyond any one of the human users. After 20 words, the human users obviously have nothing to offer. The algorithm's precision (Panel d) is also better than most of the human users in the entire range of human-recalled words, but then continues out to 50 words in the graph without losing much precision in the process.

Although our algorithm is clearly better than individual human users, using the algorithm with human input as designed has the potential to be much better than either alone.

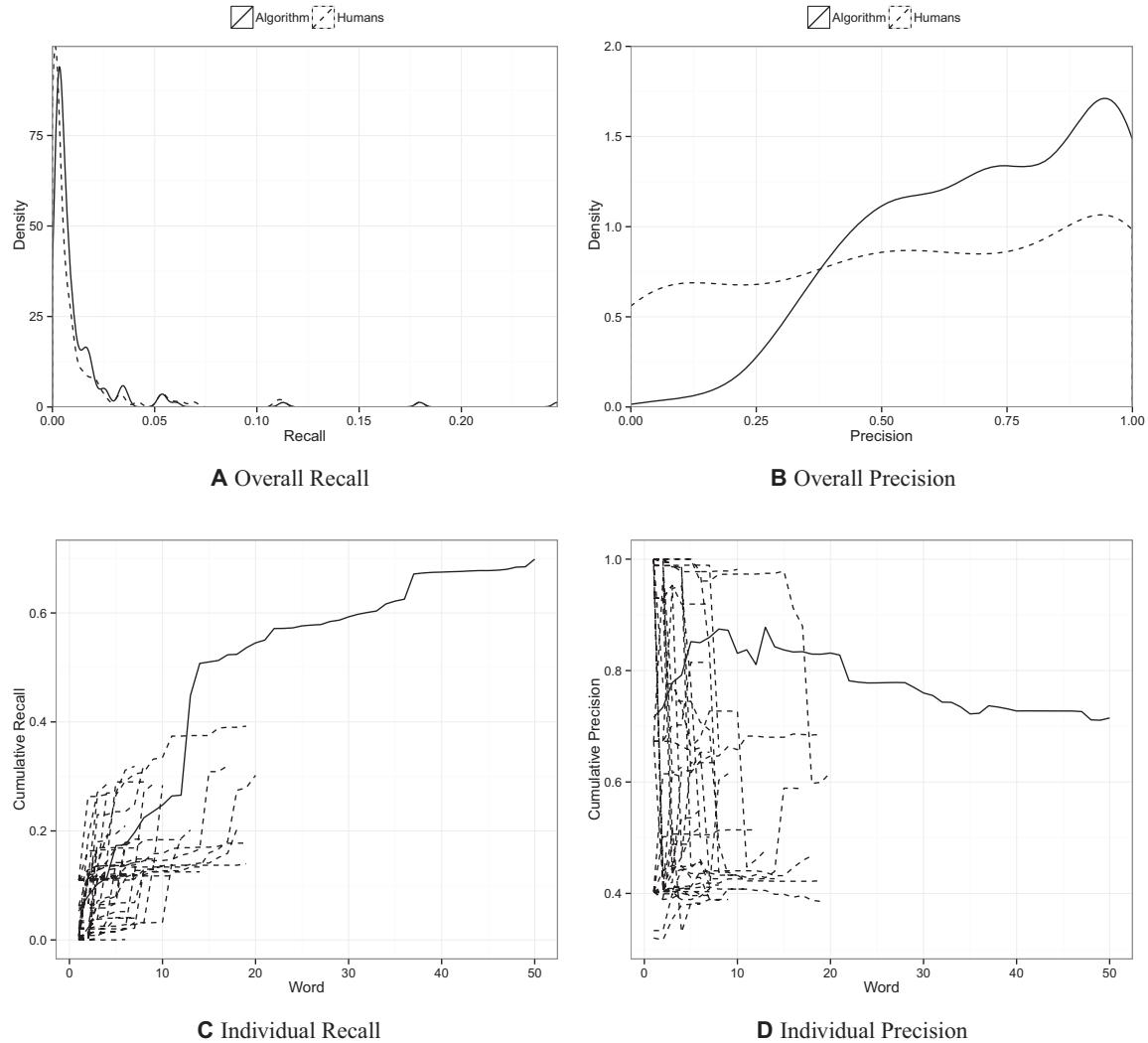
### Detecting the Language of Censorship Evasion

In what became known as the “Wang Lijun incident” in China, police chief of Chongqing Wang Lijun was abruptly demoted from his job on February 2, 2012. Rumors began circulating that Wang had fallen out of favor with his boss, party chief of Chongqing and popular political leader Bo Xilai. On February 6, 2012, Wang Lijun went to the U.S. Consulate in Chengdu, possibly to seek asylum, but after the consulate became surrounded by police, Wang agreed to leave the consulate and was detained by the Chinese government. During this time, rumors about

how the incident, perceived as treason by many in China, would affect the political prospects of Bo Xilai spread virtually across social media, culminating in Bo’s March 15 dismissal from his post. It was later revealed that Wang had fled to the consulate because he had confronted Bo that Bo and his wife, Gu Kailai, were connected to the murder of British businessman Neil Heywood, who had died in November 2011 in Chongqing. In the dramatic trials of Wang, Gu, and Bo that followed, all were convicted with lengthy prison sentences.

The Wang Lijun incident and Bo Xilai scandal were some of the most dramatic and important political events to occur in China in decades. Bo Xilai, son of famous revolutionary Bo Yibo, had gained widespread popular support in Chongqing for his crackdown on crime and promotion of Maoist culture. He was also an ambitious politician who was hoping to be promoted to higher leadership roles within the Party. Because of the scale and drama involved in the scandal, the Bo Xilai scandal was of tremendous public interest and widely discussed, but at the same time highly censored.

Social media posts that used the names “Bo Xilai,” “Gu Kailai,” and “Wang Lijun” were censored across much of the social media landscape by automated filters programmed in many social media websites. At the same time, social media users, who know about these filters, tried to write posts using creative rephrasings and neologisms so their posts would slip past the filters but still be understandable to general readers. Amid this linguistic arms race between government-controlled computers and the Chinese people, researchers trying to understand

**FIGURE 4 Comparing Recall and Precision for the Algorithm versus 43 Human Users**

Note: Panels (a) and (b) display the distribution of recall and precision for the 145 words from the humans and the top 145 words from the algorithm. Panels (c) and (d) display cumulative recall and precision for each human (dotted line) versus the first 50 target set keywords of the algorithm (solid line). Human keywords are in the order that humans thought of them.

this scandal have to scramble to keep up with these novel words and rephrasings. Missing even one may cause them to lose the thread of the conversation, bias their inferences, or make finding posts of interest difficult or impossible. We show how our algorithm can be used by researchers to find these words and the posts of interest.

We began with words widely known to be used to evade censorship for the reference set and those that were more commonly used to describe the scandal in the search set. Examples of a few of the words we discovered ap-

pear in the first column of Table 4. For example, the reference set was composed of microblogs that contained the word *bxl* (in English), the first letter of each syllable in Bo's name, during the first half of 2012, and the search set was the broader term to describe the scandal "Chongqing incident" (重庆事件). The target set picked up a variety of words related to the event, including words that netizens were using to evade censorship. For example, 王丽娟, a homophone for Wang Lijun, appeared within the top 100 of the list. *Bu xing le* (不行了, which

**TABLE 4** Words the Chinese Use to Evade Government Censors

Keyword Discovered	Reference Set	Search Set	Found In	Meaning
王丽娟	bxl	重庆事件	target set	homophone for Wang Lijun (王立军)
不行了	bxl	重庆事件	reference set	bu xing le, has same initials as Bo Xilai
护士长	王丽娟	重庆事件	reference set	“matron,” nickname for Wang Lijun
hwd	薄熙来	gkl	target set	abbreviation for Neil Heywood's last name

means “not OK,” but has the same initials as Bo Xilai) appeared within the keyword list associated with the reference set BXL. Upon reading texts with these words, we verified that both of these words were being used to evade censorship.

Based on the new words we found to evade censorship, we further revised the reference set and reran the algorithm to search for other keywords. For example, we used the homophone for Wang Lijun, 王丽娟, as the reference set and again “Chongqing incident” (重庆事件) as the search set. We discovered yet another nickname for Wang Lijun, “matron” (护士长). Using Bo’s full name 薄熙来 to define the reference set and the abbreviation for Gu Kailai’s name, “gxl,” as the search set, we also found the abbreviation for Neil Heywood’s name in the keyword target set, “hwd.”

Of course, not every word on the list was being used to evade censorship, since to be effective these words need to be rare. For example, many of the words were closely indicative of the scandal but not neologisms. However, a human user knowledgeable about the region can easily pick out the words that are being used to evade the censors from this longer list. Seeing the English abbreviation “hwd” out of a list of mostly Chinese characters automatically alerts the reader or researcher that it is being used as shorthand for another word, and knowing the context (or perusing the documents) would enable one to ascertain whether it is being used to substitute for a censored word. Similar patterns emerge in the purely Chinese words as well. The power here comes from the combination of the algorithm doing the “recalling” and the human doing the recognition of what is relevant.

## Prior Literature

Our algorithm is related to the information retrieval literature and “query expansion” methods, including algorithms that add or reweight keywords within search queries to retrieve a more representative set of documents (for a review, see Carpineto and Romano (2012), Rocchio (1971), Xu and Croft (1996)). Our approach differs in two

important ways. First, most query expansion methods retrieve new keywords by stemming the original keyword, looking for synonyms or co-occurrences, or finding related terms within the corpus defined by the original keyword (Bai et al. 2005; Schütze and Pedersen 1997). In contrast, our approach finds related keywords in external corpora that do not include the original keyword. For example, thesauri will not reveal novel hashtags or many of the terms in log tail search or those used to evade censors.

While some query expansion methods use large external corpora, such as Wikipedia, to enhance keyword retrieval (Weerkamp, Balog, and de Rijke 2012), our method allows the user to define the external corpus without any structured data aside from the sets  $R$  and  $S$ . We thus rely on the user’s expertise to define the search and reference sets from which new, related keywords will be generated.

Second, current query expansion methods often try to limit “topic drift” or are concerned with identifying keywords that are too general (Mitra, Singhal, and Buckley 1998). As a result, most of those methods implicitly focus on maximizing the precision of the documents retrieved (making sure the documents retrieved are all relevant), whereas we focus on both precision and recall (making sure to retrieve as many of the relevant documents as possible). Our method intentionally suggests both general and specific keywords and includes topic drift, not as a problem to be fixed but, at times, as the subject of the study. We instead rely on the user interaction phase of our model to refine the keyword suggestions and avoid topic drift outside the user’s interest.

Finally, most query expansion methods rely on probabilistic models of the lexical properties of text (e.g. Carpineto and Romano 2004; Voorhees 1994). Our approach uses ensembles of document classifiers to first group documents that may be of interest to the user. (A related approach is search results clustering [SRC], except with user-specified corpora of documents; see Carpineto et al. 2009 for a review.) It then retrieves keywords that are likely to appear in this document group,

but unlikely to appear in the rest of the search data set. Despite the differences between our approach and the current query expansion methods, our approach is actually a more general framework that can incorporate many of the existing methods, as we describe in a later section.

## Concluding Remarks

The human-led, computer-assisted, iterative algorithm we propose here learns from the mistakes made by automated classifiers, as well as the decisions of users in interacting with the system. In applications, it regularly produces lists of keywords that are intuitive, as well as those that would have been unlikely to have been thought of by a user working in isolation. Compared to a team of 43 human users, our algorithm has the same recall but far better precision; the algorithm also dominates individual human users on many dimensions. The algorithm discovers keywords, and associated document sets, by mining unstructured text, defined by the user, without requiring structured data. The resulting statistical framework and methods open up a range of applications for further analyses. In addition to the examples in English and Chinese, this algorithm has been useful in detecting Arabic dialects (Smith 2016), and we see no reason why it would not work on all human languages, but this would of course need to be studied further.

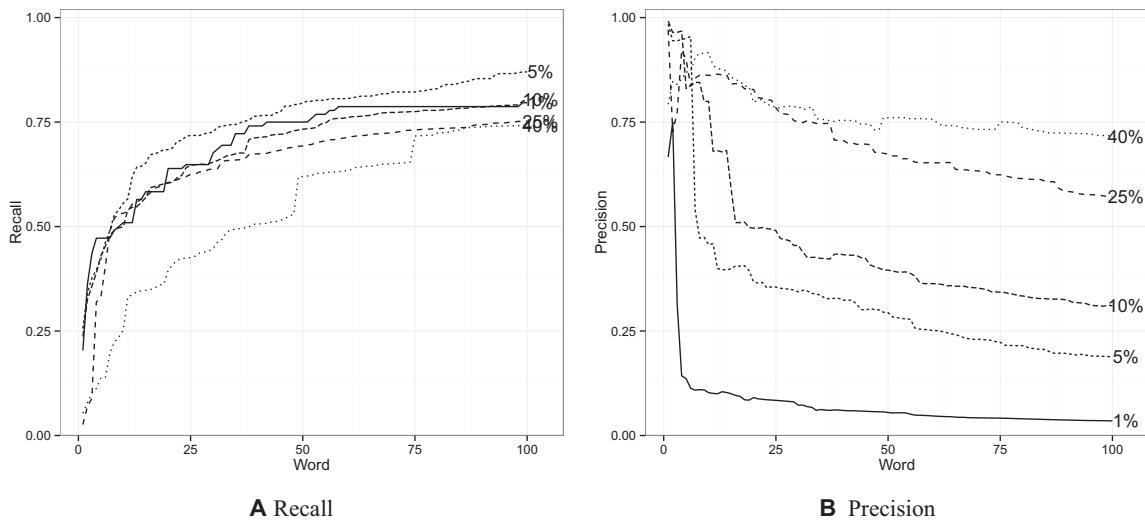
## Appendix A

### Robustness to Target Set Size

We study here how robust our keyword list discovery is as the target set size declines as a percentage of the search set. In the section “Evaluations,” the (true) target set size was about 30% of the entire search set. We now test a variety of target set proportions from 1% to 40%. We create these samples by setting the search set size to 10,000 and then randomly drawing from the coded target and nontarget sets to control the overall proportions.

Figure A1 gives cumulative recall and precision for different target set sizes. Clearly, the general trends from Figure 3 in the main text continue to hold. In addition, recall goes up and is higher for smaller target set proportions, which makes sense since fewer documents of interest need to be retrieved. Precision also follows the same downward sloping trend—higher for target sets that are a larger proportion of the search set. With more documents of interest and thus less noise in the search set, more high-quality keywords exist, and more pertinent information can be found with each retrieval relative to noise. Note that for very small target sets, the precision drops off fast after only a few words, which suggests smaller target sets in general will have many fewer words that are of high quality. In practice, human users may choose to respond to this situation by broadening the concept of interest if that is an option or, to find a needle in a large textual haystack, using small numbers of words to search document sets.

**FIGURE A1 Cumulative Recall and Precision of Target Set Keywords for Different Target Set Percentages**



## Appendix B

### Building Queries for Large Data Sets

In the section “Evaluations,” our validation example assumed a single data set that was divided into a reference and search set. The workflow in a single small data set is relatively simple. We first separate the reference set from the search set, run our algorithm, and then retrieve a list of target set keywords and nontarget set keywords. The user can then use the keywords for various applications, one of which is building a comprehensive Boolean query  $Q_{RT}$  to retrieve a set of documents of interest.  $Q_{RT}$  can be built in this setup by simply taking the initial reference query  $Q_R$  and adding target set keywords with OR operators and/or nontarget set keywords with the NOT and OR operators. For example, a query for the Boston Bombings example (assuming that the entire data set is given) could be “#bostonbombings OR (suspect OR fbi OR #prayforboston) AND NOT (sox OR celtics OR bruins),” where the words correspond to words from the reference query, target keyword list, and nontarget keyword list, respectively.

For large or potentially infinite data sources such as social media, the workflow described above is not feasible for a couple of reasons. In cases where the data set is large but finite, processing and running the algorithm over the entire data set as a search set may be infeasible computationally. For data sets of infinite size, there is no single search set that can be defined to run the algorithm. The user must define the search set manually via Boolean query or other means, a decision which then highly affects the results. We describe in more detail here a workflow alluded to in the section “Algorithm,” where the definition of the search set may be incorporated into the workflow and the algorithm run multiple times to define the comprehensive query  $Q_{RT}$ . Consider the following steps to the workflow:

1. Define reference set  $R$ .
2. Mine  $R$  for keywords  $K_R$  to expand the query or use any other query expansion method available.
3. Choose one or more words from the query expansion to add to the query by either
  - (a) adding the query expansion words to the comprehensive query  $Q_{RT}$  as is, or
  - (b) using the query expansion words to define a search set  $S$ , running our algorithm, and then adding additional words from our algorithm to refine the query expansion words for the comprehensive query  $Q_{RT}$ .
4. Repeat Step 3 multiple times.

We demonstrate this workflow in an example of gathering relevant tweets about the Paris terrorist attacks on November 13, 2015. We collected a set of tweets between November 13 and November 15 with the hashtags `#parisattacks` as our initial reference set. We show here a very simplified version of the workflow for how to collect keywords and use our algorithm to develop a comprehensive Boolean query to retrieve tweets about the Paris attacks.

1. Use `#parisattacks` to define a reference set. ( $Q_{RT}$ : `#parisattacks`)
2. Use a simple query expansion method by simply mining the entire reference set for keywords and rank them according to their document frequency. Then scan the top 100 words for ideas about expanding the query. We can alternatively include any other query expansion method in the literature here.
3. See the word `#prayforparis` in the expansion list. Through substantive knowledge, recognize that all tweets returned by `#prayforparis` are likely to be relevant, so simply add it to the query, ( $Q_{RT}$ : `#parisattacks OR #prayforparis`)
4. See the word `paris` in the expansion list. We would like to add it to the query, but not all documents retrieved by `paris` will be relevant, so we need to use the algorithm to subset further. Define and retrieve a search set with `paris` but excluding `#parisattacks` or `#prayforparis`.
5. Run the algorithm on the newly defined search set and look at the top 100 words in each list. See words that will help retrieve relevant posts from the target set list (e.g., `prayer`, `raid`, `abaaoud`, `mastermind`) and words that indicate nonrelevant posts from nontarget set list (e.g., `climate`, `change`, `conference`). Add to the comprehensive query. ( $Q_{RT}$ : `#parisattacks OR #prayforparis OR (paris AND (prayer OR raid OR abaaoud OR mastermind) AND NOT (climate OR change OR conference))`)
6. From the expansion list in Step 2, see the word `france` and use it as a search set for investigation. Repeat the algorithm with the new search set and find words that separate `france` into relevant and irrelevant posts. ( $Q_{RT}$ : `#parisattacks OR #prayforparis OR (paris AND (prayer OR raid OR abaaoud OR mastermind) AND NOT (climate OR change OR conference)) OR (france AND (suspect OR victim OR attack OR terrorist) AND NOT (air OR england OR russia OR benzema))`)
7. Repeat until satisfied.

Through this workflow that involves both human and algorithmic expertise, the user can work through a large or infinite set of documents and retrieve the relevant documents of interest by building long and comprehensive queries.

## References

- Antoun, Christopher, Chan Zhang, Frederick G. Conrad, and Michael F. Schober. 2015. "Comparisons of Online Recruitment Strategies for Convenience Samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk." *Field Methods* 28(3): 231–46.
- Bai, Jing, Dawei Song, Peter Bruza, Jian-Yun Nie, and Guihong Cao. 2005. "Query Expansion Using Term Relationships in Language Models for Information Retrieval." In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, ed. Abdur Chowdhury, Norbert Fuhr, Marc Ronthaler, Hans-Jorg Schek, Wilfried Teiken, New York: ACM Press, 688–95.
- Bauml, Karl-Heinz. 2008. "Inhibitory Processes." In *Learning and Memory: A Comprehensive Reference. Volume 2: Cognitive Psychology of Memory*, ed. Henry L. Roediger. Oxford: Elsevier, 195–220.
- Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.
- Carpinetto, Claudio, Stanislaw Osiński, Giovanni Romano, and Dawid Weiss 2009. "A Survey of Web Clustering Engines." *ACM Computing Surveys (CSUR)* 41(3): 1–38.
- Carpinetto, Claudio, and Giovanni Romano. 2004. "Exploiting the Potential of Concept Lattices for Information Retrieval with CREDO." *Journal of Universal Computer Science* 10(8): 985–1013.
- Carpinetto, Claudio, and Giovanni Romano. 2012. "A Survey of Automatic Query Expansion in Information Retrieval." *ACM Computing Surveys (CSUR)* 44(1): 1–50.
- Chen, Yifan, Gui-Rong Xue, and Yong Yu. 2008. "Advertising Keyword Suggestion Based on Concept Hierarchy." In *Proceedings of the International Conference on Web Search and Web Data Mining*. New York: ACM, 251–60.
- D'Orazio, Vito, Steven T. Landis, Glenn Palmer, and Philip Schrodt. 2014. "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines." *Political Analysis* 22(2): 224–42.
- Eshbaugh-Soha, Matthew. 2010. "The Tone of Local Presidential News Coverage." *Political Communication* 27(2): 121–40.
- Gentzkow, Matthew, and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence from US Daily Newspapers." *Econometrica* 78(1): 35–71.
- Grimmer, Justin, and Gary King. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization." *Proceedings of the National Academy of Sciences* 108(7): 2643–50.
- Hand, David J. 2006. "Classifier Technology and the Illusion of Progress." *Statistical Science* 21(1): 1–14.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd ed. New York: Springer.
- Hayes, Philip J., and Steven P. Weinstein. 1990. "CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories." *IAAI 90*: 49–64.
- Ho, Daniel E., and Kevin M. Quinn. 2008. "Measuring Explicit Political Positions of Media." *Quarterly Journal of Political Science* 3(4): 353–77.
- Hopkins, Daniel, and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1): 229–47.
- King, Gary, Patrick Lam, and Margaret E. Roberts. 2016. "Replication Data for Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." doi:10.7910/DVN/FMJDCD Harvard DataVerse, [UNF:6:56ELwemliNH+ALideeh3Q==].
- King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism But Silences Collective Expression." *American Political Science Review* 107(2): 1–18.
- Kulkarni, Sanjeev R., Gábor Lugosi, and Santosh S. Venkatesh. 1998. "Learning Pattern Classification—A Survey." *IEEE Transactions on Information Theory*, 44(6): 2178–2206.
- Letham, Benjamin, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2013. "Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model." *Annals of Applied Statistics* 9(3): 1350–71.
- Letham, Benjamin, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. "Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model." *Annals of Applied Statistics* 9(3): 1350–71.
- Mitra, Mandar, Amit Singhal, and Chris Buckley. 1998. "Improving Automatic Query Expansion." In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 206–14.
- Nielsen, Finn Årup. 2011. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs". *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*. 93–98. (CEUR Workshop Proceedings; Journal number 718).
- Puglisi, Riccardo, and James M. Snyder. 2011. "Newspaper Coverage of Political Scandals." *Journal of Politics* 73(3): 931–50.
- Rocchio, Joseph John. 1971. *Relevance Feedback in Information Retrieval*. Englewood Cliffs, NJ: Prentice-Hall.
- Roediger, Henry L., and James H. Neely. 1982. "Retrieval Blocks in Episodic and Semantic Memory." *Canadian Journal of Psychology/Revue canadienne de psychologie* 36(2): 213–42.
- Schapire, Robert E., and Yoav Freund. 2012. *Boosting: Foundations and Algorithms*. Cambridge, MA: MIT Press.
- Schütze, Hinrich, and Jan O. Pedersen. 1997. "A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval." *Information Processing & Management* 33(3): 307–18.
- Smith, Evann. 2016. *Mass Mobilization in the Middle East: Form, Perception, and Language*. PhD dissertation, Harvard University.
- Voorhees, Ellen M. 1994. "Query Expansion Using Lexical-Semantic Relations." In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and*

- Development in Information Retrieval.* New York: Springer-Verlag, 61–69.
- Weerkamp, Wouter, Krisztian Balog, and Maarten de Rijke. 2012. “Exploiting External Collections for Query Expansion.” *ACM Transactions on the Web (TWEB)* 6(4): 1–29.
- Xu, Jinxi, and W. Bruce Croft. 1996. “Query Expansion Using Local and Global Document Analysis.” In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 4–11.
- Yang, Guobin. 2009. *The Power of the Internet in China: Citizen Activism Online.* New York: Columbia University Press.
- Zhu, Xiaojin, and Andrew B. Goldberg. 2009. “Introduction to Semi-Supervised Learning.” *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3(1): 1–130.



## Editorial

## Introduction—Topic models: What they are and why they matter

---

**Abstract**

We provide a brief, non-technical introduction to the text mining methodology known as “topic modeling.” We summarize the theory and background of the method and discuss what kinds of things are found by topic models. Using a text corpus comprised of the eight articles from the special issue of *Poetics* on the subject of topic models, we run a topic model on these articles, both as a way to introduce the methodology and also to help summarize some of the ways in which social and cultural scientists are using topic models. We review some of the critiques and debates over the use of the method and finally, we link these developments back to some of the original innovations in the field of content analysis that were pioneered by Harold D. Lasswell and colleagues during and just after World War II.

© 2013 Published by Elsevier B.V.

---

### 1. Introduction

Content analysis is a technique which aims at describing, with optimum objectivity  
precision, and generalizability, what is said on a given subject in a given place  
at a given time.

*Harold Lasswell et al. (1952, p. 34),  
The Comparative Study of Symbols: An Introduction*

In this short essay, we provide a brief, non-technical introduction to the text mining methodology known as “topic modeling.” We start with the basic question, what is a topic model? We summarize the theory behind the method and then focus on the question of what exactly is a topic? (Or, to put it the other way round, we ask what does a topic model measure?) We address this issue by describing the work published here in this special issue. For each article we pose three questions: What topics have these researchers found? How have they interpreted the meaning of their topics? And how have they used them as a component within a larger research project? We turn then to briefly discuss some of the demands, dilemmas and limitations of topic models and proceed to the second question telegraphed by our title—why do topic models matter? We answer this by describing some of the ways that we think these methods can change how scholars in the social and cultural sciences approach (and use) texts and textual

analysis, and we end by taking the long view of just how topic models represent a certain kind of closure on one chapter in the history of content analysis and the beginning of another.

## 2. What is a topic model?

Topic models are a promising new class of text analysis methods that are likely to be of interest to a wide range of scholars in the social sciences, humanities and beyond.<sup>1</sup> The most distinctive feature of topic models is that they provide an automated procedure for coding the content of a corpus of texts (including very large corpora) into a set of substantively meaningful coding categories called “topics.” The algorithms can do this with a minimum of human intervention, and this makes the method more inductive than traditional approaches to text analysis in the social and human sciences.<sup>2</sup> Instead of starting with pre-defined codes or categories of meaning (like those we generate when we start to hand-code a text), the researcher begins by specifying the number of topics for the algorithm to find. The program then identifies that specified number of topics and returns the probabilities of words being used in a topic, as well as an accounting of the distribution of those topics across the corpus of texts. While not infallible, when used thoughtfully and applied carefully, the method seems to consistently yield very plausible readings of the texts, demonstrating what DiMaggio, Nag and Blei describe in this special issue as high levels of “substantive interpretability.”

### 2.1. The theory behind the method

So, how do topic models work? How does an automated procedure reliably find textual meanings that prove to be useful? A simple answer is that the method depends upon the presumption that meanings are relational ([Saussure, 1959](#)). In this case, the meanings that define a coherent topic of conversation are constructed from a set of word clusters. Thus, a topic might

---

<sup>1</sup> “New” is a relative term here. The original article on latent Dirichlet allocation (LDA) by [Blei et al. \(2003\)](#) was published a decade ago. As that article quite usefully explains, there is also a long pre-history to the method—including the early work on Latent Semantic Indexing (LSI) by [Deerwester et al. \(1990\)](#) and Hoffman’s ([1999](#)) probabilistic Latent Semantic Indexing (pLSI) approach. There is also another tradition of topic modeling using Gibbs sampling techniques that dates back to work by [Griffiths and Steyvers \(2004\)](#) (see, also [Griffiths et al., 2005; Newman et al., 2007](#)). [McNamara \(2010\)](#) provides a broad view of thirteen classes of Latent Semantic Analysis (LSA) that she describes as representing different “statistical models of semantics” (of which topic modeling is one). McNamara also traces the field back to the original work of [Osgood et al. \(1957\)](#). Nonetheless, it is still largely a new class of methods for most social scientists and humanists. There are some exceptions. A few political scientists have been quick to pick up these methods and employ them in useful ways ([Grimmer, 2010; Grimmer and King, 2011; Grimmer and Stewart, 2013](#)). Also, the Digital Humanities community is way ahead on the use of topic models; this is, in part, thanks to workshops funded by the NEA such as the “Networks and Network Analysis for the Humanities” conference held at the Institute for Pure and Applied Mathematics at UCLA (organized by Tim Tangherlini, also an author of one of the articles published here in this special issue). See, also the special issue on topic models in the *Journal of Digital Humanities*, edited by [Meeks and Weingart \(2012\)](#), as well as new books by [Jockers \(2013\)](#) and [Moretti \(2013\)](#). Among sociologists, there is the early work by [Moody and Light \(2006\)](#), and there is also some interesting new work coming out by [Bail \(forthcoming\)](#), [Mutzel \(2012\)](#) and by [Kaplan and Vakili \(2012\)](#), among others.

<sup>2</sup> From a machine learning perspective topic modeling is an unsupervised task, meaning that no prior human annotation, labeling or hand-coding is necessary to infer a model. But of course it is also important to point out that what is inductive for the analyst is deductive for the method, in the sense that LDA topic models presume a particular theory of the meaning of a text and this theory is expressed in way in which the LDA model is constructed. We say more about these assumptions below.

be thought of as the constellation of words that tend to come up in a discussion (and, thus, to co-occur more frequently than they otherwise would) whenever that (unobserved and latent) topic is being discussed. Note that topic models capture co-occurrences regardless of these words' embeddedness within other complexities of language—such as syntax, narrative, or location within the text. Instead, each document is treated as if it were a so-called “bag of words.” The goals of a topic model analysis are then to analyze these various word bags, to identify word co-occurrence patterns across the corpus of bags, and then to use these to produce a mapping of the distribution of words into the topics and of the topics into the bags.

Within the more general data science field, topic modeling is an instance of probabilistic modeling.<sup>3</sup> The simplest and most widely used model is Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003).<sup>4</sup> As DiMaggio, Nag and Blei explain in their article published in this special issue, “LDA is a statistical model of language.” The generative process behind the model is a convenient way to introduce its intuition. Each document (text) within a corpus is viewed as a *bag-of-words* produced according to a mixture of themes that the author of the text intended to discuss. Each theme (or topic) is a distribution over all observed words in the corpus, such that words that are strongly associated with the document’s dominant topics have a higher chance of being selected and placed in the document bag. Given the above distributions, the author repeatedly picks a topic, then a word and places them in the bag until a document is complete. The objective of topic modeling is to find the parameters of the LDA process that has likely generated the corpus. This is also referred to as “inference” in the LDA literature and, in essence, it is the task of reverse-engineering the intents of the author(s) in producing the corpus.<sup>5</sup>

Among the outputs of the inference is a set of per-word topic distributions associating a probability with every topic-word pair and a similar set of per-topic document distributions describing the probability of choosing a particular topic for every specific corpus document. But note again, the obtained structure is latent, which means that the learned per-word topic distributions are not associated with an explicit topic label, but instead with a set of word probabilities that, when ordered by decreasing probability, often relate closely to what a human would call a “topic” or a “theme.”

---

<sup>3</sup> Another computer science branch that deals with text is natural language processing (NLP). The latter differs from topic models in that many of the developed methods require human/expert training. It is important to note that these different model families are compatible and, in fact, could be combined to get closer to meanings in a text. In this special issue, the articles by McFarland et al., Mohr et al. and Jockers and Mimno all take this issue up explicitly. For an example of another kind of combination of topic models with other modalities of text analysis, see Diesner and Carley (2005,2008,2010).

<sup>4</sup> Since the inception of early topic models like pLSI (Hofmann, 1999) and LDA (Blei et al., 2003), a family of approaches have been proposed to address and develop some of the assumptions in the original models and make them more applicable to specific real world analysis tasks by uncovering more sophisticated structures within texts. Some extensions relax the bag-of-words assumption by modeling the word order (Griffiths et al., 2005; Wallach, 2006). Other extensions deal with dependent documents in the corpus by modeling links (Chang and Blei, 2010) and dynamic topic models incorporating a temporal order of documents within the corpus (Blei and Lafferty, 2006). The assumption of a priori known number of latent topics is addressed by Teh et al. (2006). A more complete list of extensions and new topic models is available in Blei (2011) and Jelisavcic et al. (2012). The article published in this special issue by McFarland and colleagues also provides a useful review.

<sup>5</sup> Formally, the LDA algorithms are founded on a Bayesian probabilistic model. The DiMaggio, Nag and Blei article in this special issue does a nice job of offering a simple explanation of the formal logic behind the approach. Rhody (2012) also has a useful explanation of the probabilistic logic behind a topic model in which she uses the homespun analogy of trying to guess what the proportions of vegetables were being sold at the local farmer’s market based on a post hoc examination of one’s neighbors’ shopping bags. Brett (2012) provides a non-technical overview of the broader topic modeling methodology. Other papers by Blei (2012a,b) also provide very accessible introductions.

## 2.2. *What is a topic?*

This then brings us back to the question of just what exactly is a topic? For an answer to this question, we will focus on the way that the authors published in this special issue have used the method, and we will look to see what types of topics they have found. [Table 1](#) provides a summary of the articles allowing us to see at a glance the range and diversity of topic model applications published in this special issue.

The data sources vary widely—both by type of data and by size of corpus. Ian Miller analyzes over a hundred years of the Qing Dynasty’s “Veritable Records” containing comprehensive archives of “*zouzhe*,” or messages of concern that were reported directly to the Chinese emperor. McFarland, Ramage, Chuang, Heer, Manning and Jurafsky draw on a corpus of over a million dissertation abstracts (for dissertations filed between 1980 and 2010) as a way to map out the changing contours of academic fields. DiMaggio, Nag and Blei analyze a corpus of nearly 8000 newspapers articles (published between 1986 and 1997) that were concerned with the National Endowment for the Arts (NEA) or with publicly funded art projects in general. Bonilla and Grimmer study over 51,000 news-stories (taken from both newspapers and nightly news broadcasts) sampled after days in which the color coded terror alert level had been raised by the Bush Administration. Tangherlini and Leonard analyze (among other things) more than 34,000 Danish folk legends. Jockers and Mimmo use a corpus of over 3000 British, American and Irish 19th century novels, Marshall studies more than 3000 post-war academic journal articles (written by British and French demographers) while Mohr, Wagner-Pacificci, Breiger and Bogdanov have a corpus that consists of eleven official National Security Strategy documents (containing about a half million words).

What do the topic modelers get from topic modeling all this data? Both as a way to introduce the articles and also to help us think more deeply about these methods, we ask three questions of each article—what topics have they found? What are the meanings and understandings that the authors attribute to the topics? And how are the topic data deployed to help advance a specific research agenda?

The first two articles provide broad introductions to the method. DiMaggio, Nag and Blei investigate the controversies that erupted over U.S. federal funding of the arts during the 1980s and 1990s. They use an LDA algorithm to code 7958 newspaper articles selected from five newspapers (culled for stories relevant to the subject published between 1986 and 1997). They suggest that, when applied to data of this type, topic models provide a useful way to measure what social scientists have generally called “frames.” DiMaggio, Nag and Blei define a frame as “a set of discursive cues (words, images, narrative) that suggests a particular interpretation of a person, event, organization, practice, condition, or situation.” Media frames are important because they are powerful interpretive devices that “prime particular associations or interpretations of a phenomenon in a reader,” DiMaggio and colleagues write. Different media frames are promoted by different institutional actors as a way to try to influence the course of public discourse or the shape of political debate.

DiMaggio, and colleagues ask for twelve topics when they model their corpus. Looking at their results, we see that some of their topic-frames capture what appear to be generic news discussions of the arts; for example, one concerns “all kinds of musical performances and organizations,” another describes “museum exhibits and visual arts” but other topics clearly reflect more politicized frames, such as the “NEA grant controversies” or “1990s culture wars” (for all these examples, see [Table 1](#)). By mapping out the distribution of these different topic-frames, both across types of newspapers and across time, DiMaggio and his co-authors are able to

**Table 1**  
The nature and scope of topic model applications in articles published in *Poetics* Vol. 41, no. 6 (part-1).

Authors	Discipline	Source	Size of corpus	# topics	A sampling of “topics” identified by analysis	Measured object/use of measure
Mohr and Bogdanov	7 1	Articles Published in this special issue of <i>Poetics</i> (Vol. 41, no. 6)	8 (Tot # articles) 92,260 (Tot # Words)	25	“Engaging the canon” “Forgotten-versions” “Topic model” “Authors’ Gender” . . .	Topics measure themes in research articles. TMs used to identify paper specific themes and common themes across papers and to illustrate method.
DiMaggio, Nag and Blei	7 7 1	Newspaper Articles, (if, “NEA”, “Arts Agencies”, “public funding of arts”) (Houston Chron., NY Times, Seattle Times, WSJ & Wash. Post) (1986–1997)	7958 (Tot # articles) 54,982 (Tot # terms) 3,381,574 (Tot # Words)	12	“NEA grant controversies” “Congressional deliberations” “1990s culture wars” “All kinds of musical performances & orgs” “Museum exhibits & visual arts” “Theater and dance” . . .	Topics measure media frames within a policy domain. TMs used as part of research design that focuses on the use of different frames by different types of newspapers and the more general questions about the drop in popularity of public funding for the arts.
McFarland, Ramage, Chuang, Heer, Manning and Jurafsky	7 1 1 1 4 1	Dissertation Abstracts from 240 U.S. Research Universities (Proquest) (1980–2010)	1 million+ (tot # abstracts) (here: a sub-corpus just Anthropology related abstracts)	40	“Social structures” “Physical anthropology” “Archeology” “Identity studies” “Cultural anthropology” . . .	Topics measure group language conventions. Paper reviews series of uses of TMs to understand language differentiation in academic communities. Includes summary of different types of TMs.

Table 1 (Continued)

Authors	Discipline	Source	Size of corpus	# topics	A sampling of “topics” identified by analysis	Measured object/use of measure
Miller	2	Qing dynasty veritable records (1723–1911)	—	50	“Crime” “Unrest” “Sedition” “Rebellion” “Border rebellion” “Major rebellion” ...	Topics measure how the central state (during Qing dynasty) thought about and categorized mass violence. Used here to gain new insights into the crime rates & state record-keeping practices of 18th and 19th century China.
Bonilla and Grimmer	5	News stories on nightly newscasts by ABC, CBS and NBC and Newspapers from across the country (from Lexis-Nexis) (2002–2005)	51,766 (tot # news stories)	24	“Memorial” “Local small business” “Criminal prosecution” “Iraq/World” “Local philanthropy” “Law and order” “Personal interest stories” “2004 Presidential campaign” “Iraq war” ...	Topics measure broad, thematic categories for newspaper stories. Used to show that Bush’s <i>Terror Alerts</i> raise the public’s perceived likelihood of a terror attack, but not opinions about President’s job performance, foreign intervention, or willingness to restrict civil liberties.
PhD Discipline:	1. Computer Science, 2. East Asian Languages & Civilizations, 3. English, 4. Linguistics, 5. Political Science, 6. Scandinavian, 7. Sociology, 8. Swedish Literature.					
Mohr, Wagner-Pacifici, Breiger and Bogdanov	7 7 7 1	U.S. National Security Strategy reports (1990–2010)	11 (# NSS Documents) 6102 (Tot # Agents) 572,358 (tot # Words)	15	“Terrorism” “Economic development” “Human rights” “Global security strategy” “Military operations” “Peace” ...	Topics measure dramaticistic “scenes.” Incorporated into a model for graphing the Burkean “grammar of motives” of official United States National Security Strategy texts.

Marshall	7	<i>Population Studies</i> (if: “fertility”) <i>Population</i> (if: “fecondité” “natalité”) & select newspaper: <i>Times &amp; Guardian</i> (1946–2005)	1623 (tot # Articles in <i>Pop. Studies</i> ) 1835 (tot # Articles in <i>Population</i> )	75	British: “Africa and data” “Economics & transition” “Married fertility” “Nuptiality” ... French: “housing” “war & France” “abortion & contraception” ...	Topics measure content of professional discourse. TMs used as part of a cross-national comparison of research discourse (and its impact) in the (British & French) academic discipline of demography.
Tangherlini and Leonard	6	<i>Topic Model Data</i>	1. Two books	100	1. “Social instinct” ...	Topics measure literary <i>feel</i> . Sub-corpus topic modeling (STM) is presented as a new tool for discovering meaningful
	8	1. “The Origin of Species” & “The Descent of Man” 2. Modern Breakthrough authors: Jacobsen, Schandorf & Drachman 3. Folk legends collected by Tang Kristensen 1892–1901; 1928–1939 <i>Trawl data</i> : Google Books Danish corpora (1860–1920)	2. Selections from several books 3. ~34,000 legends from Kristensen’s collection	50 100	“struggle for survival” ... 2. “A woman’s thoughts” ... “her self” “intelligence” “Men, little girls, god, black robes and shouting” ... 3. “death and churchyards” “shooting and witches” “horses & wagons” “the minister” “serpents” ...	passages in a larger corpora. 3 tests of STM trawls here: 1. Tracing the diffusion of Darwin’s ideas. 2. Finding unknown authors of the Modern Break-through. 3. Finding the <i>feel</i> of Danish folklore in other Danish literature.
Jockers and Mimmo	3	British, American & Irish works of fiction (from Chadwyck Healey collection, Project Gutenberg & the Internet archive) (1750–1899)	3346 (Tot # books)	500	“Female fashion” “Enemies” “Convents & abbeys” “Religion” ...	Topics are a measurable, data-driven proxy for literary themes. Used here to assess how meta-data (like date of pub, gender...) predict fluctuations in the use of themes and the individual word choices within themes. Tests whether this evidence is statistically significant.

PhD Discipline: 1. Computer Science, 2. East Asian Languages & Civilizations, 3. English, 4. Linguistics, 5. Political Science, 6. Scandinavian, 7. Sociology, 8. Swedish Literature.

use topic models as a tool to answer basic questions about the changing dynamics of policy debates for public support of the arts during this volatile decade. In the process they also provide what is probably one of the best introductions to the use of LDA topic modeling for social scientific research.

To help us better demonstrate these methods, we ran an LDA model on the articles published in this special issue. Of course this is a much smaller corpus than the techniques were designed for—but we think that, even at this scale, it can be a useful exercise. After exploring some alternatives, we settled on a 25-topic model. [Table 2](#) presents our results. The leftmost column lists the topics, the other columns report the probability that a word in a given article will have been drawn from the topic in each row (note that the columns sum to 1).<sup>6</sup> Reading down the first column of data, we can see that just a handful of the topics had a very high probability of occurrence in the DiMaggio, Nag and Blei article. Only five of the twenty-five topics have a probability greater than .025. Topic 14 (which we have labeled “Frames for coverage of art-news”) is the most important topic in this article (words have nearly .4 probability of being “on this topic”). As the label suggests, this is a word constellation that captures the main intellectual themes of the article; it is defined by terms like: topic, arts, assigned, Times, NEA, art, coverage, grants, York, frames, prevalence, culture, funding, newspapers, government, and controversial.<sup>7</sup> Notice that none of the other articles in the special issue discuss Topic 14 (Bonilla and Grimmer, the only other newspaper study, has the highest probability at just over .025). This illustrates an important (but not surprising) result of our analysis: most of the topics that we have identified are unique to a specific paper.

In fact, just a few of the topics are shared across the articles and the only topic that is shared across *all* of the article is the subject of this special issue. Described by words such as: topic, topics, words, model, analysis, corpus, time, texts, terms, related, models, modeling, documents, social, results, word, number, document—Topic 8 captures the topic of topic modeling itself. Words chosen for the Marshall article (which devotes extra attention to the question of how to go about choosing the proper number of topics) have more than .4 probability of being generated (introduced into the paper) from Topic 8. At the other end of this scale is the article by Bonilla and Grimmer, with a probability of .1574 for Topic 8, a reflection of the fact that much of that article is not concerned with topic models at all but rather with survey data (that were serendipitously collected during the same time periods and which Bonilla and Grimmer brilliantly use as a way to assess the effects of the media framing that they show—using topic models—is linked to the escalating terror alerts).

Words in the DiMaggio et al. article have a probability of .2720 of being on the topic of topic modeling. They also have .2091 probability of being linked to Topic 16. This is more curious since it is labeled “Studying the media effects of terror alerts,” suggesting that it is the topic that captures the thematic focus of the Bonilla and Grimmer article (and, in fact, words in that article have a probability of .7197 of being on this topic). But if we dig a little deeper into the list, we also see words like: percent, arts, attention, support (Note: the top 8 words for each topic are listed in [Table A.1](#) of the Appendix). Looking further down the list adds: media, Bush, increase, stories, articles, news, figure, effect, policy, percentage, terms, surveys. Having traced these words and

---

<sup>6</sup> To be precise, since the number of texts (eight) is small, we trained the model first by running it with each paragraph in the corpus as a separate document. Then we ran each of the eight complete documents against this existing LDA model, asking for the probability of each topic occurring in the eight whole documents.

<sup>7</sup> Here and elsewhere, likely capitalizations of the words have been added by us—the actual terms used in the analysis were not case specific.

**Table 2**  
Topic distribution across articles published in *Poetics* Vol. 41, no. 6. (Articles listed by first author).

Topic and its Description	DiMaggio	McFarland	Miller	Bonilla	Mohr	Marshall	Tangherlini	Jockers
<i>Doc Word Count (tot)</i> 92,260 =	18,440	8,394	12,013	8,285	12,756	12,174	12,345	7,853
T-1 Engaging the canon	0.0008	0.0030	0.0001	0.0005	0.0002	0.0039	<b>0.0466</b>	0.0019
T-2 Predicting economic expectations	0.0168	0.0002	0.0001	<b>0.0441</b>	0.0013	0.0019	0.0001	0.0019
T-3 Archives & struggles	0.0001	0.0055	0.0001	0.0002	0.0002	0.0010	<b>0.0430</b>	0.0025
T-4 Identification & extraction of names	0.0003	0.0012	0.0008	0.0005	0.0035	0.0022	<b>0.0308</b>	0.0002
T-5 Computer models of language	0.0001	<b>0.5225</b>	0.0009	0.0002	<b>0.0527</b>	0.0001	0.0001	0.0002
T-6 Forgotten versions	0.0002	0.0023	0.0023	0.0002	0.0039	0.0121	<b>0.0244</b>	0.0005
T-7 Earlier efforts	0.0003	0.0005	0.0070	0.0008	0.0002	0.0072	<b>0.0181</b>	0.0002
T-8 Topic models	<b>0.2720</b>	<b>0.3492</b>	<b>0.2546</b>	<b>0.1574</b>	<b>0.2201</b>	<b>0.4186</b>	0.1976	<b>0.1789</b>
T-9 Bauditz largely deliberately missing	0.0001	0.0080	0.0027	0.0002	0.0002	0.0086	<b>0.0428</b>	0.0022
T-10 Anniversary result	0.0122	0.0002	0.0020	<b>0.0300</b>	0.0028	0.0001	0.0001	0.0029
T-11 themes, authors, gender	0.0010	0.0062	0.0003	0.0002	0.0125	0.0001	0.0019	<b>0.2307</b>
T-12 Original, arbitrary and begrudgingly famous	0.0003	0.0023	0.0008	0.0005	0.0024	0.0001	<b>0.0302</b>	0.0005
T-13 Author turn began	0.0005	0.0002	0.0033	0.0011	0.0006	0.0031	<b>0.0466</b>	0.0015
T-14 Frames for coverage of Art news	<b>0.3896</b>	0.0027	0.0001	0.0254	0.0106	0.0013	0.0001	0.0019
T-15 Great Britain, WWII & uninformative topics	0.0003	0.0002	0.0029	0.0002	0.0024	<b>0.0471</b>	0.0005	0.0005
T-16 Study of the media effects of terror alerts	0.2091	0.0002	0.0010	<b>0.7197</b>	0.0103	0.0001	0.0003	0.0008
T-17 Films & meanings	<b>0.0706</b>	0.0012	0.0001	0.0018	0.0088	0.0004	<b>0.0001</b>	0.0002
T-18 Novels as bags of character names	0.0010	0.0116	0.0057	0.0002	0.0140	0.0147	0.0157	<b>0.0328</b>
T-19 Standard, relational, predicted, occurrences	<b>0.0212</b>	0.0002	0.0001	0.0133	0.0144	0.0001	0.0001	0.0015
T-20 Authors' Gender	0.0003	0.0119	0.0038	0.0015	0.0088	0.0004	0.0037	<b>0.5275</b>
T-21 Crime, banditry, unrest & rebellion	0.0002	0.0002	<b>0.6503</b>	0.0005	0.0062	0.0004	0.0001	0.0002
T-22 Communities of authors: Research on literary passages & demography journals	0.0003	0.0350	0.0048	0.0002	0.0013	<b>0.4525</b>	0.4417	0.0002
T-23 Honor, position & conscience	0.0009	0.0005	0.0079	0.0008	0.0002	0.0025	<b>0.0288</b>	0.0002
T-24 Banditry as an ontological question	0.0001	0.0002	<b>0.0477</b>	0.0002	0.0002	0.0101	0.0230	0.0002
T-25 Texts, meaning & national security	0.0023	0.0347	0.0008	0.0005	<b>0.6224</b>	0.0109	0.0037	0.0103
<b>Column Total</b>	<b>1.0006</b>	<b>.9999</b>	<b>-1.0001</b>	<b>1.0002</b>	<b>1.0002</b>	<b>.9995</b>	<b>1.0001</b>	<b>1.0004</b>
	x ≥ .25	.25 > x ≥ .10	.10 > x ≥ .020	RowLargest				

the identified paragraphs back into the articles leads us to suggest that the hybridity of this topic (the mixing of paragraphs from the DiMaggio and the Bonilla articles) reflects the fact that, beyond its first few words concerning the public terror alerts, Topic 16 is also capturing a broader, shared discussion on “media effects research.” In other words, the algorithm is finding common passages about the use of news-stories, studied statistically, that are linked to the study of public policy. When seen from this perspective, the connection between the DiMaggio and the Bonilla articles makes sense.

The DiMaggio article is also linked to Topic 17 (.0706). The most important terms here include: film, solutions, produced, films, museum, Hollywood, meanings, core, solution, percent, observed, appendix, cases, today, university, independent. This is an interesting case because the topic is capturing an extended discussion in the article (that continues into an Appendix) about how topic models respond to the difficult analytic problem of polysemy. DiMaggio, Nag and Blei explain that the word “film” is used in several of their topics but that the word has different meanings in the different thematic contexts, thereby helping to validate the power of the LDA method.<sup>8</sup> The last topic of any note is Topic 19, (with a probability of just .02). The first four words—standard, relational, predicted, occurrences—suggests a kind of mantra for the style of formalization being discussed in this article (and elsewhere—note that both the Mohr et al. and the Bonilla and Grimmer articles resonate with this theme).<sup>9</sup>

The McFarland, Ramage, Chuang, Heer, Manning and Jurafsky article reports on a stream of work that the group has published over the last few years employing various types of topic model, as well as other text mining methodologies. Their article provides another useful introduction to topic models by focusing on some alternative types and applications of the method. The main research described here analyzes 30 years of dissertation abstracts (1980–2010) drawn from the ProQuest database (of 240 U.S. research universities). They use topic models to identify intellectual streams in this data. A topic identifies constellations of words that co-occur inside the discourse of an intellectual sub-field. For example, looking just at the data from anthropology, their model identifies one topic (they label it “Archeology”), which is associated with these (stemmed) terms of art: site, archeology, period, region, evid, pattern, late, popul, settlement, materi, suggest, valley, earli, prehistory, etc. A different topic/discourse frame (labeled “Identity studies”) is defined by terms such as: ident, practice, discours, culture, nation, construct, global, etc. By observing the flow of these topics across the data, McFarland and colleagues are able to track the differentiation and blending of academic disciplines across time.

From [Table 2](#) we can see that words in the McFarland et al. article are likely to be distributed into just two main topics. More than a third of the words are sorted into the topic model topic (Topic 8). A bit more than half are sorted into Topic 5, which we have labeled “Computer models of language” to capture the main themes of their paper. This topic is defined by words such as: language, LDA, models, field, document, fields, labels, identified, knowledge, domains, label, Ramage, supervised, categories, work, identify, subject, labeled, applied, and anthropology. There are also low probabilities of words being sorted into Topic 22, “Communities of authors: Research on literary passages and demography journals,” and Topic 25, “Texts, meaning & national security” (both of which we will discuss presently).

---

<sup>8</sup> In each case, we have gone back to the actual texts (and textual contexts) and inspected the mapping of words and topics in order to facilitate our interpretation of these topic word lists. Readers can approximate this experience themselves, as they read along in the special issue, by searching for the particular word constellations described here.

<sup>9</sup> Note that the top terms for this topic includes the word “box,” which refers to George Box, the statistician, and also to a box in a figure containing topic model probabilities.

The next three articles focus on the use of topic models as a method of studying forms of state discourse. The paper by Miller uses topic models to analyze an archive of official notes sent to the Chinese emperor during the years of the Qing dynasty (1723–1911). At the time, as Miller explains, terminology for different categories of illegitimate public violence was fluid and the distinctions had great significance up and down the chain of command. Thus the official records on violent social episodes are ambiguous and difficult to interpret. Miller, in a very Foucault-like maneuver, uses topic models to sort through the corpus of “*zouzhe*” to identify constellations of words (actually, Chinese characters) that capture coherent classifications of imperial concern, including socially constructed categories of violent offenders.

As Miller puts it in this special issue, “Of the fifty topics in this model, six are clearly related to areas where the dynastic apparatus encountered illegitimate violence.” He labels these: crime, unrest, sedition, rebellion, border rebellion, and major rebellion. The topic of “crime” is defined by a constellation of terms that include: crime, case, punishment/sentence, board/ministry, try/interrogate, precedent/sub-statute and behead. Miller describes this as a topic that captures the regular pattern of court proceedings. It differs from the topic/social category of “sedition,” which is defined by words like: capture, investigate, confession, case, and teach religion. Miller explains that the latter describes a collection of crimes such as heresy, the printing of banned books and the cutting off of queues. Research applications for these measures are primarily historical. Miller graphs the frequency of these topics across time, and in doing so, he is able to contribute to our understanding of (at least) three specific issues—crime rates (and their reporting) during 18th–19th century China, the social processes affecting the cultural construction (and prosecution) of the crime of sedition, and the responses of the state to rebellions (especially during the later years of the Qing dynasty).

Looking at [Table 2](#), only three of our topics apply to Miller’s paper. The highest probability topic (.6503) is T21, “Crime, banditry, unrest & rebellion,” which well captures the main themes of Miller’s article. The second most important is the topic model topic at .2546. Third is T24, “Banditry as an ontological phenomenon” (.0477), which is an ambitious, engaging and clearly articulated sub-theme in Miller’s article (that interestingly, also resonates with a similar theme running through Tangherlini and Leonard’s article in this special issue).

Bonilla and Grimmer is the second article concerned with the study of state discourse practices. Their research concerns the Bush administration’s color-coded terror alert system and its effects on news coverage of terrorism and on public opinion more broadly. They sample front-page news-stories and nightly news broadcasts for the two days before, the day of and then two days after each of the terror alert escalation events. Like the DiMaggio paper, Bonilla and Grimmer primarily see topic models as a way to identify media frames in news-stories but they modify the standard LDA model by forcing every news-story to have just one topic. This allows them to more easily calculate when the terror alerts becomes a central focus. They search for twenty-four topics in their corpus which results in stories (or media frames) such as: “local small business,” “law and order,” “the Iraq war,” as well as one topic which is focused on terror alerts. This research design allows them to directly measure the effect of the terror alert announcements on the content of what is subsequently reported in the news. Then, by drawing on a series of surveys that happened to be conducted during the same time windows, they are able also to directly assess the impacts of the terror alerts on public opinion regarding matters such as economic expectations and support for Bush administration policy agendas.

In terms of our topic model, the Bonilla and Grimmer article is mostly focused on Topic 16 (.7197), “Studying the media effects of terror alerts” which captures the main thematic focus of

their article and is defined by words such as: alert, terror, public, percent, arts, attention, support, media, Bush, increase, stories, etc. (Note, this is the same topic that we looked at earlier because it overlaps with the DiMaggio et al. article). Next is the topic model topic (at .1574), followed by two topics that capture more specific sub-themes in the article, Topic 2 (.0441), “Predicting economic expectations,” and Topic 10 (.0300), “Anniversary result” (which looks at the effects of 9/11 anniversary commemoration events on news coverage and public opinion).

The last article in this section on the study of the state is by Mohr, Wagner-Pacific, Breiger and Bogdanov. Data come from a series of publications by the U.S. Office of the President regarding the National Security Strategy of the United States (1990–2010). The goal of the article is to better understand the rhetoric that is used by the U.S. state for describing and characterizing the strategic situation of the world (and the U.S. posture there). To do this, Mohr and colleagues draw upon the dramatistic theory of rhetoric developed by the literary theorist Kenneth Burke (a half century ago). Topic models are used here as part of a suite of text mining methods that are applied to measure the different elements within Burke’s theory of motives. Specifically, topic models are used to measure Burke’s concept of a scene, which he defined as the setting in which a dramatistic act occurs. For the Mohr et al. article, then, topic models are used to capture the kinds of thematic scenes that tend to re-occur again and again in the U.S. strategy discussions that unfold around global security. It is within dramatistic scenes that the other elements of Burke’s grammar of motives—the acts, actors, agencies and purposes—are combined and combusted. Terrorism is one such thematic scene that emerges from the corpus. Others topic-scenes include economic development, human rights, and military operations.

According to [Table 2](#), most of the Mohr et al. article is focused on just two topics—the topic model topic (.2201) and Topic 25, “Texts, meaning and national security” (.6224), which is defined by the following words: text, states, security, texts, united, scene, national, act, figure, meaning, motives, terms, semantic, documents, basic, terrorism, coding, strategy, acts, and automated. There is also some overlap (.0527) with the McFarland et al. article over the use of the “Computer models of language” theme.

The last three articles in the special issue have in common their use of topic models as a strategy for measuring academic and literary fields. In a wonderful paper reminiscent of a classic style of work in the sociology of knowledge, Emily Marshall compares two academic communities—one French, one British, all demographers—by analyzing the intellectual ideas that they use for constructing theories about the world. Specifically, she explores the differential embrace by the French of the theory of the “demographic revolution” in contrast to the British demographers’ commitment to the “demographic transition” theory. To test out the implications of this difference, Marshall collects all articles published between 1946 and 2005 on the subject of “fertility” (“fecondité” or “natalité” in French) in the British and French flagship demography journals. Then, using correlated topic models (CTM),<sup>10</sup> she identifies 75 topics (or intellectual frames) in both the French and in the British corpora. She hand codes these topics (by closely combing through the texts and passages identified with each topic) to discern which reflect high-fertility subjects (like family planning programs, a preoccupation of the British) or low-fertility subjects (like working mothers, a concern of the French). She then maps these categories of topics across time and context (supplemented with a second set of topic models of newspapers over the same period) to demonstrate the tangible persistence of intellectual

---

<sup>10</sup> In contrast to LDA models (which assume topics are not correlated across documents), CTM is a variation of topic models that assumes that topics are correlated across texts ([Blei and Lafferty, 2007](#)).

frameworks (or logics) in academic communities and the way in which those frames endure even in the face of demographic (e.g., objective) trends that challenge them.

In our model, the Marshall paper is split into two themes. Reflecting her attention to the question of how to find the best number of topics, the words in Marshall's article have a .4186 probability of being in the topic model topic. And the most important (.4525) is T22, which we have labeled "Community of authors: Research on literary passages and demography journals." What is most intriguing about this topic is that it is shared almost equally with Tangherlini and Leonard (.4417). After reading through the two texts and relevant (identified) passages again, we weren't that surprised.

The Tangherlini and Leonard article is the second in this section focused on measuring academic and literary fields. Like the McFarland et al. article, it reports on several topic modeling projects, in this case, three demonstrations of a procedure Tangherlini and Leonard call "sub-corpus topic modeling" (STM). Their idea is to take a small, well-understood corpus of texts and to use them to provide a training logic that can then be applied to larger, less well understood corpora in order to identify examples of textual passages containing similar literary forms. They describe this as a kind of targeted fishing expedition. In the article, Tangherlini and Leonard offer three different STM experiments.

First, they train their algorithm on two of Charles Darwin's books (*The Origin of Species* and *The Descent of Man*), which they use as "bait" to trawl through the Google Danish books corpus (1860–1920) looking for matches. What they "catch" is a splendid array of fish in which Darwinian ideas are woven into unexpected literary passages. As Tangherlini and Leonard explain, these borrowings are in no sense innocent because Darwin's ideas played a critical role in a bitter intellectual dispute waged in late 19th century Denmark between a progressive looking Naturalism (which admired Darwin's works and held them up as an ideal) and a reactionary Romanticism that held to a theocentric scientism and a conservative political vision. The focused trawl of the STM enables Tangherlini and Leonard to pull entirely unknown works of literature up for display and examination in a way that begins to fill in a much broader and non-canonical history of these intellectual movements.

In their second example, Tangherlini and Leonard again trawl for unknown authors, but this time they start out by training their algorithm on a collection of works by three canonical figures (Jacobsen, Schandorf and Drachman) in the Danish literary movement known as the Modern Breakthrough. The STM analysis enables Tangherlini and Leonard to locate a variety of (non-canonical) authors (often women) who were also early innovators in the literary form, all of whom had become lost to modern scholarship. And, finally, in what is perhaps their most intriguing experiment, Tangherlini and Leonard go one step beyond Propp (1958) by applying STM to search for the "feel of the rural." This time they trained their algorithm on 34,000 Danish folk tales and returned a series of topics such as "death and churchyards," "shooting and witches," "horses & wagons," "serpents," and "the minister" that turn up in all kinds of interesting ways across a range of other literary genres.

In terms of our analysis, Tangherlini and Leonard are unusual for being the source of so many unique topics (T1, T3, T4, T6, T9, T12, T13, and T23). Some of this reflects the fact that the article is broken into a series of three discrete experiments, each with its own scholarly context and set of intellectual problems (but we suspect we may also be encountering variations in the sensitivity of topic models to different intellectual and rhetorical styles, especially with a small corpus such as this). For Tangherlini and Leonard, T22 (which is the theme shared with Marshall) is also their most important topic (.4417). The hybridity between the two articles was a surprise at first, though it makes sense when we recognize that both articles use topic models to identify

communities of authors who share common ways of writing and common styles of thinking, that both articles focus on national communities of authors, and that both address matters concerning alternative styles of scientific thought. Beyond this, however, they are substantively far apart and this is apparent in the key words that are braided together here—fertility, literary, demographic, passages, British, Breakthrough, journal, research, population, work, articles, Danish, modern, literature, French, works, Darwin, authors, language, academic, etc.

The last article in this section (and the last in the special issue) is by Matthew Jockers and David Mimno. More than the other articles collected here, Jockers and Mimno are especially focused on calibrating the methodology. Using a clear and intellectually precise research design, they take a corpus of over 3000 British, American and Irish novels (published between 1750–1899) and sort them into three groups—novels with male authors, novels with female authors and those by authors of unknown gender. Topic models identify coherent literary themes across the corpus (they ask for 500 topics), and then Jockers and Mimno explore the ways in which the gender of an author affects the selection of particular themes. Rather than just looking at simple distributions, however, Jockers and Mimno develop a series of formal assessments—a permutation test, a bootstrap test and a classification test—to assess the reliability of inferences from meta-data for all of these kinds of models. Table 2 shows that the most important themes for this paper are T20 (.5275), which we have labeled “Author’s gender” and T11, “Themes, authors, gender.”

Fig. 1 summarizes the information we have presented so far. We have constructed a graph of the articles with arcs drawn to represent shared topics (excluding T8, which was shared by all the articles). With the exception of T22, whenever one article has a higher proportion of a given topic, we have represented this with an asymmetric arrow. At the dyad level it is interesting to see

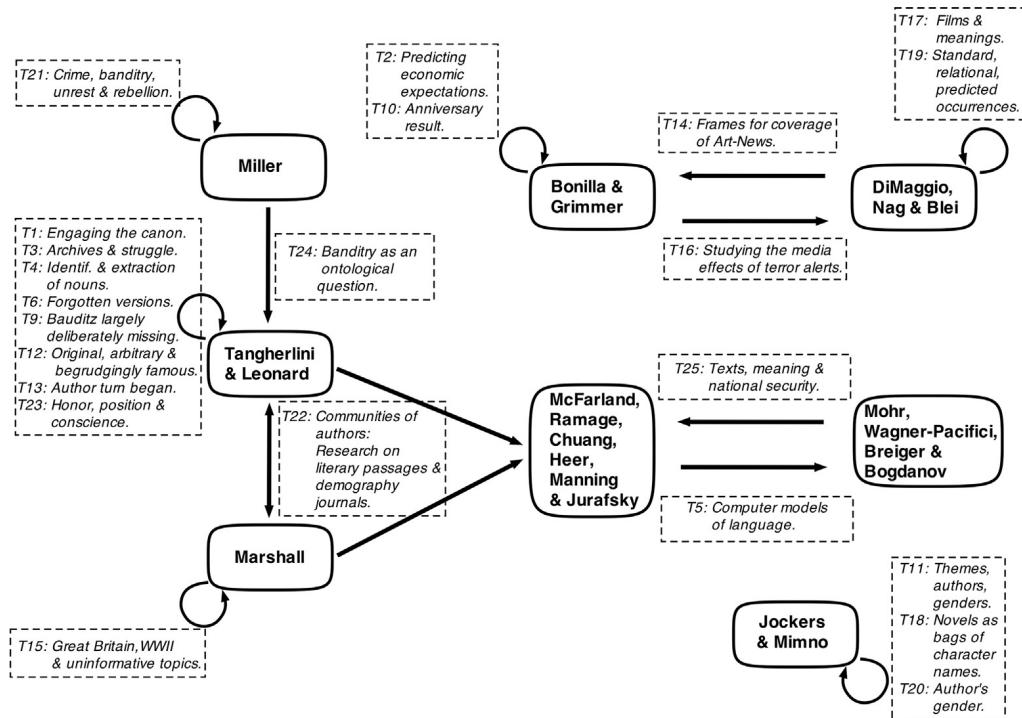


Fig. 1. Summary of shared topics among the articles published in *Poetics*, Vol. 41, no. 6.

that in the both the pairings of Bonilla/DiMaggio and McFarland/Mohr., there is a balance to the topic sharing (with each article sharing its main topic with the other article). It is also useful to see the overall mapping of the articles. The Tangherlini/Marshall/McFarland articles are tied together around the communities of authors topic. The Bonilla/DiMaggio pair shares a common focus on newspaper frames and media effects. The McFarland/Mohr dyad has a common focus on applying a range of text-mining tools to tackle problems in the social sciences.

### 3. Some demands, dilemmas and limitations of the method

Content analysis should begin where traditional modes of research end. The man who wishes to use content analysis for a study of the propaganda of some political party, for example, should steep himself in that propaganda. Before he begins to count, he should read it to detect characteristic mechanisms and devices. He should study the vocabulary and format. He should know the party organization and personnel. From this knowledge he should organize his hypotheses and predictions. At this point, in a conventional study, he would start writing. At this point, in a content analysis, he is, instead, ready to set up his categories, to pretest them, and then to start counting.

Harold Lasswell et al. (1952, p. 65),  
*The Comparative Study of Symbols: An Introduction*

The most common complaint that is heard about topic models is that they rely upon the “bag-of-words” assumption, disregarding the order of words within a text (Meeks and Weingart, 2012). To many, it seems hard to believe that one can discard all of that critical information and not be left with a severely hobbled analysis of meaning. While this is true in some literal sense, it strikes us as being a critique that misses the point because the real genius of topic models is precisely that, for this specific type and level of meaningful content, it appears as though relationality trumps syntax. It turns out that you *can* remove all of that other information from the analysis and still get robust results. So a better question to ask is what sort of a trade-off shall we make in terms of surrendering this more localized (syntactic) information in order to realize gains in information of the sort that topic models can afford us?

But this does point to why topic models will be good for some kinds of meaning measurement projects but will be a poor choice for others. So, for example, as scholars from a variety of disciplines have now demonstrated, narratives can be very usefully modeled as tie-based networks—for example, see Bearman and Stovel (2000) for a network analysis of Nazi life-stories, Franco Moretti’s (2013) network analysis of Shakespearian plays, or even the work of David Herman (2004) who develops a logic model for narratives. All of this suggests that topic models, since they have discarded this type of localized relational information, would be much less useful for studying narrativity.<sup>11</sup>

To us, a more worrisome concern that has also been expressed is the utter simplicity of topic models as a textual analysis method (Grimmer and Stewart, 2013; Schmidt, 2012). One might be forgiven for imagining that one needs nothing more than a text to analyze and a copy of a software program like *Mallet* to produce brilliant cultural research. In truth, as with any scholarly pursuit, the quality of the knowledge about the case and the clarity of thinking about the phenomena determine the utility and the richness of the analysis regardless of the sophistication

---

<sup>11</sup> Then again, we might have also predicted the same would be true of a genre such as poetry, but Rhody (2012) finds differently.

of the methods employed. With topic models, researchers are responsible for knowing enough about the phenomena under investigation to be able to understand what the discourse field is about. They must pick a corpus that has substantively meaningful content within the field under investigation and be familiar enough with that corpus to have a good sense of how the text reads and how its contents will address the analytic problem at hand. Moreover, researchers need to be able to make sense of the topic word clusters that are produced by the algorithm and to be able to recognize when a set of topics are worthless or misleading (because, perhaps, there are no well organized topics in the corpus or because the number of topics asked for by the researcher doesn't match the actual number of topics in the corpus, etc.), and when the topics are indeed capturing word clusters that makes good sense, to a well informed observer (a subject-area specialist) who understands the discursive context of the corpus.<sup>12</sup> Of course, there are also technical requirements—analysts need to prepare the text (this might include, for example, removing stop-words, etc.) and to meet all the formal assumptions of the model.<sup>13</sup> Researchers must also interpret the topic model output, probably iteratively, so that a best fit can be found between the number of topics and an overall level of interpretability. And finally, all of this new topic model data must be fitted into a well-informed, explanatory or substantively meaningful analysis of the social phenomena under investigation.

Seen in this light, it is useful to think about topic models not as providing an automatic text analysis program but rather as providing a lens that allows researchers working on a problem to view a relevant textual corpus in a different light and at a different scale. In this special issue, DiMaggio, Nag and Blei use this metaphor and, as they note, it is a frame with its roots deep in the analytic process itself, “(f)inding the right lens is different than evaluating a statistical model based on a population sample. The point is not to estimate population parameters correctly, but to identify the lens through which one can see the data most clearly.”

One implication is that well informed interpretive work—hermeneutic work—is still required in order to read and to interpret the meanings that operate within a textual corpus, even when one is peering through the lens of a topic model. It is not the need for a deep understanding of one's textual corpus that has changed, it's the place where this style of knowledge comes into play. We began this section with a quote from Lasswell et al. (1952) about the importance of obtaining a deep historical and contextual understanding of one's corpus before beginning to count. With topic models, this is inverted. One counts, and then one begins to interpret. In this sense, what topic models and other types of automated text analysis tools do for cultural researchers is to shift the locus of subjectivity within the methodological program—interpretation is still required, but from the perspective of the actual modeling of the data, the more subjective moment of the procedure has been shifted over to the post-modeling phase of the analysis.<sup>14</sup>

---

<sup>12</sup> Some progress is being made on developing more formal decision rules for goodness of fit of different levels of topic. Emily Marshall presents some ideas about this in her contribution to this special issue. See also the suggestions in this special issue by Bonilla and Grimmer.

<sup>13</sup> Another assumption is that documents within the corpus are independent—i.e., each text is generated by the author without the knowledge/reference or temporal dependencies to the rest of the documents in the corpus. While addressing such assumptions is essential in certain analysis endeavors, they also keep the models simple and general and do not require expert/human knowledge or additional information.

<sup>14</sup> Topic models are certainly not the first to do this. The social sciences have used many such types of methods in the past—starting back with Lazarsfeld's theory of latent factor analysis, and the use of methods such as factor analysis, LISREL, multi-dimensional scaling or Multiple Correspondence Analysis (MCA), as used in Europe.

#### 4. Why do topic models matter?

Content analysis will not tell us whether a given work is good literature; it *will tell us* whether the style is varied. It will not tell us whether a paper is subversive; it *will tell us* if its contents change with the party line. It will not tell us how to convince the Russians; it *will tell us* what are the most frequent themes of Soviet propaganda.

Harold Lasswell et al. (1952, p. 45),  
*The Comparative Study of Symbols: An Introduction*

Topic models matter for a lot of reasons. Most obviously, they matter because they provide a way for researchers to obtain reasonable automated content coding of large text corpora. As social and cultural scientists become increasingly engaged with what is now being called “Big Data”—large-scale data streams taken from the Internet, social media sites, or archives like Google books—having tools that scale becomes increasingly important. Thus, topic models matter because they enable us to take the measure of large-scale social phenomena that we could not have previously been able to do. Whether our goal is to study attitude change in twitter feeds (Ramage et al., 2010) or genre shifts in literary fields (Jockers, 2013; Moretti, 2013), topic models matter because they enable researchers to study phenomena of the sort that can only be viewed through a macroscopic lens.

But topic models also matter because they can be used for viewing small-scale text corpora. In this special issue’s article by Mohr, Wagner-Pacificci, Breiger and Bogdanov, topic models are one of three text analysis methodologies that are combined to study a relatively small corpus (of a half million words) that is well within reach of a traditional “close reading” by experts in hermeneutics and relevant subject areas. Here, formalization supplements (rather than displaces) a close reading of the corpus. So, again, topic models matter because they provide new lenses for new projects.

As we have sought to highlight here, topic models also matter because they facilitate a fundamental shift in the locus of methodological subjectivity—from pre-counting to post-counting. This is another major reason why topic models matter, and just to emphasize this quality, we turn to one last set of examples that can usefully illustrate this contribution. This comes from a study of a corpus of 20,000 newspaper editorials sampled from ten major newspapers (in five countries) over a sixty-year period. Researchers identify a number of topics but we will focus here on just two. The first is “International Violence.” It is defined by the terms: war, combat, battle, weapons, enemy, front, trench, foxhole, prisoners, soldiers. The next is labeled “Domestic Violence” and it includes the terms: riot, murder, strike, disorder, pickets, suicide, prison, jail, lynching, gangs (Lasswell et al., 1952, p. 68).

The example comes from the research done by Harold Lasswell and his colleagues at Stanford in the years just after World War II. The project had the goal of identifying “trends in the key symbols of modern politics” between the years of 1890 and 1945 (Lasswell et al., 1952, p. iii).<sup>15</sup> No computers were used to identify these topics. Instead, using methodologies that Lasswell had

<sup>15</sup> In his classic essay “Why be quantitative?” Lasswell (1949) describes his frustrations at seeing so many otherwise interesting and important detailed analyses of texts that were nonetheless suspect precisely because “...we are left in the dark about why he quotes one paper one day or week and omits it the next time. Even if we assume that his judgment is good, it is permissible to ask if such arbitrary selection procedures create a properly balanced picture, or whether they result in special pleading based, if not on deliberate deception, then on unconscious bias” (Lasswell, 1949, p. 44). Rogers (1994) has a useful review of Lasswell’s career.

pioneered as director of “the experimental division for the study of wartime communications, established at the Library of Congress during World War II” (Lasswell and Leites, 1949; Lasswell et al., 1952, p. 40), he and his colleagues assembled a team of human coders and (to insure consistent coding of the text), they created a coding protocol (they called it a rulebook) that channeled the interpretive focus of the coders down to a narrow range of explicitly pre-considered choices and clear decision rules.

Their method was highly dependent upon adequate pre-specification of the “key symbols” that were to be coded. Lasswell divided these into three types—those referring “to persons or groups (symbols of identification), to preferences and volitions (symbols of demand), and to the assumption of facts (symbols of expectation)” (Lasswell et al., 1952, p. 15). Coders were instructed to check each editorial for “the presence of any of 416 symbols which constituted our symbol list. Of these, 206 were the names of national or similar units: countries, national minorities, continents, etc.; and 210 were key symbols of the major ideologies which have been contending in world politics during the past half-century. These included, to cite the ‘N’s’ as an example, Nationalism, Nazism, Neutrality, and Nonintervention” (Lasswell et al., 1952, p. 43).

As we have seen in the passage cited earlier, Lasswell and his colleagues worried a lot about the processes, time and effort that went into establishing their lists of key symbols. They had good reason to worry. Once the coding categories were negotiated, pre-tested, written into the rulebook and the team had begun to code the corpus, there was no going back—no chance to re-code, re-compile or re-run. This meant that Lasswell and his team had to establish a deep knowledge of the case and do so well before the counting began. To do this, they crafted a six-step process that culminates in “...constructing our tentative symbol list. This we can do partly on the a priori basis of our reflections on the past and present, partly on the empirical basis of our preliminary scrutiny of the media to be analyzed” (Lasswell et al., 1952, p. 68).<sup>16</sup>

But to what end? After orchestrating a text analysis project of this magnitude and precision, what did Lasswell and his colleagues want? Interestingly, it seems that what they really wanted was something like a topic model. Consider the project described here. We noted that Lasswell started by having his coders track 416 key symbols. On closer inspection, we find that the key symbol list begins to look a lot like processed lists of terms in which stop-words have been removed, stemming has occurred, and synonymous terms have been collapsed together.<sup>17</sup> And once the corpus had been coded according to this scheme, what next? On this, Lasswell and colleagues are clear. Although they lament the lack of viable theoretical models for understanding how to model idea structures, they insisted on the importance of advancing on the problem with empirical research,

Today we have no models at all and, therefore, no basis for predicting how symbols will behave under specified conditions...We do have some models of attitude formation, propaganda effects, and ideological behavior...But we should not confuse theories about

---

<sup>16</sup> The six steps are as follows: “First, decide which segments of the population we wish to test for this particular change in symbolic behavior...Next, select...a representative medium of symbolic behavior...Third...estimate roughly the period to be covered...we should next set up a tentative scheme of periodization...fifth...state our hypothesis with sufficient definiteness to enable us to construct the list of symbols which would index it...With these propositions before us, we can take the sixth step of constructing our tentative symbol list” (Lasswell et al., 1952, pp. 67–68).

<sup>17</sup> Indeed, those techniques, matched with a named entity recognition (NER) program might be able to provide a list of key symbols that come pretty close to what Lasswell was after. The Mohr, Wagner-Pacificis, Breiger and Bogdanov article (in this special issue) explains NER processors in more detail.

ideas with theories about symbols. . . Ideas are expressed by symbols. Their manifest form is nothing more than a conglomeration of symbols. . . We need models of how symbols operate to produce the configurations called ideas, attitudes, ideologies. . . And our knowledge of symbolic behavior can be advanced only if we learn how ideas take form out of the symbolic elements through which they are expressed. (Lasswell et al., 1952, pp. 64–65)

In other words, for Lasswell and his colleagues, the reason to have teams of human coders track 416 symbols across ten newspapers (from five countries across sixty years) was so that these data could be used to identify larger structures of meaning that could then be linked back to broader research agendas.

If the list of symbols is sufficiently extensive, it will be found that groups of symbols follow common patterns. It will be possible, in other words, to apply a sort of factorial analysis to the list, which one will find that the large number of symbols occurring do not each represent an independent variable, but that groups of symbols form constellations, certain words appearing together. The independent factor is an idea to which the group of symbols refers and whose fluctuations it indexes. (Lasswell et al., 1952, p. 55)<sup>18</sup>

Thus, for Lasswell and his team members, the goal was to find ways to measure *ideas* which were latent constructs indexed by constellations of word *symbols*. They worked at developing a model for capturing this process, but they didn't succeed. "Criteria for the validity of a list may now be stated more formally, although the statistical working out of the procedure remains to be done" (Lasswell et al., 1952, p. 56). They concluded somewhat optimistically,

Symbolic behavior seems to be prone to factorial analysis, since a limited number of unit ideas fall into complex constellations. It seems unlikely that the probability of the appearance of symbols with respect to each other and over time could be represented by a regular surface. We hope that statisticians will address themselves to testing these hunches and resolving some of these problems. (Lasswell et al., 1952, p. 57)

Too bad for Lasswell, he was born half a century too early to be able to make use of LDA models to analyze his corpus. What then did he and his colleagues do? They used their best guesses. "It should be noted that, for certain studies, an *a priori* list remains most appropriate." And so, in fact, our last set of topic models examples—"International Violence" and "Domestic Violence"—were not models at all, but a set of best guesses about how to go about assembling a set of index measures by hand from a dataset about which the researchers already know a great deal. After having gone through six steps of preparation, Lasswell and his colleagues write, "we can quickly think up several dozen symbols which most Americans associate with violence. Here are two groups which occurred to these writers, by free association, within a few minutes" (Lasswell et al., 1952, p. 68).

What is striking to us today is just how much Lasswell, at the very beginning of the modern field of content analysis, began with the goal of assembling a set of text analysis measures that end up looking a lot like what topic models deliver. In that sense, we might say the creators of topic models have stepped up to Lasswell's challenge. But of course, topic models do a lot more than solve Lasswell's statistical problem. In fact, with topic models the entire process of creating

---

<sup>18</sup> Lasswell et al. go on to sketch the basic model of a latent factor analysis model for content analysis by drawing on Lazarsfeld's ideas about latent factor analysis in survey research.

the code lists, writing the rulebook as well as the actual coding of the corpus itself are replaced by a set of automated algorithmic procedures. One might say this puts content analysis back on an equal footing with traditional modes of scientific research, no longer must content analysis start where all other methods end (as Lasswell had warned). But, as we have also tried to emphasize here, topic models do not remove the scholarly or the hermeneutic work from the project of analyzing a textual corpus, topic models simply move the bulk of this labor over to the other side of the data modeling procedure. And so one last way that topic models matter is that they—in this sense—represent something of a symbolic ending to a first chapter in the history of content analysis methodologies and the beginning of another.

## 5. Conclusion

...the amount of wasted effort will be much less with adequate preparation of the sort we recommend.

Harold Lasswell et al. (1952, p. 66),  
*The Comparative Study of Symbols: An Introduction*

In this essay, we have described what we see as the important features of topic models for scholars in the social and cultural sciences who might want to use this method. Of course, there are also limitations and caveats (and we have discussed some of those here), but it is clear that to the extent that topic models prove to be an effective way of coding the meanings inside text corpora, then these are methods that can provide a way to analyze texts (including “Big Data” texts) that is substantively quicker, more efficient and more objective than traditional methods of content analysis in the social and cultural sciences.

We have gone to some lengths to trace out the parallels between the work of the founder of modern content analysis methods, Harold Lasswell, and new developments in this emerging field of contemporary topic modeling. While the two projects may have initially seemed quite far removed from one another, we have sought to demonstrate that they, in fact, are perfect bookends to one period in the history of modern content analysis methodologies, a period that got its start (as so many other modern social scientific methodological programs) in the crucible of applied social science during World War II.<sup>19</sup>

What Lasswell and his colleagues initially invented as a set of procedures for human beings has now been fully supplanted by a set of algorithms. And though this really does—in a profound sense—change everything, many of Lasswell’s precautions and concerns remain with us still. We still need to have learned well about the case. We still need to think clearly and analytically about the connections between the measure of textual content and the way in which these measures articulate into other types of social structures. And in this, notice that the ambition of content analysis researchers continues unabated. Lasswell and his colleagues saw this as the real end and ultimate goal of content analysis, and they described this style of work as “interaction analysis.” They write, “The aim of interaction analysis is to associate the flow of symbols directly with the flow of events. In the ideal case, fluctuations with respect to a single type of event... could be correlated with fluctuations in treatment of a single type of symbol” (Lasswell et al., 1952, p. 38). They also warn us, “This is the most difficult use of content analysis” (Lasswell et al., 1952, p. 38). We agree, but we also believe that, ultimately, the goal of modern content analysis should be

---

<sup>19</sup> Mohr and Rawlings (2010) discuss some of the ways that other formal models of culture emerged during this historical period. See also Platt (1996) for a broader historical review.

to emphasize this very kind of interaction (or duality) analysis and, in so doing, to help to rebalance the social sciences, by bringing the formal study of culture and meaning back into some form of parity with the quantitative study of social structures and material logics that have generally been ascendant since about the time that Lasswell and his colleagues were writing ([Mohr, 1998](#)).

### Acknowledgements

We thank Timothy Dowd for his superb editorial support both on this essay as well as on all of the articles collected in this special issue (and for his patience). We also would like to thank Paul DiMaggio for his reading of the article and the members of the Department of Sociology at the University of Barcelona, where we presented an early version of this paper. Special thanks to José A. Rodríguez, José Luis C. Bosch, Liliana Arroyo and Jesús de Miguel for their useful feedback and suggestions.

### Appendix. Additional information of the articles in the special issue

See [Table A.1](#).

Table A.1  
Top 8 words per topic.

		W-1	W-2	W-3	W-4	W-5	W-6	W-7	W-8
<i>Part 1</i>									
T-1 Engaging the canon	Representative	Understood	Showing	Canonical	Engaging	Gennembruds	Bugge	Accepts	
	0.008605	0.007753	0.006049	0.005197	0.004345	0.004345	0.003493	0.003493	
T-2 Predicting economic expectations	Expectations	Sufficient	Manipulations	Focuses	Constant	Overlap	Uncontroversial	Limited	
	0.006762	0.006762	0.005928	0.005928	0.005093	0.005093	0.004258	0.004258	
T-3 Archives and struggles	Struggle	Archive	Collections	Revealed	Position	Descriptions	Captured	Feel	
	0.006889	0.005188	0.004337	0.004337	0.004337	0.004337	0.004337	0.004337	
T-4 Identification and extraction of nouns	Identification	Remove	Extract	Nouns	Understood	Live	Skram	Health	
	0.005878	0.005878	0.003951	0.003951	0.003951	0.003951	0.003951	0.002987	
T-5 Computer models of language	Language	Ida	Models	Field	Document	Fields	Labels	Identified	
	0.020458	0.012451	0.011650	0.010850	0.008848	0.008848	0.008447	0.006446	
T-6 Forgotten versions	Forgotten	Versions	Closely	Half	Independence	Budgets	Interpretive	Pair	
	0.006375	0.005477	0.003681	0.003681	0.003681	0.003681	0.003681	0.003681	
T-7 earlier efforts	Existing	Hundreds	Efforts	Earlier	Unrest	Publication	Accepted	Slightly	
	0.008785	0.006855	0.004924	0.004924	0.003958	0.003958	0.003958	0.003958	
T-8 topic models	Topic	Topics	Words	Model	Analysis	Corpus	Time	Texts	
	0.059637	0.051214	0.026614	0.020812	0.015402	0.015178	0.012278	0.012222	
T-9 Bauditz largely deliberately missing	Bauditz	Missing	Largely	Deliberately	Urban	Contents	Small	Aspects	
	0.009259	0.006757	0.005923	0.005088	0.005088	0.004254	0.004254	0.004254	
T-10 Anniversary result	Examining	Result	perspective	Anniversary	Ensuring	Capture	Fall	Interpretation	
	0.007472	0.006651	0.006651	0.005830	0.005009	0.005009	0.005009	0.004188	
T-11 themes, authors, gender	Themes	Thematic	Corpus	Author	Gender	Work	Century	Theme	
	0.031632	0.019601	0.014588	0.012583	0.011580	0.011079	0.010076	0.008572	
T-12 Original, arbitrary and begrudgingly famous	Original	Instinct	Arbitrary	Begrudgingly	Famous	Lens	Tales	Serpents	
	0.005855	0.005855	0.004895	0.003935	0.003935	0.003935	0.003935	0.003935	
T-13 Author turn began	Author	Turn	Began	Represented	Interests	Suggest	Mechlenburg	Moretti	
	0.011393	0.004436	0.004436	0.003566	0.003566	0.003566	0.003566	0.003566	

*Part 2*

	Topic	Arts	Assigned	Times	Nea	Art	Coverage	Grants
T-14 frames for coverage of art news	0.048580	0.027056	0.016984	0.015467	0.011741	0.009672	0.009396	0.008016
T-15 Great Britain, WWII and uninformative topics	Frequently	Great	WWII	Discuss	Treated	Program	Uninformative	Probability
T-16 studying the media effects of terror alerts	0.006631	0.005697	0.005697	0.004763	0.004763	0.004763	0.004763	0.004763
T-17 Films and meanings	Alert	Terror	Alerts	Public	Percent	Arts	Attention	Support
T-18 novels as bags of character names	Film	Solutions	Produced	Films	Museum	Hollywood	Meanings	Core
T-19 standard, relational, predicted, occurrences	Names	Novels	Frequently	Character	Bag	Reach	Fiction	Influence
T-20 Authors' Gender	0.017345	0.007584	0.006833	0.006833	0.006833	0.006082	0.005331	0.005331
T-21 Crime, banditry, unrest and rebellion	Standard	Relational	Predicted	Occurrences	London	Contributes	Confidence	Box
T-22 communities of authors: Research on literary passages and demography journals	0.007113	0.005357	0.005357	0.004478	0.004478	0.003600	0.003600	0.003600
T-23 Honor, position and conscience	Female	Male	Authors	Figure	Data	Topic	Novels	Word
T-24 Banditry as an ontological question	0.029062	0.027337	0.026475	0.023025	0.021588	0.020726	0.017276	0.015551
T-25 Texts, meaning and national security	Rebellion	Crime	Unrest	Violence	State	Records	Bandits	Major
	0.029128	0.019148	0.018317	0.017277	0.011872	0.009585	0.009169	0.008753
	Fertility	Literary	Demographic	Passages	British	Breakthrough	Journal	Research
	0.020368	0.012460	0.011006	0.010188	0.010097	0.009916	0.009916	0.009825
	Honor	Societal	Corpus	Position	Conscience	Original	Precise	Tested
	0.009242	0.006497	0.005582	0.004667	0.004667	0.003752	0.003752	0.003752
	Textual	Historical	Phenomenon	Robber	Concept	Accounts	Bandit	Ontological
	0.012484	0.011105	0.008346	0.006966	0.006277	0.005587	0.005587	0.005587
	Text	States	Security	Texts	United	Scene	National	Act
	0.019047	0.013456	0.013456	0.011965	0.010474	0.009356	0.009356	0.008238

## References

- Bail, C.A., forthcoming. Measuring culture with big data. *Theory and Society* 43.
- Bearman, P., Stovel, K., 2000. *Becoming a Nazi: a model for narrative networks*. *Poetics* 27, 69–90.
- Blei, D.M., 2011. Introduction to Probabilistic Topic Models. Computer Science Department, Princeton University <http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf> (accessed 15.11.13).
- Blei, D.M., 2012a. Topic modeling and digital humanities. *Journal of Digital Humanities* 2 (1) 8–11.
- Blei, D.M., 2012b. Probabilistic topic models. *Communications of the ACM* 55 (4) 77–84.
- Blei, D.M., Lafferty, J., 2006. Dynamic topic models. In: International Conference on Machine Learning, ACM, NY, pp. 113–120.
- Blei, D.M., Lafferty, J.D., 2007. A correlated topic model of Science. *Annals of Applied Statistics* 1 (1) 17–35.
- Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022.
- Brett, M.R., 2012. Topic modeling: a basic introduction. *Journal of Digital Humanities* 2 (1) 12–16.
- Chang, J., Blei, D.M., 2010. Hierarchical relational models for document networks. *Annals of Applied Statistics* 4 (1) 124–150.
- Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R., 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41 (6) 391–407.
- Diesner, J., Carley, K.M., 2005. Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. In: Narayanan, V.K., Armstrong, D.J. (Eds.), *Causal Mapping for Research in Information Technology*. Idea Group Publishing, Hershey, PA, pp. 81–108.
- Diesner, J., Carley, K.M., 2008. Conditional random fields for entity extraction and ontological text coding. *Journal of Computational and Mathematical Organization Theory* 14, 248–262.
- Diesner, J., Carley, K.M., 2010. A methodology for integrating network theory and topic modeling and its application to innovation diffusion. In: Proceedings of Social Computing (SocialCom). IEEE Second International Conference on Social Computing, Minneapolis, MN, pp. 687–692.
- Griffiths, T., Steyvers, M., 2004. Finding scientific topics. *National Academy of Sciences* 101 (Suppl. 1) 5228–5235.
- Griffiths, T., Steyvers, M., Blei, D.M., Tenenbaum, J., 2005. Integrating topics and syntax. In: Saul, L.K., Weiss, Y., Bottou, L. (Eds.), *Advances in Neural Information Processing Systems* 17. MIT Press, Cambridge, MA, pp. 537–544.
- Grimmer, J., 2010. A Bayesian hierarchical topic model for political texts: measuring expressed agendas in senate press releases. *Political Analysis* 18 (1) 1–35.
- Grimmer, J., King, G., 2011. General purpose computer-assisted clustering and conceptualization. In: Proceedings of the National Academy of Sciences, <http://dx.doi.org/10.1073/pnas.1018067108>.
- Grimmer, J., Stewart, B., 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political documents. *Political Analysis* 21 (3) 267–297.
- Herman, D., 2004. *Story Logic*. University of Nebraska Press, Lincoln, NB.
- Hofmann, T., 1999. Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-99), Morgan Kaufmann, San Francisco, pp. 289–296.
- Jelisavcic, V., Furlan, B., Protic, J., Milutinovic, V., 2012. Topic models and advanced algorithms for profiling of knowledge in scientific papers. In: MIPRO, 2012 Proceedings of the 35th International Convention. pp. 1030–1035.
- Jockers, M.L., 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, Urbana, IL.
- Kaplan, S., Vakili, K., 2012. Breakthrough Innovations: Using Topic Modeling to Distinguish the Cognitive from the Economic. Rotman School of Management, University of Toronto (unpublished manuscript).
- Lasswell, H.D., 1949. Why be quantitative? In: Lasswell, H.D., Leites, N. (Eds.), *Language of Politics: Studies in Quantitative Semantics*. George W. Stewart, NY, pp. 40–52.
- Lasswell, H.D., Leites, N. (Eds.), 1949. *Language of Politics: Studies in Quantitative Semantics..* George W. Stewart, NY.
- Lasswell, H.D., Lerner, D., Pool, I.d.S., 1952. *The Comparative Study of Symbols: An Introduction*. Stanford University Press, Palo Alto, CA.
- McNamara, D.S., 2010. Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science* 3 (1) 3–17.
- Meeks, E., Weingart, S., 2012. The digital humanities contribution to topic modeling. *Journal of Digital Humanities* 2 (1) 1–6.
- Mohr, J.W., 1998. Measuring meaning structures. *Annual Review of Sociology* 24, 345–370.
- Mohr, J.W., Rawlings, C., 2010. Formal models of culture. In: Hall, J., Grindstaff, L., Lo, M.-C. (Eds.), *A Handbook of Cultural Sociology*. Routledge, New York, pp. 118–128.

- Moody, J., Light, R., 2006. A view from above: the evolving sociological landscape. *American Sociologist* 37 (2) 67–86.
- Moretti, F., 2013. *Distant Reading*. Verso, London.
- Mutzel, S., 2012. Newness and Collaborative Category Construction from Stories. Social Science Research Center Berlin (unpublished manuscript).
- Osgood, C.E., Suci, G., Tannenbaum, P., 1957. *The Measurement of Meaning*. University of Illinois Press, Urbana, IL.
- Newman, D., Hagedorn, K., Chemudugunta, C., Smyth, P., 2007. Subject metadata enrichment using statistical topic models. In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, ACM, Vancouver, BC, pp. 366–375.
- Platt, J., 1996. *A History of Sociological Research Methods in America, 1920–1960*. Cambridge University Press, Cambridge, UK.
- Propp, V., 1958. *Morphology of the Folktale*. University of Texas Press, Austin, TX.
- Ramage, D., Dumais, S., Liebling, D., 2010. Characterizing microblogs with topic models. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence, pp. 130–137.
- Rhody, L., 2012. Topic modeling and figurative language. *Journal of Digital Humanities* 2 (1) 19–35.
- Rogers, E., 1994. *A History of Communication Study: A Biographical Approach*. The Free Press, New York, NY.
- Saussure, F., 1959. *Course in General Linguistics*. McGraw-Hill, New York.
- Schmidt, B.M., 2012. Words alone: dismantling topic models in the Humanities. *Journal of Digital Humanities* 2 (1) 49–65.
- Teh, T., Jordan, M., Beal, M., Blei, D., 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101 (476) 1566–1581.
- Wallach, H., 2006. Topic modeling: beyond bag of words. In: Proceedings of the 23rd International Conference on Machine Learning, ACM, New York, pp. 977–984.

**John W. Mohr** (Ph.D., Yale University) is Professor in the Department of Sociology at the University of California, Santa Barbara and the Director of the UCSB Social Science Survey Research Center. He has long been interested in using formal methods to analyze texts. His work can be seen at [www.soc.ucsb.edu/ct](http://www.soc.ucsb.edu/ct).

**Petko Bogdanov** is a Postdoctoral Researcher at University of California at Santa Barbara. He received his B.Eng. from Technical University of Sofia, Bulgaria and his M.S. and Ph.D. degrees from University of California at Santa Barbara. His current research interests are in network science and database and data mining methods, with a focus on graph data arising in social networks, biology and the humanities.

John W. Mohr\*

*Department of Sociology, 3103 Social Sciences & Media Studies,  
University of California Santa Barbara, Santa Barbara, CA 93106-9430, USA*

Petko Bogdanov

*Department of Computer Science, University of California Santa Barbara,  
CA 93106-5110, USA*

\*Corresponding author E-mail addresses: [mohr@soc.ucsb.edu](mailto:mohr@soc.ucsb.edu) (J.W. Mohr)  
[petko@cs.ucsb.edu](mailto:petko@cs.ucsb.edu) (P. Bogdanov)

---

# Reading Tea Leaves: How Humans Interpret Topic Models

---

**Jonathan Chang \***  
Facebook  
1601 S California Ave.  
Palo Alto, CA 94304  
jonchang@facebook.com

**Jordan Boyd-Graber \***  
Institute for Advanced Computer Studies  
University of Maryland  
jbg@umiacs.umd.edu

**Sean Gerrish, Chong Wang, David M. Blei**  
Department of Computer Science  
Princeton University  
{sgerrish, chongw, blei}@cs.princeton.edu

## Abstract

Probabilistic topic models are a popular tool for the unsupervised analysis of text, providing both a predictive model of future text and a latent topic representation of the corpus. Practitioners typically assume that the latent space is semantically meaningful. It is used to check models, summarize the corpus, and guide exploration of its contents. However, whether the latent space is interpretable is in need of quantitative evaluation. In this paper, we present new quantitative methods for measuring semantic meaning in inferred topics. We back these measures with large-scale user studies, showing that they capture aspects of the model that are undetected by previous measures of model quality based on held-out likelihood. Surprisingly, topic models which perform better on held-out likelihood may infer less semantically meaningful topics.

## 1 Introduction

Probabilistic topic models have become popular tools for the unsupervised analysis of large document collections [1]. These models posit a set of latent *topics*, multinomial distributions over words, and assume that each document can be described as a mixture of these topics. With algorithms for fast approximate posterior inference, we can use topic models to discover both the topics and an assignment of topics to documents from a collection of documents. (See Figure 1.)

These modeling assumptions are useful in the sense that, empirically, they lead to good models of documents. They also anecdotally lead to semantically meaningful decompositions of them: topics tend to place high probability on words that represent concepts, and documents are represented as expressions of those concepts. Perusing the inferred topics is effective for model verification and for ensuring that the model is capturing the practitioner's intuitions about the documents. Moreover, producing a human-interpretable decomposition of the texts can be a goal in itself, as when browsing or summarizing a large collection of documents.

In this spirit, much of the literature comparing different topic models presents examples of topics and examples of document-topic assignments to help understand a model's mechanics. Topics also can help users discover new content via corpus exploration [2]. The presentation of these topics serves, either explicitly or implicitly, as a qualitative evaluation of the latent space, but there is no explicit *quantitative* evaluation of them. Instead, researchers employ a variety of metrics of model fit, such as perplexity or held-out likelihood. Such measures are useful for evaluating the predictive model, but do not address the more explanatory goals of topic modeling.

---

\*Work done while at Princeton University.

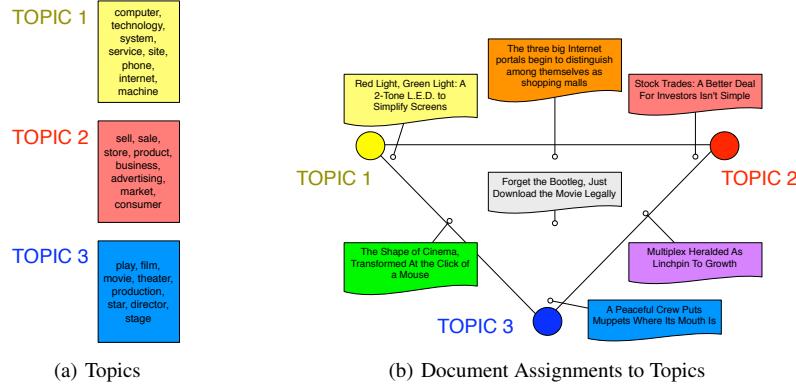


Figure 1: The latent space of a topic model consists of topics, which are distributions over words, and a distribution over these topics for each document. On the left are three topics from a fifty topic LDA model trained on articles from the New York Times. On the right is a simplex depicting the distribution over topics associated with seven documents. The line from each document's title shows the document's position in the topic space.

In this paper, we present a method for measuring the interpretability of a topic model. We devise two human evaluation tasks to explicitly evaluate both the quality of the topics inferred by the model and how well the model assigns topics to documents. The first, *word intrusion*, measures how semantically “cohesive” the topics inferred by a model are and tests whether topics correspond to natural groupings for humans. The second, *topic intrusion*, measures how well a topic model’s decomposition of a document as a mixture of topics agrees with human associations of topics with a document. We report the results of a large-scale human study of these tasks, varying both modeling assumptions and number of topics. We show that these tasks capture aspects of topic models not measured by existing metrics and—surprisingly—models which achieve better predictive perplexity often have less interpretable latent spaces.

## 2 Topic models and their evaluations

Topic models posit that each document is expressed as a mixture of topics. These topic proportions are drawn once per document, and the topics are shared across the corpus. In this paper we will consider topic models that make different assumptions about the topic proportions. Probabilistic Latent Semantic Indexing (pLSI) [3] makes no assumptions about the document topic distribution, treating it as a distinct parameter for each document. Latent Dirichlet allocation (LDA) [4] and the correlated topic model (CTM) [5] treat each document’s topic assignment as a multinomial random variable drawn from a symmetric Dirichlet and logistic normal prior, respectively.

While the models make different assumptions, inference algorithms for all of these topic models build the same type of latent space: a collection of topics for the corpus and a collection of topic proportions for each of its documents. While this common latent space has explored for over two decades, its interpretability remains unmeasured.

### Pay no attention to the latent space behind the model

Although we focus on probabilistic topic models, the field began in earnest with latent semantic analysis (LSA) [6]. LSA, the basis of pLSI’s probabilistic formulation, uses linear algebra to decompose a corpus into its constituent themes. Because LSA originated in the psychology community, early evaluations focused on replicating human performance or judgments using LSA: matching performance on standardized tests, comparing sense distinctions, and matching intuitions about synonymy (these results are reviewed in [7]). In information retrieval, where LSA is known as latent semantic indexing (LSI) [8], it is able to match queries to documents, match experts to areas of expertise, and even generalize across languages given a parallel corpus [9].

The reticence to look under the hood of these models has persisted even as models have moved from psychology into computer science with the development of pLSI and LDA. Models either use measures based on held-out likelihood [4, 5] or an external task that is independent of the topic space such as sentiment detection [10] or information retrieval [11]. This is true even for models engineered to have semantically coherent topics [12].

For models that use held-out likelihood, Wallach et al. [13] provide a summary of evaluation techniques. These metrics borrow tools from the language modeling community to measure how well the information learned from a corpus applies to unseen documents. These metrics generalize easily and allow for likelihood-based comparisons of different models or selection of model parameters such as the number of topics. However, this adaptability comes at a cost: these methods only measure the probability of observations; the internal representation of the models is ignored.

Griffiths et al. [14] is an important exception to the trend of using external tasks or held-out likelihood. They showed that the number of topics a word appears in correlates with how many distinct senses it has and reproduced many of the metrics used in the psychological community based on human performance. However, this is still not a deep analysis of the structure of the latent space, as it does not examine the structure of the topics themselves.

We emphasize that not measuring the internal representation of topic models is at odds with their presentation and development. Most topic modeling papers display qualitative assessments of the inferred topics or simply assert that topics are semantically meaningful, and practitioners use topics for model checking during the development process. Hall et al. [15], for example, used latent topics deemed historically relevant to explore themes in the scientific literature. Even in production environments, topics are presented as themes: Rexa (<http://rexa.info>), a scholarly publication search engine, displays the topics associated with documents. This implicit notion that topics have semantic meaning for users has even motivated work that attempts to automatically label topics [16]. Our goal is to measure the success of interpreting topic models across number of topics and modeling assumptions.

### 3 Using human judgments to examine the topics

Although there appears to be a longstanding assumption that the latent space discovered by topic models is meaningful and useful, evaluating such assumptions is difficult because discovering topics is an unsupervised process. There is no gold-standard list of topics to compare against for every corpus. Thus, evaluating the latent space of topic models requires us to gather exogenous data.

In this section we propose two tasks that create a formal setting where humans can evaluate the two components of the latent space of a topic model. The first component is the makeup of the topics. We develop a task to evaluate whether a topic has human-identifiable semantic coherence. This task is called *word intrusion*, as subjects must identify a spurious word inserted into a topic. The second task tests whether the association between a document and a topic makes sense. We call this task *topic intrusion*, as the subject must identify a topic that was not associated with the document by the model.

#### 3.1 Word intrusion

To measure the coherence of these topics, we develop the *word intrusion* task; this task involves evaluating the latent space presented in Figure 1(a). In the word intrusion task, the subject is presented with six randomly ordered words. The task of the user is to find the word which is out of place or does not belong with the others, i.e., the *intruder*. Figure 2 shows how this task is presented to users.

When the set of words minus the intruder makes sense together, then the subject should easily identify the intruder. For example, most people readily identify *apple* as the intruding word in the set {*dog*, *cat*, *horse*, *apple*, *pig*, *cow*} because the remaining words, {*dog*, *cat*, *horse*, *pig*, *cow*} make sense together—they are all animals. For the set {*car*, *teacher*, *platypus*, *agile*, *blue*, *Zaire*}, which lacks such coherence, identifying the intruder is difficult. People will typically choose an intruder at random, implying a topic with poor coherence.

In order to construct a set to present to the subject, we first select at random a topic from the model. We then select the five most probable words from that topic. In addition to these words, an intruder

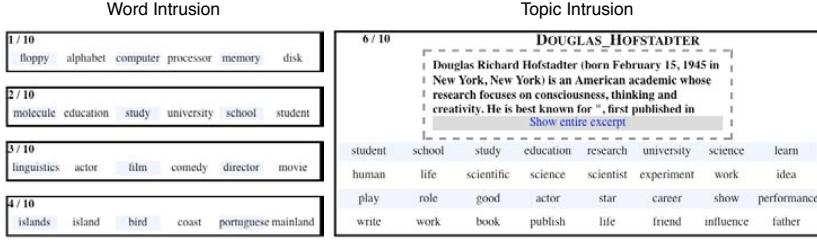


Figure 2: Screenshots of our two human tasks. In the word intrusion task (left), subjects are presented with a set of words and asked to select the word which does not belong with the others. In the *topic intrusion* task (right), users are given a document’s title and the first few sentences of the document. The users must select which of the four groups of words does not belong.

word is selected at random from a pool of words with low probability in the current topic (to reduce the possibility that the intruder comes from the same semantic group) but high probability in some other topic (to ensure that the intruder is not rejected outright due solely to rarity). All six words are then shuffled and presented to the subject.

### 3.2 Topic intrusion

The *topic intrusion* task tests whether a topic model’s decomposition of documents into a mixture of topics agrees with human judgments of the document’s content. This allows for evaluation of the latent space depicted by Figure 1(b). In this task, subjects are shown the title and a snippet from a document. Along with the document they are presented with four topics (each topic is represented by the eight highest-probability words within that topic). Three of those topics are the highest probability topics assigned to that document. The remaining *intruder topic* is chosen randomly from the other low-probability topics in the model.

The subject is instructed to choose the topic which does not belong with the document. As before, if the topic assignment to documents were relevant and intuitive, we would expect that subjects would select the topic we randomly added as the topic that did not belong. The formulation of this task provides a natural way to analyze the quality of document-topic assignments found by the topic models. Each of the three models we fit explicitly assigns topic weights to each document; this task determines whether humans make the same association.

Due to time constraints, subjects do not see the entire document; they only see the title and first few sentences. While this is less information than is available to the algorithm, humans are good at extrapolating from limited data, and our corpora (encyclopedia and newspaper) are structured to provide an overview of the article in the first few sentences. The setup of this task is also meaningful in situations where one might be tempted to use topics for corpus exploration. If topics are used to find relevant documents, for example, users will likely be provided with similar views of the documents (e.g. title and abstract, as in Rexa).

For both the word intrusion and topic intrusion tasks, subjects were instructed to focus on the meanings of words, not their syntactic usage or orthography. We also presented subjects with the option of viewing the “correct” answer after they submitted their own response, to make the tasks more engaging. Here the “correct” answer was determined by the model which generated the data, presented as if it were the response of another user. At the same time, subjects were encouraged to base their responses on their own opinions, not to try to match other subjects’ (the models’) selections. In small experiments, we have found that this extra information did not bias subjects’ responses.

## 4 Experimental results

To prepare data for human subjects to review, we fit three different topic models on two corpora. In this section, we describe how we prepared the corpora, fit the models, and created the tasks described in Section 3. We then present the results of these human trials and compare them to metrics traditionally used to evaluate topic models.

#### 4.1 Models and corpora

In this work we study three topic models: probabilistic latent semantic indexing (pLSI) [3], latent Dirichlet allocation (LDA) [4], and the correlated topic model (CTM) [5], which are all mixed membership models [17]. The number of latent topics,  $K$ , is a free parameter in each of the models; here we explore this with  $K = 50, 100$  and  $150$ . The remaining parameters –  $\beta_k$ , the topic multinomial distribution for topic  $k$ ; and  $\theta_d$ , the topic mixture proportions for document  $d$  – are inferred from data. The three models differ in how these latent parameters are inferred.

**pLSI** In pLSI, the topic mixture proportions  $\theta_d$  are a parameter for each document. Thus, pLSI is not a fully generative model, and the number of parameters grows linearly with the number of documents. We fit pLSI using the EM algorithm [18] but regularize pLSI’s estimates of  $\theta_d$  using pseudo-count smoothing,  $\alpha = 1$ .

**LDA** LDA is a fully generative model of documents where the mixture proportions  $\theta_d$  are treated as a random variable drawn from a Dirichlet prior distribution. Because the direct computation of the posterior is intractable, we employ variational inference [4] and set the symmetric Dirichlet prior parameter,  $\alpha$ , to 1.

**CTM** In LDA, the components of  $\theta_d$  are nearly independent (i.e.,  $\theta_d$  is statistically neutral). CTM allows for a richer covariance structure between topic proportions by using a logistic normal prior over the topic mixture proportions  $\theta_d$ . For each topic,  $k$ , a real  $\gamma$  is drawn from a normal distribution and exponentiated. This set of  $K$  non-negative numbers are then normalized to yield  $\theta_d$ . Here, we train the CTM using variational inference [5].

We train each model on two corpora. For each corpus, we apply a part of speech tagger [19] and remove all tokens tagged as proper nouns (this was for the benefit of the human subjects; success in early experiments required too much encyclopedic knowledge). Stop words [20] and terms occurring in fewer than five documents are also removed. The two corpora we use are 1.) a collection of 8447 articles from the *New York Times* from the years 1987 to 2007 with a vocabulary size of 8269 unique types and around one million tokens and 2.) a sample of 10000 articles from *Wikipedia* (<http://www.wikipedia.org>) with a vocabulary size of 15273 unique types and three million tokens.

#### 4.2 Evaluation using conventional objective measures

There are several metrics commonly used to evaluate topic models in the literature [13]. Many of these metrics are *predictive* metrics; that is, they capture the model’s ability to predict a *test set* of unseen documents after having learned its parameters from a *training set*. In this work, we set aside 20% of the documents in each corpus as a test set and train on the remaining 80% of documents. We then compute predictive rank and predictive log likelihood.

To ensure consistency of evaluation across different models, we follow Teh et al.’s [21] approximation of the predictive likelihood  $p(\mathbf{w}_d|D_{\text{train}})$  using  $p(\mathbf{w}_d|D_{\text{train}}) \approx p(\mathbf{w}_d|\hat{\theta}_d)$ , where  $\hat{\theta}_d$  is a point estimate of the posterior topic proportions for document  $d$ . For pLSI  $\hat{\theta}_d$  is the MAP estimate; for LDA and CTM  $\hat{\theta}_d$  is the mean of the variational posterior. With this information, we can ask what words the model believes will be in the document and compare it with the document’s actual composition. Given document  $\mathbf{w}_d$ , we first estimate  $\hat{\theta}_d$  and then for every word in the vocabulary, we compute  $p(w|\hat{\theta}_d) = \sum_z p(w|z)p(z|\hat{\theta}_d)$ . Then we compute the average rank for the terms that actually appeared in document  $\mathbf{w}_d$  (we follow the convention that lower rank is better).

The average word likelihood and average rank across all documents in our test set are shown in Table 1. These results are consistent with the values reported in the literature [4, 5]; in most cases CTM performs best, followed by LDA.

#### 4.3 Analyzing human evaluations

The tasks described in Section 3 were offered on Amazon Mechanical Turk (<http://www.mturk.com>), which allows workers (our pool of prospective subjects) to perform small jobs for a fee through a Web interface. No specialized training or knowledge is typically expected of the workers. Amazon Mechanical Turk has been successfully used in the past to develop gold-standard data for natural language processing [22] and to label images [23]. For both the word intrusion and topic intrusion

Table 1: Two predictive metrics: predictive log likelihood/predictive rank. Consistent with values reported in the literature, CTM generally performs the best, followed by LDA, then pLSI. The bold numbers indicate the best performance in each row.

CORPUS	TOPICS	LDA	CTM	PLSI
NEW YORK TIMES	50	<b>-7.3214 / 784.38</b>	-7.3335 / 788.58	-7.3384 / 796.43
	100	<b>-7.2761 / 778.24</b>	<b>-7.2647 / 762.16</b>	-7.2834 / 785.05
	150	<b>-7.2477 / 777.32</b>	<b>-7.2467 / 755.55</b>	<b>-7.2382 / 770.36</b>
WIKIPEDIA	50	<b>-7.5257 / 961.86</b>	-7.5332 / <b>936.58</b>	-7.5378 / 975.88
	100	-7.4629 / 935.53	<b>-7.4385 / 880.30</b>	-7.4748 / 951.78
	150	-7.4266 / 929.76	<b>-7.3872 / 852.46</b>	-7.4355 / 945.29

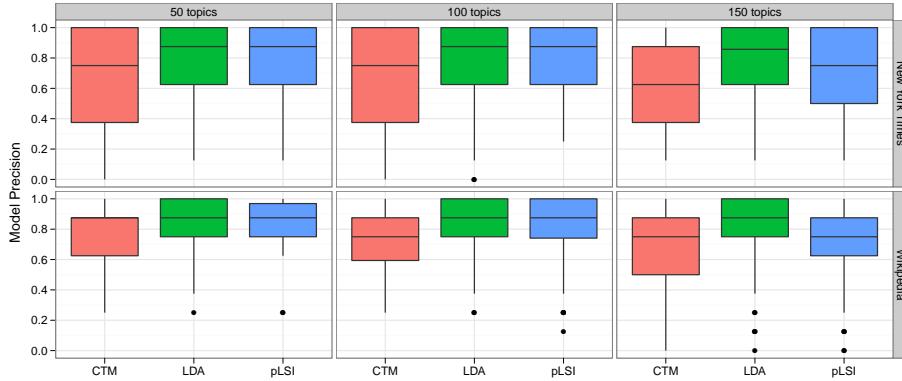


Figure 3: The model precision (Equation 1) for the three models on two corpora. Higher is better. Surprisingly, although CTM generally achieves a better predictive likelihood than the other models (Table 1), the topics it infers fare worst when evaluated against human judgments.

tasks, we presented each worker with jobs containing ten of the tasks described in Section 3. Each job was performed by 8 separate workers, and workers were paid between \$0.07 – \$0.15 per job.

**Word intrusion** As described in Section 3.1, the word intrusion task measures how well the inferred topics match human concepts (using *model precision*, i.e., how well the intruders detected by the subjects correspond to those injected into ones found by the topic model).

Let  $\omega_k^m$  be the index of the intruding word among the words generated from the  $k^{th}$  topic inferred by model  $m$ . Further let  $i_{k,s}^m$  be the intruder selected by subject  $s$  on the set of words generated from the  $k^{th}$  topic inferred by model  $m$  and let  $S$  denote the number of subjects. We define model precision by the fraction of subjects agreeing with the model,

$$\text{MP}_k^m = \sum_s \mathbb{1}(i_{k,s}^m = \omega_k^m) / S. \quad (1)$$

Figure 3 shows boxplots of the precision for the three models on the two corpora. In most cases LDA performs best. Although CTM gives better predictive results on held-out likelihood, it does not perform as well on human evaluations. This may be because CTM finds correlations between topics and correlations within topics are confounding factors; the intruder for one topic might be selected from another highly correlated topic. The performance of pLSI degrades with larger numbers of topics, suggesting that overfitting [4] might affect interpretability as well as predictive power.

Figure 4 (left) shows examples of topics with high and low model precisions from the NY Times data fit with LDA using 50 topics. In the example with high precision, the topic words all coherently express a painting theme. For the low precision example, “taxis” did not fit in with the other political words in the topic, as 87.5% of subjects chose “taxis” as the intruder.

The relationship between model precision,  $\text{MP}_k^m$ , and the model’s estimate of the likelihood of the intruding word in Figure 5 (top row) is surprising. The highest probability did not have the best interpretability; in fact, the trend was the opposite. This suggests that as topics become more fine-grained in models with larger number of topics, they are less useful for humans. The downward

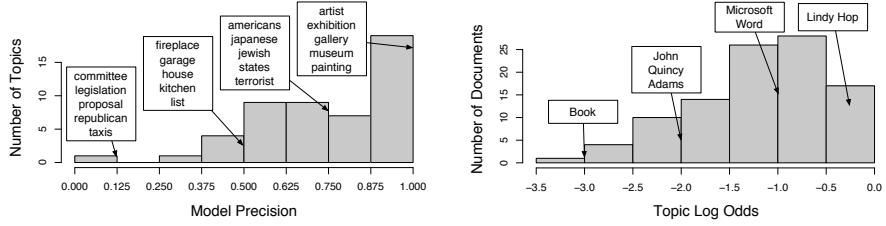


Figure 4: A histogram of the model precisions on the New York Times corpus (left) and topic log odds on the Wikipedia corpus (right) evaluated for the fifty topic LDA model. On the left, example topics are shown for several bins; the topics in bins with higher model precision evince a more coherent theme. On the right, example document titles are shown for several bins; documents with higher topic log odds can be more easily decomposed as a mixture of topics.

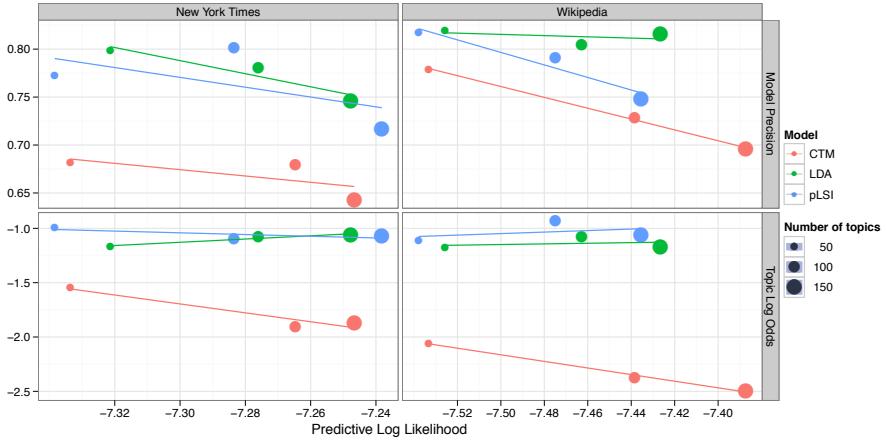


Figure 5: A scatter plot of model precision (top row) and topic log odds (bottom row) vs. predictive log likelihood. Each point is colored by model and sized according to the number of topics used to fit the model. Each model is accompanied by a regression line. Increasing likelihood does not increase the agreement between human subjects and the model for either task (as shown by the downward-sloping regression lines).

sloping trend lines in Figure 5 implying that the models are often trading improved likelihood for lower interpretability.

The model precision showed a negative correlation (Spearman's  $\rho = -0.235$  averaged across all models, corpora, and topics) with the number of senses in WordNet of the words displayed to the subjects [24] and a slight positive correlation ( $\rho = 0.109$ ) with the average pairwise Jiang-Conrath similarity of words<sup>1</sup> [25].

**Topic intrusion** In Section 3.2, we introduced the topic intrusion task to measure how well a topic model assigns topics to documents. We define the *topic log odds* as a quantitative measure of the agreement between the model and human judgments on this task. Let  $j_d^m$  denote model  $m$ 's point estimate of the topic proportions vector associated with document  $d$  (as described in Section 4.2). Further, let  $j_{ds}^m \prec 1/K$  be the intruding topic selected by subject  $s$  for document  $d$  on model  $m$  and let  $j_d^m$  denote the “true” intruder, i.e., the one generated by the model. We define the topic log odds as the log ratio of the probability mass assigned to the true intruder to the probability mass

<sup>1</sup>Words without entries in WordNet were ignored; polysemy was handled by taking the maximum over all senses of words. To handle words in the same synset (e.g. “fight” and “battle”), the similarity function was capped at 10.0.

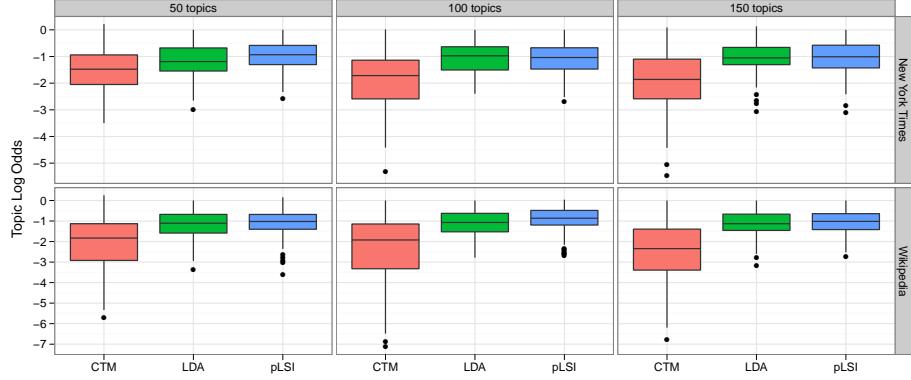


Figure 6: The topic log odds (Equation 2) for the three models on two corpora. Higher is better. Although CTM generally achieves a better predictive likelihood than the other models (Table 1), the topics it infers fare worst when evaluated against human judgments.

assigned to the intruder selected by the subject,

$$TLO_d^m = (\sum_s \log \hat{\theta}_{d,j_{d,*}^m}^m - \log \hat{\theta}_{d,j_{d,s}^m}^m) / S. \quad (2)$$

The higher the value of  $TLO_d^m$ , the greater the correspondence between the judgments of the model and the subjects. The upper bound on  $TLO_d^m$  is 0. This is achieved when the subjects choose intruders with a mixture proportion no higher than the true intruder's.

Figure 6 shows boxplots of the topic log odds for the three models. As with model precision, LDA and pLSI generally outperform CTM. Again, this trend runs counter to CTM's superior performance on predictive likelihood. A histogram of the TLO of individual Wikipedia documents is given in Figure 4 (right) for the fifty-topic LDA model. Documents about very specific, unambiguous concepts, such as "Lindy Hop," have high TLO because it is easy for both humans and the model to assign the document to a particular topic. When documents express multiple disparate topics, human judgments diverge from those of the model. At the low end of the scale is the article "Book" which touches on diverse areas such as history, science, and commerce. It is difficult for LDA to pin down specific themes in this article which match human perceptions.

Figure 5 (bottom row) shows that, as with model precision, increasing predictive likelihood does not imply improved topic log odds scores. While the topic log odds are nearly constant across all numbers of topics for LDA and pLSI, for CTM topic log odds and predictive likelihood are negatively correlated, yielding the surprising conclusion that higher predictive likelihoods do not lead to improved model interpretability.

## 5 Discussion

We presented the first validation of the assumed coherence and relevance of topic models using human experiments. For three topic models, we demonstrated that traditional metrics do not capture whether topics are coherent or not. Traditional metrics are, indeed, negatively correlated with the measures of topic quality developed in this paper. Our measures enable new forms of model selection and suggest that practitioners developing topic models should thus focus on evaluations that depend on real-world task performance rather than optimizing likelihood-based measures.

In a more qualitative vein, this work validates the use of topics for corpus exploration and information retrieval. Humans appreciate the semantic coherence of topics and can associate the same documents with a topic that a topic model does. An intriguing possibility is the development of models that explicitly seek to optimize the measures we develop here either by incorporating human judgments into the model-learning framework or creating a computational proxy that simulates human judgments.

## Acknowledgements

David M. Blei is supported by ONR 175-6343, NSF CAREER 0745520 and grants from Google and Microsoft. We would also like to thank Dan Osherson for his helpful comments.

## References

- [1] Blei, D., J. Lafferty. *Text Mining: Theory and Applications*, chap. Topic Models. Taylor and Francis, 2009.
- [2] Mimno, D., A. McCallum. Organizing the OCA: learning faceted subjects from a library of digital books. In *JCDL*. 2007.
- [3] Hofmann, T. Probabilistic latent semantic analysis. In *UAI*. 1999.
- [4] Blei, D., A. Ng, M. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.
- [5] Blei, D. M., J. D. Lafferty. Correlated topic models. In *NIPS*. 2005.
- [6] Landauer, T., S. Dumais. Solutions to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 2(104):211–240, 1997.
- [7] Landauer, T. K. On the computational basis of learning and cognition: Arguments from LSA. *The Psychology of Learning and Motivation*, 41:43–84, 2002.
- [8] Deerwester, S., S. Dumais, T. Landauer, et al. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
- [9] Berry, M. W., S. T. Dumais, T. A. Letsche. Computational methods for intelligent information access. In *Supercomputing*. 1995.
- [10] Titov, I., R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *HLT*. 2008.
- [11] Wei, X., B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR*. 2006.
- [12] Boyd-Graber, J. L., D. M. Blei, X. Zhu. Probabalistic walks in semantic hierarchies as a topic model for WSD. In *HLT*. 2007.
- [13] Wallach, H. M., I. Murray, R. Salakhutdinov, et al. Evaluation methods for topic models. In *ICML*. 2009.
- [14] Griffiths, T., M. Steyvers. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, W. Kintsch, eds., *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2006.
- [15] Hall, D., D. Jurafsky, C. D. Manning. Studying the history of ideas using topic models. In *EMNLP*. 2008.
- [16] Mei, Q., X. Shen, C. Zhai. Automatic labeling of multinomial topic models. In *KDD*. 2007.
- [17] Erosheva, E., S. Fienberg, J. Lafferty. Mixed-membership models of scientific publications. *PNAS*, 101(Suppl 1):5220 — 5227, 2004.
- [18] Dempster, A., N. Laird, D. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.
- [19] Schmid, H. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*. 1994.
- [20] Loper, E., S. Bird. NLTK: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*. 2002.
- [21] Teh, Y. W., K. Kurihara, M. Welling. Collapsed variational inference for HDP. In *NIPS*. 2008.
- [22] Snow, R., B. O’Connor, D. Jurafsky, et al. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP*. 2008.
- [23] Deng, J., W. Dong, R. Socher, et al. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*. 2009.
- [24] Miller, G. A. Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*, 3(4):245–264, 1990.
- [25] Jiang, J. J., D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*. 1997.

## Latent Dirichlet Allocation

**David M. Blei**

*Computer Science Division  
University of California  
Berkeley, CA 94720, USA*

BLEI@CS.BERKELEY.EDU

**Andrew Y. Ng**

*Computer Science Department  
Stanford University  
Stanford, CA 94305, USA*

ANG@CS.STANFORD.EDU

**Michael I. Jordan**

*Computer Science Division and Department of Statistics  
University of California  
Berkeley, CA 94720, USA*

JORDAN@CS.BERKELEY.EDU

**Editor:** John Lafferty

### Abstract

We describe *latent Dirichlet allocation* (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

### 1. Introduction

In this paper we consider the problem of modeling text corpora and other collections of discrete data. The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments.

Significant progress has been made on this problem by researchers in the field of information retrieval (IR) (Baeza-Yates and Ribeiro-Neto, 1999). The basic methodology proposed by IR researchers for text corpora—a methodology successfully deployed in modern Internet search engines—reduces each document in the corpus to a vector of real numbers, each of which represents ratios of counts. In the popular *tf-idf* scheme (Salton and McGill, 1983), a basic vocabulary of “words” or “terms” is chosen, and, for each document in the corpus, a count is formed of the number of occurrences of each word. After suitable normalization, this term frequency count is compared to an inverse document frequency count, which measures the number of occurrences of a

word in the entire corpus (generally on a log scale, and again suitably normalized). The end result is a term-by-document matrix  $X$  whose columns contain the  $tf\text{-}idf$  values for each of the documents in the corpus. Thus the  $tf\text{-}idf$  scheme reduces documents of arbitrary length to fixed-length lists of numbers.

While the  $tf\text{-}idf$  reduction has some appealing features—notably in its basic identification of sets of words that are discriminative for documents in the collection—the approach also provides a relatively small amount of reduction in description length and reveals little in the way of inter- or intra-document statistical structure. To address these shortcomings, IR researchers have proposed several other dimensionality reduction techniques, most notably *latent semantic indexing (LSI)* (Deerwester et al., 1990). LSI uses a singular value decomposition of the  $X$  matrix to identify a linear subspace in the space of  $tf\text{-}idf$  features that captures most of the variance in the collection. This approach can achieve significant compression in large collections. Furthermore, Deerwester et al. argue that the derived features of LSI, which are linear combinations of the original  $tf\text{-}idf$  features, can capture some aspects of basic linguistic notions such as synonymy and polysemy.

To substantiate the claims regarding LSI, and to study its relative strengths and weaknesses, it is useful to develop a generative probabilistic model of text corpora and to study the ability of LSI to recover aspects of the generative model from data (Papadimitriou et al., 1998). Given a generative model of text, however, it is not clear why one should adopt the LSI methodology—one can attempt to proceed more directly, fitting the model to data using maximum likelihood or Bayesian methods.

A significant step forward in this regard was made by Hofmann (1999), who presented the *probabilistic LSI (pLSI)* model, also known as the *aspect model*, as an alternative to LSI. The pLSI approach, which we describe in detail in Section 4.3, models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of “topics.” Thus each word is generated from a single topic, and different words in a document may be generated from different topics. Each document is represented as a list of mixing proportions for these mixture components and thereby reduced to a probability distribution on a fixed set of topics. This distribution is the “reduced description” associated with the document.

While Hofmann’s work is a useful step toward probabilistic modeling of text, it is incomplete in that it provides no probabilistic model at the level of documents. In pLSI, each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers. This leads to several problems: (1) the number of parameters in the model grows linearly with the size of the corpus, which leads to serious problems with overfitting, and (2) it is not clear how to assign probability to a document outside of the training set.

To see how to proceed beyond pLSI, let us consider the fundamental probabilistic assumptions underlying the class of dimensionality reduction methods that includes LSI and pLSI. All of these methods are based on the “bag-of-words” assumption—that the order of words in a document can be neglected. In the language of probability theory, this is an assumption of *exchangeability* for the words in a document (Aldous, 1985). Moreover, although less often stated formally, these methods also assume that documents are exchangeable; the specific ordering of the documents in a corpus can also be neglected.

A classic representation theorem due to de Finetti (1990) establishes that any collection of exchangeable random variables has a representation as a mixture distribution—in general an infinite mixture. Thus, if we wish to consider exchangeable representations for documents and words, we need to consider mixture models that capture the exchangeability of both words and documents.

This line of thinking leads to the *latent Dirichlet allocation (LDA)* model that we present in the current paper.

It is important to emphasize that an assumption of exchangeability is not equivalent to an assumption that the random variables are independent and identically distributed. Rather, exchangeability essentially can be interpreted as meaning “*conditionally* independent and identically distributed,” where the conditioning is with respect to an underlying latent parameter of a probability distribution. Conditionally, the joint distribution of the random variables is simple and factored while marginally over the latent parameter, the joint distribution can be quite complex. Thus, while an assumption of exchangeability is clearly a major simplifying assumption in the domain of text modeling, and its principal justification is that it leads to methods that are computationally efficient, the exchangeability assumptions do not necessarily lead to methods that are restricted to simple frequency counts or linear operations. We aim to demonstrate in the current paper that, by taking the de Finetti theorem seriously, we can capture significant intra-document statistical structure via the mixing distribution.

It is also worth noting that there are a large number of generalizations of the basic notion of exchangeability, including various forms of partial exchangeability, and that representation theorems are available for these cases as well (Diaconis, 1988). Thus, while the work that we discuss in the current paper focuses on simple “bag-of-words” models, which lead to mixture distributions for single words (unigrams), our methods are also applicable to richer models that involve mixtures for larger structural units such as  $n$ -grams or paragraphs.

The paper is organized as follows. In Section 2 we introduce basic notation and terminology. The LDA model is presented in Section 3 and is compared to related latent variable models in Section 4. We discuss inference and parameter estimation for LDA in Section 5. An illustrative example of fitting LDA to data is provided in Section 6. Empirical results in text modeling, text classification and collaborative filtering are presented in Section 7. Finally, Section 8 presents our conclusions.

## 2. Notation and terminology

We use the language of text collections throughout the paper, referring to entities such as “words,” “documents,” and “corpora.” This is useful in that it helps to guide intuition, particularly when we introduce latent variables which aim to capture abstract notions such as topics. It is important to note, however, that the LDA model is not necessarily tied to text, and has applications to other problems involving collections of data, including data from domains such as collaborative filtering, content-based image retrieval and bioinformatics. Indeed, in Section 7.3, we present experimental results in the collaborative filtering domain.

Formally, we define the following terms:

- A *word* is the basic unit of discrete data, defined to be an item from a vocabulary indexed by  $\{1, \dots, V\}$ . We represent words using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, using superscripts to denote components, the  $v$ th word in the vocabulary is represented by a  $V$ -vector  $w$  such that  $w^v = 1$  and  $w^u = 0$  for  $u \neq v$ .
- A *document* is a sequence of  $N$  words denoted by  $\mathbf{w} = (w_1, w_2, \dots, w_N)$ , where  $w_n$  is the  $n$ th word in the sequence.
- A *corpus* is a collection of  $M$  documents denoted by  $D = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$ .

We wish to find a probabilistic model of a corpus that not only assigns high probability to members of the corpus, but also assigns high probability to other “similar” documents.

### 3. Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.<sup>1</sup>

LDA assumes the following generative process for each document  $\mathbf{w}$  in a corpus  $D$ :

1. Choose  $N \sim \text{Poisson}(\xi)$ .
2. Choose  $\theta \sim \text{Dir}(\alpha)$ .
3. For each of the  $N$  words  $w_n$ :
  - (a) Choose a topic  $z_n \sim \text{Multinomial}(\theta)$ .
  - (b) Choose a word  $w_n$  from  $p(w_n | z_n, \beta)$ , a multinomial probability conditioned on the topic  $z_n$ .

Several simplifying assumptions are made in this basic model, some of which we remove in subsequent sections. First, the dimensionality  $k$  of the Dirichlet distribution (and thus the dimensionality of the topic variable  $z$ ) is assumed known and fixed. Second, the word probabilities are parameterized by a  $k \times V$  matrix  $\beta$  where  $\beta_{ij} = p(w^j = 1 | z^i = 1)$ , which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed. Furthermore, note that  $N$  is independent of all the other data generating variables ( $\theta$  and  $\mathbf{z}$ ). It is thus an ancillary variable and we will generally ignore its randomness in the subsequent development.

A  $k$ -dimensional Dirichlet random variable  $\theta$  can take values in the  $(k - 1)$ -simplex (a  $k$ -vector  $\theta$  lies in the  $(k - 1)$ -simplex if  $\theta_i \geq 0$ ,  $\sum_{i=1}^k \theta_i = 1$ ), and has the following probability density on this simplex:

$$p(\theta | \alpha) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \cdots \theta_k^{\alpha_k-1}, \quad (1)$$

where the parameter  $\alpha$  is a  $k$ -vector with components  $\alpha_i > 0$ , and where  $\Gamma(x)$  is the Gamma function. The Dirichlet is a convenient distribution on the simplex — it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution. In Section 5, these properties will facilitate the development of inference and parameter estimation algorithms for LDA.

Given the parameters  $\alpha$  and  $\beta$ , the joint distribution of a topic mixture  $\theta$ , a set of  $N$  topics  $\mathbf{z}$ , and a set of  $N$  words  $\mathbf{w}$  is given by:

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta), \quad (2)$$

---

1. We refer to the latent multinomial variables in the LDA model as topics, so as to exploit text-oriented intuitions, but we make no epistemological claims regarding these latent variables beyond their utility in representing probability distributions on sets of words.

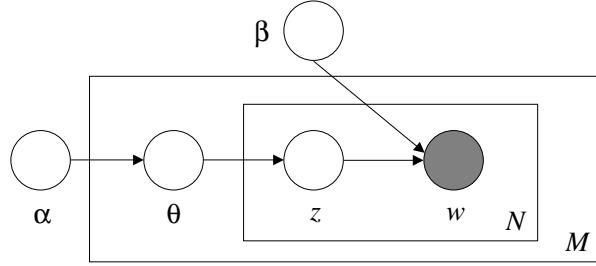


Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

where  $p(z_n | \theta)$  is simply  $\theta_i$  for the unique  $i$  such that  $z_n^i = 1$ . Integrating over  $\theta$  and summing over  $z$ , we obtain the marginal distribution of a document:

$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left( \prod_{n=1}^N \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right) d\theta. \quad (3)$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(D | \alpha, \beta) = \prod_{d=1}^M \int p(\theta_d | \alpha) \left( \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right) d\theta_d.$$

The LDA model is represented as a probabilistic graphical model in Figure 1. As the figure makes clear, there are three levels to the LDA representation. The parameters  $\alpha$  and  $\beta$  are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables  $\theta_d$  are document-level variables, sampled once per document. Finally, the variables  $z_{dn}$  and  $w_{dn}$  are word-level variables and are sampled once for each word in each document.

It is important to distinguish LDA from a simple Dirichlet-multinomial clustering model. A classical clustering model would involve a two-level model in which a Dirichlet is sampled once for a corpus, a multinomial clustering variable is selected once for each document in the corpus, and a set of words are selected for the document conditional on the cluster variable. As with many clustering models, such a model restricts a document to being associated with a single topic. LDA, on the other hand, involves three levels, and notably the topic node is sampled *repeatedly* within the document. Under this model, documents can be associated with multiple topics.

Structures similar to that shown in Figure 1 are often studied in Bayesian statistical modeling, where they are referred to as *hierarchical models* (Gelman et al., 1995), or more precisely as *conditionally independent hierarchical models* (Kass and Steffey, 1989). Such models are also often referred to as *parametric empirical Bayes models*, a term that refers not only to a particular model structure, but also to the methods used for estimating parameters in the model (Morris, 1983). Indeed, as we discuss in Section 5, we adopt the empirical Bayes approach to estimating parameters such as  $\alpha$  and  $\beta$  in simple implementations of LDA, but we also consider fuller Bayesian approaches as well.

### 3.1 LDA and exchangeability

A finite set of random variables  $\{z_1, \dots, z_N\}$  is said to be *exchangeable* if the joint distribution is invariant to permutation. If  $\pi$  is a permutation of the integers from 1 to  $N$ :

$$p(z_1, \dots, z_N) = p(z_{\pi(1)}, \dots, z_{\pi(N)}).$$

An infinite sequence of random variables is *infinitely exchangeable* if every finite subsequence is exchangeable.

De Finetti's representation theorem states that the joint distribution of an infinitely exchangeable sequence of random variables is as if a random parameter were drawn from some distribution and then the random variables in question were *independent* and *identically distributed*, conditioned on that parameter.

In LDA, we assume that words are generated by topics (by fixed conditional distributions) and that those topics are infinitely exchangeable within a document. By de Finetti's theorem, the probability of a sequence of words and topics must therefore have the form:

$$p(\mathbf{w}, \mathbf{z}) = \int p(\theta) \left( \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n) \right) d\theta,$$

where  $\theta$  is the random parameter of a multinomial over topics. We obtain the LDA distribution on documents in Eq. (3) by marginalizing out the topic variables and endowing  $\theta$  with a Dirichlet distribution.

### 3.2 A continuous mixture of unigrams

The LDA model shown in Figure 1 is somewhat more elaborate than the two-level models often studied in the classical hierarchical Bayesian literature. By marginalizing over the hidden topic variable  $z$ , however, we can understand LDA as a two-level model.

In particular, let us form the word distribution  $p(w | \theta, \beta)$ :

$$p(w | \theta, \beta) = \sum_z p(w | z, \beta) p(z | \theta).$$

Note that this is a random quantity since it depends on  $\theta$ .

We now define the following generative process for a document  $\mathbf{w}$ :

1. Choose  $\theta \sim \text{Dir}(\alpha)$ .
2. For each of the  $N$  words  $w_n$ :
  - (a) Choose a word  $w_n$  from  $p(w_n | \theta, \beta)$ .

This process defines the marginal distribution of a document as a continuous mixture distribution:

$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left( \prod_{n=1}^N p(w_n | \theta, \beta) \right) d\theta,$$

where  $p(w_n | \theta, \beta)$  are the mixture components and  $p(\theta | \alpha)$  are the mixture weights.

Figure 2 illustrates this interpretation of LDA. It depicts the distribution on  $p(w | \theta, \beta)$  which is induced from a particular instance of an LDA model. Note that this distribution on the  $(V - 1)$ -simplex is attained with only  $k + kV$  parameters yet exhibits a very interesting multimodal structure.

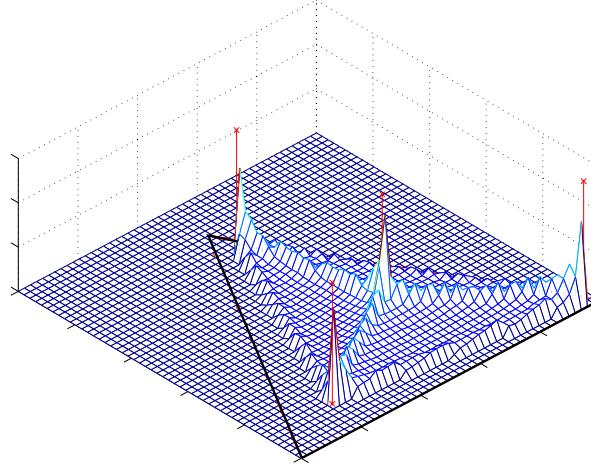


Figure 2: An example density on unigram distributions  $p(w|\theta, \beta)$  under LDA for three words and four topics. The triangle embedded in the x-y plane is the 2-D simplex representing all possible multinomial distributions over three words. Each of the vertices of the triangle corresponds to a deterministic distribution that assigns probability one to one of the words; the midpoint of an edge gives probability 0.5 to two of the words; and the centroid of the triangle is the uniform distribution over all three words. The four points marked with an x are the locations of the multinomial distributions  $p(w|z)$  for each of the four topics, and the surface shown on top of the simplex is an example of a density over the  $(V - 1)$ -simplex (multinomial distributions of words) given by LDA.

#### 4. Relationship with other latent variable models

In this section we compare LDA to simpler latent variable models for text—the unigram model, a mixture of unigrams, and the pLSI model. Furthermore, we present a unified geometric interpretation of these models which highlights their key differences and similarities.

##### 4.1 Unigram model

Under the unigram model, the words of every document are drawn independently from a single multinomial distribution:

$$p(\mathbf{w}) = \prod_{n=1}^N p(w_n).$$

This is illustrated in the graphical model in Figure 3a.

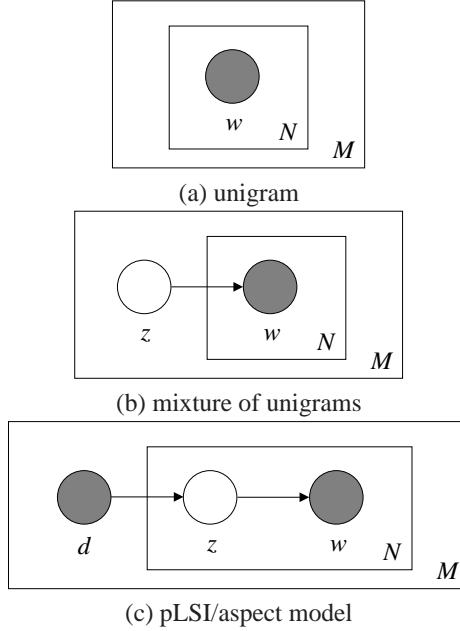


Figure 3: Graphical model representation of different models of discrete data.

#### 4.2 Mixture of unigrams

If we augment the unigram model with a discrete random topic variable  $z$  (Figure 3b), we obtain a *mixture of unigrams* model (Nigam et al., 2000). Under this mixture model, each document is generated by first choosing a topic  $z$  and then generating  $N$  words independently from the conditional multinomial  $p(w|z)$ . The probability of a document is:

$$p(\mathbf{w}) = \sum_z p(z) \prod_{n=1}^N p(w_n|z).$$

When estimated from a corpus, the word distributions can be viewed as representations of topics under the assumption that each document exhibits exactly one topic. As the empirical results in Section 7 illustrate, this assumption is often too limiting to effectively model a large collection of documents.

In contrast, the LDA model allows documents to exhibit multiple topics to different degrees. This is achieved at a cost of just one additional parameter: there are  $k - 1$  parameters associated with  $p(z)$  in the mixture of unigrams, versus the  $k$  parameters associated with  $p(\theta|\alpha)$  in LDA.

#### 4.3 Probabilistic latent semantic indexing

Probabilistic latent semantic indexing (pLSI) is another widely used document model (Hofmann, 1999). The pLSI model, illustrated in Figure 3c, posits that a document label  $d$  and a word  $w_n$  are

conditionally independent given an unobserved topic  $z$ :

$$p(d, w_n) = p(d) \sum_z p(w_n | z) p(z | d).$$

The pLSI model attempts to relax the simplifying assumption made in the mixture of unigrams model that each document is generated from only one topic. In a sense, it does capture the possibility that a document may contain multiple topics since  $p(z | d)$  serves as the mixture weights of the topics for a particular document  $d$ . However, it is important to note that  $d$  is a dummy index into the list of documents in the *training set*. Thus,  $d$  is a multinomial random variable with as many possible values as there are training documents and the model learns the topic mixtures  $p(z | d)$  only for those documents on which it is trained. For this reason, pLSI is not a well-defined generative model of documents; there is no natural way to use it to assign probability to a previously unseen document.

A further difficulty with pLSI, which also stems from the use of a distribution indexed by training documents, is that the number of parameters which must be estimated grows linearly with the number of training documents. The parameters for a  $k$ -topic pLSI model are  $k$  multinomial distributions of size  $V$  and  $M$  mixtures over the  $k$  hidden topics. This gives  $kV + kM$  parameters and therefore linear growth in  $M$ . The linear growth in parameters suggests that the model is prone to overfitting and, empirically, overfitting is indeed a serious problem (see Section 7.1). In practice, a tempering heuristic is used to smooth the parameters of the model for acceptable predictive performance. It has been shown, however, that overfitting can occur even when tempering is used (Popescul et al., 2001).

LDA overcomes both of these problems by treating the topic mixture weights as a  $k$ -parameter hidden *random variable* rather than a large set of individual parameters which are explicitly linked to the training set. As described in Section 3, LDA is a well-defined generative model and generalizes easily to new documents. Furthermore, the  $k + kV$  parameters in a  $k$ -topic LDA model do not grow with the size of the training corpus. We will see in Section 7.1 that LDA does not suffer from the same overfitting issues as pLSI.

#### 4.4 A geometric interpretation

A good way of illustrating the differences between LDA and the other latent topic models is by considering the geometry of the latent space, and seeing how a document is represented in that geometry under each model.

All four of the models described above—unigram, mixture of unigrams, pLSI, and LDA—operate in the space of distributions over words. Each such distribution can be viewed as a point on the  $(V - 1)$ -simplex, which we call the word simplex.

The unigram model finds a single point on the word simplex and posits that all words in the corpus come from the corresponding distribution. The latent variable models consider  $k$  points on the word simplex and form a sub-simplex based on those points, which we call the topic simplex. Note that any point on the topic simplex is also a point on the word simplex. The different latent variable models use the topic simplex in different ways to generate a document.

- The mixture of unigrams model posits that for each document, one of the  $k$  points on the word simplex (that is, one of the corners of the topic simplex) is chosen randomly and all the words of the document are drawn from the distribution corresponding to that point.

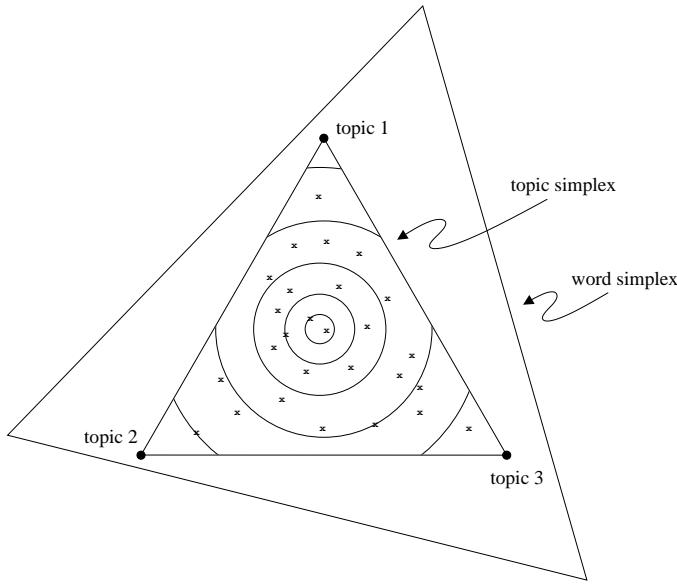


Figure 4: The topic simplex for three topics embedded in the word simplex for three words. The corners of the word simplex correspond to the three distributions where each word (respectively) has probability one. The three points of the topic simplex correspond to three different distributions over words. The mixture of unigrams places each document at one of the corners of the topic simplex. The pLSI model induces an empirical distribution on the topic simplex denoted by x. LDA places a smooth distribution on the topic simplex denoted by the contour lines.

- The pLSI model posits that each word of a *training* document comes from a randomly chosen topic. The topics are themselves drawn from a document-specific distribution over topics, i.e., a point on the topic simplex. There is one such distribution for each document; the set of training documents thus defines an empirical distribution on the topic simplex.
- LDA posits that each word of both the observed and unseen documents is generated by a randomly chosen topic which is drawn from a distribution with a randomly chosen parameter. This parameter is sampled once per document from a smooth distribution on the topic simplex.

These differences are highlighted in Figure 4.

## 5. Inference and Parameter Estimation

We have described the motivation behind LDA and illustrated its conceptual advantages over other latent topic models. In this section, we turn our attention to procedures for inference and parameter estimation under LDA.

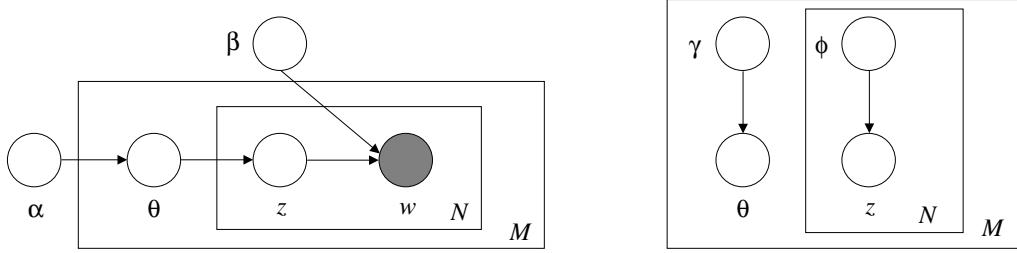


Figure 5: (Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA.

### 5.1 Inference

The key inferential problem that we need to solve in order to use LDA is that of computing the posterior distribution of the hidden variables given a document:

$$p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)}{p(\mathbf{w} | \alpha, \beta)}.$$

Unfortunately, this distribution is intractable to compute in general. Indeed, to normalize the distribution we marginalize over the hidden variables and write Eq. (3) in terms of the model parameters:

$$p(\mathbf{w} | \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left( \prod_{i=1}^k \theta_i^{\alpha_i - 1} \right) \left( \prod_{n=1}^N \sum_{i=1}^k \prod_{j=1}^V (\theta_i \beta_{ij})^{w_n^j} \right) d\theta,$$

a function which is intractable due to the coupling between  $\theta$  and  $\beta$  in the summation over latent topics (Dickey, 1983). Dickey shows that this function is an expectation under a particular extension to the Dirichlet distribution which can be represented with special hypergeometric functions. It has been used in a Bayesian context for censored discrete data to represent the posterior on  $\theta$  which, in that setting, is a random parameter (Dickey et al., 1987).

Although the posterior distribution is intractable for exact inference, a wide variety of approximate inference algorithms can be considered for LDA, including Laplace approximation, variational approximation, and Markov chain Monte Carlo (Jordan, 1999). In this section we describe a simple convexity-based variational algorithm for inference in LDA, and discuss some of the alternatives in Section 8.

### 5.2 Variational inference

The basic idea of convexity-based variational inference is to make use of Jensen's inequality to obtain an adjustable lower bound on the log likelihood (Jordan et al., 1999). Essentially, one considers a family of lower bounds, indexed by a set of *variational parameters*. The variational parameters are chosen by an optimization procedure that attempts to find the tightest possible lower bound.

A simple way to obtain a tractable family of lower bounds is to consider simple modifications of the original graphical model in which some of the edges and nodes are removed. Consider in particular the LDA model shown in Figure 5 (left). The problematic coupling between  $\theta$  and  $\beta$

arises due to the edges between  $\theta$ ,  $\mathbf{z}$ , and  $\mathbf{w}$ . By dropping these edges and the  $\mathbf{w}$  nodes, and endowing the resulting simplified graphical model with free variational parameters, we obtain a family of distributions on the latent variables. This family is characterized by the following variational distribution:

$$q(\theta, \mathbf{z} | \gamma, \phi) = q(\theta | \gamma) \prod_{n=1}^N q(z_n | \phi_n), \quad (4)$$

where the Dirichlet parameter  $\gamma$  and the multinomial parameters  $(\phi_1, \dots, \phi_N)$  are the free variational parameters.

Having specified a simplified family of probability distributions, the next step is to set up an optimization problem that determines the values of the variational parameters  $\gamma$  and  $\phi$ . As we show in Appendix A, the desideratum of finding a tight lower bound on the log likelihood translates directly into the following optimization problem:

$$(\gamma^*, \phi^*) = \arg \min_{(\gamma, \phi)} D(q(\theta, \mathbf{z} | \gamma, \phi) \| p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta)). \quad (5)$$

Thus the optimizing values of the variational parameters are found by minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior  $p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta)$ . This minimization can be achieved via an iterative fixed-point method. In particular, we show in Appendix A.3 that by computing the derivatives of the KL divergence and setting them equal to zero, we obtain the following pair of update equations:

$$\phi_{ni} \propto \beta_{iw_n} \exp\{\mathbb{E}_q[\log(\theta_i) | \gamma]\} \quad (6)$$

$$\gamma_i = \alpha_i + \sum_{n=1}^N \phi_{ni}. \quad (7)$$

As we show in Appendix A.1, the expectation in the multinomial update can be computed as follows:

$$\mathbb{E}_q[\log(\theta_i) | \gamma] = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^k \gamma_j\right), \quad (8)$$

where  $\Psi$  is the first derivative of the  $\log\Gamma$  function which is computable via Taylor approximations (Abramowitz and Stegun, 1970).

Eqs. (6) and (7) have an appealing intuitive interpretation. The Dirichlet update is a posterior Dirichlet given expected observations taken under the variational distribution,  $\mathbb{E}[z_n | \phi_n]$ . The multinomial update is akin to using Bayes' theorem,  $p(z_n | w_n) \propto p(w_n | z_n)p(z_n)$ , where  $p(z_n)$  is approximated by the exponential of the expected value of its logarithm under the variational distribution.

It is important to note that the variational distribution is actually a conditional distribution, varying as a function of  $\mathbf{w}$ . This occurs because the optimization problem in Eq. (5) is conducted for fixed  $\mathbf{w}$ , and thus yields optimizing parameters  $(\gamma^*, \phi^*)$  that are a function of  $\mathbf{w}$ . We can write the resulting variational distribution as  $q(\theta, \mathbf{z} | \gamma^*(\mathbf{w}), \phi^*(\mathbf{w}))$ , where we have made the dependence on  $\mathbf{w}$  explicit. Thus the variational distribution can be viewed as an approximation to the posterior distribution  $p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta)$ .

In the language of text, the optimizing parameters  $(\gamma^*(\mathbf{w}), \phi^*(\mathbf{w}))$  are document-specific. In particular, we view the Dirichlet parameters  $\gamma^*(\mathbf{w})$  as providing a representation of a document in the topic simplex.

```

(1) initialize  $\phi_{ni}^0 := 1/k$  for all  $i$  and  $n$ 
(2) initialize  $\gamma_i := \alpha_i + N/k$  for all  $i$ 
(3) repeat
(4)   for  $n = 1$  to  $N$ 
(5)     for  $i = 1$  to  $k$ 
(6)        $\phi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$ 
(7)       normalize  $\phi_n^{t+1}$  to sum to 1.
(8)      $\gamma^{t+1} := \alpha + \sum_{n=1}^N \phi_n^{t+1}$ 
(9)   until convergence

```

Figure 6: A variational inference algorithm for LDA.

We summarize the variational inference procedure in Figure 6, with appropriate starting points for  $\gamma$  and  $\phi_n$ . From the pseudocode it is clear that each iteration of variational inference for LDA requires  $O((N+1)k)$  operations. Empirically, we find that the number of iterations required for a single document is on the order of the number of words in the document. This yields a total number of operations roughly on the order of  $N^2k$ .

### 5.3 Parameter estimation

In this section we present an empirical Bayes method for parameter estimation in the LDA model (see Section 5.4 for a fuller Bayesian approach). In particular, given a corpus of documents  $D = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$ , we wish to find parameters  $\alpha$  and  $\beta$  that maximize the (marginal) log likelihood of the data:

$$\ell(\alpha, \beta) = \sum_{d=1}^M \log p(\mathbf{w}_d | \alpha, \beta).$$

As we have described above, the quantity  $p(\mathbf{w} | \alpha, \beta)$  cannot be computed tractably. However, variational inference provides us with a tractable lower bound on the log likelihood, a bound which we can maximize with respect to  $\alpha$  and  $\beta$ . We can thus find approximate empirical Bayes estimates for the LDA model via an alternating *variational EM* procedure that maximizes a lower bound with respect to the variational parameters  $\gamma$  and  $\phi$ , and then, for fixed values of the variational parameters, maximizes the lower bound with respect to the model parameters  $\alpha$  and  $\beta$ .

We provide a detailed derivation of the variational EM algorithm for LDA in Appendix A.4. The derivation yields the following iterative algorithm:

1. (E-step) For each document, find the optimizing values of the variational parameters  $\{\gamma_d^*, \phi_d^* : d \in D\}$ . This is done as described in the previous section.
2. (M-step) Maximize the resulting lower bound on the log likelihood with respect to the model parameters  $\alpha$  and  $\beta$ . This corresponds to finding maximum likelihood estimates with expected sufficient statistics for each document under the approximate posterior which is computed in the E-step.

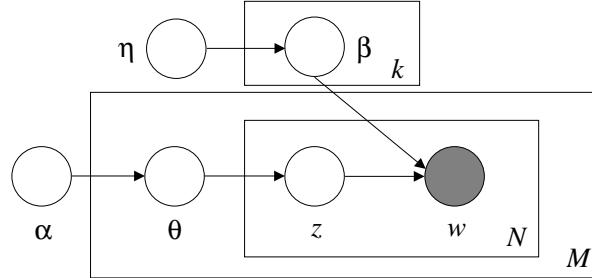


Figure 7: Graphical model representation of the smoothed LDA model.

These two steps are repeated until the lower bound on the log likelihood converges.

In Appendix A.4, we show that the M-step update for the conditional multinomial parameter  $\beta$  can be written out analytically:

$$\beta_{ij} \propto \sum_{d=1}^M \sum_{n=1}^{N_d} \phi_{dn}^* w_{dn}^j. \quad (9)$$

We further show that the M-step update for Dirichlet parameter  $\alpha$  can be implemented using an efficient Newton-Raphson method in which the Hessian is inverted in linear time.

#### 5.4 Smoothing

The large vocabulary size that is characteristic of many document corpora creates serious problems of sparsity. A new document is very likely to contain words that did not appear in any of the documents in a training corpus. Maximum likelihood estimates of the multinomial parameters assign zero probability to such words, and thus zero probability to new documents. The standard approach to coping with this problem is to “smooth” the multinomial parameters, assigning positive probability to all vocabulary items whether or not they are observed in the training set (Jelinek, 1997). Laplace smoothing is commonly used; this essentially yields the mean of the posterior distribution under a uniform Dirichlet prior on the multinomial parameters.

Unfortunately, in the mixture model setting, simple Laplace smoothing is no longer justified as a maximum a posteriori method (although it is often implemented in practice; cf. Nigam et al., 1999). In fact, by placing a Dirichlet prior on the multinomial parameter we obtain an intractable posterior in the mixture model setting, for much the same reason that one obtains an intractable posterior in the basic LDA model. Our proposed solution to this problem is to simply apply variational inference methods to the extended model that includes Dirichlet smoothing on the multinomial parameter.

In the LDA setting, we obtain the extended graphical model shown in Figure 7. We treat  $\beta$  as a  $k \times V$  random matrix (one row for each mixture component), where we assume that each row is independently drawn from an exchangeable Dirichlet distribution.<sup>2</sup> We now extend our inference procedures to treat the  $\beta_i$  as random variables that are endowed with a posterior distribution,

---

2. An exchangeable Dirichlet is simply a Dirichlet distribution with a single scalar parameter  $\eta$ . The density is the same as a Dirichlet (Eq. 1) where  $\alpha_i = \eta$  for each component.

conditioned on the data. Thus we move beyond the empirical Bayes procedure of Section 5.3 and consider a fuller Bayesian approach to LDA.

We consider a variational approach to Bayesian inference that places a separable distribution on the random variables  $\beta$ ,  $\theta$ , and  $\mathbf{z}$  (Attias, 2000):

$$q(\beta_{1:k}, \mathbf{z}_{1:M}, \theta_{1:M} | \lambda, \phi, \gamma) = \prod_{i=1}^k \text{Dir}(\beta_i | \lambda_i) \prod_{d=1}^M q_d(\theta_d, \mathbf{z}_d | \phi_d, \gamma_d),$$

where  $q_d(\theta, \mathbf{z} | \phi, \gamma)$  is the variational distribution defined for LDA in Eq. (4). As is easily verified, the resulting variational inference procedure again yields Eqs. (6) and (7) as the update equations for the variational parameters  $\phi$  and  $\gamma$ , respectively, as well as an additional update for the new variational parameter  $\lambda$ :

$$\lambda_{ij} = \eta + \sum_{d=1}^M \sum_{n=1}^{N_d} \phi_{dn}^* w_{dn}^j.$$

Iterating these equations to convergence yields an approximate posterior distribution on  $\beta$ ,  $\theta$ , and  $\mathbf{z}$ .

We are now left with the hyperparameter  $\eta$  on the exchangeable Dirichlet, as well as the hyperparameter  $\alpha$  from before. Our approach to setting these hyperparameters is again (approximate) empirical Bayes—we use variational EM to find maximum likelihood estimates of these parameters based on the marginal likelihood. These procedures are described in Appendix A.4.

## 6. Example

In this section, we provide an illustrative example of the use of an LDA model on real data. Our data are 16,000 documents from a subset of the TREC AP corpus (Harman, 1992). After removing a standard list of stop words, we used the EM algorithm described in Section 5.3 to find the Dirichlet and conditional multinomial parameters for a 100-topic LDA model. The top words from some of the resulting multinomial distributions  $p(w | z)$  are illustrated in Figure 8 (top). As we have hoped, these distributions seem to capture some of the underlying topics in the corpus (and we have named them according to these topics).

As we emphasized in Section 4, one of the advantages of LDA over related latent variable models is that it provides well-defined inference procedures for previously unseen documents. Indeed, we can illustrate how LDA works by performing inference on a held-out document and examining the resulting variational posterior parameters.

Figure 8 (bottom) is a document from the TREC AP corpus which was not used for parameter estimation. Using the algorithm in Section 5.1, we computed the variational posterior Dirichlet parameters  $\gamma$  for the article and variational posterior multinomial parameters  $\phi_n$  for each word in the article.

Recall that the  $i$ th posterior Dirichlet parameter  $\gamma_i$  is approximately the  $i$ th prior Dirichlet parameter  $\alpha_i$  plus the expected number of words which were generated by the  $i$ th topic (see Eq. 7). Therefore, the prior Dirichlet parameters subtracted from the posterior Dirichlet parameters indicate the expected number of words which were allocated to each topic for a particular document. For the example article in Figure 8 (bottom), most of the  $\gamma_i$  are close to  $\alpha_i$ . Four topics, however, are significantly larger (by this, we mean  $\gamma_i - \alpha_i \geq 1$ ). Looking at the corresponding distributions over words identifies the topics which mixed to form this document (Figure 8, top).

Further insight comes from examining the  $\phi_n$  parameters. These distributions approximate  $p(z_n | \mathbf{w})$  and tend to peak towards one of the  $k$  possible topic values. In the article text in Figure 8, the words are color coded according to these values (i.e., the  $i$ th color is used if  $q_n(z_n^i = 1) > 0.9$ ). With this illustration, one can identify how the different topics mixed in the document text.

While demonstrating the power of LDA, the posterior analysis also highlights some of its limitations. In particular, the bag-of-words assumption allows words that should be generated by the same topic (e.g., “William Randolph Hearst Foundation”) to be allocated to several different topics. Overcoming this limitation would require some form of extension of the basic LDA model; in particular, we might relax the bag-of-words assumption by assuming partial exchangeability or Markovianity of word sequences.

## 7. Applications and Empirical Results

In this section, we discuss our empirical evaluation of LDA in several problem domains—document modeling, document classification, and collaborative filtering.

In all of the mixture models, the expected complete log likelihood of the data has local maxima at the points where all or some of the mixture components are equal to each other. To avoid these local maxima, it is important to initialize the EM algorithm appropriately. In our experiments, we initialize EM by seeding each conditional multinomial distribution with five documents, reducing their effective total length to two words, and smoothing across the whole vocabulary. This is essentially an approximation to the scheme described in Heckerman and Meila (2001).

### 7.1 Document modeling

We trained a number of latent variable models, including LDA, on two text corpora to compare the generalization performance of these models. The documents in the corpora are treated as unlabeled; thus, our goal is density estimation—we wish to achieve high likelihood on a held-out test set. In particular, we computed the *perplexity* of a held-out test set to evaluate the models. The perplexity, used by convention in language modeling, is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates better generalization performance.<sup>3</sup> More formally, for a test set of  $M$  documents, the perplexity is:

$$\text{perplexity}(D_{\text{test}}) = \exp \left\{ -\frac{\sum_{d=1}^M \log p(\mathbf{w}_d)}{\sum_{d=1}^M N_d} \right\}.$$

In our experiments, we used a corpus of scientific abstracts from the C. Elegans community (Avery, 2002) containing 5,225 abstracts with 28,414 unique terms, and a subset of the TREC AP corpus containing 16,333 newswire articles with 23,075 unique terms. In both cases, we held out 10% of the data for test purposes and trained the models on the remaining 90%. In preprocessing the data,

---

3. Note that we simply use perplexity as a figure of merit for comparing models. The models that we compare are all unigram (“bag-of-words”) models, which—as we have discussed in the Introduction—are of interest in the information retrieval context. We are *not* attempting to do language modeling in this paper—an enterprise that would require us to examine trigram or other higher-order models. We note in passing, however, that extensions of LDA could be considered that involve Dirichlet-multinomial over trigrams instead of unigrams. We leave the exploration of such extensions to language modeling to future work.

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

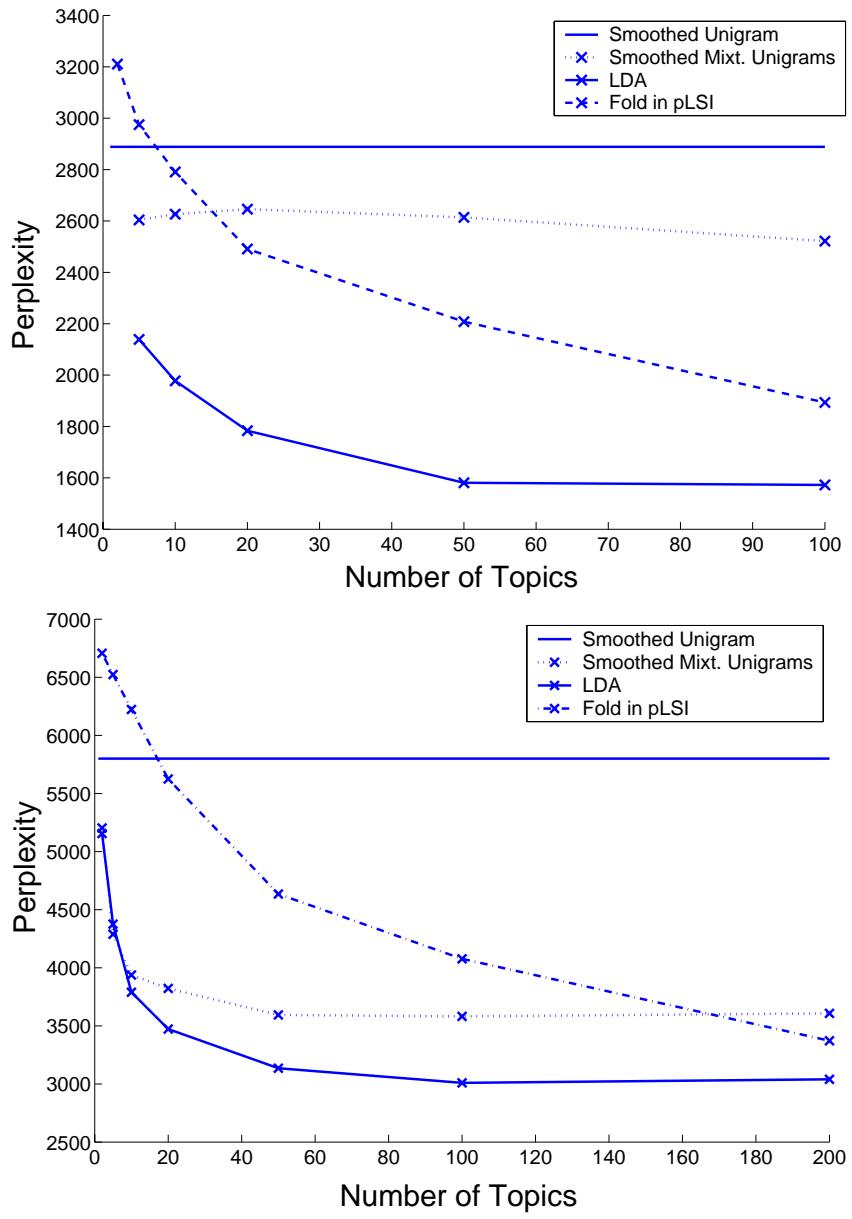


Figure 9: Perplexity results on the nematode (Top) and AP (Bottom) corpora for LDA, the unigram model, mixture of unigrams, and pLSI.

Num. topics ( $k$ )	Perplexity (Mult. Mixt.)	Perplexity (pLSI)
2	22,266	7,052
5	$2.20 \times 10^8$	17,588
10	$1.93 \times 10^{17}$	63,800
20	$1.20 \times 10^{22}$	$2.52 \times 10^5$
50	$4.19 \times 10^{106}$	$5.04 \times 10^6$
100	$2.39 \times 10^{150}$	$1.72 \times 10^7$
200	$3.51 \times 10^{264}$	$1.31 \times 10^7$

Table 1: Overfitting in the mixture of unigrams and pLSI models for the AP corpus. Similar behavior is observed in the nematode corpus (not reported).

we removed a standard list of 50 stop words from each corpus. From the AP data, we further removed words that occurred only once.

We compared LDA with the unigram, mixture of unigrams, and pLSI models described in Section 4. We trained all the hidden variable models using EM with exactly the same stopping criteria, that the average change in expected log likelihood is less than 0.001%.

Both the pLSI model and the mixture of unigrams suffer from serious overfitting issues, though for different reasons. This phenomenon is illustrated in Table 1. In the mixture of unigrams model, overfitting is a result of peaked posteriors in the training set; a phenomenon familiar in the supervised setting, where this model is known as the naive Bayes model (Rennie, 2001). This leads to a nearly deterministic clustering of the training documents (in the E-step) which is used to determine the word probabilities in each mixture component (in the M-step). A previously unseen document may best fit one of the resulting mixture components, but will probably contain at least one word which did not occur in the training documents that were assigned to that component. Such words will have a very small probability, which causes the perplexity of the new document to explode. As  $k$  increases, the documents of the training corpus are partitioned into finer collections and thus induce more words with small probabilities.

In the mixture of unigrams, we can alleviate overfitting through the variational Bayesian smoothing scheme presented in Section 5.4. This ensures that all words will have some probability under every mixture component.

In the pLSI case, the hard clustering problem is alleviated by the fact that each document is allowed to exhibit a different proportion of topics. However, pLSI only refers to the training documents and a different overfitting problem arises that is due to the dimensionality of the  $p(z|d)$  parameter. One reasonable approach to assigning probability to a previously unseen document is by marginalizing over  $d$ :

$$p(\mathbf{w}) = \sum_d \prod_{n=1}^N \sum_z p(w_n | z) p(z | d) p(d).$$

Essentially, we are integrating over the empirical distribution on the topic simplex (see Figure 4).

This method of inference, though theoretically sound, causes the model to overfit. The document-specific topic distribution has some components which are close to zero for those topics that do not appear in the document. Thus, certain words will have very small probability in the estimates of

each mixture component. When determining the probability of a new document through marginalization, only those training documents which exhibit a similar proportion of topics will contribute to the likelihood. For a given training document’s topic proportions, any word which has small probability in all the constituent topics will cause the perplexity to explode. As  $k$  gets larger, the chance that a training document will exhibit topics that cover all the words in the new document decreases and thus the perplexity grows. Note that pLSI does not overfit as quickly (with respect to  $k$ ) as the mixture of unigrams.

This overfitting problem essentially stems from the restriction that each future document exhibit the same topic proportions as were seen in one or more of the training documents. Given this constraint, we are not free to choose the most likely proportions of topics for the new document. An alternative approach is the “folding-in” heuristic suggested by Hofmann (1999), where one ignores the  $p(z|d)$  parameters and refits  $p(z|d_{\text{new}})$ . Note that this gives the pLSI model an unfair advantage by allowing it to refit  $k - 1$  parameters to the test data.

LDA suffers from neither of these problems. As in pLSI, each document can exhibit a different proportion of underlying topics. However, LDA can easily assign probability to a new document; no heuristics are needed for a new document to be endowed with a different set of topic proportions than were associated with documents in the training corpus.

Figure 9 presents the perplexity for each model on both corpora for different values of  $k$ . The pLSI model and mixture of unigrams are suitably corrected for overfitting. The latent variable models perform better than the simple unigram model. LDA consistently performs better than the other models.

## 7.2 Document classification

In the text classification problem, we wish to classify a document into two or more mutually exclusive classes. As in any classification problem, we may wish to consider generative approaches or discriminative approaches. In particular, by using one LDA module for each class, we obtain a generative model for classification. It is also of interest to use LDA in the discriminative framework, and this is our focus in this section.

A challenging aspect of the document classification problem is the choice of features. Treating individual words as features yields a rich but very large feature set (Joachims, 1999). One way to reduce this feature set is to use an LDA model for dimensionality reduction. In particular, LDA reduces any document to a fixed set of real-valued features—the posterior Dirichlet parameters  $\gamma^*(w)$  associated with the document. It is of interest to see how much discriminatory information we lose in reducing the document description to these parameters.

We conducted two binary classification experiments using the Reuters-21578 dataset. The dataset contains 8000 documents and 15,818 words.

In these experiments, we estimated the parameters of an LDA model on all the documents, without reference to their true class label. We then trained a support vector machine (SVM) on the low-dimensional representations provided by LDA and compared this SVM to an SVM trained on all the word features.

Using the SVMLight software package (Joachims, 1999), we compared an SVM trained on all the word features with those trained on features induced by a 50-topic LDA model. Note that we reduce the feature space by 99.6 percent in this case.

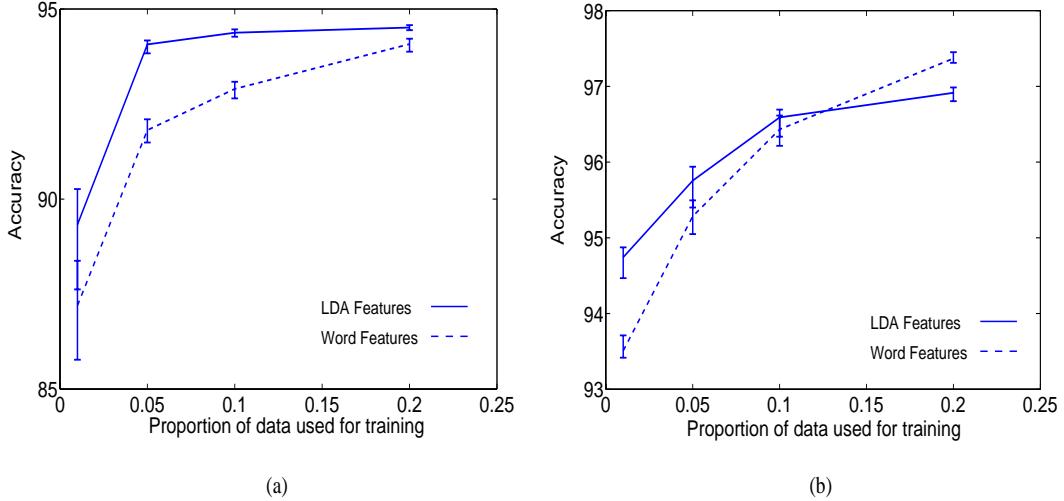


Figure 10: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN vs. NOT EARN. Graph (b) is GRAIN vs. NOT GRAIN.

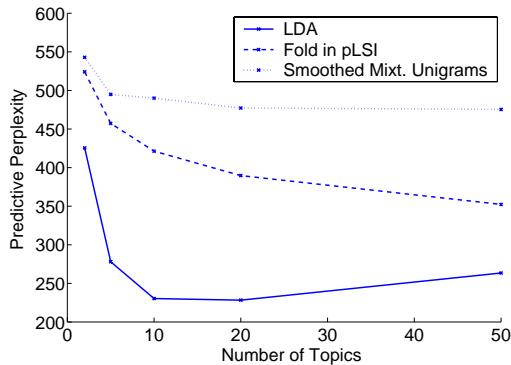


Figure 11: Results for collaborative filtering on the EachMovie data.

Figure 10 shows our results. We see that there is little reduction in classification performance in using the LDA-based features; indeed, in almost all cases the performance is improved with the LDA features. Although these results need further substantiation, they suggest that the topic-based representation provided by LDA may be useful as a fast filtering algorithm for feature selection in text classification.

### 7.3 Collaborative filtering

Our final experiment uses the EachMovie collaborative filtering data. In this data set, a collection of users indicates their preferred movie choices. A user and the movies chosen are analogous to a document and the words in the document (respectively).

The collaborative filtering task is as follows. We train a model on a fully observed set of users. Then, for each unobserved user, we are shown all but one of the movies preferred by that user and are asked to predict what the held-out movie is. The different algorithms are evaluated according to the likelihood they assign to the held-out movie. More precisely, define the predictive perplexity on  $M$  test users to be:

$$\text{predictive-perplexity}(D_{\text{test}}) = \exp \left\{ -\frac{\sum_{d=1}^M \log p(w_{d,N_d} | \mathbf{w}_{d,1:N_d-1})}{M} \right\}.$$

We restricted the EachMovie dataset to users that positively rated at least 100 movies (a positive rating is at least four out of five stars). We divided this set of users into 3300 training users and 390 testing users.

Under the mixture of unigrams model, the probability of a movie given a set of observed movies is obtained from the posterior distribution over topics:

$$p(w|\mathbf{w}_{\text{obs}}) = \sum_z p(w|z)p(z|\mathbf{w}_{\text{obs}}).$$

In the pLSI model, the probability of a held-out movie is given by the same equation except that  $p(z|\mathbf{w}_{\text{obs}})$  is computed by folding in the previously seen movies. Finally, in the LDA model, the probability of a held-out movie is given by integrating over the posterior Dirichlet:

$$p(w|\mathbf{w}_{\text{obs}}) = \int \sum_z p(w|z)p(z|\theta)p(\theta|\mathbf{w}_{\text{obs}})d\theta,$$

where  $p(\theta|\mathbf{w}_{\text{obs}})$  is given by the variational inference method described in Section 5.2. Note that this quantity is efficient to compute. We can interchange the sum and integral sign, and compute a linear combination of  $k$  Dirichlet expectations.

With a vocabulary of 1600 movies, we find the predictive perplexities illustrated in Figure 11. Again, the mixture of unigrams model and pLSI are corrected for overfitting, but the best predictive perplexities are obtained by the LDA model.

## 8. Discussion

We have described latent Dirichlet allocation, a flexible generative probabilistic model for collections of discrete data. LDA is based on a simple exchangeability assumption for the words and topics in a document; it is therefore realized by a straightforward application of de Finetti's representation theorem. We can view LDA as a dimensionality reduction technique, in the spirit of LSI, but with proper underlying generative probabilistic semantics that make sense for the type of data that it models.

Exact inference is intractable for LDA, but any of a large suite of approximate inference algorithms can be used for inference and parameter estimation within the LDA framework. We have presented a simple convexity-based variational approach for inference, showing that it yields a fast

algorithm resulting in reasonable comparative performance in terms of test set likelihood. Other approaches that might be considered include Laplace approximation, higher-order variational techniques, and Monte Carlo methods. In particular, Leisink and Kappen (2002) have presented a general methodology for converting low-order variational lower bounds into higher-order variational bounds. It is also possible to achieve higher accuracy by dispensing with the requirement of maintaining a bound, and indeed Minka and Lafferty (2002) have shown that improved inferential accuracy can be obtained for the LDA model via a higher-order variational technique known as expectation propagation. Finally, Griffiths and Steyvers (2002) have presented a Markov chain Monte Carlo algorithm for LDA.

LDA is a simple model, and although we view it as a competitor to methods such as LSI and pLSI in the setting of dimensionality reduction for document collections and other discrete corpora, it is also intended to be illustrative of the way in which probabilistic models can be scaled up to provide useful inferential machinery in domains involving multiple levels of structure. Indeed, the principal advantages of generative models such as LDA include their modularity and their extensibility. As a probabilistic module, LDA can be readily embedded in a more complex model—a property that is not possessed by LSI. In recent work we have used pairs of LDA modules to model relationships between images and their corresponding descriptive captions (Blei and Jordan, 2002). Moreover, there are numerous possible extensions of LDA. For example, LDA is readily extended to continuous data or other non-multinomial data. As is the case for other mixture models, including finite mixture models and hidden Markov models, the “emission” probability  $p(w_n | z_n)$  contributes only a likelihood value to the inference procedures for LDA, and other likelihoods are readily substituted in its place. In particular, it is straightforward to develop a continuous variant of LDA in which Gaussian observables are used in place of multinomials. Another simple extension of LDA comes from allowing mixtures of Dirichlet distributions in the place of the single Dirichlet of LDA. This allows a richer structure in the latent topic space and in particular allows a form of document clustering that is different from the clustering that is achieved via shared topics. Finally, a variety of extensions of LDA can be considered in which the distributions on the topic variables are elaborated. For example, we could arrange the topics in a time series, essentially relaxing the full exchangeability assumption to one of partial exchangeability. We could also consider partially exchangeable models in which we condition on exogenous variables; thus, for example, the topic distribution could be conditioned on features such as “paragraph” or “sentence,” providing a more powerful text model that makes use of information obtained from a parser.

## Acknowledgements

This work was supported by the National Science Foundation (NSF grant IIS-9988642) and the Multidisciplinary Research Program of the Department of Defense (MURI N00014-00-1-0637). Andrew Y. Ng and David M. Blei were additionally supported by fellowships from the Microsoft Corporation.

## References

- M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions*. Dover, New York, 1970.

- D. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, pages 1–198. Springer, Berlin, 1985.
- H. Attias. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*, 2000.
- L. Avery. Caenorhabditis genetic center bibliography. 2002. URL <http://elegans.swmed.edu/wli/cgcbbib>.
- R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.
- D. Blei and M. Jordan. Modeling annotated data. Technical Report UCB//CSD-02-1202, U.C. Berkeley Computer Science Division, 2002.
- B. de Finetti. *Theory of probability. Vol. 1-2*. John Wiley & Sons Ltd., Chichester, 1990. Reprint of the 1975 translation.
- S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
- P. Diaconis. Recent progress on de Finetti’s notions of exchangeability. In *Bayesian statistics, 3 (Valencia, 1987)*, pages 111–125. Oxford Univ. Press, New York, 1988.
- J. Dickey. Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, 78:628–637, 1983.
- J. Dickey, J. Jiang, and J. Kadane. Bayesian methods for censored categorical data. *Journal of the American Statistical Association*, 82:773–781, 1987.
- A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman & Hall, London, 1995.
- T. Griffiths and M. Steyvers. A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 2002.
- D. Harman. Overview of the first text retrieval conference (TREC-1). In *Proceedings of the First Text Retrieval Conference (TREC-1)*, pages 1–20, 1992.
- D. Heckerman and M. Meila. An experimental comparison of several clustering and initialization methods. *Machine Learning*, 42:9–29, 2001.
- T. Hofmann. Probabilistic latent semantic indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference*, 1999.
- F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, 1997.
- T. Joachims. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*. M.I.T. Press, 1999.
- M. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

- M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.
- R. Kass and D. Steffey. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84(407):717–726, 1989.
- M. Leisink and H. Kappen. General lower bounds based on computer generated higher order expansions. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference*, 2002.
- T. Minka. Estimating a Dirichlet distribution. Technical report, M.I.T., 2000.
- T. P. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Uncertainty in Artificial Intelligence (UAI)*, 2002.
- C. Morris. Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–65, 1983. With discussion.
- K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.
- K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.
- C. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala. Latent semantic indexing: A probabilistic analysis. pages 159–168, 1998.
- A. Popescul, L. Ungar, D. Pennock, and S. Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference*, 2001.
- J. Rennie. Improving multi-class text classification with naive Bayes. Technical Report AITR-2001-004, M.I.T., 2001.
- G. Ronning. Maximum likelihood estimation of Dirichlet distributions. *Journal of Statistical Computation and Simulation*, 34(4):215–221, 1989.
- G. Salton and M. McGill, editors. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

## Appendix A. Inference and parameter estimation

In this appendix, we derive the variational inference procedure (Eqs. 6 and 7) and the parameter maximization procedure for the conditional multinomial (Eq. 9) and for the Dirichlet. We begin by deriving a useful property of the Dirichlet distribution.

### A.1 Computing $E[\log(\theta_i | \alpha)]$

The need to compute the expected value of the log of a single probability component under the Dirichlet arises repeatedly in deriving the inference and parameter estimation procedures for LDA. This value can be easily computed from the natural parameterization of the exponential family representation of the Dirichlet distribution.

Recall that a distribution is in the exponential family if it can be written in the form:

$$p(x|\eta) = h(x) \exp \{ \eta^T T(x) - A(\eta) \},$$

where  $\eta$  is the natural parameter,  $T(x)$  is the sufficient statistic, and  $A(\eta)$  is the log of the normalization factor.

We can write the Dirichlet in this form by exponentiating the log of Eq. (1):

$$p(\theta | \alpha) = \exp \left\{ \left( \sum_{i=1}^k (\alpha_i - 1) \log \theta_i \right) + \log \Gamma \left( \sum_{i=1}^k \alpha_i \right) - \sum_{i=1}^k \log \Gamma(\alpha_i) \right\}.$$

From this form, we immediately see that the natural parameter of the Dirichlet is  $\eta_i = \alpha_i - 1$  and the sufficient statistic is  $T(\theta_i) = \log \theta_i$ . Furthermore, using the general fact that the derivative of the log normalization factor with respect to the natural parameter is equal to the expectation of the sufficient statistic, we obtain:

$$E[\log \theta_i | \alpha] = \Psi(\alpha_i) - \Psi \left( \sum_{j=1}^k \alpha_j \right)$$

where  $\Psi$  is the digamma function, the first derivative of the log Gamma function.

### A.2 Newton-Raphson methods for a Hessian with special structure

In this section we describe a linear algorithm for the usually cubic Newton-Raphson optimization method. This method is used for maximum likelihood estimation of the Dirichlet distribution (Ronning, 1989, Minka, 2000).

The Newton-Raphson optimization technique finds a stationary point of a function by iterating:

$$\alpha_{\text{new}} = \alpha_{\text{old}} - H(\alpha_{\text{old}})^{-1} g(\alpha_{\text{old}})$$

where  $H(\alpha)$  and  $g(\alpha)$  are the Hessian matrix and gradient respectively at the point  $\alpha$ . In general, this algorithm scales as  $O(N^3)$  due to the matrix inversion.

If the Hessian matrix is of the form:

$$H = \text{diag}(h) + \mathbf{1}\mathbf{1}^T, \quad (10)$$

where  $\text{diag}(h)$  is defined to be a diagonal matrix with the elements of the vector  $h$  along the diagonal, then we can apply the matrix inversion lemma and obtain:

$$H^{-1} = \text{diag}(h)^{-1} - \frac{\text{diag}(h)^{-1} \mathbf{1}\mathbf{1}^T \text{diag}(h)^{-1}}{z^{-1} + \sum_{j=1}^k h_j^{-1}}$$

Multiplying by the gradient, we obtain the  $i$ th component:

$$(H^{-1} g)_i = \frac{g_i - c}{h_i}$$

where

$$c = \frac{\sum_{j=1}^k g_j / h_j}{z^{-1} + \sum_{j=1}^k h_j^{-1}}.$$

Observe that this expression depends only on the  $2k$  values  $h_i$  and  $g_i$  and thus yields a Newton-Raphson algorithm that has linear time complexity.

### A.3 Variational inference

In this section we derive the variational inference algorithm described in Section 5.1. Recall that this involves using the following *variational distribution*:

$$q(\theta, \mathbf{z} | \gamma, \phi) = q(\theta | \gamma) \prod_{n=1}^N q(z_n | \phi_n) \quad (11)$$

as a surrogate for the posterior distribution  $p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)$ , where the *variational parameters*  $\gamma$  and  $\phi$  are set via an optimization procedure that we now describe.

Following Jordan et al. (1999), we begin by bounding the log likelihood of a document using Jensen's inequality. Omitting the parameters  $\gamma$  and  $\phi$  for simplicity, we have:

$$\begin{aligned} \log p(\mathbf{w} | \alpha, \beta) &= \log \int \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) d\theta \\ &= \log \int \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) q(\theta, \mathbf{z})}{q(\theta, \mathbf{z})} d\theta \\ &\geq \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) d\theta - \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log q(\theta, \mathbf{z}) d\theta \\ &= E_q[\log p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)] - E_q[\log q(\theta, \mathbf{z})]. \end{aligned} \quad (12)$$

Thus we see that Jensen's inequality provides us with a lower bound on the log likelihood for an arbitrary variational distribution  $q(\theta, \mathbf{z} | \gamma, \phi)$ .

It can be easily verified that the difference between the left-hand side and the right-hand side of the Eq. (12) is the KL divergence between the variational posterior probability and the true posterior probability. That is, letting  $L(\gamma, \phi; \alpha, \beta)$  denote the right-hand side of Eq. (12) (where we have restored the dependence on the variational parameters  $\gamma$  and  $\phi$  in our notation), we have:

$$\log p(\mathbf{w} | \alpha, \beta) = L(\gamma, \phi; \alpha, \beta) + D(q(\theta, \mathbf{z} | \gamma, \phi) \| p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)). \quad (13)$$

This shows that maximizing the lower bound  $L(\gamma, \phi; \alpha, \beta)$  with respect to  $\gamma$  and  $\phi$  is equivalent to minimizing the KL divergence between the variational posterior probability and the true posterior probability, the optimization problem presented earlier in Eq. (5).

We now expand the lower bound by using the factorizations of  $p$  and  $q$ :

$$\begin{aligned} L(\gamma, \phi; \alpha, \beta) &= E_q[\log p(\theta | \alpha)] + E_q[\log p(\mathbf{z} | \theta)] + E_q[\log p(\mathbf{w} | \mathbf{z}, \beta)] \\ &\quad - E_q[\log q(\theta)] - E_q[\log q(\mathbf{z})]. \end{aligned} \quad (14)$$

Finally, we expand Eq. (14) in terms of the model parameters  $(\alpha, \beta)$  and the variational parameters  $(\gamma, \phi)$ . Each of the five lines below expands one of the five terms in the bound:

$$\begin{aligned}
L(\gamma, \phi; \alpha, \beta) = & \log \Gamma\left(\sum_{j=1}^k \alpha_j\right) - \sum_{i=1}^k \log \Gamma(\alpha_i) + \sum_{i=1}^k (\alpha_i - 1) (\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)) \\
& + \sum_{n=1}^N \sum_{i=1}^k \phi_{ni} (\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)) \\
& + \sum_{n=1}^N \sum_{i=1}^k \sum_{j=1}^V \phi_{ni} w_n^j \log \beta_{ij} \\
& - \log \Gamma\left(\sum_{j=1}^k \gamma_j\right) + \sum_{i=1}^k \log \Gamma(\gamma_i) - \sum_{i=1}^k (\gamma_i - 1) (\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)) \\
& - \sum_{n=1}^N \sum_{i=1}^k \phi_{ni} \log \phi_{ni},
\end{aligned} \tag{15}$$

where we have made use of Eq. (8).

In the following two sections, we show how to maximize this lower bound with respect to the variational parameters  $\phi$  and  $\gamma$ .

### A.3.1 VARIATIONAL MULTINOMIAL

We first maximize Eq. (15) with respect to  $\phi_{ni}$ , the probability that the  $n$ th word is generated by latent topic  $i$ . Observe that this is a constrained maximization since  $\sum_{i=1}^k \phi_{ni} = 1$ .

We form the Lagrangian by isolating the terms which contain  $\phi_{ni}$  and adding the appropriate Lagrange multipliers. Let  $\beta_{iv}$  be  $p(w_n^v = 1 | z^i = 1)$  for the appropriate  $v$ . (Recall that each  $w_n$  is a vector of size  $V$  with exactly one component equal to one; we can select the unique  $v$  such that  $w_n^v = 1$ ):

$$L_{[\phi_{ni}]} = \phi_{ni} (\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)) + \phi_{ni} \log \beta_{iv} - \phi_{ni} \log \phi_{ni} + \lambda_n (\sum_{j=1}^k \phi_{nj} - 1),$$

where we have dropped the arguments of  $L$  for simplicity, and where the subscript  $\phi_{ni}$  denotes that we have retained only those terms in  $L$  that are a function of  $\phi_{ni}$ . Taking derivatives with respect to  $\phi_{ni}$ , we obtain:

$$\frac{\partial L}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j) + \log \beta_{iv} - \log \phi_{ni} - 1 + \lambda.$$

Setting this derivative to zero yields the maximizing value of the variational parameter  $\phi_{ni}$  (cf. Eq. 6):

$$\phi_{ni} \propto \beta_{iv} \exp(\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)). \tag{16}$$

### A.3.2 VARIATIONAL DIRICHLET

Next, we maximize Eq. (15) with respect to  $\gamma_i$ , the  $i$ th component of the posterior Dirichlet parameter. The terms containing  $\gamma_i$  are:

$$\begin{aligned} L_{[\gamma]} = & \sum_{i=1}^k (\alpha_i - 1) (\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)) + \sum_{n=1}^N \phi_{ni} (\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)) \\ & - \log \Gamma(\sum_{j=1}^k \gamma_j) + \log \Gamma(\gamma_i) - \sum_{i=1}^k (\gamma_i - 1) (\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)). \end{aligned}$$

This simplifies to:

$$L_{[\gamma]} = \sum_{i=1}^k (\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)) (\alpha_i + \sum_{n=1}^N \phi_{ni} - \gamma_i) - \log \Gamma(\sum_{j=1}^k \gamma_j) + \log \Gamma(\gamma_i).$$

We take the derivative with respect to  $\gamma_i$ :

$$\frac{\partial L}{\partial \gamma_i} = \Psi'(\gamma_i) (\alpha_i + \sum_{n=1}^N \phi_{ni} - \gamma_i) - \Psi'(\sum_{j=1}^k \gamma_j) \sum_{j=1}^k (\alpha_j + \sum_{n=1}^N \phi_{nj} - \gamma_j).$$

Setting this equation to zero yields a maximum at:

$$\gamma_i = \alpha_i + \sum_{n=1}^N \phi_{ni}. \quad (17)$$

Since Eq. (17) depends on the variational multinomial  $\phi$ , full variational inference requires alternating between Eqs. (16) and (17) until the bound converges.

### A.4 Parameter estimation

In this final section, we consider the problem of obtaining empirical Bayes estimates of the model parameters  $\alpha$  and  $\beta$ . We solve this problem by using the variational lower bound as a surrogate for the (intractable) marginal log likelihood, with the variational parameters  $\phi$  and  $\gamma$  fixed to the values found by variational inference. We then obtain (approximate) empirical Bayes estimates by maximizing this lower bound with respect to the model parameters.

We have thus far considered the log likelihood for a single document. Given our assumption of exchangeability for the documents, the overall log likelihood of a corpus  $D = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$  is the sum of the log likelihoods for individual documents; moreover, the overall variational lower bound is the sum of the individual variational bounds. In the remainder of this section, we abuse notation by using  $L$  for the total variational bound, indexing the document-specific terms in the individual bounds by  $d$ , and summing over all the documents.

Recall from Section 5.3 that our overall approach to finding empirical Bayes estimates is based on a variational EM procedure. In the variational E-step, discussed in Appendix A.3, we maximize the bound  $L(\gamma, \phi; \alpha, \beta)$  with respect to the variational parameters  $\gamma$  and  $\phi$ . In the M-step, which we describe in this section, we maximize the bound with respect to the model parameters  $\alpha$  and  $\beta$ . The overall procedure can thus be viewed as coordinate ascent in  $L$ .

#### A.4.1 CONDITIONAL MULTINOMIALS

To maximize with respect to  $\beta$ , we isolate terms and add Lagrange multipliers:

$$L_{[\beta]} = \sum_{d=1}^M \sum_{n=1}^{N_d} \sum_{i=1}^k \sum_{j=1}^V \phi_{dn} w_{dn}^j \log \beta_{ij} + \sum_{i=1}^k \lambda_i \left( \sum_{j=1}^V \beta_{ij} - 1 \right).$$

We take the derivative with respect to  $\beta_{ij}$ , set it to zero, and find:

$$\beta_{ij} \propto \sum_{d=1}^M \sum_{n=1}^{N_d} \phi_{dn} w_{dn}^j.$$

#### A.4.2 DIRICHLET

The terms which contain  $\alpha$  are:

$$L_{[\alpha]} = \sum_{d=1}^M \left( \log \Gamma \left( \sum_{j=1}^k \alpha_j \right) - \sum_{i=1}^k \log \Gamma(\alpha_i) + \sum_{i=1}^k ((\alpha_i - 1) (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^k \gamma_{dj}))) \right)$$

Taking the derivative with respect to  $\alpha_i$  gives:

$$\frac{\partial L}{\partial \alpha_i} = M (\Psi(\sum_{j=1}^k \alpha_j) - \Psi(\alpha_i)) + \sum_{d=1}^M (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^k \gamma_{dj}))$$

This derivative depends on  $\alpha_j$ , where  $j \neq i$ , and we therefore must use an iterative method to find the maximal  $\alpha$ . In particular, the Hessian is in the form found in Eq. (10):

$$\frac{\partial L}{\partial \alpha_i \alpha_j} = \delta(i, j) M \Psi'(\alpha_i) - \Psi'(\sum_{j=1}^k \alpha_j),$$

and thus we can invoke the linear-time Newton-Raphson algorithm described in Appendix A.2.

Finally, note that we can use the same algorithm to find an empirical Bayes point estimate of  $\eta$ , the scalar parameter for the exchangeable Dirichlet in the smoothed LDA model in Section 5.4.

DOI:10.1145/2133806.2133826

## Surveying a suite of algorithms that offer a solution to managing large document archives.

BY DAVID M. BLEI

# Probabilistic Topic Models

AS OUR COLLECTIVE knowledge continues to be digitized and stored—in the form of news, blogs, Web pages, scientific articles, books, images, sound, video, and social networks—it becomes more difficult to find and discover what we are looking for. We need new computational tools to help organize, search, and understand these vast amounts of information.

Right now, we work with online information using two main tools—search and links. We type keywords into a search engine and find a set of documents related to them. We look at the documents in that set, possibly navigating to other linked documents. This is a powerful way of interacting with our online archive, but something is missing.

Imagine searching and exploring documents based on the themes that run through them. We might “zoom in” and “zoom out” to find specific or broader themes; we might look at how those themes changed through time or how they are connected to each other. Rather than finding documents through keyword search alone, we might first find the theme that we are interested in, and then examine the documents related to that theme.

For example, consider using themes to explore the complete history of the New York Times. At a broad level, some of the themes might correspond to the sections of the newspaper—foreign policy, national affairs, sports. We could zoom in on a theme of interest, such as foreign policy, to reveal various aspects of it—Chinese foreign policy, the conflict in the Middle East, the U.S.’s relationship with Russia. We could then navigate through time to reveal how these specific themes have changed, tracking, for example, the changes in the conflict in the Middle East over the last 50 years. And, in all of this exploration, we would be pointed to the original articles relevant to the themes. The thematic structure would be a new kind of window through which to explore and digest the collection.

But we do not interact with electronic archives in this way. While more and more texts are available online, we simply do not have the human power to read and study them to provide the kind of browsing experience described above. To this end, machine learning researchers have developed *probabilistic topic modeling*, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information. Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over

### » key insights

- Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.
- Topic modeling algorithms can be applied to massive collections of documents. Recent advances in this field allow us to analyze streaming collections, like you might find from a Web API.
- Topic modeling algorithms can be adapted to many kinds of data. Among other applications, they have been used to find patterns in genetic data, images, and social networks.

time. (See, for example, Figure 3 for topics found by analyzing the *Yale Law Journal*.) Topic modeling algorithms do not require any prior annotations or labeling of the documents—the topics emerge from the analysis of the original texts. Topic modeling enables us to organize and summarize electronic archives at a scale that would be impossible by human annotation.

### Latent Dirichlet Allocation

We first describe the basic ideas behind *latent Dirichlet allocation* (LDA), which is the simplest topic model.<sup>8</sup> The intuition behind LDA is that documents exhibit multiple topics. For example, consider the article in Figure 1. This article, entitled “Seeking Life’s Bare (Genetic) Necessities,” is about using data analysis to determine the number of genes an organism needs to survive (in an evolutionary sense).

By hand, we have highlighted different words that are used in the article. Words about *data analysis*, such as “computer” and “prediction,” are highlighted in blue; words about *evolutionary biology*, such as “life” and “organism,” are highlighted in pink; words about *genetics*, such as “sequenced” and

“genes,” are highlighted in yellow. If we took the time to highlight every word in the article, you would see that this article blends genetics, data analysis, and evolutionary biology in different proportions. (We exclude words, such as “and” “but” or “if,” which contain little topical content.) Furthermore, knowing that this article blends those topics would help you situate it in a collection of scientific articles.

LDA is a statistical model of document collections that tries to capture this intuition. It is most easily described by its generative process, the imaginary random process by which the model assumes the documents arose. (The interpretation of LDA as a probabilistic model is fleshed out later.)

We formally define a *topic* to be a distribution over a fixed vocabulary. For example, the *genetics* topic has words about genetics with high probability and the *evolutionary biology* topic has words about evolutionary biology with high probability. We assume that these topics are specified before any data has been generated.<sup>a</sup> Now for each

<sup>a</sup> Technically, the model assumes that the topics are generated first, before the documents.

document in the collection, we generate the words in a two-stage process.

► Randomly choose a distribution over topics.

► For each word in the document

- Randomly choose a topic from the distribution over topics in step #1.

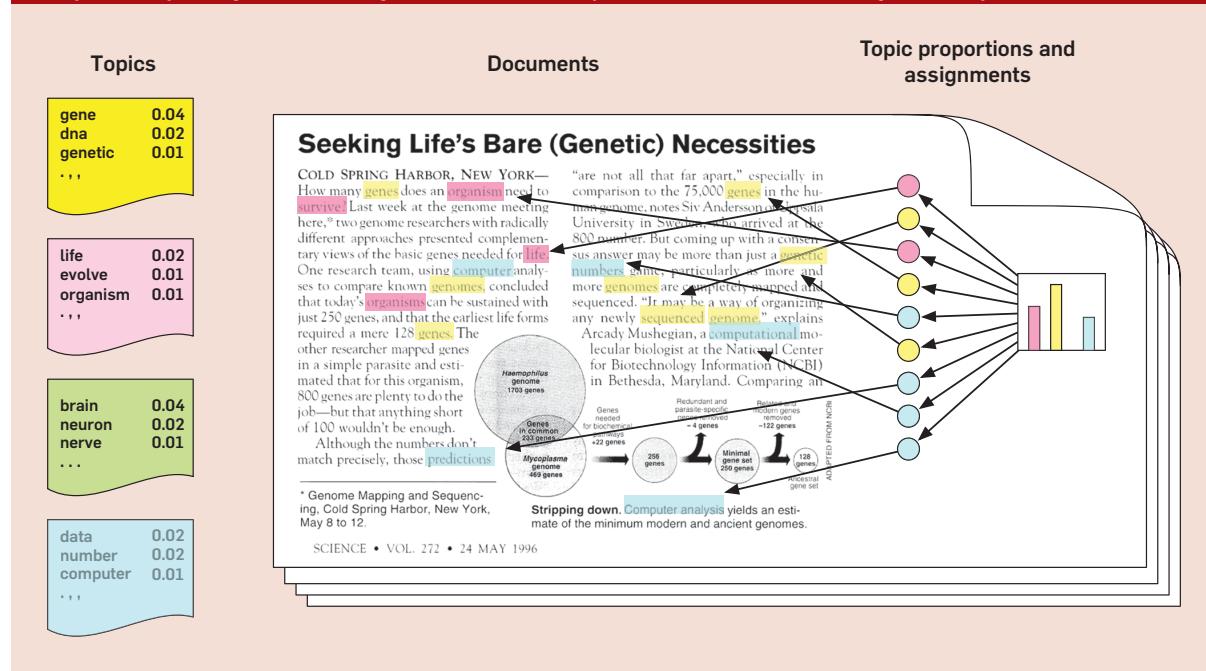
- Randomly choose a word from the corresponding distribution over the vocabulary.

This statistical model reflects the intuition that documents exhibit multiple topics. Each document exhibits the topics in different proportion (step #1); each word in each document is drawn from one of the topics (step #2b), where the selected topic is chosen from the per-document distribution over topics (step #2a).<sup>b</sup>

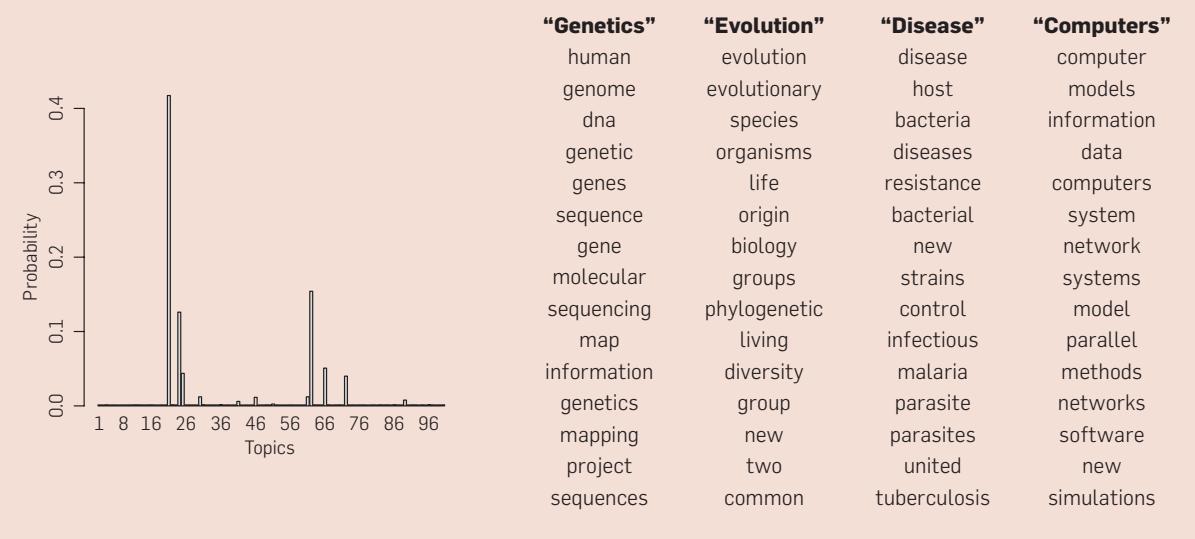
In the example article, the distribution over topics would place probability on *genetics*, *data analysis*, and

<sup>b</sup> We should explain the mysterious name, “latent Dirichlet allocation.” The distribution that is used to draw the per-document topic distributions in step #1 (the cartoon histogram in Figure 1) is called a *Dirichlet distribution*. In the generative process for LDA, the result of the Dirichlet is used to *allocate* the words of the document to different topics. Why *latent*? Keep reading.

**Figure 1. The intuitions behind latent Dirichlet allocation.** We assume that some number of “topics,” which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



**Figure 2. Real inference with LDA.** We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



*evolutionary biology*, and each word is drawn from one of those three topics. Notice that the next article in the collection might be about *data analysis* and *neuroscience*; its distribution over topics would place probability on those two topics. This is the distinguishing characteristic of latent Dirichlet allocation—all the documents in the collection share the same set of topics, but each document exhibits those topics in different proportion.

As we described in the introduction, the goal of topic modeling is to automatically discover the topics from a collection of documents. The documents themselves are observed, while the topic structure—the topics, per-document topic distributions, and the per-document per-word topic assignments—is *hidden structure*. The central computational problem for topic modeling is to use the observed documents to infer the hidden topic structure. This can be thought of as “reversing” the generative process—what is the hidden structure that likely generated the observed collection?

Figure 2 illustrates example inference using the same example document from Figure 1. Here, we took 17,000 articles from *Science* magazine and used a topic modeling algorithm to infer the hidden topic structure. (The

algorithm assumed that there were 100 topics.) We then computed the inferred topic distribution for the example article (Figure 2, left), the distribution over topics that best describes its particular collection of words. Notice that this topic distribution, though it can use any of the topics, has only “activated” a handful of them. Further, we can examine the most probable terms from each of the most probable topics (Figure 2, right). On examination, we see that these terms are recognizable as terms about genetics, survival, and data analysis, the topics that are combined in the example article.

We emphasize that the algorithms have no information about these subjects and the articles are not labeled with topics or keywords. The interpretable topic distributions arise by computing the hidden structure that likely generated the observed collection of documents.<sup>c</sup> For example, Figure 3 illustrates topics discovered from *Yale Law Journal*. (Here the number of topics was set to be 20.) Topics

about subjects like genetics and data analysis are replaced by topics about discrimination and contract law.

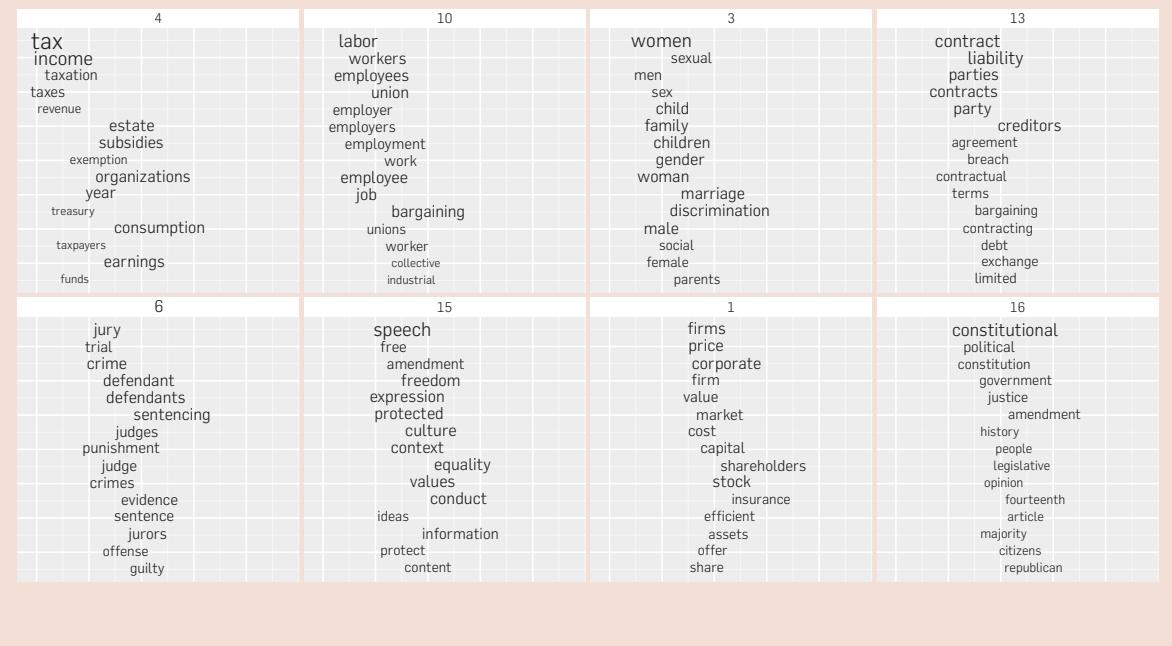
The utility of topic models stems from the property that the inferred hidden structure resembles the thematic structure of the collection. This interpretable hidden structure annotates each document in the collection—a task that is painstaking to perform by hand—and these annotations can be used to aid tasks like information retrieval, classification, and corpus exploration.<sup>d</sup> In this way, topic modeling provides an algorithmic solution to managing, organizing, and annotating large archives of texts.

**LDA and probabilistic models.** LDA and other topic models are part of the larger field of *probabilistic modeling*. In generative probabilistic modeling, we treat our data as arising from a generative process that includes *hidden variables*. This generative process defines a *joint probability distribution* over both the observed and hidden random variables. We perform data analysis by using that joint distribution to compute the *conditional distribution* of the hidden variables given the

<sup>c</sup> Indeed calling these models “topic models” is retrospective—the topics that emerge from the inference algorithm are interpretable for almost any collection that is analyzed. The fact that these look like topics has to do with the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of LDA.

<sup>d</sup> See, for example, the browser of *Wikipedia* built with a topic model at <http://www.sccs.swarthmore.edu/users/08/ajb/tmve/wiki100k/browse/topic-list.html>.

**Figure 3.** A topic model fit to the *Yale Law Journal*. Here, there are 20 topics (the top eight are plotted). Each topic is illustrated with its top-most frequent words. Each word's position along the x-axis denotes its specificity to the documents. For example “estate” in the first topic is more specific than “tax.”



observed variables. This conditional distribution is also called the *posterior distribution*.

LDA falls precisely into this framework. The observed variables are the words of the documents; the hidden variables are the topic structure; and the generative process is as described here. The computational problem of inferring the hidden topic structure from the documents is the problem of computing the posterior distribution, the conditional distribution of the hidden variables given the documents.

We can describe LDA more formally with the following notation. The topics are  $\beta_{1:K}$ , where each  $\beta_k$  is a distribution over the vocabulary (the distributions over words at left in Figure 1). The topic proportions for the  $d$ th document are  $\theta_d$ , where  $\theta_{d,k}$  is the topic proportion for topic  $k$  in document  $d$  (the cartoon histogram in Figure 1). The topic assignments for the  $d$ th document are  $z_d$ , where  $z_{d,n}$  is the topic assignment for the  $n$ th word in document  $d$  (the colored coin in Figure 1). Finally, the observed words for document  $d$  are  $w_d$ , where  $w_{d,n}$  is the  $n$ th word in document  $d$ , which is an element from the fixed vocabulary.

With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables,

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D p(\theta_d) \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right). \quad (1)$$

Notice that this distribution specifies a number of dependencies. For example, the topic assignment  $z_{d,n}$  depends on the per-document topic proportions  $\theta_d$ . As another example, the observed word  $w_{d,n}$  depends on the topic assignment  $z_{d,n}$  and all of the topics  $\beta_{1:K}$ . (Operationally, that term is defined by looking up as to which topic  $z_{d,n}$  refers to and looking up the probability of the word  $w_{d,n}$  within that topic.)

These dependencies define LDA. They are encoded in the statistical assumptions behind the generative process, in the particular mathematical form of the joint distribution, and—in a third way—in the *probabilistic graphical model* for LDA. Probabilistic graphical models provide a graphical

language for describing families of probability distributions.<sup>e</sup> The graphical model for LDA is in Figure 4. These three representations are equivalent ways of describing the probabilistic assumptions behind LDA.

In the next section, we describe the inference algorithms for LDA. However, we first pause to describe the short history of these ideas. LDA was developed to fix an issue with a previously developed probabilistic model *probabilistic latent semantic analysis* (pLSI).<sup>21</sup> That model was itself a probabilistic version of the seminal work on *latent semantic analysis*,<sup>14</sup> which revealed the utility of the singular value decomposition of the document-term matrix. From this matrix factorization perspective, LDA can also be seen as a type of principal component analysis for discrete data.<sup>11, 12</sup>

**Posterior computation for LDA.** We now turn to the computational

<sup>e</sup> The field of graphical models is actually more than a language for describing families of distributions. It is a field that illuminates the deep mathematical links between probabilistic independence, graph theory, and algorithms for computing with probability distributions.<sup>35</sup>

problem, computing the conditional distribution of the topic structure given the observed documents. (As we mentioned, this is called the *posterior*.) Using our notation, the posterior is

$$\frac{p(\beta_{1:D}, \theta_{1:D}, z_{1:D} | w_{1:D})}{p(w_{1:D})}. \quad (2)$$

The numerator is the joint distribution of all the random variables, which can be easily computed for any setting of the hidden variables. The denominator is the *marginal probability* of the observations, which is the probability of seeing the observed corpus under any topic model. In theory, it can be computed by summing the joint distribution over every possible instantiation of the hidden topic structure.

That number of possible topic structures, however, is exponentially large; this sum is intractable to compute.<sup>f</sup> As for many modern probabilistic models of interest—and for much of modern Bayesian statistics—we cannot compute the posterior because of the denominator, which is known as the *evidence*. A central research goal of modern probabilistic modeling is to develop efficient methods for approximating it. Topic modeling algorithms—like the algorithms used to create Figures 1 and 3—are often adaptations of general-purpose methods for approximating the posterior distribution.

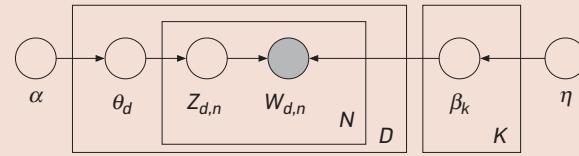
Topic modeling algorithms form an approximation of Equation 2 by adapting an alternative distribution over the latent topic structure to be close to the true posterior. Topic modeling algorithms generally fall into two categories—sampling-based algorithms and variational algorithms.

Sampling-based algorithms attempt to collect samples from the posterior to approximate it with an empirical distribution. The most commonly used sampling algorithm for topic modeling is *Gibbs sampling*, where we construct a *Markov chain*—a sequence of random variables, each dependent on the previous—whose

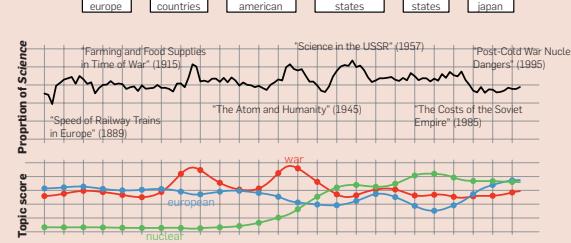
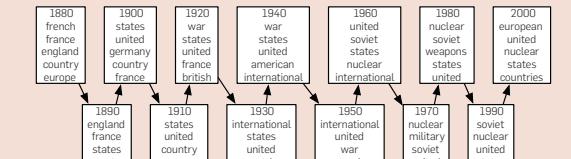
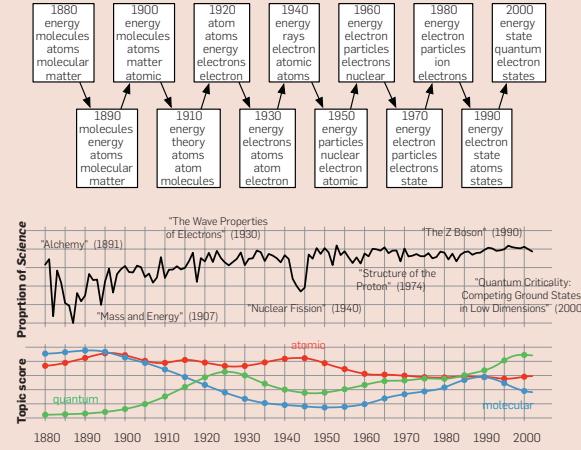
limiting distribution is the posterior. The Markov chain is defined on the hidden topic variables for a particular corpus, and the algorithm is to run the chain for a long time, collect samples

from the limiting distribution, and then approximate the distribution with the collected samples. (Often, just one sample is collected as an approximation of the topic structure with

**Figure 4. The graphical model for latent Dirichlet allocation.** Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes—the topic proportions, assignments, and topics—are unshaded. The observed nodes—the words of the documents—are shaded. The rectangles are “plate” notation, which denotes replication. The  $N$  plate denotes the collection words within documents; the  $D$  plate denotes the collection of documents within the collection.



**Figure 5. Two topics from a dynamic topic model.** This model was fit to *Science* from 1880 to 2002. We have illustrated the top words at each decade.



<sup>f</sup> More technically, the sum is over all possible ways of assigning each observed word of the collection to one of the topics. Document collections usually contain observed words at least on the order of millions.

maximal probability.) See Steyvers and Griffiths<sup>33</sup> for a good description of Gibbs sampling for LDA, and see <http://CRAN.R-project.org/package=lda> for a fast open-source implementation.

Variational methods are a deterministic alternative to sampling-based algorithms.<sup>22,35</sup> Rather than approximating the posterior with samples, variational methods posit a parameterized family of distributions over the hidden structure and then find the member of that family that is closest to the posterior.<sup>g</sup> Thus, the inference problem is transformed to an optimization problem. Variational methods open the door for innovations in optimization to have practical impact in probabilistic modeling. See Blei et al.<sup>8</sup> for a coordinate ascent variational inference algorithm for LDA; see Hoffman et al.<sup>20</sup> for a much faster online algorithm (and open-source software) that easily handles millions of documents and can accommodate streaming collections of text.

Loosely speaking, both types of algorithms perform a search over the topic structure. A collection of documents (the observed random variables in the model) are held fixed and serve as a guide toward where to search. Which approach is better depends on the particular topic model being used—we have so far focused on LDA, but see below for other topic models—and is a source of academic debate. For a good discussion of the merits and drawbacks of both, see Asuncion et al.<sup>1</sup>

#### **Research in Topic Modeling**

The simple LDA model provides a powerful tool for discovering and exploiting the hidden thematic structure in large archives of text. However, one of the main advantages of formulating LDA as a probabilistic model is that it can easily be used as a module in more complicated models for more complicated goals. Since its introduction, LDA has been extended and adapted in many ways.

**Relaxing the assumptions of LDA.** LDA is defined by the statistical assumptions it makes about the

<sup>g</sup> Closeness is measured with *Kullback–Leibler divergence*, an information theoretic measurement of the distance between two probability distributions.

**One direction for topic modeling is to develop evaluation methods that match how the algorithms are used.**  
**How can we compare topic models based on how interpretable they are?**

corpus. One active area of topic modeling research is how to relax and extend these assumptions to uncover more sophisticated structure in the texts.

One assumption that LDA makes is the “bag of words” assumption, that the order of the words in the document does not matter. (To see this, note that the joint distribution of Equation 1 remains invariant to permutation of the words of the documents.) While this assumption is unrealistic, it is reasonable if our only goal is to uncover the coarse semantic structure of the texts.<sup>h</sup> For more sophisticated goals—such as language generation—it is patently not appropriate. There have been a number of extensions to LDA that model words nonexchangeably. For example, Wallach<sup>36</sup> developed a topic model that relaxes the bag of words assumption by assuming that the topics generate words conditional on the previous word; Griffiths et al.<sup>18</sup> developed a topic model that switches between LDA and a standard HMM. These models expand the parameter space significantly but show improved language modeling performance.

Another assumption is that the order of documents does not matter. Again, this can be seen by noticing that Equation 1 remains invariant to permutations of the ordering of documents in the collection. This assumption may be unrealistic when analyzing long-running collections that span years or centuries. In such collections, we may want to assume that the *topics* change over time. One approach to this problem is the dynamic topic model<sup>5</sup>—a model that respects the ordering of the documents and gives a richer posterior topical structure than LDA. Figure 5 shows a topic that results from analyzing all of *Science* magazine under the dynamic topic model. Rather than a single distribution over words, a topic is now a sequence of distributions over words. We can find an underlying theme of the collection and track how it has changed over time.

A third assumption about LDA is that the number of topics is assumed

<sup>h</sup> As a thought experiment, imagine shuffling the words of the article in Figure 1. Even when shuffled, you would be able to glean that the article has something to do with genetics.

known and fixed. The Bayesian nonparametric topic model<sup>34</sup> provides an elegant solution: the number of topics is determined by the collection during posterior inference, and furthermore, new documents can exhibit previously unseen topics. Bayesian nonparametric topic models have been extended to hierarchies of topics, which find a tree of topics, moving from more general to more concrete, whose particular structure is inferred from the data.<sup>3</sup>

There are still other extensions of LDA that relax various assumptions made by the model. The correlated topic model<sup>6</sup> and pachinko allocation machine<sup>24</sup> allow the occurrence of topics to exhibit correlation (for example, a document about *geology* is more likely to also be about *chemistry* than it is to be about *sports*); the spherical topic model<sup>28</sup> allows words to be *unlikely* in a topic (for example, “wrench” will be particularly unlikely in a topic about *cats*); sparse topic models enforce further structure in the topic distributions;<sup>37</sup> and “bursty” topic models provide a more realistic model of word counts.<sup>15</sup>

**Incorporating metadata.** In many text analysis settings, the documents contain additional information—such as author, title, geographic location, links, and others—that we might want to account for when fitting a topic model. There has been a flurry of research on adapting topic models to include metadata.

The author-topic model<sup>29</sup> is an early success story for this kind of research. The topic proportions are attached to authors; papers with multiple authors are assumed to attach each word to an author, drawn from a topic drawn from his or her topic proportions. The author-topic model allows for inferences about authors as well as documents. Rosen-Zvi et al. show examples of author similarity based on their topic proportions—such computations are not possible with LDA.

Many document collections are linked—for example, scientific papers are linked by citation or Web pages are linked by hyperlink—and several topic models have been developed to account for those links when estimating the topics. The *relational topic model* of Chang and Blei<sup>13</sup> assumes that each document is modeled as in LDA and that the links

between documents depend on the distance between their topic proportions. This is both a new topic model and a new network model. Unlike traditional statistical models of networks, the relational topic model takes into account node attributes (here, the words of the documents) in modeling the links.

Other work that incorporates metadata into topic models includes models of linguistic structure,<sup>10</sup> models that account for distances between corpora,<sup>38</sup> and models of named entities.<sup>26</sup> General-purpose methods for incorporating metadata into topic models include Dirichlet-multinomial regression models<sup>25</sup> and supervised topic models.<sup>7</sup>

**Other kinds of data.** In LDA, the topics are distributions over words and this discrete distribution generates observations (words in documents). One advantage of LDA is that these choices for the topic parameter and data-generating distribution can be adapted to other kinds of observations with only small changes to the corresponding inference algorithms. As a class of models, LDA can be thought of as a *mixed-membership model* of grouped data—rather than associating each group of observations (document) with one component (topic), each group exhibits multiple components in different proportions. LDA-like models have been adapted to many kinds of data, including survey data, user preferences, audio and music, computer code, network logs, and social networks. We describe two areas where mixed-membership models have been particularly successful.

In population genetics, the same probabilistic model was independently invented to find ancestral populations (for example, originating from Africa, Europe, the Middle East, among others) in the genetic ancestry of a sample of individuals.<sup>27</sup> The idea is that each individual’s genotype descends from one or more of the ancestral populations. Using a model much like LDA, biologists can both characterize the genetic patterns in those populations (the “topics”) and identify how each individual expresses them (the “topic proportions”). This model is powerful because the genetic patterns in ancestral populations can be hypothesized, even when “pure” samples from them are not available.

LDA has been widely used and adapted in computer vision, where the

inference algorithms are applied to natural images in the service of image retrieval, classification, and organization. Computer vision researchers have made a direct analogy from images to documents. In document analysis, we assume that documents exhibit multiple topics and the collection of documents exhibits the same set of topics. In image analysis, we assume that each image exhibits a combination of visual patterns and that the same visual patterns recur throughout a collection of images. (In a preprocessing step, the images are analyzed to form collections of “visual words.”) Topic modeling for computer vision has been used to classify images,<sup>16</sup> connect images and captions,<sup>4</sup> build image hierarchies,<sup>2,23,31</sup> and other applications.

## Future Directions

Topic modeling is an emerging field in machine learning, and there are many exciting new directions for research.

**Evaluation and model checking.** There is a disconnect between how topic models are evaluated and why we expect topic models to be useful. Typically, topic models are evaluated in the following way. First, hold out a subset of your corpus as the test set. Then, fit a variety of topic models to the rest of the corpus and approximate a measure of model fit (for example, probability) for each trained model on the test set. Finally, choose the model that achieves the best held-out performance.

But topic models are often used to organize, summarize, and help users explore large corpora, and there is no technical reason to suppose that held-out accuracy corresponds to better organization or easier interpretation. One open direction for topic modeling is to develop evaluation methods that match how the algorithms are used. How can we compare topic models based on how interpretable they are?

This is the *model checking* problem. When confronted with a new corpus and a new task, which topic model should I use? How can I decide which of the many modeling assumptions are important for my goals? How should I move between the many kinds of topic models that have been developed? These questions have been given some attention by statisticians,<sup>9,30</sup> but they have been scrutinized less for the scale

of problems that machine learning tackles. New computational answers to these questions would be a significant contribution to topic modeling.

#### Visualization and user interfaces.

Another promising future direction for topic modeling is to develop new methods of interacting with and visualizing topics and corpora. Topic models provide new exploratory structure in large collections—how can we best exploit that structure to aid in discovery and exploration?

One problem is how to display the topics. Typically, we display topics by listing the most frequent words of each (see Figure 2), but new ways of labeling the topics—by either choosing different words or displaying the chosen words differently—may be more effective. A further problem is how to best display a document with a topic model. At the document level, topic models provide potentially useful information about the structure of the document. Combined with effective topic labels, this structure could help readers identify the most interesting parts of the document. Moreover, the hidden topic proportions implicitly connect each document to the other documents (by considering a distance measure between topic proportions). How can we best display these connections? What is an effective interface to the whole corpus and its inferred topic structure?

These are user interface questions, and they are essential to topic modeling. Topic modeling algorithms show much promise for uncovering meaningful thematic structure in large collections of documents. But making this structure useful requires careful attention to information visualization and the corresponding user interfaces.

#### Topic models for data discovery.

Topic models have been developed with information engineering applications in mind. As a statistical model, however, topic models should be able to tell us something, or help us form a hypothesis, about the data. What can we learn about the language (and other data) based on the topic model posterior? Some work in this area has appeared in political science,<sup>19</sup> bibliometrics,<sup>17</sup> and psychology.<sup>32</sup> This kind of research adapts topic models to measure an external variable of interest, a

difficult task for unsupervised learning that must be carefully validated.

In general, this problem is best addressed by teaming computer scientists with other scholars to use topic models to help explore, visualize, and draw hypotheses from their data. In addition to scientific applications, such as genetics and neuroscience, one can imagine topic models coming to the service of history, sociology, linguistics, political science, legal studies, comparative literature, and other fields, where texts are a primary object of study. By working with scholars in diverse fields, we can begin to develop a new interdisciplinary computational methodology for working with and drawing conclusions from archives of texts.

#### Summary

We have surveyed *probabilistic topic models*, a suite of algorithms that provide a statistical solution to the problem of managing large archives of documents. With recent scientific advances in support of unsupervised machine learning—flexible components for modeling, scalable algorithms for posterior inference, and increased access to massive datasets—topic models promise to be an important component for summarizing and understanding our growing digitized archive of information. □

#### References

- Asuncion, A., Welling, M., Smyth, P., Teh, Y. On smoothing and inference for topic models. In *Uncertainty in Artificial Intelligence* (2009).
- Bart, E., Welling, M., Perona, P. Unsupervised organization of image collections: Taxonomies and beyond. *Trans. Pattern Recognit. Mach. Intell.* 33, 11 (2010) (2301–2315).
- Blei, D., Griffiths, T., Jordan, M. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM* 57, 2 (2010), 1–30.
- Blei, D., Jordan, M. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2003), ACM Press, 127–134.
- Blei, D., Lafferty, J. Dynamic topic models. In *International Conference on Machine Learning* (2006), ACM, New York, NY, USA, 113–120.
- Blei, D., Lafferty, J. A correlated topic model of Science. *Ann. Appl. Stat.*, 1, 1 (2007), 17–35.
- Blei, D., McAuliffe, J. Supervised topic models. In *Neural Information Processing Systems* (2007).
- Blei, D., Ng, A., Jordan, M. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3 (January 2003), 993–1022.
- Box, G. Sampling and Bayes' inference in scientific modeling and robustness. *J. Roy. Stat. Soc. Ser. B* 143, 4 (1980), 383–430.
- Boyd-Graber, J., Blei, D. Syntactic topic models. In *Neural Information Processing Systems* (2009).
- Buntine, W. Variational extensions to EM and multinomial PCA. In *European Conference on Machine Learning* (2002).
- Buntine, W., Jakulin, A. Discrete component analysis. *Subspace, Latent Structure and Feature Selection*. C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, Eds. Springer, 2006.
- Chang, J., Blei, D. Hierarchical relational models for document networks. *Ann. Appl. Stat.* 4, 1 (2010).
- Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inform. Sci.* 41, 6 (1990), 391–407.
- Doyle, G., Elkan, C. Accounting for burstiness in topic models. In *International Conference on Machine Learning* (2009), ACM, 281–288.
- Fei-Fei, L., Perona, P. A Bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Vision and Pattern Recognition* (2005), 524–531.
- Gerrish, S., Blei, D. A language-based approach to measuring scholarly impact. In *International Conference on Machine Learning* (2010).
- Griffiths, T., Steyvers, M., Blei, D., Tenenbaum, J. Integrating topics and syntax. *Advances in Neural Information Processing Systems* 17. L. K. Saul, Y. Weiss, and L. Bottou, eds. MIT Press, Cambridge, MA, 2005, 537–544.
- Grimmer, J. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Polit. Anal.* 18, 1 (2010), 1.
- Hoffman, M., Blei, D., Bach, F. On-line learning for latent Dirichlet allocation. In *Neural Information Processing Systems* (2010).
- Hofmann, T. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence (UAI)* (1999).
- Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L. Introduction to variational methods for graphical models. *Mach. Learn.* 37 (1999), 183–233.
- Li, J., Wang, C., Lim, Y., Blei, D., Fei-Fei, L., Building and using a semantivisual image hierarchy. In *Computer Vision and Pattern Recognition* (2010).
- Li, W., McCallum, A. Pachinko allocation: DAG-structured mixture models of topic correlations. In *International Conference on Machine Learning* (2006), 577–584.
- Mimno, D., McCallum, A. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence* (2008).
- Newman, D., Chemudugunta, C., Smyth, P. Statistical entity-topic models. In *Knowledge Discovery and Data Mining* (2006).
- Pritchard, J., Stephens, M., Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* 155 (June 2000), 945–959.
- Reisinger, J., Waters, A., Silverthorn, B., Mooney, R. Spherical topic models. In *International Conference on Machine Learning* (2010).
- Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smith, P. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (2004), AUAI Press, 487–494.
- Rubin, D. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.* 12, 4 (1984), 1151–1172.
- Sivic, J., Russell, B., Zisserman, A., Freeman, W., Efros, A. Unsupervised discovery of visual object class hierarchies. In *Conference on Computer Vision and Pattern Recognition* (2008).
- Socher, R., Gershman, S., Perotte, A., Sederberg, P., Blei, D., Norman, K. A Bayesian analysis of dynamics in free recall. In *Advances in Neural Information Processing Systems* 22. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds. 2009.
- Steyvers, M., Griffiths, T. Probabilistic topic models. *Latent Semantic Analysis: A Road to Meaning*. T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, Eds. Lawrence Erlbaum, 2006.
- Teh, Y., Jordan, M., Beal, M., Blei, D. Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.* 101, 476 (2006), 1566–1581.
- Wainwright, M., Jordan, M. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* 1(1–2) (2008), 1–305.
- Wallach, H. Topic modeling: Beyond bag of words. In *Proceedings of the 23rd International Conference on Machine Learning* (2006).
- Wang, C., Blei, D. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. *Advances in Neural Information Processing Systems* 22. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds. 2009, 1982–1989.
- Wang, C., Thiesson, B., Meek, C., Blei, D. Markov topic models. In *Artificial Intelligence and Statistics* (2009).

**David M. Blei** (blei@cs.princeton.edu) is an associate professor in the computer science department of Princeton University, Princeton, N.J.

© 2012 ACM 0001-0782/12/04 \$10.00

# A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases

Justin Grimmer

*Department of Government, Harvard University, 1737 Cambridge Street,  
Cambridge, MA 02138*  
*e-mail: jgrimmer@fas.harvard.edu (corresponding author)*

Political scientists lack methods to efficiently measure the priorities political actors emphasize in statements. To address this limitation, I introduce a statistical model that attends to the structure of political rhetoric when measuring expressed priorities: statements are naturally organized by author. The expressed agenda model exploits this structure to simultaneously estimate the topics in the texts, as well as the attention political actors allocate to the estimated topics. I apply the method to a collection of over 24,000 press releases from senators from 2007, which I demonstrate is an ideal medium to measure how senators explain their work in Washington to constituents. A set of examples validates the estimated priorities and demonstrates their usefulness for testing theories of how members of Congress communicate with constituents. The statistical model and its extensions will be made available in a forthcoming free software package for the R computing language.

## 1 Introduction

I introduce a statistical model to measure the priorities political actors articulate in texts, which I apply to measure how legislators explain their work to constituents. The *expressed agenda model* incorporates information about the authors of texts and other covariates to create a method explicitly designed to measure how legislators articulate priorities to constituents. A Bayesian approach, coupled with the use of a deterministic method for estimating complex posteriors, makes estimation and inference straightforward, as well as inference about quantities of interest derived from the priorities. I apply the model to an original collection of over 24,000 Senate press releases, collected from each Senate office in 2007, demonstrating that the press release data and statistical model facilitate comprehensive tests of theories about how legislators communicate with their constituents.

Members of Congress invest substantial resources to communicate with constituents, issuing thousands of statements, press releases, and speeches during each legislative term.

---

*Author's note:* I thank the Center for American Political Studies and the Institute for Quantitative Social Science for financial support. I have benefited from conversations with Ken Benoit, Matt Blackwell, Daniel Carpenter, Jacqueline Chattopadhyay, Andrew Coe, Brian Feinstein, Rob Franzese, Claudine Gay, Jeff Gill, David Hadley, Frank Howland, Emily Hickey, D. Sunshine Hillygus, Daniel Hopkins, Michael Kellerman, Gary King, Burt Monroe, Clayton Nall, Stephen Purpura, Kevin Quinn, Brandon Stewart, seminar participants at Harvard University, participants at the 2008 Summer Political Methodology meeting, and 2009 Southern Political Science Association meeting.

© The Author 2009. Published by Oxford University Press on behalf of the Society for Political Methodology.  
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

In spite of the recognized importance of this communication to understanding political representation and legislative behavior (Mayhew 1974; Fenno 1978), political scientists know surprisingly little about the content of these statements and how legislators translate their activities in Washington into statements to constituents. This is due, in large part, to the difficulty in collecting and analyzing the multitude of statements from members of Congress. Most studies of congressional communication employ methods that are too expensive and time consuming to apply to each member of Congress or even large samples of members (Fenno 1978; Lipinski 2004). As a result, much of our knowledge about how legislators explain their work to constituents is derived from observations made about a few members of Congress and a small subset of statements made to constituents.

As an alternative to manual coding, political scientists have recently turned to unsupervised learning methods to analyze attention in large text corpora: methods that simultaneously estimate the categories in a collection of texts and sorts documents into the estimated categories (Quinn et al. forthcoming).<sup>1</sup>

These methods, however, cannot be directly applied to measure the priorities articulated by representatives. When analyzing the attention senators (or other political actors) dedicate to issues there is a hierarchical structure: political statements, at the bottom of the hierarchy, are organized according to their author, at the top of the hierarchy. Previously developed unsupervised learning methods ignore this hierarchy and instead focus upon assigning documents to topics (Banerjee et al. 2005) or introduce structure designed to answer a different (but still important) question about how attention varies over time in the whole legislature (Quinn et al. forthcoming).

I accommodate the hierarchical structure when measuring author attention with the expressed agenda model. The method simultaneously discovers the topics in the data, assigns documents to their likely topic, and measures the attention a set of authors dedicate to the estimated topics. Like other unsupervised learning methods, the expressed agenda model does not require any pre-read documents, estimating the topics in the press releases. The use of Bayesian inference and a recently developed approach to estimation of complex posteriors, variational inference, makes fully Bayesian inference straightforward, whereas previous statistical models for political text assume that all estimates are known with certainty (Quinn et al. forthcoming). Using the new data set and statistical model, I demonstrate that the expressed agenda model is able to identify substantively important topics. A series of validations also demonstrates that the expressed agenda model provides substantively interesting estimates of senators expressed priorities and facilitates tests of important hypotheses across all members of a legislature—a previously infeasible task.

## 2 Expressed Agendas in Legislatures and Politics

By measuring how legislators explain their work in Washington to constituents, the model and data set provide powerful tools for understanding how political representation operates in America.

Legislators employ the resources of their office to portray how they are *responding* to the priorities and concerns of their constituents (and to distract attention from areas where

---

<sup>1</sup>There is a burgeoning literature that analyzes communication with *supervised* methods (Hopkins and King forthcoming; Hillard, Purpura, and Wilkerson 2008). Supervised methods provide hand-coded documents and pre-defined categories to a method in order to teach—supervise—the method how to place documents into categories. This approach is not employed here because there is still substantial uncertainty about the topics legislators could raise in conversation with constituents.

they appear less responsive) (Mayhew 1974; Fenn 1978; Kingdon 1989; Arnold 1992). Understanding the contents of these portrayals are critical to understanding *home style* and are inherently important to explaining how legislators maintain their connection with constituents (Fenn 1978; Kingdon 1989; Arnold 1992; Sulkin 2005), assessing deliberative standards of democracy (Gutmann and Thompson 1996; Mansbridge 2003), identifying the causes of the incumbency advantage (Gelman and King 1990), determining how legislators claim credit for resources secured for their state (Mayhew 1974), and understanding how legislators interact with the media (Arnold 2004). The importance of understanding legislators' statements to constituents is most clearly articulated by Fenn when he remarks "empirical theories of representation will always be incomplete without theories that explain explaining" (Fenn 1978, 162).

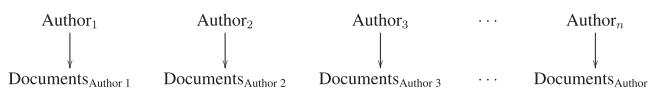
In this paper, I address a critical component of home style (Fenn 1978): *the issues legislators' emphasize in communication with constituents*. This quantity is of theoretical interest for studies of Congressional communication, ranging from qualitative studies of explanation (Mayhew 1974; Fenn 1978) to quantitative studies of senators' communicated issue priorities (Schiller 2000; Sulkin 2005). Although the expressed priorities of legislators (and other political actors) appear in numerous studies, it lacks a common name across applications. Therefore, I call the attention a senator allocates to issues in public statements her *expressed agenda*. It is an *agenda* because it measures the priorities of each senator, as articulated in press releases. Importantly, it is an *expressed agenda* because the attention dedicated to issues in communication are the issues that senators express as their priorities, not necessarily those issues that receive the most attention from senators while in Washington or from their staff. (But as I demonstrate, a senator's expressed agenda is closely tied to other indicators of legislative activity.)

### 2.1 Substantive Structure and Unsupervised Learning Methods

While intended to measure the priorities legislators and other political actors express in public statements, the expressed agenda model is also a new model for document *clustering*. A large literature in statistics and computer science advocate using clustering methods as an effective method for grouping together documents of similar content (Ng, Jordan, and Weiss 2002; Blei et al. 2003; Quinn et al. forthcoming; Manning et al. 2008). Although each method has performed well on a subset of problems, it is well known that the optimal application of any clustering method requires that the method be tuned to a particular substantive problem. The need for problem-specific methods arises because classification is only possible by making assumptions (a result known as the "ugly duckling theorem" [Watanabe 1969]) and because all clustering methods have the same average performance across all possible problems (a no-free lunch theorem [Wolpert and Macready 1997]). Therefore, arguments about whether to adopt a statistical or algorithmic approach to clustering are misplaced: the debate should focus on whether a specific class of clustering methods are well tuned to discover substantively interesting clusters from a particular collection of documents.

One approach to developing problem-specific clustering methods is to construct a hierarchical model, where the hierarchy includes additional information about each of the documents (Blei et al. 2003; Blei and Lafferty 2006; Teh et al. 2006; Mimno and McCallum 2008).<sup>2</sup> A particularly novel example of this approach is advanced in Quinn et al.

<sup>2</sup>Note, that this is distinct from hierarchical clustering, which creates a dendrogram describing a series of partitions in the data that obey some organizing rule (Hastie, Tibshirani, and Friedman 2001).



**Fig. 1** Hierarchical structure in political texts exploited in expressed agenda model: Documents organized by J.G.

(forthcoming), which analyzes Senate floor-speeches and includes information about the day a speech was made on the Senate floor. This method is therefore tuned to measure how attention varies over time *in the entire legislature*. But this hierarchical structure is ill-suited for the quantity of interest in this paper: *how attention to issues varies across legislators*, because it ignores the authors of particular documents.

The expressed agenda model explicitly includes information about the author of the documents and therefore is designed to address the priorities legislators articulate to constituents. Figure 1 shows the general structure that underlies the model: many statements from each author, with several authors in the collection. While applied to study legislative home style in press releases here, this same structure is employed anytime the *quantities of interest are the priorities a set of actors allocate to issues* and therefore the expressed agenda model has wide range of applications in political science. This includes examinations of issue ownership in political campaigns, where the interest is in comparing the issues Democrat and Republican candidates emphasize during campaigns (Petrocik 1996; Simon 2002; Sigelman and Buell 2004). Likewise, scholars of the news media ask which stories are afforded attention in different newspapers (Armstrong et al. 2006). Scholars of the presidency are often interested in the priorities presidents communicate in their public statements (Lee 2008), and deliberative democrats are interested in exploring the explanations are offered for new policies (Gutmann and Thompson 1996). Table 1 highlights these and other potential applications of the expressed agenda model.

### 3 Press Releases and Measuring Expressed Agendas

Press releases are an ideal medium for measuring how legislators present themselves to constituents. There are essentially no *formal* constraints imposed upon a press release's content, and they comprise a critical component of how senators explain activities in Washington to constituents (Yiannakis 1982). In contrast, use of media coverage to measure politicians' issue agenda conflates the issues politicians discuss and the media outlet's choice to cover an issue (Sulkin 2005). Surveys of Senate staffers require the strong (and

**Table 1** Potential applications of the expressed agenda model

<i>Question</i>	<i>Example study</i>
How do campaigns affect attention in congress?	Sulkin (2005)
What do senators discuss in floor statements?	Hill and Hurley (2002)
Do competing candidates emphasize the same issues?	Petrocik (1996)
What do presidents address in daily speeches?	Lee (2008)
What reasons are offered to justify policy?	Gutmann and Thompson (1996)
Do competing political elites discuss the same issues?	Gabel and Scheve (2007)
What issues receive attention from newspapers?	McCombs (2004)

often violated) assumption that the staffer is able to offer an unbiased recollection of a politician's stated priorities (Cook 1988).

Press releases are also ideal because they are regularly used by each Senate office. In 2007, the average Senate office released four and a half press releases per week, and the Senate, as a whole, issued an average of 66 press releases per day. The frequency of press releases stands in contrast to newsletters legislators send to constituents, which are only sent occasionally during a legislative session and have extremely limited space (Lipinski 2004).

### 3.1 Press Releases and Newspaper Coverage

Press releases are also important because their content reaches citizens through local newspapers. Newspapers—particularly local papers—often have only a small budget dedicated toward covering what representatives do while in Congress (Vinson 2002). To fill this gap, newspaper editors rely upon wire service stories and press releases from Congressional offices (Cook 1989; Arnold 2004; Schaffner 2006).

### 3.2 “Ventriloquism” and Press Releases

Press secretaries know that they will have high levels of success in generating news coverage with press releases (Cook 1989; Schaffner 2006). In fact, some press releases are run *almost verbatim* in papers: Table 2 collects three press releases (left-hand column) that were subsequently repeated in newspapers (right-hand column) (much like a ventriloquist's dummy). The italic text in Table 2 identifies the duplicated content. Printing of press releases with little modification appears to be common in small-town newspapers. For example, Richard Lugar (R-IN) issued a press release on July 17, 2007, describing why Reynolds, IN was selected to receive federal funding for alternative fuels research. His explanation was repeated, *almost exactly*, on July 18, 2007, in The Times—a local newspaper in heavily Democratic northwest Indiana. The repetition of press releases also occurs in major metropolitan newspapers. On May 30, 2008, Senator Dick Durbin's (D-IL) office issued a press release about funding secured for hybrid buses in Chicago. The Chicago Tribune, which has the fifth largest circulation among American newspapers, used Durbin's release with only slight modification on May 31, 2008 (second row of Table 2). The third example shows that a joint press release from Susan Collins (R-ME) and Olympia Snowe (R-ME) announcing funds secured for laid-off factory workers was reprinted—essentially unchanged—in the Bangor Daily News.

### 3.3 Measuring the Coverage Rate of Senate Press Releases in Newspapers

Press releases can influence how a newspaper covers a member of Congress without being plagiarized directly by providing a source for statements on a representative's position or drawing attention toward funds secured for a district. To more systematically analyze how often press releases translate into *coverage* in local newspapers, I measured the percentage of press releases from 10 Senate offices (identified in column 2, Table 2) that were quoted, paraphrased, or plagiarized in six local newspapers (identified in column 1). The total number of press releases from a senate office that were covered in a given newspaper is contained in column 3, whereas column 4 identifies the percentage of a senator's total press releases that a newspaper covered.

To determine if a press release from a Senate office was used in a newspaper, I first collected all newspaper stories from 2007 and January 2008 that contained the relevant senator's name.<sup>3</sup> I then used publicly available *cheating detection* software to uncover press releases and newspaper stories that had extremely similar content (Bloomfield 2008). The software provides an efficient method for searching over the 1,069,430 potential newspaper-press release pairs that must be checked to generate Table 3. I then took the pairs of newspaper stories and press releases that the software identified with similar content and manually validated that the newspaper article contained a quote from the press release.

Overall, Table 3 indicates that press releases are regularly used by local newspapers. For example, both Orrin Hatch (R-UT) and Bob Bennett (R-UT) had over a quarter of their press releases covered in the Deseret Morning News. This should not be surprising: the Deseret Morning News has a history of financial problems that constrain its ability to cover politics, and the editor-in-chief, Joseph Cannon, is the former chair of the Utah Republican party. Other newspapers use press releases at a similar rate: Byron Dorgan (D-ND) had 54 of his press releases used in the Bismarck Tribune, and 74 press releases from Susan Collins (R-ME) were used in the Bangor Daily News. Even the San Francisco Chronicle used press releases from both Diane Feinstein (D-CA) and Barbara Boxer (D-CA). The high percentage of press releases used from each Senate office is particularly striking given that Senate offices write press releases for an entire state. Therefore, a sizable portion of press releases from a Senate office is irrelevant to a local paper.

This analysis, although limited to a subset of senators and newspapers, shows that press releases are a common source of information for newspapers. Press releases are regularly quoted in local newspapers—such as the Bismarck Tribune—and are used in major metropolitan papers—like the San Francisco Chronicle. This confirms that press releases are an important medium that legislators use to communicate with constituents, as the messages in press releases are likely to reach constituents.

#### 4 Preparing the Texts for Analysis

Using the thousands of press releases from all Senate offices, the expressed agenda model measures the priorities senators communicate to their constituents through press releases. To perform this analysis, a set of preprocessing steps are performed on the press releases, all of which are well established in the literature on the statistical analysis of text (Manning et al. 2008). The first step discards the order of words in the press release, leaving an unordered set of words remaining (Hopkins and King forthcoming; Quinn et al. forthcoming). Although one might expect the order of words to be crucial to understanding the sentiment expressed in a text, identifying the topic of a press release should be invariant to permutations of word order. Certain topics, such as the Iraq war, should result in specific words appearing with

<sup>3</sup>Stories were collected from the Lexis-Nexis database of newspaper stories. The newspapers selected are not a random sample of papers. I intentionally selected newspapers that I conjectured would display a great deal of variation in their use of press releases. The Deseret Morning News (Salt Lake City, Utah), The Bismarck Tribune (Bismarck, North Dakota), and the Bangor Daily News (Bangor, Maine) are all local newspapers that were likely to be highly reliant on the information senators provided. The Salt Lake Tribune (Salt Lake City, Utah), The Pioneer Press (St Paul, Minnesota), and the San Francisco Chronicle (San Francisco, California) are all large newspapers with greater capacity for covering political news. I also selected a mix of Republican, Democrat, and split delegations; senators from big and small states; and senators who issued a great deal of press releases and senators who issue only a few. Given the nature of the sample selection, inference to a broader population is inappropriate from these data, but is the subject of future research.

**Table 2** “Ventriloquism”: press releases in local newspapers

<i>Richard Lugar (R-IN), 7/17/2007</i> The Town of <i>Reynolds</i> was selected in 2005 to demonstrate to the nation and the world that a community's energy needs can be fully met through locally produced renewable sources, including electricity, natural gas replacement, and vehicular fuel (Lugar 2007).	<i>The Times (IN), 7/18/07</i> <i>Reynolds</i> , located about 20 miles north of Lafayette, was chosen in 2005 to demonstrate that a community's energy needs can be fully met through locally produced renewable sources, including electricity, natural gas replacement, and vehicular fuel (AP 2007).
<i>Dick Durbin (D-IL), 5/30/2008</i> U.S. Senator Dick Durbin (D-IL) announced today that the U.S. Department of Transportation (DOT) has awarded a \$9.6 million grant to the city of Chicago that will allow the Chicago Transit Authority to purchase approximately 13 additional articulated diesel hybrid buses. Hybrid buses are quieter, cleaner, burn less gas, and run more smoothly than conventional diesel. (Durbin 2008).	<i>Chicago Tribune (IL), 5/31/2008</i> U.S. Senator Dick Durbin says the city of Chicago will receive \$9.6 million from the federal government to buy hybrid buses. Durbin said Friday that the grant from the U.S. Department of Transportation will allow the Chicago Transportation Authority to buy about 13 more articulated diesel hybrid buses. In March, the CTA announced plans to lease 150 hybrid buses at the cost of \$13.4 million a year. Hybrid buses burn less gas than conventional diesel buses (AP 2008).
<i>Susan Collins (R-ME), 11/1/2007</i> U.S. Senators Olympia J. Snowe and Susan Collins today announced that the U.S. Department of Labor has approved their request for \$894,918 in National Emergency Grant funding for Domtar and Fraser Mill workers. Last month Senators Snowe and Collins sent Secretary Chao a letter urging the Department of Labor to quickly review and approve the NEG funding request for the 300 workers who lost their jobs at Domtar Industries in Baileyville and Fraser Papers of Madawaska. “This is great news for 300 workers in Northern and Eastern Maine who lost their jobs through no fault of their own” said Senators Snowe and Collins (Collins 2007).	<i>Bangor Daily News (ME), 11/2/2007</i> U.S. Senators Olympia J. Snowe and Susan Collins Thursday announced that the U.S. Department of Labor has approved their request for \$894,918 in National Emergency Grant funding for Domtar and Fraser Mill workers. Last month, the senators sent Secretary Elaine Chao a letter urging the Department of Labor to quickly review and approve the NEG funding request for the 300 workers who lost their jobs at Domtar Industries in Baileyville and Fraser Papers of Madawaska. “This is great news for 300 workers in Northern and Eastern Maine who lost their jobs through no fault of their own,” said the senators (Staff 2007).

high frequency (troop, war, iraqi) irrespective of whether the senator supports or opposes the war.

Next, all the words are placed into lower case and all punctuations are removed. Then, I applied the Porter stemming algorithm to each word (Porter 1980). The stemming algorithm takes as an input a word and returns the word's basic building block, or *stem*. For example, the stemming algorithm takes the words family, families and returns famili.

After stemming the words in each document, I counted the number of occurrences of each word in the *corpus*, the total set of press releases. All words that do not occur in at least 0.5% of press releases were removed (Quinn et al. forthcoming). Finally, I removed all *stop* words (e.g., around, whereas, why, whether), along with any word that appears in over 90%

**Table 3** Measuring the coverage rate in Senate press releases

Newspaper	Senator	Number quoted	Percent of press releases
Deseret Morning	Bennett (R-UT)	35	32.4
Deseret Morning	Hatch (R-UT)	67	27.2
Salt Lake Tribune	Bennett (R-UT)	21	19.4
Bangor Daily	Collins (R-ME)	74	18.2
Salt Lake Tribune	Hatch (R-UT)	43	17.4
Bismarck Tribune	Dorgan (D-ND)	54	16.8
Bismarck Tribune	Conrad (D-ND)	33	16.3
Pioneer Press	Klobuchar (D-MN)	29	13.1
Pioneer Press	Coleman (R-MN)	32	12.2
Bangor Daily	Snowe (R-ME)	44	11.9
San Francisco Chronicle	Boxer (D-CA)	11	7.2
San Francisco Chronicle	Feinstein (D-CA)	24	6.3

*Note.* This table presents the coverage rate of press releases in local newspapers and shows that constituents are likely to read the contents of their representative's press releases in local newspapers. The first column contains the name of the newspaper and the second column identifies which senator's press releases were used. The third column presents the number of press releases that had content appear in a story in the local newspaper. To compute this number, I used freely available cheating detection software to uncover sentences that were the same or highly similar (Bloomfield 2008). The fourth column presents the percentage of press releases from a Senate office that was covered in the newspaper.

of any individual senator's press releases. This ensures that each senator's press releases are not grouped together based upon language unique to each senator, yet unrelated to the topic of the document.

After preprocessing the press releases, 1988 unique stems remain, along with 3,715,293 stem observations in the 24,236 press releases. Each document is represented as a  $w \times 1$  vector, where  $w$  are the number of stems that remain after the preprocessing (in this example,  $w = 1988$ ).

## 5 A Statistical Model for Expressed Agendas

When measuring the attention political actors allocate toward topics in texts, the data are naturally organized hierarchically, with press releases grouped according to the Senate office that authored the statement. At the top of the hierarchy we suppose that there are a set of senators, indexed by  $i = 1, \dots, n = 100$ . Each senator decides how much attention to dedicate to each topic  $k$  ( $k = 1, \dots, K$ ) present in her press releases. The vector describing the attention a senator dedicates to each topic is her *expressed agenda* and probabilistically determines how often each of the  $K$  topics appear in her press releases.

At the bottom of the hierarchy are each senator's press releases. Represent press release  $j$  ( $j = 1, \dots, D_i$ ) from senator  $i$  with the  $w \times 1$  vector  $y_{ij}$ . Typical element of  $y_{ij}$ ,  $y_{ijz}$ , measures the number of times the  $z$ th stem occurs in the  $j$ th document from the  $i$ th senator. To connect the senator's priorities with the content of her press releases, suppose that each press release has only one topic. Although a common assumption in statistical topic models, this assumption is particularly appropriate for Senate press releases. Press releases are written in a style similar to short news stories, designed to draw attention to one particular aspect of a senator's activities in Washington. Thus, most press releases address one particular topic. The topic of each press release is a random draw, with the probability of a specific topic occurring determined by the attention senator  $i$  dedicates to the issue.

Conditional upon this sampled topic, a press release's content is drawn from a distribution that is specific to each topic. Formally, the expressed agenda model is a hierarchical mixture model where the mixture weights (senators' expressed agendas) are allowed to vary across senators, but the components of the mixture (topics) are fixed across authors to ensure that the priorities of senators are comparable (see Section 6 below). To complete our preliminary notation, suppose that there are a total of  $D = \sum_{i=1}^{100} D_i$  press releases and collect all the press releases into the  $D \times w$  matrix  $\mathbf{Y}$ .

### 5.1 Senator-Level Parameters: Senators' Expressed Agendas

The expressed agenda for each senator determines the probability that topics appear in documents. Call the attention senator  $i$  allocates to issue  $k$ ,  $\pi_{ik}$ . Equivalently,  $\pi_{ik}$  represents the expected probability that a press release is generated by the  $k$ th topic. Each senator's expressed agenda,  $\boldsymbol{\pi}_i$ , is then defined as the  $K \times 1$  vector describing the attention she dedicates to each topic,  $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{iK})$ . In order for  $\boldsymbol{\pi}_i$  to be interpreted as the probability of each topic appearing in a press release, its elements must sum to one,  $\sum_{k=1}^K \pi_{ik} = 1$ , and every entry must be greater than zero,  $\pi_{ik} \geq 0$  for each  $k = 1, \dots, K$ . Substantively, this assumption implies that senators are resource constrained when allocating attention to issues and cannot distract from an issue any more than not issuing a press release on the issue.

### 5.2 Document-Level Parameters: Topics and Words

Conditional on a senator's expressed agenda,  $\boldsymbol{\pi}_i$ , we draw the topic of each press release. Represent press release  $y_{ij}$ 's topic with the  $K \times 1$  indicator vector  $\boldsymbol{\tau}_{ij}$ : if press release  $y_{ij}$  was generated by the  $k$ th topic, then  $\tau_{ijk} = 1$  and the other  $K - 1$  elements of  $\boldsymbol{\tau}_{ij}$  are equal to 0.<sup>4</sup>

The topic of each press release  $\boldsymbol{\tau}_{ij}$  is a draw from a multinomial distribution,

$$\boldsymbol{\tau}_{ij} | \boldsymbol{\pi}_i \sim \text{Multinomial}(1, \boldsymbol{\pi}_i). \quad (5.1)$$

Equation (5.1) connects the topics of press releases to a senator's expressed agenda. The expected proportion of senator  $i$ 's press releases allocated to the  $k$ th topic is  $\pi_{ik}$ .

Conditional on the sampled topic,  $\boldsymbol{\tau}_{ij}$ , we draw the content (words) of each press release. One possibility would be to model the contents of each press release  $y_{ij}$  directly as a draw from a normal distribution (Fraley and Raftery 2002). But using normal distributions to cluster documents will tend to group press releases together based upon the number of words used in the document or the length of  $y_{ij}$  (Banerjee et al. 2005). If the length of a document does not contain information about the topic of a document, then using the normal distribution is inappropriate.

To eliminate the influence of word count when clustering press releases, I normalize each press release to have unit length. The unit length representation of  $y_{ij}$  is given by  $\mathbf{y}_{ij}^*$ , with  $\mathbf{y}_{ij}^* = \frac{\mathbf{y}_{ij}}{\|\mathbf{y}_{ij}\|}$  where  $\|\cdot\|$  is defined as the Euclidean norm,  $\|\mathbf{y}_{ij}\| = (\mathbf{y}_{ij}' \mathbf{y}_{ij})^{1/2}$ .  $\mathbf{y}_{ij}^*$ , now measures the relative rate words, are used in each press release rather than the total number of times each stem is used in a document.

After normalizing each press release, we suppose that  $\mathbf{y}_{ij}^*$  is a draw from a distribution that is defined only the set of unit-length vectors (or a unit hypersphere): the von Mises-Fisher

---

<sup>4</sup>Collect the indicator vectors for all of senator  $i$ 's press releases into the  $D_i \times K$  matrix  $\boldsymbol{\tau}_i$ .

(vMF) distribution (Banerjee et al. 2005). The vMF distribution is characterized by a  $w \times 1$  vector that governs the distribution's center,  $\boldsymbol{\mu}$ , and a scalar that determines the distribution's dispersion,  $\kappa$ .  $\boldsymbol{\mu}$  points to the location on the unit hypersphere, where the vMF distribution reaches its mode.  $\kappa$  is an inverse dispersion parameter: as  $\kappa \rightarrow 0$  the vMF density approaches the uniform distribution on a sphere, as  $\kappa \rightarrow \infty$  the vMF converges upon a spike at the center,  $\boldsymbol{\mu}$ .

Suppose that there are  $K$  vMF distributions and represent the center and dispersion parameter for the  $k$ th vMF distribution as  $\boldsymbol{\mu}_k, \kappa_k$ .  $\boldsymbol{\mu}_k$  can be thought of as the prototype document for the  $k$ th topic. A press release's topic,  $\tau_{ij}$ , determines the vMF distribution used to generate the content of a press release. Formally, if  $\tau_{ijk} = 1$ , then

$$\mathbf{y}_{ij}^* | (\tau_{ijk} = 1), \boldsymbol{\mu}_k, \kappa \sim \text{vonMises-Fisher}_w(\boldsymbol{\mu}_k, \kappa). \quad (5.2)$$

The vMF distribution has sampling density  $f(\mathbf{y}_{ij}^* | \boldsymbol{\mu}_k, \kappa_k) = c(\kappa_k)_w \exp(\kappa_k \boldsymbol{\mu}_k' \mathbf{y}_{ij}^*)$ ,  $c(\kappa_k)_w$  is a normalizing constant given by  $c_w(\kappa_k) = \frac{\kappa_k^{w/2-1}}{(2\pi)^{w/2} I_{w/2-1}(\kappa_k)}$  and  $I_{w/2-1}$  is a modified Bessel function of the first kind.<sup>5</sup> It will be convenient to collect the center of each topic's vMF distribution into the  $w \times K$  matrix,  $\boldsymbol{\mu}$ , and the inverse dispersion parameter for each topic into the  $K \times 1$  vector  $\kappa$ .

### 5.3 Priors for the Expressed Agenda Model

I place a prior on each senator's expressed agenda,  $\boldsymbol{\pi}_i$ , to partially pool information across senators to allow for more efficient inferences (Gelman and Hill 2007). Suppose that each  $\boldsymbol{\pi}_i$  is a draw from a Dirichlet distribution,

$$\boldsymbol{\pi}_i | \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \quad (5.3)$$

where  $\boldsymbol{\alpha}$  is a  $K \times 1$  vector of shape parameters which govern the Dirichlet distribution.<sup>6</sup> Rather than assume specific values of  $\boldsymbol{\alpha}$ , we estimate the parameters to determine the amount of pooling from the data. We suppose that each  $\alpha_k$  is a draw from a Gamma distribution and assume parametric values of the Gamma distribution to limit the amount of pooling across senators.<sup>7</sup>

#### 5.3.1 Including covariates in the prior

In Appendix B.6, I modify the prior on senators' priorities to allow for the inclusion of covariates, using a Dirichlet-multinomial regression, a modified version of the prior

<sup>5</sup>Alternatively, the model could be developed using a multinomial distribution for the words in documents, and this option is available in the statistical package.

<sup>6</sup>Any distribution on the simplex will suffice for the general setup of the model. The Dirichlet distribution was selected because it makes inference straightforward and it has limited influence on the results. The Dirichlet distribution assumes that there is a negative covariance between the attention dedicated to each topic (Aitchison 1986). This assumption is dangerous only if certain components of the composition have a large, positive covariance. As an alternative, a logistic normal distribution could be used to pool the expressed agenda across senators. A version of the model with a logistic normal distribution is available in the software package. Tests with the logistic normal prior indicate that the more general model does not yield different estimates of expressed priorities than the model with the Dirichlet prior.

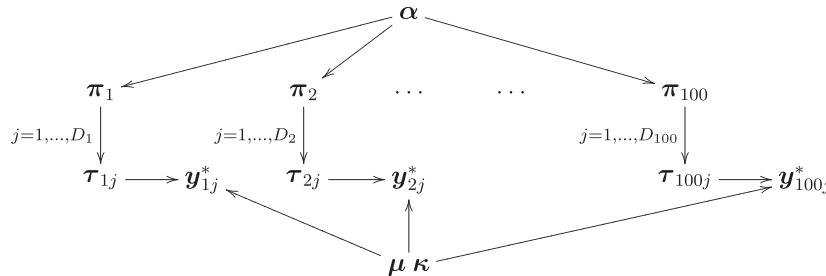
<sup>7</sup>Suppose that the Gamma distribution has sampling density,  $g(\alpha_k | \lambda, \delta) = \alpha_k^{\delta-1} \exp(-\frac{\alpha_k}{\lambda}) / \lambda^\delta \Gamma(\delta)$ . We set  $\lambda = 1$  and  $\delta = 1$ , which ensures that the value of each  $\alpha_k$  remains relatively small, limiting the pooling.

introduced in Mimno and McCallum (2008). These covariates allow for the inclusion of additional information that allows for smoothing across groups of senators who share similar characteristics—such as a senator’s political party or a dummy variable for a specific senator who served in different years. For expository purposes, this paper proceeds with a model that does not include additional covariate information, but this model is available in the software package.<sup>8</sup>

I fix all  $K \kappa$ ’s to a single value in the results below,  $\kappa$  is set to 100 (Zhong and Ghosh 2003).<sup>9</sup> Conditional on  $\kappa$  we assume a conjugate prior for the center of vMF distributions  $\mu_k | \kappa \sim \text{vMF}_w(\eta, \kappa)$ . The center of the prior vMF distribution,  $\eta$ , has typical element  $\frac{1}{\sqrt{w}}$ .<sup>10</sup>

#### 5.4 Posterior Distribution for the Expressed Agenda Model

Figure 2 provides a graphical display of the complete expressed agenda model: the directed acyclic graph consistent with the model and priors that comprise the expressed agenda model. The arrows in the graph depict the parameters that each random variable’s density is dependent upon. For example, the directed edge  $\alpha \rightarrow \pi^i$  denotes that the sampling density of senator  $i$ ’s expressed agenda,  $\pi^i$ , depends upon  $\alpha$ . Notice the hierarchical structure



**Fig. 2** Bayesian graph of expressed agenda model. This figure presents the expressed agenda model. We assume that each senator’s expressed agenda  $\pi_i$  is a draw from a Senate-wide Dirichlet distribution  $\text{Dirichlet}(\alpha)$ . Conditional on  $\pi_i$ , each press release’s topic is a draw from a Multinomial(1,  $\pi_i$ ). Conditional upon this draw, we assume that each document is then drawn from a vMF distribution, with center  $\mu_j$  and inverse dispersion parameter  $\kappa_j$ . Note, that all senators select from the same set of topics to ensure that their priorities are comparable. Further, notice the hierarchical structure inherent in the model, with press releases organized according to their author.

<sup>8</sup>The exclusion of covariates in the model does not introduce omitted variable bias as we might expect when using regression to make causal inferences. Rather, the inclusion of covariates improves the information that is borrowed across senators during smoothing. We might expect additional covariates to improve the performance of the model, and therefore, presenting the model with no covariates represents a disadvantage against the expressed agenda model in the evaluations performed below.

<sup>9</sup>Fixing  $\kappa$  across clusters is similar to the approach in Zhong and Ghosh (2003) for estimating mixtures of vMF distributions. The model has been estimated with  $\kappa$  ranging from 50 to 500, and the substantive results remain unchanged.

<sup>10</sup>This is the least informative conjugate prior on  $\mu$  that treats all coordinates of the vMF distribution identically because the vMF distribution measures the relative rate at which words occur. To see this, suppose we choose an arbitrarily small quantity  $\epsilon > 0$  for each component. The length of  $\epsilon = (\epsilon, \dots, \epsilon)$  is  $\|\epsilon\| = \sqrt{\omega} \times \epsilon$ . Then the normalized vector  $\epsilon^* = (\frac{1}{\sqrt{\omega}}, \dots, \frac{1}{\sqrt{\omega}})$ .

present in the model: press releases organized by their author. Appendix B provides the full posterior distribution.

### 5.5 Inference for the Expressed Agenda Model

Due to the large number of components necessary to capture the variety of topics in press releases, computationally intensive approaches to inference—such as MCMC—are prohibitively slow. Sampling-based methods also face difficulty because permutations of the cluster labels result in the same height of the posterior density, greatly complicating simulation-based inference. One proposed solution is to constrain the parameter space, but this hinders the convergence of the Markov chain (McLachlan and Peel 2000). As an alternative, current best practice recommends running the chains without constraints and then post-processing, identifying the same clusters using a clustering algorithm on the output. This is a useful approach with a small number of mixture components, but MCMC methods have difficulty exploring the posterior as the number of mixture components increases.

Alternatively, one could employ the expectation maximization (EM) algorithm to generate maximum a posteriori (MAP) estimates (McLachlan and Krishnan 1997). This is a reasonable method for inference when MCMC is infeasible, but generating uncertainty estimates from the results of an EM algorithm can be computationally challenging in large mixture models, due to the large number of parameters (McLachlan and Peel 2000). As a result, estimates from an EM model are often analyzed under the assumption that there is no uncertainty in the estimates (Quinn et al. forthcoming). This is an unattractive assumption for two reasons. First, some uncertainty is always present when measuring senator's expressed agendas. Second, the amount of information present about a senator's priorities varies considerably by Senate office. As a result, uncertainty about a senator's expressed agenda should vary by senator as well.

To avoid the difficulties associated with sampling methods and to estimate the entire posterior distribution on each senator's expressed agenda, I use a *variational approximation* to derive an analytical—rather than computational—approximation to the posterior distribution for each senator's expressed agenda (Jordan et al. 1999). Like EM algorithms, variational methods avoid the identification problem because optimization occurs according to a deterministic algorithm based upon starting values and the posterior distribution. Rather than generating MAP parameter estimates, variational methods analytically estimate the entire posterior distribution on each senator's expressed agenda. To perform this estimation, we first restrict the model to a simpler family of distributions. Then, we use the *calculus of variations* to select the member of this distributional family that is closest to the true posterior distribution, where proximity between the distributions is measured using the *Kullback-Leibler (KL) divergence* (Bishop 2006). In Appendix B, I derive the update equations used to estimate the posterior distributions.

### 5.6 Details of Estimation

The results presented in this paper use the variational algorithm derived in Appendix B and assume there are 43 topics present in the data. I varied the number of assumed topic from only five topics, up to 85 different topics. Assuming too few topics resulted in distinct issue being lumped together, whereas too many topics results in several clusters referring to the same issues. During my tests, 43 issues represented a decent middle ground. I corroborated

this number of clusters using a nonparametric model for text clustering, based upon the Dirichlet process prior. This model identified 40–45 clusters in the data set under a wide range of hyperpriors (Blei and Lafferty 2006). The variational algorithm described in Appendix 12 was randomly restarted 100 times, and the analysis was performed on the “best” run.<sup>11</sup>

## 6 Comparison to Ad-Hoc Approaches to Measuring Author Priorities

Existing methods for unsupervised learning and text clustering are designed to assign documents to topics or to measure the attention in an entire collection of documents—ignoring the information about authors. As a result, these methods are either unable to measure author-specific attention or would require ad-hoc modifications that fail to have the many benefits of the expressed agenda model.

### 6.1 Clustering Each Senator’s Press Releases Separately

To use existing clustering algorithms to measure senators’ expressed priorities, one could apply a clustering algorithm to each senator’s press releases separately and equate senator attention with the proportion of press releases assigned to each topic.

This method would fail, however, because the estimated topics would be different across senators and the set of topics must be fixed across senators to allow for priorities to be comparable. If a senator issues a press release about a topic (say the Iraq war) only occasionally, an unsupervised learning method will lump together press releases about a topic with other press releases about similar, though distinct, topics (defense spending and veteran affairs, perhaps). The clustering solution for a senator who allocates a great deal of attention to the issue, however, will identify the Iraq war as a distinct topic. As a result, a press release with identical content issued from two different senators could give the impression that the two senators are focusing upon *different* issues.

To demonstrate this problem, consider the press releases of two senators with similar explanatory styles: Robert Menendez (D-NJ) and Frank Lautenberg (D-NJ). I used a mixture of vMF distributions to separately cluster Lautenberg’s and Menendez’s press releases (Banerjee et al. 2005). To show that two press releases with identical content can be allocated to different topics, I used a *joint press* release—a press release from two senators with identical text—issued by Lautenberg and Menendez on July 31, 2007, that touted the senators’ efforts to improve reporting standards about toxic waste disposal (Lautenberg 2007; Menendez 2007). The clustering result from Lautenberg’s press releases placed the joint press release in a bureaucratic regulation cluster, with identifying stems push, require, law, bureau.<sup>12</sup> The clustering solution from Menendez assigned the *same document* to a cluster about economic growth (with stems economi, future, econom, studi, growth), because Menendez dedicates considerably less attention to bureaucratic regulation than Lautenberg. This shows that *the same press release can create the appearance that two senators are focusing on different issues* if applied to each senator’s press releases separately. The expressed agenda model avoids this problem by fixing the topics across senators.

<sup>11</sup>In Appendix B, I show that optimization occurs by increasing a lower bound on the marginal log-posterior of the data. We analyze the run that had the highest lower bound.

<sup>12</sup>The stems were identified using the mutual information between words and documents. See Section 7.2.

## 6.2 The Inadequacy of Ad-Hoc Modifications of Existing Methods

Ad-hoc modifications of existing clustering models could provide estimates of authors' priorities. For example, one could run an off-the-shelf clustering method on the entire collection of press releases from each Senate office, then tally the proportion of a senator's press releases that fall into each of the topics. This would create a measure of author-specific attention where the topics are fixed across senators.

This ad-hoc approach, however, is inadequate for several reasons. Most importantly, ad-hoc modifications are unable to provide uncertainty estimates about author-specific attention and subsequently, uncertainty about auxiliary quantities of interest. In contrast, the expressed agenda model estimates the posterior distribution on each author's priorities and is easily extended to posterior distributions on other quantities of interests derived from priorities.

An ad-hoc modification of existing methods also fails to exploit the additional information available to the analyst: the author of each press release. The expressed agenda model uses this additional information to aide in the discovery of topics, assign documents to topics, and measure the priorities authors express relative to the topics. A generative statistical model makes clear the assumptions of the statistical model and how the model could be extended to include across senator and over-time dependence. The statistical model also facilitates the borrowing of information across senators, allowing for efficient inference (Gelman and Hill 2007).

## 7 Labeling and Validating Topics

An advantage of the expressed agenda model is that the analyst does not need to prespecify the topics in the data. Rather the topics are estimated from the texts. In order to sensibly interpret the expressed agenda of each author, we must reliably label each of the topics and also validate that we are estimating reasonable topics from the data. I use three approaches to perform this evaluation: reading a subset of randomly chosen documents to provide a label, automatically generating distinctive stems to label clusters using the *mutual information* between stems and a topic, and exploiting over time variation in salience to check the reasonableness of cluster labels.

### 7.1 Labeling Clusters through Manual Document Checking

As a first step to assess the validity of the topics and to generate labels for topics, I randomly selected 10 documents from each topic with a high posterior probability of belonging to that topic (Quinn et al. forthcoming). I then read each of the 10 documents to generate the label found in the first column of Table 4.

On the whole, the clusters seemed to group together documents that referred to the same topic. For example, one group of texts discussed judicial nominations. Press releases in this category include releases from senate delegations to "Announce Recommendations for Eastern District Federal Judgeships" (Webb 2007), from members of the Judiciary committee who publicize that the "Senate Approves Kyl-Feinstein Provision Adding Judgeship to Ninth Circuit" (Kyl 2007), or declare that "The United States Senate unanimously confirmed Norman Randy Smith today to serve on the Ninth Circuit Court of Appeals" (Craig 2007). Another randomly chosen set of press releases dealt with energy policy. Among the press releases selected from this category is an announcement from a group of senators who "introduced legislation that will increase American drivers' access to ethanol at fuel

**Table 4** The topics estimated by the expressed agenda model

Description	Stems	Identifier
FEMA	disast,fema,storm,damag,declar,emerg,flood,recoveri,rebuild,reco	
Food safety	food,fda,agricultur,contamin,recal,inspect,product,nutrit,drug,consum	
Worker rights	worker,employe,wage,employ,labor,workplac,job,minimum,fair,workforc	
AG/justice	gener,justic,gonzal,judiciari,confirm,resign,investig,million,polit,nomine	
Agriculture	farmer,agricultur,crop,produ,rancher,usda,livestock,nutrit,conserv,food	
SCHIP	children,insur,uninsur,schip,kid,entrol,chip,reauthor,parent,incom	
Public land	land forest,manag,fish,wildif,public,recreat,area,natur,speci	
Pres. Veto/SOTU	presid,bush,veto,depart,iraq,announc,speech,union,democrat,facil	
Loan crisis	mortgag,loan,lender,borrow,homeown,lend,bank,crisi,rate,market	
Border security	border,homeland,immigr,patrol,secur,cross,agent,mexico,illeg,dh	
Illegal immigr.	immigr,border,illeg,reform,legal,debat,enfocr,broken,alien,citizenship	
Honorary	honor,provid,renemb,friend,program,celebr,depart,prayer,tribut,legaci	
Global warming	climat,warm,emiss,greenhous,global,carbon,chang,pollut,reduce,environment	
Science	scienc,math,competiti,compet,technolog,inno,engin,research,global,edg	
Higher edu.	colleg,higher,graduat,loan,univers,maximum,aid,school,grant,afford	
Iraq war	iraq,troop,iraqi,war,withdraw,polit,militari,strategi,petraeu,baghdad	
Veterans' affairs	veteran,affair,medic,mental,wound,war,desery,traumat,militari,afghanistan	
Tax policy	tax,relief,taxpay,deduct,incom,perm,revenu,credit,code,minimum	
Prescrip. drugs	drug,prescrip,fda,medicin,medicar,food,patient,market,medic,consum	
Energy policy	energi,fuel,oil,renew,sourc,gallon,ethanol,depend,biofuel,effici	
Nat/Coast Guard	guard,deploy,mission,militari,duti,soldier,defens,coast,command,iraq	
NCLB/School	school,teacher,district,classroom,academ,child,elementari,grade,children,teach	
Air Force	air,forc,aircraft,base,wing,mission,airlin,militari,plane,defens	
Women's issues	women,sexual,violenc,woman,assault,victim,domest,awar,prevent,abus	
Consumer sfty.	consum,product,recal,commiss,manufactur,store,danger,regul,ban,lead	
Judicial nom.	judg,nomin,confirm,nomine,circuit,judici,district,judiciari,appeal,legal	
Stem cells/research	research,diseas,cur,institut,univers,scientif,scienc,scientist,cell,stem	
Intl. trade	trade,china,agreement,market,export,manufactur,unfair,worker,product,intern	

*Continued*

**Table 4** (continued)

<i>Description</i>	<i>Items</i>	<i>Identifier</i>
Gov.Reg/Ethics Ref.	govern,rule,reform,transpar,democraci,account,program,report,elect,foreign	
Wounded soldiers	militari,defens,soldier,wound,armi,warrior,arm,veteran,men,walter	
Approp: Def. Proj.	million,appropi,defens,project,militari,fiscal,navi,research,air,armi	
Approp: Water Proj.	water,corp,wrd,engin,project,flood,drink,armi,navig,restor	
Approp: Econ. Dev.	econom,develop,grant,announc,growth,invest,job,economi,rural,award	
Approp: Home State	000,project,500,univers,appropi,hospit,youth,colleg,labor,human	
Approp: Firefight	firefight,homeland,grant,award,volunt,respond,afg,Equip,afgp,depart	
Approp: Airport	airport,aviat,faa,transport,dot,tourist,announc,travel,aircraft,air	
Approp: Public Works	project,appropi,fiscal,omnibus,approv,transport,hous,announc,signatur,develop	
Approp: DHS	secur,homeland,terrorist,dh,threat,attack,terror,1,1,risk,respond	
Approp: School Grants	program,school,youth,particip,grant,children,provide,success,reauthor,teach	
Approp: Health Care	patient,medicat,hospt,medic,qualiti,medicaid,access,doctor,insur,healthcar	
Approp: Crime	crime,enfore,justic,law,crimin,polic,violent,proseuct,gang,local	
Approp: HUD	hous,urban,hud,afford,homeless,incom,low,technolog,rehabilit,develop	
Approp: Transp.	transport,rail,transit,commut,congest,traffic,corridor,infrastructur,railroad,passeng	

pumps" (Harkin 2007) or Saxby Chambliss (R-GA) stating that he "addressed members of the Governor's Ethanol Coalition" (Chambliss 2007), a summary of an investigation into oil companies' attempt to "prohibit or strongly discourage the sale of alternative fuels" (Grassley 2007), and legislation introduced to "dramatically expand renewable fuel sources" (Bingaman 2007). These press releases all deal with energy—and in particular biofuel as an alternative fuel source.

### 7.2 An Automatic Cluster Labeling Method

A second approach to applying labels to topics uses the output from the model to identify words that distinguish the documents in a particular topic. The goal is to identify words that are common among documents that discuss the same topic and rare in documents that were generated by another topic. To identify the set of words that satisfy these properties, I select 10 words with the highest mutual information with a topic to label the clusters, which provides a principled method for cluster labeling appropriate for any unsupervised learning technique.

The mutual information between a topic and word measures the amount of information a word provides about whether a topic generated a document randomly chosen from the corpus. Suppose that after estimating the topics using the expressed agenda model, we want to compute the probability that a randomly chosen document  $y_{ij} \in Y$  was generated by topic  $k$ . Define the event that the document was generated by topic  $k$  as  $\zeta = I(\tau_k = 1)$ , and  $\Pr(\zeta = 1)$  is the probability that topic  $k$  generated the randomly chosen document. We can summarize our uncertainty about this classification by calculating the *entropy* that  $k$  generated a document,  $H(k)$  (MacKay 2003),

$$H(k) = - \sum_{t=0}^1 \Pr(\zeta = t) \log_2 \Pr(\zeta = t), \quad (7.1)$$

where  $\log_2$  is used because uncertainty is usually measured in bits. Entropy encodes uncertainty about whether a topic generated a document. It reaches a minimum if all the mass of the probability distribution is centered upon one value (all documents assigned to the same cluster) and reaches a maximum if the probability mass is evenly spread over the possible events (the documents are spread evenly across topics, MacKay 2003).

Conditioning upon additional information, such as a word  $w$ , can reduce the uncertainty about whether a topic generated a document. To represent the uncertainty after conditioning upon the additional information, first define the event that a word,  $w$ , appears in a document  $y_{ij}$  as  $\omega = I(w \in y)$  and the probability that a word  $w$  appears in a randomly chosen document is given by  $\Pr(\omega = 1)$ . We can now define the entropy for a topic, conditional on word  $w$ ,  $H(k|w)$ , as

$$H(k|w) = - \sum_{t=0}^1 \sum_{s=0}^1 \Pr(\zeta = t, \omega = s) \log_2 \Pr(\zeta = t | \omega = s). \quad (7.2)$$

As one would expect  $H(k) \geq H(k|w)$  for all  $k$  and  $w$ , with equality only if  $w$  provides no information about the clustering, or if the distribution of words in the cluster and outside of the cluster is identical (MacKay 2003).

To generate labels for each topic, we select stems that provide a great deal of information about whether a randomly chosen document belongs to a topic. Intuitively, we want to measure how much a stem reduces the uncertainty in  $H(k)$ , which we can compute as the

difference between equations (7.1) and (7.2). Define this difference as the mutual information for topic  $k$  with stem  $w$ , and denote this quantity with  $I(k|w) = H(k) - H(k|w)$  (MacKay 2003). If a word  $w$  provides no information about whether a topic generated a document, then  $H(k) = H(k|w)$  and  $I(k|w) = 0$ . But, if word  $w$  removes all uncertainty about whether a document was generated by topic  $k$ , then  $H(k|w) = 0$  and  $I(k|w)$  obtain its maximum possible value,  $H(k)$ . Further, as the information a word provides about the probability a document was generated by topic  $k$  increases,  $I(k|w)$  will increase as well (until reaching its maximum). Thus, the stems with the highest mutual information with each topic provide effective labels for a topic. In Appendix C, I provide the formula used to evaluate the mutual information.

In column 2 of Table 4, I have placed the stems with the 10 largest mutual information with each of the 43 categories. The words identified using the mutual information indicate that the expressed agenda model has uncovered well-defined topics. For example, stems with a high mutual information with the FEMA topic include *disast*, *FEMA*, *storm*, *damag*, *declar*, *emerg*, *flood*, *recoveri*, *rebuild*, *recov*. The Veteran Affairs topic has a high mutual information with stems *veteran*, *affair*, *medic*, *mental*, *wound*, *war*, *deserve*, *traumat*, *militari*, *afghanistan*.

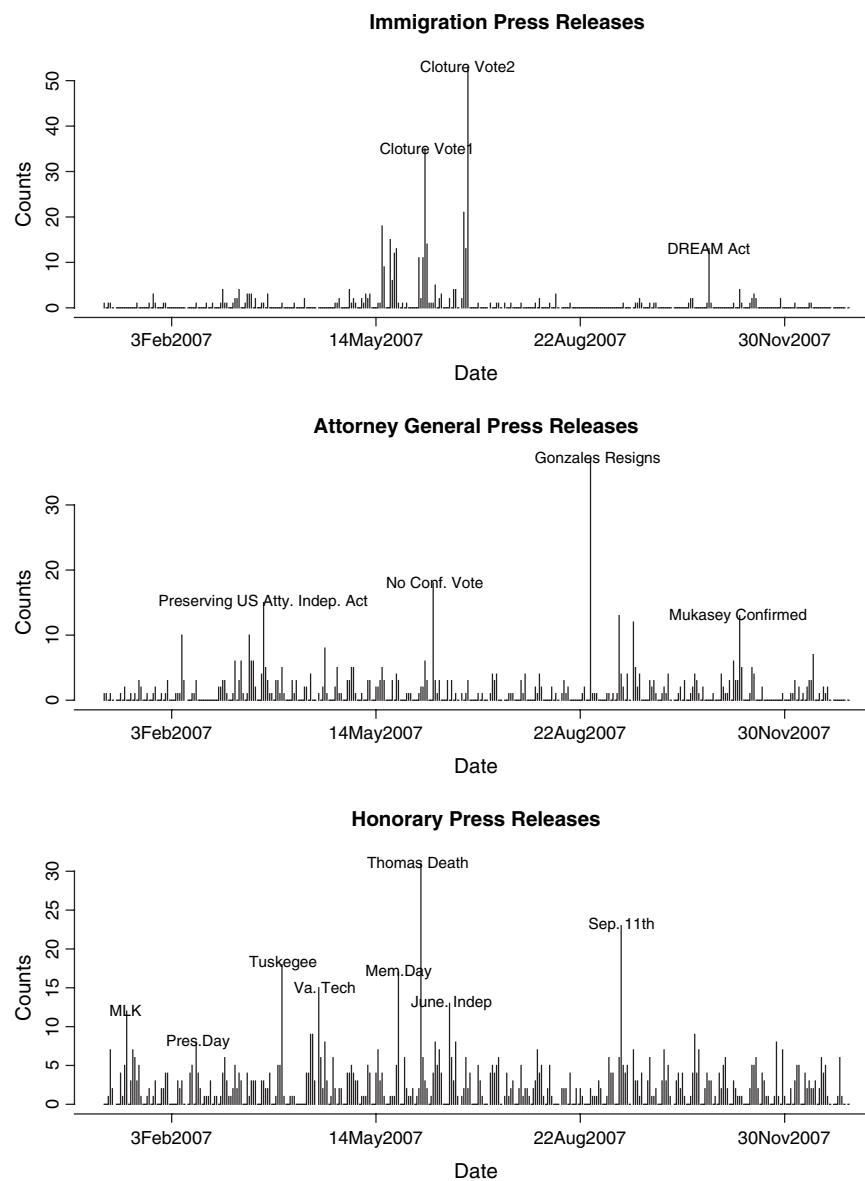
In addition to these formal validation methods, a heuristic look at Table 4 suggests the model was able to identify important issues in press releases from senators. The model estimated categories of press releases discussing the Walter Reed scandal and the subsequent Wounded Warrior legislation, the Iraq war, illegal immigration, global warming, the mortgage crisis, and a topic for press releases written to honor constituents and historic events. This suggests that the expressed agenda model was able to recover substantively interesting topics from the data. The final column of Table 4 provides the unique identifier that will be used for each topic throughout the paper.

### 7.3 Using Senate Debates and External Events to Validate Topics

Following a validation outlined in Quinn et al. (forthcoming), we can use the daily number of press releases generated by each topic as another validity check on the estimated topics. Consider the debate around the Comprehensive Immigration Reform Act of 2007 (S. 1348). President Bush's proposed immigration reforms were met with fierce resistance in the Senate and failed on two separate occasions. Both cloture votes in the Senate were high profile events, garnering a large amount of media and public attention. If the expressed agenda model captures meaningful communications from senators, we should expect to see a spike in the number of press releases about immigration around the cloture votes.

The top plot in Fig. 3 shows the number of press releases placed in the immigration category over 2007.<sup>13</sup> The two days with largest number of press releases about immigration correspond with the two cloture votes in the Senate. The model also detects the debate about the Development, Relief and Education for Alien Minors Act (DREAM) act that would have allowed the children of illegal immigrants to be eligible for college scholarships and enlist in the military. The other two plots in Fig. 3 further illustrate that the model is accurately capturing the content of press releases. Daily press releases about the Attorney General spike during the no-confidence vote for Alberto Gonzales and his subsequent resignation. Honorary press releases—press releases that discuss holidays and honor the

<sup>13</sup>The topic of press release  $i$  from senator  $j$  was assumed to be the largest element of  $\tau_{ij}$ .



**Fig. 3** Senate debates and external events explain spikes in the daily press releases from each topic.

recently deceased—also have spikes corresponding to national holidays, unforeseen tragedies, and the death of Senator Craig Thomas (R-WY).

## 8 Assessing Validity of Estimated Priorities

To validate the estimated expressed agendas from Senate press releases, I use a set of well-established facts about legislative behavior that also have intuitive appeal. If the expressed

agenda model agrees with these patterns first observed in smaller scale qualitative studies, we can have more confidence in applying the results of the model to test more contentious theories of legislative home style. These examples also demonstrate how easily the expressed agenda model can be used to assess how political actors explain work to constituents by incorporating information from every member of a legislature.

### 8.1 Validation 1: Committee Leaders Focus on their Committee's Issues

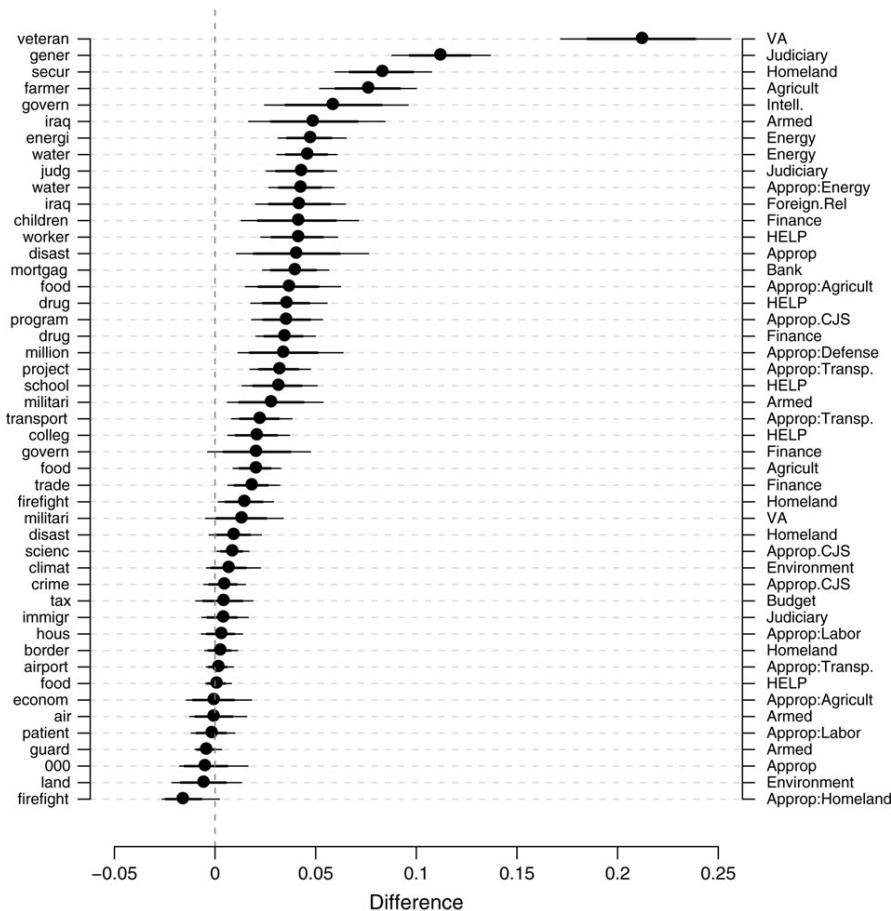
Members of Congress have strong incentives to emphasize their positions of power within the legislature. Fenno explains, "House members explain their use of power in Congress because they believe it will help them win renomination and reelection" (1978, 139). Elected officials also portray themselves as powerful to be perceived as creating effective policy (Fenno 1978), and legislators are likely to have strong personal interest in the issues that come before committeees they lead (Fenno 1973). An implication of Fenno's (1978) argument is that we should observe leaders of Senate committees—chairs and ranking members—allocate more attention to issues that fall under the jurisdiction of their committee than other senators. This straightforward explanation provides an ideal test of the validity of the estimated expressed agenda model.

Employing the results from the expressed agenda model, Fig. 4 carries out the comparison between prominent committee leaders and the rest of the Senate.<sup>14</sup> In Fig. 4, committee leaders' average attention dedicated to an issue under their committee's jurisdiction is compared with the average attention among the other 98 senators for 47 committee-topic pairs.<sup>15</sup> The left-hand vertical axis denotes the topics that were used for the comparison, and the right-hand vertical axis contains an abbreviated committee or appropriations subcommittee name. The solid dot represents the expected difference between committee leaders and the rest of the Senate, the thick lines are 80% and 95% highest posterior density (HPD) intervals, respectively. If committee leaders discuss issues related to their committee more often, then the estimates should be to the right of the vertical dotted line at zero.

Figure 4 shows that committee leaders allocate more attention to issues under their committee's jurisdiction than the average senator. In all but seven instances committee leaders allocate more attention to the issues under their committee's jurisdiction than other senators, and in some instances, leaders of Senate committees allocate substantially more attention to issues under their jurisdiction than other senators. For example, Joseph Lieberman (ID-CT) and Susan Collins (R-ME), chair and ranking member of the Homeland Security and Governmental Affairs committee, each allocate almost 10 percentage points more attention to Homeland Security issues than other senators, on average. The largest difference between committee leaders and the rest of the Senate corresponds to the Veterans' Affairs committee whose chairman, Daniel Akaka (D-HI), discusses Veterans' issues in 36% of his press releases—20 percentage points more than the closest senator. This example demonstrates that the Expressed Agenda Model is able to retrieve Fenno's (1978) observation that legislators will attempt to highlight their position of power in communications.

<sup>14</sup>In addition to committee leaders on standing committees, I also included subcommittee chairs on the Appropriations committee, due to the prominence of committee membership and the large and diverse nature of the legislation considered by this committee.

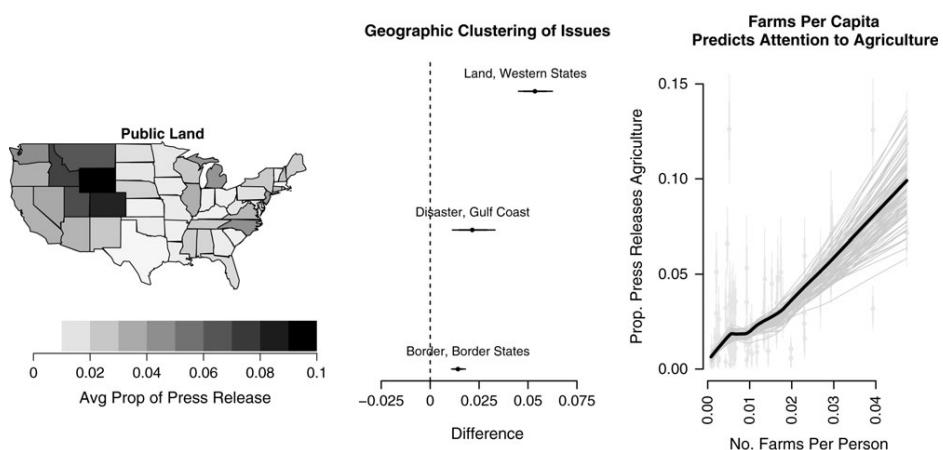
<sup>15</sup>Veterans' affairs were calculated with Richard Burr (R-NC) as the ranking member and not Larry Craig (R-ID); the results remain unchanged if Craig is used in place of Burr.



**Fig. 4** Chairman and ranking members of committees allocate more attention to issues under their committees' jurisdiction than other senators. This figure compares the attention that Senate committee leaders—chairs or ranking members—dedicate to topics under their committee jurisdictions to the attention allocated by the rest of the Senate. The solid dots represent the expected difference, the thick lines are 80% credible intervals, and the thin lines are 95% intervals. Along the left-hand vertical axis, the topics are listed, and on the right-hand side, the corresponding committee names are listed. In all but seven cases, the dot is to the right of the zero line, indicating that leaders of committees discuss issues that highlight their power in the institution more often than other senators.

### 8.2 Validation 2: Expressed Agendas Cluster Geographically

Studies of legislative behavior have found that the priorities legislators pursue in Washington and emphasize to constituents vary by location. Arnold argues that this variation occurs because of geographic specific costs and concentrated benefits to many of the policies enacted by Congress (1992, 26). This provides legislators an incentive to emphasize issues persistently important to their constituents. For example, Fenno describes a senator from a Western state seeking a seat on a committee with jurisdiction over issues important to the “public-land” states (1973, 139–40). If the expressed agenda model and



**Fig. 5** Attention to issues follows expected geographic patterns. This figure demonstrates that senators' expressed agendas are grouped geographically. The left-hand plot shows that senators from western states allocate substantial attention to public-land issues. Darker shades indicate that the average expected attention from the state's delegation to public-land issues is larger. The center plot carries out a comparison of three different regional issues: public-land and western states (top estimate), hurricanes and gulf coast states (middle estimate), and border-security and states that share an international border (bottom estimate). The point in each plot represents the expected difference between the attention to senators in a geographic area allocate to an issue and the attention senators from other areas of the country dedicate to the same issue. The thick and thin lines are 80% and 95% HPD intervals for this difference. Each point is to the right of the zero, indicating that the issues receive more attention in the geographic areas we would expect. The right-hand plot shows that senators from states with a large number of farms per person also tend to allocate more attention to agriculture issues. The horizontal axis represents the number of farms per resident of the state (one measure of agriculture's importance to a state), and the vertical axis indicates the proportion of press releases allocated to agricultural issues. The gray lines are lowess curves indicates the relationship between the number of farms per capita and the attention to agriculture, whereas the black line is the average relationship.

press release data are recovering valid estimates of legislative behavior, then we should observe this geographic clustering along some issues in the estimated expressed agendas.

The left-hand plot in Fig. 5 shows that this clustering is found in expressed agendas. This plot demonstrates that senators from Western states allocate substantial attention toward public-land topics—indicating a concern with this issue similar to the Western senator in Fenno (1973). The color of each state represents the average expected attention the state's delegation allocated to public land issues. The darker the state, the more attention to the issue and we see that the Western states are nearly black. A manual check shows that western delegations allocate substantial attention to public land. Wyoming's Senate delegation (John Barasso [R-WY] and Mike Enzi [R-WY]) dedicate an average of 18% of their releases to discussions of public land issues and Colorado's delegation (Ken Salazar [D-CO] and Wayne Allard [R-CO]) allocate 14.3% of their releases to land.

The center plot of Fig. 5 carries out the comparison between the attention western and non-western delegations allocate to public-land directly, along with two other geographic comparisons. This plot exhibits the geographic clustering we would intuitively expect from qualitative studies. The top-point represents the expected difference between the attention

to public-land issues for Western senators and the attention to public-land issues among other senators, whereas the thick and thin lines are 80% and 95% HPD intervals for the difference.<sup>16</sup> This shows that there is a very high-posterior probability that senators from Western states allocate more attention to public-land issues than senators from other parts of the country, corroborating an expected geographic comparison. The next two points indicate two other kinds of geographic clustering: senators from the Gulf coast states allocate more attention to disaster (hurricane)-related issues than other senators, and senators from states that share a border with Canada and Mexico issue a larger proportion of press releases about Border security (separate from immigration).

States that do not share borders may bear similar costs or receive similar benefits from policies. As a result, senators from these states with similar interests should attend to similar issues. For example, numerous states have a high-density of farms, but these states are not necessarily grouped in one location. Nonetheless, senators from the high-density agriculture states may be expected to address farm-related issues more than other senators. The right-hand plot shows that this is the case: senators from agricultural states allocate more attention to farming than other senators. The horizontal axis represents the number of farms per resident of the state (one measure of agriculture's importance to a state), and the vertical axis indicates the proportion of press releases allocated to agricultural issues.<sup>17</sup> The light gray lines are lowess curves indicating the relationship between the number of farms per capita and the attention to agriculture. Each gray line represents this relationship for one draw from each senators expressed agenda, whereas the solid black line indicates the average relationship between farms per capita and the proportion of press releases allocated to agriculture.<sup>18</sup> The gray lines slope upwards quickly, demonstrating that senators from states with a high concentration of farms also tend to invest attention in highlighting agricultural issues.

Taken together, the three plots in Fig. 5 demonstrate that the expressed agenda model is able to retrieve geographic and interest-based clustering in expressed agendas: an intuitive property of explanations well established in the qualitative literature on Congressional communication.

### 8.3 Validation 3: Attention to Appropriations Predicts Opposition to Earmark Reform

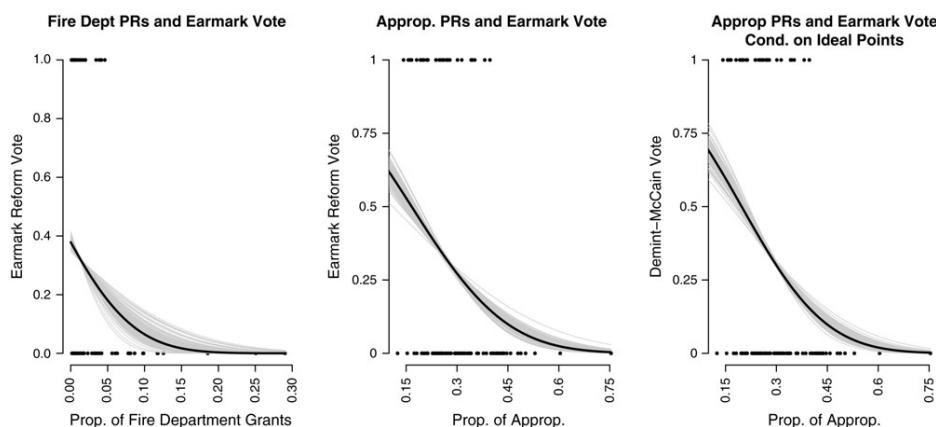
Senators who rely upon appropriations secured for their state in press releases have strong incentive to support institutions that allow them to continue to secure particularistic goods (Mayhew 1974). Senators who regularly tout appropriations secured for a state are likely to view these appropriations as essential to their electoral security (Fenno 1978; Cain, Ferejohn, and Fiorina 1987; King 1991). Senators may also feel pressure to ensure that their actions in Washington are consistent with the priorities emphasized to constituents, lest the legislator be portrayed as a hypocrite in future elections (Fenno 1978). In this section, I use a unique vote in the US Senate to show that the results of the expressed agenda model predict aspects of legislative behavior beyond ideal points.

On March 13, 2008, the Senate voted on the Demint-McCain amendment: a proposal introduced by Jim Demint (R-SC) and John McCain (R-AZ) to place a 1-year moratorium on earmarks in senate appropriations bills. Given the incentives to support institutions

<sup>16</sup>Western senators were identified using the region classification from the census bureau.

<sup>17</sup>The numbers of farms per state were obtained from the U.S. Department of Agriculture.

<sup>18</sup>The gray points in the background represent each senator's expected attention to farming, whereas the thick and thin lines are 50% and 90% HPD intervals for this quantity.



**Fig. 6** Senators who dedicate more attention to appropriations were more likely to oppose Demint-McCain. This figure shows that senators who dedicate more attention to appropriations in their press releases are more likely to oppose the Demint-McCain amendment. The vertical axis plots the vote on the amendment, and along the horizontal axis is the average proportion of press releases dedicate to discussing appropriations secured for fire departments. To generate the light gray lines, I took 100 draws from each senator's posterior expressed agenda and then regressed the earmark vote on the draw from the posterior. The gray lines represent the expected probability of supporting the Demint-McCain amendment, and the solid black line is the expected value of the relationship, averaged over the draws from the posterior distribution on the expressed agenda. The left-hand figure shows that senators who discuss fire department grants more often were more likely to oppose the Demint-McCain amendment, and the center plot shows that this relationship was even stronger for an aggregate appropriations category. The right-hand plot shows that the relationship remains even after conditioning upon estimated ideal points of senators, suggesting that consistency explains components of voting behavior beyond ideal point estimates.

essential to maintaining their incumbency advantage and to remain consistent, senators who allocate a large proportion of their press releases toward discussions of appropriations secured for the home state should be more likely to oppose the Demint-McCain amendment.

Figure 6 displays the relationship between senators' vote on the Demint-McCain amendment and two components of the expressed agenda: the proportion of press releases allocated to discussing fire department grants (left-hand plot) and a composite measure of appropriations (center- and right-hand plots). In the left-hand plot in Fig. 6, each senator's vote on the Demint-McCain amendment is predicted using the proportion of press releases dedicated to discussing grants secured for local fire departments—one measure of how often a senator discusses appropriations with constituents.<sup>19</sup> The vertical axis plots the vote on the amendment, and the horizontal axis represents the expected proportion of press releases discussing fire department grants. The gray lines account for the uncertainty inherent in measuring the legislators' priorities by taking 100 draws from each senator's posterior expressed agenda and then regressing the earmark vote on the draws using a probit regression. The black lines represent the average relationship over 1000 draws.

<sup>19</sup>The Demint-McCain amendment was defeated 29-71. I did not include Roger Wicker (R-MS) and Trent Lott (R-MS) due to the change in senate seat after the 2007 session.

Figure 6 shows that senators' votes on the Demint-McCain amendment tended to be consistent with the priorities articulated to constituents. In the left-hand plot, as the proportion of press releases dedicated to fire department grants increases, senators were less likely to support the moratorium on earmarks. The center plot exhibits the relationship between the Demint-McCain vote and an aggregated appropriations category (constituted of the bottom 13 topics from Table 4).<sup>20</sup>

This shows an even stronger relationship: senators who allocate more attention to appropriations were much less likely to vote for the Demint-McCain amendment. The right-hand plot in Fig. 6 shows that the results of the expressed agenda model provides predictive power beyond low-dimensional summaries of previous roll-call votes: the relationship between a senator's vote on Demint-McCain and the proportion of press releases discussing appropriations is still strong and negative, even after conditioning upon a senator's ideal point.<sup>21</sup> Taken together, these three plots show that the results of the expressed agenda model relate as expected to votes on the Senate floor. This provides another validation that the expressed agenda model estimates quantities of theoretical interest.

## 9 Applying the Expressed Agenda Model

In this section, I show that the estimated expressed agendas are ideal to address theoretically important questions about legislators' home styles. The use of Bayesian inference allows for direct inference about quantities of interest derived from the estimated expressed agendas. Further, by efficiently using all the press releases from each Senate office, the expressed agenda model allows comprehensive tests of hypotheses, in contrast to the limited tests that had been previously carried out in the literature.

### 9.1 Partners Not Rivals in the Senate

The structure of representation in the Senate is distinctive from other legislative bodies, with each state allocated two senators. Schiller argues that the dual representation in the Senate forces senators representing the same state to articulate distinctive priorities due to the persistent competition for media and public attention (2000, 65). Schiller (2000) provides evidence and a persuasive argument for this novel hypothesis, but is hindered by the existing methods and data, only comparing the statements of a handful of legislators from newspapers. While newspaper stories are an excellent measure of the kind of information available to citizens, newspaper stories conflate senators' priorities with the depictions offered by news writers. The expressed agenda model and the Senate press releases allow a direct and comprehensive test of whether senators from the same state articulate a distinctive set of priorities in press releases.

Schiller's (2000) argument asserts that senators from the same state respond to each other's priorities by advocating a different set of issues, which implies that we should observe senators from the same state articulating a distinctive set of issues. But, Schiller's

<sup>20</sup>Two methods were used to label these topics. First, I used topics with appropriations-related labels and increased attention around the passage of an appropriation bill. Second, I used a topic hierarchy to identify groups of issues that clearly referred to appropriations.

<sup>21</sup>Ideal points were estimated using a one-dimensional item-response theory model, as implemented in MCMCpack (Martin and Quinn 2008). The regressions in the right-hand plot incorporate uncertainty from both the priorities and the ideal points. In each plot, the senators' votes were regressed on draws from the posterior distribution on the priorities and the ideal points for each senator. The simulated lines were generated by varying the attention allocated to appropriations.

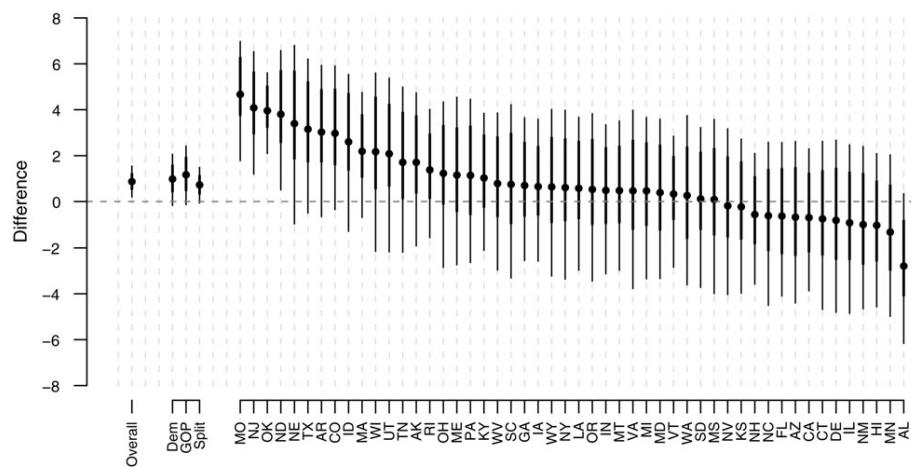
(2000) argument could be incorrect and we could still observe some differences in the stated priorities of same-state senators, due to idiosyncratic differences between the two senators in a delegation: such as distinct personal interests, divergent backgrounds, support among different constituencies located in the same state, and different partisanship. All these factors are unrelated to the strategic considerations outlined in Schiller (2000). Therefore, the critical test is not whether two senators articulate a different set of priorities: all senators will have some differences in their stated priorities. Rather, a test of Schiller's (2000) hypothesis depends upon whether senators from the same state have priorities that are *more distinctive* than a comparison group of senators who have no incentive to intentionally articulate different priorities. To perform this test, I compare the differences in priorities among senators who represent the same state to the differences in priorities among senators who represent different states. Under Schiller's (2000) hypothesis, senators who represent different states do not have incentive to carve out distinctive expressed agendas, and therefore, senators who represent different states provide a reasonable group to compare the differences that should be expected due to idiosyncratic variation.

To measure the distance between two expressed agendas, I use the distance metric on the simplex defined in Billheimer et al. (2001), which generalizes intuition about properties of distance in Euclidean space to the simplex. In Appendix 14 I define this metric. Define the distance between two expressed agendas  $\boldsymbol{\pi}^j, \boldsymbol{\pi}^i$ ,  $\text{Distance}_{i,j} = g(\boldsymbol{\pi}^j, \boldsymbol{\pi}^i)$  where  $g(\cdot, \cdot)$  is the distance metric developed in Billheimer et al. (2001). To test Schiller's (2000) hypothesis, I compare the average distance between expressed agendas of senators who represent different states to the average distance between expressed agendas of senators who represent the same state.<sup>22</sup> If Schiller's (2000) hypothesis is correct, average distance between expressed agendas of senators who represent different states to the average distance between expressed agendas of senators who represent the same state should be negative: implying that senators from the same state tend to have expressed agendas further apart than senators from different states.

The left-most line in Fig. 7 presents this quantity estimated from Senate press releases. This shows that senators who represent the same state have expressed agendas that are *more similar* than senators who represent different states. The solid dot in Fig. 7 represents the expected value of average distance between expressed agendas of senators who represent different states to the average distance between expressed agendas of senators who represent the same state and is above the horizontal dotted line, indicating that senators from the same state have more similar priorities, on average, than senators who represent different states. The thick lines and thin lines are 50% and 90% HPD intervals for the difference and both fail to intersect the zero line, indicating that there is a high posterior probability that senators from the same state tend to emphasize similar issues. This holds regardless of the partisanship of the state delegation: the next three lines show that the expressed agendas of split, Republican, and Democratic delegations

<sup>22</sup>To derive the comparison between the expressed agendas of senators from the same state and different states, collect the  $\binom{100}{2} = 4950$  pairs of senators into the set  $\mathcal{P}$ . For example, one pair of senators in this set is Grassley (R-IA), Murray (D-WA)). Define the set  $\mathcal{S}$  as the set of 50 pairs of senators who represent the same state, such as (Bayh (D-IN), Lugar (R-IN))  $\in \mathcal{S}$ . And define  $\mathcal{S}' = \mathcal{P} \setminus \mathcal{S}$  as the 4900 pairs of senators who represent different states, for example, (McCain (R-AZ), Obama (D-IL))  $\in \mathcal{S}'$ . Formally,

$$\text{Diff}(\mathcal{S}', \mathcal{S}) = \sum_{(i,j) \in \mathcal{S}'} \frac{g(\boldsymbol{\pi}^j, \boldsymbol{\pi}^i)}{4900} - \sum_{(k,m) \in \mathcal{S}} \frac{g(\boldsymbol{\pi}^m, \boldsymbol{\pi}^k)}{50} \quad (9.1)$$



**Fig. 7** Senators who represent the same state have more similar expressed agendas than senators from other states. This figure compares the average distance among senators who represent different states to the average distance of senators who represent the same states. The solid dots represent the expected difference, and the thick and thin lines are 50% and 90% HPD intervals, respectively. If senate delegations have more similar expressed agendas than senators who represent different states, then the estimates should be above the horizontal dashed zero-line. The first line compares the average distance between expressed agendas from senators from different states with the average distance between expressed agendas of senators from the same state, showing that senators from the same state communicate a more similar set of priorities than senators from other states. This same pattern holds regardless if the delegation is split, Republican, or Democrat. Further, most states' delegations have more similar expressed agendas than senators who represent different states.

are closer, on average, than the average distance between priorities for senators who represent different states. The final set of lines compare the distance in each senate delegation with the average distance between the expressed agendas of senators who represent different states, and the lines are color-coded according to the partisanship of the delegation. This shows that the majority of state's delegations tend to be more similar than the average distance between the priorities of senators who represent different states, although there is substantial variation in this quantity across states.

This shows that contrary to the prediction's from Schiller's (2000) theory, senators from the same state emphasize a *more similar set of priorities than senators who represent different states*, in press releases from 2007. This similarity could occur because senators from the same state may rely upon similar groups as part of their “reelection” constituency (Fenno 1978), subsequently leading senators to identify a similar set of priorities to please this constituency. Alternatively, senators might be able to multiply the effectiveness of their own communication by coupling their efforts with the other senator from their state—forming a partnership to help ensure reelection for both senators.

## 10 Conclusions and Future Work

This paper has introduced a new method for analyzing the expressed priorities of political actors, as articulated in political texts: the expressed agenda model. This method is capable

of handling thousands of texts from hundreds of political actors to estimate the topics in a data set, assign documents to topics, and measure the proportion of press releases each political actor dedicates to the topics. Using a Bayesian model and a recently developed estimation procedure allows for efficient inference about each senator's priorities. I apply this method to an original collection of press releases from Senate offices and show that the expressed agenda model is capable of retrieving a theoretically relevant set of topics and that press releases are an ideal medium for measuring how senators portray themselves to constituents. Through a series of applications I validate the estimated priorities and topics and show that the model facilitate tests of theoretically important questions about congressional communication.

The statistical model developed in this paper is applicable to a variety of political situations beyond the study of home style and therefore has broad implications for the way political scientists study political communication. The expressed agenda model is ideal whenever scholars are interested in comparing the priorities that authors articulate in text, an important problem in large literatures studying campaign strategy (Petrocik 1996), media-content (Armstrong et al. 2006), and presidential communication (Lee 2008). The expressed agenda model can also be applied to study other forms of Congressional communication, like the attention allocated to issues in Senate floor speeches (Quinn et al. forthcoming) or the issues raised during Senate committee hearings. The forthcoming software package (implemented in the R computing language) makes applying the expressed agenda model straightforward.

The press release data used to analyze senator's expressed agendas provide a comprehensive collection of statements senators make to constituents, which facilitates testing a number of theories. For example, the press releases, coupled with stories from local newspapers, suggest a new approach to studying the connection between politicians and the media. Previous studies of this interaction have relied upon time-series regressions to measure how the priorities politicians articulate covaries with the issues discussed in the media (Bartels 1996). This method provides only suggestive evidence of how politicians and the press interact. In contrast, the press release coverage rate provides a direct measure of how politicians ensure that their message is repeated (and amplified) by the press. Expanding upon this analysis is an important topic for future research. Measuring how often newspapers cover elite statements would provide an answer to a number of theoretically important questions, including how reliant local newspapers are on information from Senate offices, identifying the role of partisanship in determining how often a newspaper prints a legislator's message, and determining how a newspaper's reliance upon information from Congressional offices influences the incumbency advantage.

### Appendix A. Collecting the Press Releases

The data set used in this paper contains all the Senate press releases from the 2007 calendar year or the first session of the 110th Congress. Due to the large number of press releases, they require a large expenditure of resources to analyze manually (Yiannakis 1982; Lipinski 2004). I overcome this problem by using automatic data collection methods: I wrote a set of "screen scraping" scripts in the Python computing language. Each script collects all the press releases from a senator's Web site, removes any extraneous content unrelated to the text of the press release, and then stores the text. The result of this automated collection process is a data set of 24,236 press releases for 2007.

## Appendix B. Deriving the Update Equations for Variational Inference

Given the model and priors outlined above the posterior is given by

$$\begin{aligned} \alpha_k | \delta, \lambda &\sim \text{Gamma}(\lambda, \delta) \quad \text{for all } k = 1, \dots, K \\ \pi_i | \alpha &\sim \text{Dirichlet}(\alpha) \quad \text{for all } i = 1, \dots, n \\ \tau_{ij} | \pi_i &\sim \text{Multinom}(\pi_i) \quad \text{for all } j = 1, \dots, D_i; i = 1, \dots, n \\ \mu_k | \eta_k, \kappa &\sim \text{vMF}_w(\eta_k, \kappa) \quad \text{for all } k = 1, \dots, K \\ y_{ij}^* | \mu, \kappa, \tau_{di,j} &= 1 \sim \text{vMF}_w(\mu_j, \kappa) \quad \text{for all } j = 1, \dots, D_i; i = 1, \dots, n \end{aligned}$$

with parametric form

$$\begin{aligned} p(\pi, \mu, \alpha, \tau | Y) &\propto \prod_{k=1}^K \exp(-\alpha_k) \exp(\kappa \eta' \mu_k) \times \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K (\alpha_k)} \\ &\times \prod_{i=1}^{100} \left[ \prod_{k=1}^K (\pi_{ik})^{\alpha_k - 1} \prod_{j=1}^{D_i} \prod_{k=1}^K [\pi_{ik} \exp(\kappa \mu' y_{ij}^*)]^{\tau_{ijk}} \right]^{-1} \end{aligned} \quad (\text{B.1})$$

In the supplemental notes, I provide the model and derive the estimation algorithm for the expressed agenda model with a multinomial distribution used to model document content.

### B.1. Approximating Distribution

I adopt a standard *mean-field* approach to estimation of equation (B.1).<sup>23</sup> Specifically, I approximate the full posterior with a family of distributions that contain additional independence assumptions *but no specific parametric forms are assumed* and then select the member of this distributional family that minimizes the Kullback-Leibler divergence between the true posterior and the approximating distribution. Call the approximating distribution  $q(\pi, \tau, \mu, \alpha)$  and assume that this distribution factors into  $q(\pi)q(\tau)q(\mu)q(\alpha)$ . We will estimate the full posterior for topics  $q(\tau)$  and senators' priorities  $q(\pi)$  and then obtain *Maximum a Posteriori* estimates for  $\mu_k$  and  $\alpha$ .<sup>24</sup> This implies that we can write the approximating distribution as  $q(\pi, \tau, \mu, \alpha) = \prod_{i=1}^N q(\pi)_i \prod_{i=1}^N \prod_{j=1}^{D_i} q(\tau_{ij}) \prod_{k=1}^K \delta_{\mu_k^*} \delta_{\alpha^*}$  where  $\delta_{(.)}$  is the Dirac delta function,  $\mu_k^*$  represents the MAP estimates for the  $k$ th category and  $\alpha^*$  represents the MAP estimates for  $\alpha$ .<sup>25</sup>

### B.2. Minimizing KL Divergence

The standard approach to minimizing the KL divergence between the true posterior and the approximating distribution in variational approximations is to solve an equivalent problem:

<sup>23</sup>The derivation throughout this appendix is a fairly standard in the application of variational inference to mixture models and therefore should have similarities to the derivations in Bishop (2006) and Blei et al. (2003). Note, that I provide these derivations because variational inference in political science is nonstandard.

<sup>24</sup>I estimate the full posterior (with distributions) for the model with multinomial distributions—the integral with vMF distributions are difficult to compute. Furthermore, not much is gained by maintaining a full posterior on the components of the mixture because of the large number of stems used in the analysis.

<sup>25</sup>Recall that the Dirac delta function is a probability distribution that places all of the mass on a single number, given by the term in the subscript.

maximizing a lower bound on the *evidence* or the marginal probability of the data. To derive the lower bound, first write the log evidence as,

$$\log p(\mathbf{Y}) = \log \sum_{\tau} \int \int \int p(\mathbf{Y}, \pi, \alpha, \mu) d\pi d\mu d\alpha.$$

Insert the approximating distribution  $q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha})$  by multiplying by 1,

$$\log p(\mathbf{Y}) = \log \sum_{\tau} \int \int \int \frac{q(\pi, \tau, \mu, \alpha)}{q(\pi, \tau, \mu, \alpha)} p(\mathbf{Y}, \pi, \alpha, \mu) d\pi d\mu d\alpha.$$

Applying Jensen's inequality yields the lower bound

$$\log p(\mathbf{Y}) \geq \sum_{\tau} \int \int \int q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha}) \log \left\{ \frac{p(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\mu})}{q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha})} \right\} d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\alpha}. \quad (\text{B.2})$$

We will define the right-hand side of Inequality (B.2) as  $\mathcal{L}(q)$ . A straightforward proof (in supplemental notes) shows that maximizing  $\mathcal{L}(q)$  with respect to  $q$  is equivalent to minimizing the KL divergence between the approximating and true posterior.<sup>26</sup> This is the lower bound used to evaluate convergence of the model as well.

### B.3. Distributional Forms

To maximize  $\mathcal{L}(q)$  with respect to  $q$ , we need to obtain the parametric form of the approximating distribution and select the correct member of that family (maximize the parameters for a given distribution). Either from direct derivation or by applying results on the use of mean-field approximations to exponential families (Jordan et al. 1999), we can obtain the functional forms.<sup>27</sup> This derivation shows that  $q(\boldsymbol{\pi})_i$  is a Dirichlet distribution and represents the  $K \times 1$  vector of shape parameters that characterize this distribution  $\boldsymbol{\theta}_i$ . The same derivation shows that  $q(\boldsymbol{\tau})_{ij}$  is a multinomial distribution and call  $\mathbf{r}_{ij}$  the  $K \times 1$  vector of parameters for  $j$ th document from senator  $i$ .

### B.4. Iterative Algorithm for Estimation

Each iteration of the estimation algorithm proceeds in several steps. Define the values of the parameters from the previous iteration as  $\boldsymbol{\mu}^{\text{old}}, \boldsymbol{\alpha}^{\text{old}}, \boldsymbol{\theta}^{\text{old}}, \mathbf{r}^{\text{old}}$ . In each step we update the parameters to maximize the lower bound  $\mathcal{L}(q)$  with respect to each independent component of the approximating distribution. The following describes each step in more detail.

<sup>26</sup>Note that  $\mathcal{L}(q)$  is a functional: an operator that maps from a space of functions to the real line (Bishop 2006). In the case of exponential family models, the lower bound is convex in the approximating distribution—facilitating iterative (EM-like) algorithms for estimation.

<sup>27</sup>This derivation is standard in variational inferences, see Bishop (2006).

#### B.4.1. Update step for $\mathbf{r}_{ij}$

Typical element of senator  $\mathbf{r}_{ij}$  for senator  $i$ ,  $r_{ijg}^{\text{new}}$  is equal to

$$r_{ijg}^{\text{new}} = \frac{\exp\left[\Psi(\theta_{ig}^{\text{old}}) - \Psi\left(\sum_{k=1}^K \theta_{ik}^{\text{old}}\right)\right] \exp[\kappa \boldsymbol{\mu}_g^{\text{old}} \mathbf{y}_{ij}^*]}{\sum_{k=1}^K \left( \exp\left[\Psi(\theta_{ik}^{\text{old}}) - \Psi\left(\sum_{j=1}^K \theta_{ij}^{\text{old}}\right)\right] \exp[\kappa \boldsymbol{\mu}_k^{\text{old}} \mathbf{y}_{ij}^*] \right)}. \quad (\text{B.3})$$

where  $\Psi(\cdot)$  is the digamma function.

#### B.4.2. Update step for $\boldsymbol{\theta}^i$

Typical element  $\theta_{ig}$  of  $\boldsymbol{\theta}_i$  has update step (Blei et al. 2003),

$$\theta_{ig}^{\text{new}} = \alpha_g^{\text{old}} + \sum_{j=1}^{D_i} r_{ijk}^{\text{new}}. \quad (\text{B.4})$$

#### B.4.3. Update step for $\boldsymbol{\mu}_k$

The update step for  $\boldsymbol{\mu}_k^{\text{new}}$  is given by Banerjee et al. (2005),

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{\boldsymbol{\eta} + \sum_{i=1}^N \sum_{j=1}^{D_i} r_{ijk}^{\text{new}} \mathbf{y}_{ij}^*}{\left\| \boldsymbol{\eta} + \sum_{i=1}^N \sum_{j=1}^{D_i} r_{ijk}^{\text{new}} \mathbf{y}_{ij}^* \right\|}. \quad (\text{B.5})$$

#### B.4.4. Update step for $\boldsymbol{\alpha}$

Unfortunately, a closed form for the shape parameters  $\boldsymbol{\alpha}$  does not exist, so we use a Newton-Raphson algorithm, developed in Blei et al. (2003) to perform the optimization.

### B.5. Using the Model

This estimation algorithm is deterministic and therefore easy to implement in a standard package. This version of the expressed agenda model, along with various extensions, is available in the free R software package expAgenda, which is forthcoming.

### B.6. Generalizing the Expressed Agenda Model: Including Covariates

Suppose that we observe an  $M \times 1$  set of covariates for each author,  $\mathbf{X}_i$  (including an intercept term as well). The following extends the Dirichlet-multinomial regression suggested in Mimno and McCallum (2008) to allow for the inclusion of covariates to facilitate more efficient smoothing. Specifically, we modify the model to include a regression at the top of the hierarchy,

$$\begin{aligned} \boldsymbol{\beta}_k &\sim \text{Normal}(0, \sigma^2 I) \quad \text{for all } k = 1, \dots, K \\ \alpha_{ik} &= \exp(\mathbf{X}'_i \boldsymbol{\beta}_k) \\ \boldsymbol{\pi}_i | \boldsymbol{\alpha}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}_i) \quad \text{for all } i = 1, \dots, N. \\ \tau_{ij} | \boldsymbol{\pi}_i &\sim \text{Multinom}(1, \boldsymbol{\pi}_i) \quad \text{for all } j = 1, \dots, D_i; \quad i = 1, \dots, N \\ \boldsymbol{\mu}_k | \boldsymbol{\eta}_k, \kappa &\sim \text{vMF}_w(\boldsymbol{\eta}_k, \kappa) \quad \text{for all } k = 1, \dots, K \\ \mathbf{y}_{ij}^* | \boldsymbol{\mu}, \kappa, \tau_{d_{i,j}} &= 1 \sim \text{vMF}_w(\boldsymbol{\mu}_j, \kappa) \quad \text{for all } j = 1, \dots, D_i; \quad i = 1, \dots, n \end{aligned}$$

where  $\sigma^2$  represents the prior variance on the regression coefficients. The inclusion of covariates allows the model to identify subsets of senators who express similar priorities and therefore include additional information in the model that can aid in classification.

#### B.6.1. Modifying the variational approximation

In this section, I show how the algorithm in Appendix B can be extended to include the regression at the top of the hierarchy.<sup>28</sup> The first modification is an update step for the regression coefficients for each topic  $\beta_k$ . Collect the coefficient vectors into the  $M \times K$  matrix  $\beta$ . We focus upon MAP estimates for the coefficients, and a closed form update for the regression coefficients is unavailable. Therefore, we apply a BFGS algorithm to maximize the following,

$$\begin{aligned} f(\beta) = & -\sum_{k=1}^K \frac{1}{2\sigma^2} (\beta_k' \beta_k) + \sum_{i=1}^N \left[ \log \Gamma \left( \sum_{k=1}^K \alpha_{ik} \right) - \sum_{k=1}^K \Gamma(\alpha_{ik}) \right] \\ & + \sum_{i=1}^N \sum_{k=1}^K \left[ (\exp(X_{ikt} \beta_k) - 1)(\Psi(\gamma_{ik}) - \Psi \sum_{k=1}^K \gamma_{ik}) \right]. \end{aligned}$$

The only other modification to the update step for  $q(\pi_i)$  to include the additional information in the prior  $\alpha_{ik}$ ,

$$\gamma_{ik} = \alpha_{ik} + \sum_{j=1}^{D_i} r_{ijk}.$$

The algorithm otherwise remains unchanged.

### Appendix C. Deriving an Expression for Mutual Information

To derive an expression for mutual information, we apply the definitions of  $H(k)$  and  $H(k|w)$  to obtain

$$H(k) - H(k|w) = \sum_{t=0}^1 \sum_{s=0}^1 \Pr(\zeta = t, \omega = s) \log_2 \frac{\Pr(\zeta = t, \omega = s)}{\Pr(\zeta = t) \Pr(\omega = s)}. \quad (\text{C.1})$$

To evaluate equation (C.1), we compute the necessary probabilities. Define the number of documents in which word  $w_j$  appears as  $n_j = \sum_{i=1}^D \omega_j^i$  and the number of documents where  $w_j$  does not appear as  $n_{-j} = D - n_j$ . Define the effective number of documents assigned to cluster  $k$  and the effective number of documents not in cluster  $k$  as  $n_k = \sum_{i=1}^D r_{i,k}$  and  $n_{-k} = D - n_k$ . To finish the relevant counts, we need to attend to the four possible joint counts of words and topics,

$$n_{j,k} = \sum_{i=1}^D r_{i,k} \omega_j^i; n_{j,-k} = \sum_{i=1}^D (1 - r_{i,k}) \omega_j^i; n_{-j,k} = \sum_{i=1}^D r_{i,k} (1 - \omega_j^i); n_{-j,-k} = \sum_{i=1}^D (1 - r_{i,k}) (1 - \omega_j^i).$$

<sup>28</sup>Mimno and McCallum (2008) suggest stochastic EM to estimate a mixture model with the Dirichlet-multinomial regression prior. To my knowledge, this is the first suggestion of a variational-maximization approach.

The probabilities are then defined as,

$$\begin{aligned} \Pr(\zeta = 1, \omega_j = 1) &= \frac{n_{j,k}}{D}; \quad \Pr(\zeta = 1, \omega_j = 0) = \frac{n_{j,-k}}{D}; \quad \Pr(\zeta = 0, \omega_j = 1) \\ &= \frac{n_{-j,k}}{D} \Pr(\zeta = 0, \omega_j = 0) = \frac{n_{-j,-k}}{D}; \quad \Pr(\zeta = 1) = \frac{n_k}{D} \Pr(\zeta = 0) = \frac{n_{-k}}{D} \Pr(\omega_j = 1) \\ &= \frac{n_j}{D} \Pr(\omega_j = 0) = \frac{n_{-j}}{D}. \end{aligned}$$

This implies the following formula for  $I(k|w_j)$  (Manning et al. 2008),

$$I(k|w_j) = \frac{n_{j,k}}{D} \log_2 \frac{n_{j,k}D}{n_j n_k} + \frac{n_{j,-k}}{D} \log_2 \frac{n_{j,-k}D}{n_j n_{-k}} + \frac{n_{-j,k}}{D} \log_2 \frac{n_{-j,k}D}{n_{-j} n_k} + \frac{n_{-j,-k}}{D} \log_2 \frac{n_{-j,-k}D}{n_{-j} n_{-k}}.$$

## Appendix D. Defining Distance on the Simplex

In Section 9, I rely upon the distance metric on a simplex developed in Billheimer et al. (2001). In this appendix, I define this metric. Define the *composition* operator,  $\mathcal{C}(\mathbf{p}) = \left( \frac{p_1}{\sum_{i=1}^k p_i}, \dots, \frac{p_k}{\sum_{i=1}^k p_i} \right)$  and define  $\Delta^{k-1}$  as the  $k - 1$  dimensional simplex. Define the additive logistic map  $\phi : \Delta^{k-1} \rightarrow \mathfrak{R}^{k-1}$   $\phi(\mathbf{c}) = \left( \log\left(\frac{c_1}{c_k}\right), \dots, \log\left(\frac{c_{k-1}}{c_k}\right) \right)$ , where  $\mathbf{c} \in \Delta^{k-1}$  (Aitchison 1986). Suppose that  $\mathcal{N}_{k-1}^{-1} = \mathbf{I}_{k-1} - \frac{1}{k}\mathbf{1}\mathbf{1}'$  and that  $\mathbf{I}_{k-1}$  is a  $k - 1 \times k - 1$  identity matrix and  $\mathbf{1}$  is a vector of 1's. For two points in a simplex,  $\boldsymbol{\pi}^j, \boldsymbol{\pi}^i \in \Delta^{k-1}$ , define  $g : \Delta^{k-1} \times \Delta^{k-1} \rightarrow \mathfrak{R}_+$ ,  $g(\boldsymbol{\pi}^j, \boldsymbol{\pi}^i) = \phi\left(\mathcal{C}\left(\frac{\pi_1^j}{\pi_k^j}, \dots, \frac{\pi_k^j}{\pi_k^j}\right)\right) \mathcal{N}_{k-1}^{-1} \phi\left(\mathcal{C}\left(\frac{\pi_1^i}{\pi_k^i}, \dots, \frac{\pi_k^i}{\pi_k^i}\right)\right)$ .

## References

- Aitchison, John. 1986. *The statistical analysis of compositional data*. New York: Chapman and Hall.
- Armstrong, Elizabeth, Daniel Carpenter, and Marie Hojnacki. 2006. "Whose deaths matter? Mortality, advocacy, and attention to disease in the mass media." *Journal of Health Politics, Policy and Law* 31(4):729–72.
- Arnold, R. Douglas. 1992. *The logic of congressional action*. New Haven, CT: Yale University Press.
- . 2004. *Congress, the press, and political accountability*. Princeton, NJ: Princeton Press.
- Associated Press. 2007. "'Biotown' receives federal grant." *Times of Northwest Indiana* (accessed May 15, 2008).
- . 2008. "Chicago to receive 9.6 million for hybrid buses". *Chicago Tribune* (accessed June 10, 2008).
- Banerjee, Arindam, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. "Clustering on the unit hypersphere using von Mises-Fisher distributions." *Journal of Machine Learning Research* 6:1345–82.
- Bartels, Larry. 1996. "Politicians and the press: Who leads, who follows?" *Presentation at the Annual Meeting of APSA*, San Francisco, CA.
- Billheimer, D., Peter Guttorp, and William F. Fagan. 2001. "Statistical interpretation of species composition." *Journal of the American Statistical Association* 96(456):1205–15.
- Bingaman, Sen. Jeff. 2007. "Bingaman and Domenici introduce legislation to dramatically expand renewable fuel sources." <http://bingaman.senate.gov/> (accessed January 1, 2008).
- Bishop, Christopher. 2006. *Pattern recognition and machine learning*. New York: Springer.
- Blei, David, and John Lafferty. 2006. "Dynamic topic models." *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, June 25–29, 2006. 113–20.
- Blei, David, Andrew Y. Ng, and Michael Jordan. 2003. "Latent Dirichlet allocation." *Journal of Machine Learning and Research* 3:993–1022.
- Bloomfield, Louis. 2008. "WCopyFind." Software. <http://plagiarism.phys.virginia.edu/Wsoftware.html> (accessed June 1, 2008).

- Cain, Bruce, John Ferejohn, and Morris Fiorina. 1987. *The personal vote: Constituency service and electoral independence*. Cambridge, MA: Harvard University Press.
- Chambliss, Sen. Saxby 2007. "Chambliss Touts focus on BioFuels in Next Farm Bill." (accessed January 1, 2008).
- Collins, Sen. Susan 2007. "Senator Collins announces \$894,918 for Domtar, Fraser mill workers." <http://collins.senate.gov/public/> (accessed January 1, 2008).
- Cook, Timothy. 1988. "Press secretaries and media strategies in the House of Representatives: Deciding whom to pursue." *American Journal of Political Science* 32(4):1047–69.
- \_\_\_\_\_. 1989. *Making laws and making news: Media strategies in the US House of Representatives*. Washington, DC: Brookings.
- Craig, Sen. Larry 2007. "Senate confirms Randy Smith." <http://craig.senate.gov/> (accessed January 1, 2008).
- Durbin, Sen. Richard 2008. "Durbin announces a 9.6 million DOT grant for CTA hybrid buses." <http://durbin.senate.gov/> (accessed June 10, 2008).
- Fenno, Richard. 1973. *Congressmen in committees*. Boston: Little Brown and Company.
- \_\_\_\_\_. 1978. *Home style: House members in their districts*. Boston: Addison Wesley.
- Fraley, Chris, and Adrian Raftery. 2002. "Model-based clustering, discriminant analysis, and density estimation." *Journal of the American Statistical Association* 97(458):611.
- Gabel, Mathew, and Kenneth Scheve. 2007. "Estimating the effect of elite communications on public opinion." *American Journal of Political Science* 51(4):1013–28.
- Gelman, Andrew, and Gary King. 1990. "Estimating incumbency advantage without bias." *American Journal of Political Science* 34(4):1142–64.
- Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
- Grassley, Sen. Chuck. 2007. "Grassley questions big oil's commitment to lessening US dependence on foreign oil." <http://grassley.senate.gov/> (accessed January 1, 2008).
- Gutmann, Amy, and Dennis Thompson. 1996. *Democracy and disagreement*. Cambridge, MA: Harvard University Press.
- Harkin, Sen. Tom 2007. "Lawmakers make renewable fuels availability, energy efficiency a top priority for Congress." <http://harkin.senate.gov/> (accessed January 1, 2008).
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The elements of statistical learning*. New York: Springer.
- Hill, Kim Quaile, and Patricia Hurley. 2002. "Symbolic speeches in the US Senate and their representational implications." *Journal of Politics* 64(1):219–31.
- Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2008. "Computer-assisted topic classification for mixed-methods social science research." *Journal of Information Technology and Politics* 4(4):31–46.
- Hopkins, Daniel, and Gary King. Forthcoming. "Extracting systematic social science meaning from text." *American Journal of Political Science*.
- Jordan, Michael, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul. 1999. "An Introduction to variational methods for graphical models." *Machine Learning* 37:183–233.
- King, Gary. 1991. "Constituency service and the incumbency advantage." *British Journal of Politics* 21(1): 119–28.
- Kingdon, John. 1989. *Congressmen's voting decisions*. Ann Arbor: University of Michigan.
- Kyl, Sen. John. 2007. "Senate approves Kyl Feinstein provision adding judgeship." <http://kyl.senate.gov/> (accessed January 1, 2008).
- Lautenberg, Sen. Frank 2007. "Lautenberg Bill to reverse Bush administration's weakening of toxic releases reporting," Press Release.
- Lee, Frances. 2008. "Dividers, not uniters: Presidential leadership and Senate partisanship, 1981–2004." *Journal of Politics* 70(4):914–28.
- Lipinski, Daniel. 2004. *Congressional communication: Content and consequences*. Ann Arbor: University of Michigan Press.
- Lugar, Sen. Richard 2007. "Biotown awarded 1.71 million USDA grant." <http://lugar.senate.gov/> (accessed January 1, 2008).
- MacKay, David. 2003. *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.
- Manning, Christopher, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press.
- Mansbridge, Jane. 2003. "Rethinking representation." *American Political Science Review* 97(4):515–28.
- Martin, Andrew, and Kevin Quinn. 2008. "Markov chain Monte Carlo package (MCMCpack)." Software, R Package.

- Mayhew, David. 1974. *Congress: The electoral connection*. New Haven, CT: Yale University Press.
- McCombs, Maxwell. 2004. *Setting the agenda: The mass media and public opinion*. Cambridge: Polity.
- McLachlan, Geoffrey, and David Peel. 2000. *Finite mixture models*. New York: John Wiley & Sons.
- McLachlan, Geoffrey, and Thriyambakam Krishnan. 1997. *The EM algorithm and extensions*. New York: Wiley.
- Menendez, Sen. Robert, 2007. "Lautenberg Bill to reverse Bush administration's weakening of toxic releases reporting," Press Release.
- Mimno, David, and Andrew McCallum. 2008. "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression." *Conference on Uncertainty in Artificial Intelligence*. Plenary Presentation, Helsinki, Finland.
- Ng, Andrew, Michael Jordan, and Yair Weiss. 2002. "On spectral clustering: Analysis and an algorithm." *Advances in Neural Information Processing Systems 14: Proceedings of the 2002 Conference*, Vancouver, Canada.
- Petrocik, John. 1996. "Issue ownership in presidential elections, with a 1980 case study." *American Journal of Political Science* 40(3):825–50.
- Porter, Martin. 1980. "An algorithm for suffix stripping." *Program* 14(3):130–7.
- Quinn, Kevin, Burt Monroe, Michael Colaresi, Michael Crespin, and Dragomir Radev. Forthcoming. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science*.
- Schaffner, Brian. 2006. "Local news coverage and the incumbency advantage in the US house." *Legislative Studies Quarterly* 31(4):491–511.
- Schiller, Wendy. 2000. *Partners and rivals: Representation in US Senate delegations*. Princeton, NJ: Princeton University Press.
- Sigelman, Lee, and Emmitt Buell. 2004. "Avoidance or engagement? Issue convergence in US presidential campaigns, 1960–2000." *American Journal of Political Science* 48(4):650–61.
- Simon, Adam. 2002. *The winning message: Candidate behavior, campaign discourse, and democracy* Cambridge, UK: Cambridge University Press.
- Staff, 2007. "Sens. Snowe, Collins announce NEG Funding." *Bangor Daily News*, November 2, 2007 (accessed June 15, 2008).
- Sulkin, Tracy. 2005. *Issue politics in congress*. Cambridge: Cambridge University Press.
- Teh, Y., M. Jordan, M. Beal, and D. Blei. 2006. "Hierarchical Dirichlet processes." *Journal of the American Statistical Association* 101(476):1566–81.
- Vinson, Danielle. 2002. *Through local eyes: Local media coverage of congress*. Creskill, NJ: Hampton.
- Watanabe, Satosi. 1969. *Knowing and guessing: A quantitative study of inference and information*. New York: Wiley.
- Webb, Sen. Jim 2007. "Senators Warner and Webb announce recommendations for judgeships." <http://webb.senate.gov> (accessed January 1, 2008).
- Wolpert, D. H., and W. G. Macready. 1997. "No free lunch theorems for optimization." *IEEE Transactions on Evolutionary Computation* 1(1):67–82.
- Yiannakis, Diana Evans 1982. "House members' communication styles: Newsletter and press releases." *Journal of Politics* 44(4):1049–71.
- Zhong, Shi, and Joydeep Ghosh. 2003. "A unified framework for model-based clustering." *Journal of Machine Learning* 4(Nov.):1001–37.

# How to Analyze Political Attention with Minimal Assumptions and Costs

**Kevin M. Quinn** University of California, Berkeley

**Burt L. Monroe** The Pennsylvania State University

**Michael Colaresi** Michigan State University

**Michael H. Crespin** University of Georgia

**Dragomir R. Radev** University of Michigan

*Previous methods of analyzing the substance of political attention have had to make several restrictive assumptions or been prohibitively costly when applied to large-scale political texts. Here, we describe a topic model for legislative speech, a statistical learning model that uses word choices to infer topical categories covered in a set of speeches and to identify the topic of specific speeches. Our method estimates, rather than assumes, the substance of topics, the keywords that identify topics, and the hierarchical nesting of topics. We use the topic model to examine the agenda in the U.S. Senate from 1997 to 2004. Using a new database of over 118,000 speeches (70,000,000 words) from the Congressional Record, our model reveals speech topic categories that are both distinctive and meaningfully interrelated and a richer view of democratic agenda dynamics than had previously been possible.*

**W**hat are the subjects of political conflict and attention? How does the mix of topic attention change over time? How do we know? These questions are fundamental to much of political science, including studies of legislative representation (Lowi 1964; Mayhew 1974; Riker 1986), policy agenda change (Baumgartner, Green-Pedersen, and Jones 2006; Baumgartner and Jones 1993; Kingdon 1995), and issue evo-

lution (Carmines and Stimson 1989; Wolbrecht 2000). Conventional approaches to the problem of identifying and coding topic attention have used trained human coders to read documents. The careful and systematic use of human-coder techniques has helped to produce impressive data collections such as the Policy Agendas and Congressional Bills projects in American Politics (Adler and Wilkerson 2006; Jones, Wilkerson, and Baumgartner

---

Kevin M. Quinn is Professor of Law, University of California, Berkeley, 490 Simon #7200, Berkeley, CA 94720-7200 ([kquinn@law.berkeley.edu](mailto:kquinn@law.berkeley.edu)). Burt L. Monroe is Associate Professor of Political Science and Director of the Quantitative Social Science Initiative, The Pennsylvania State University, 230 Pond Lab, University Park, PA 16802-6200 ([burtmonroe@psu.edu](mailto:burtmonroe@psu.edu)). Michael Colaresi is Associate Professor of Political Science, Michigan State University, 303 South Kedzie Hall, East Lansing, MI 48824 ([colaresi@msu.edu](mailto:colaresi@msu.edu)). Michael H. Crespin is Assistant Professor of Political Science, University of Georgia, 407 Baldwin Hall, Athens, GA 30602 ([crespin@uga.edu](mailto:crespin@uga.edu)). Dragomir R. Radev is Associate Professor, School of Information and Department of Electrical Engineering and Computer Science, University of Michigan, 3310 EECS Building, 1301 Beal Avenue, Ann Arbor, MI 48109-2122 ([radev@umich.edu](mailto:radev@umich.edu)).

An earlier version of this article was presented to the Midwest Political Science Association and was awarded the 2006 Harold Gosnell Prize for Excellence in Political Methodology. We would like to thank Steven Abney, Scott Adler, Scott Ainsworth, Frank Baumgartner, Ken Bickers, David Blei, Jake Bowers, Janet Box-Steffensmeier, Patrick Brandt, Barry Burden, Suzie Linn, John Freeman, Ed Hovy, Will Howell, Simon Jackman, Brad Jones, Bryan Jones, Kris Kanthak, Gary King, Glen Krutz, Frances Lee, Bob Luskin, Chris Manning, Andrew Martin, Andrew McCallum, Iain McLean, Nate Monroe, Becky Morton, Stephen Purpura, Phil Schrodт, Gisela Sin, Betsy Sinclair, Michael Ward, John Wilkerson, Dan Wood, Chris Zorn, and seminar participants at UC Davis, Harvard University, the University of Michigan, the University of Pittsburgh, the University of Rochester, Stanford University, the University of Washington, and Washington University in St. Louis for their comments on earlier versions of the article. We would like to give special thanks to Cheryl Monroe for her contributions toward development of the Congressional corpus in specific and our data collection procedures in general. We would also like to thank Jacob Balazer (Michigan) and Tony Fader (Michigan) for research assistance. In addition, Quinn thanks the Center for Advanced Study in the Behavioral Sciences for its hospitality and support. This article is based upon work supported by the National Science Foundation under grants BCS 05-27513 and BCS 07-14688. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation. Supplementary materials, including web appendices and a replication archive with data and R package, can be found at <http://www.legislativespeech.org>.

*American Journal of Political Science*, Vol. 54, No. 1, January 2010, Pp. 209–228

©2010, Midwest Political Science Association

ISSN 0092-5853

209

n.d.) and the Comparative Manifesto Project in comparative politics (Budge et al. 2001; Klingemann et al. 2006). The impact and usefulness of these data sources to political science is difficult to overstate.<sup>1</sup> The great benefit of human-coder techniques is that the mapping of words in a text to a topic category is allowed to be highly complicated and contingent. The downside of human-coder techniques is that reliability can be a challenge, per-document costs are generally high, and it assumes that both the substance of topics and rules that govern tagging documents with a specific topic are known *a priori*.

Related tasks in political science have also been addressed using computer-checked dictionaries or, more recently, hybrid human/computer ("supervised learning") techniques. For example, event data coding in international relations has benefited enormously from the automated coding of news wire feeds using dictionaries created by the Kansas Event Data system (Gerner et al. 1994), and the Policy Agendas and Congressional Bills Projects have moved toward the use of supervised learning techniques to supplement human coding (Hillard, Purpura, and Wilkerson 2007, 2008). When automated approaches substitute computers for humans, the costs of coding are reduced and the reliability is increased (King and Lowe 2003). As with human coding, dictionary methods, and hybrid human/computer classification approaches, both assume that the substance of topics and the features that identify a particular topic are known *a priori*.

Here, we describe a statistical method to topic-code political texts over time that provides a reliable and replicable mapping of words into topics. However, unlike most extant approaches, our method estimates both the keywords that identify particular topics, as well as the division of topics from observed data, rather than assuming these features are known with certainty. Previously, if a researcher was interested in tracking topic attention over time within a set of documents, that researcher needed to bring a great deal of information into the analysis. The researcher first needed to define the substance, number, and subdivisions of each topic. Second, the researcher was required to codify a set of rules or keywords that would allow human coders or a computer to place documents into the researcher-created taxonomy of topics. In contrast, our statistical method of topic-coding text does not require a researcher to know the underlying taxonomy of categories with certainty. In-

<sup>1</sup>As outlined in the cited books and websites, each of these has inspired expansive research programs with books and papers too numerous to cite here.

stead, the division of topics and keywords that identify each topic are estimated from the text. Our statistical topic-coding method opens up the exciting possibility of tracking attention within lengthy political corpora that would be prohibitively expensive for human coders. The only additional input required from the investigator is the total number of categories into which texts should be grouped.

To illustrate the usefulness of our approach, we use our statistical model to topic-code the Congressional record for the 105th to the 108th U.S. Senate. The estimates provide (1) an ontology of topic categories and language choice and (2) a daily data series of attention to different topics in the U.S. Senate from 1997 to 2004. We believe this is the most extensive, temporally detailed map of legislative issue attention that has ever been systematically constructed. We evaluate the validity of our approach by examining (a) the extent to which there is common substantive meaning underlying the keywords *within* a topic, (b) the semantic relationships *across* topics, (c) the extent to which our daily measures of topic attention covary with roll calls and hearings on the topic of interest, (d) the relationships between exogenous events (such as 9/11 or the Iraq War) that are widely perceived to have shifted the focus of attention in particular ways, and (e) the usefulness of the produced data for testing hypotheses of substantive and theoretical interest.

## Categorizing Texts: Methods, Assumptions, and Costs

Each method for analyzing textual content imposes its own particular set of assumptions and, as a result, has particular advantages and weaknesses for any given question or set of texts. We focus our attention here on the basic problem of categorizing texts—placing texts into discrete target categories or bins.<sup>2</sup> Methods of text categorization vary along at least five dimensions: (1) whether they take the target categories as known or unknown, (2) whether the target categories have any known or unknown relationships with one another, (3) whether the relevant textual features (e.g., words, nouns, phrases, etc.) are known or unknown, (4) whether the mapping from features to categories is known or unknown, and (5) whether

<sup>2</sup>An equally interesting problem is placing texts, or their authors, in a continuous space, the problem addressed by such techniques as WORDSCORES (Laver, Benoit, and Garry 2003; Lowe 2008), WordFish (Slapin and Proksch 2008), and rhetorical ideal point estimation (Monroe and Maeda 2004; Monroe et al. 2007).

**TABLE 1 A Summary of Common Assumptions and Relative Costs Across Different Methods of Discrete Text Categorization**

A. Assumptions	Method				
	Reading	Human Coding	Dictionaries	Supervised Learning	Topic Model
Categories are known	No	Yes	Yes	Yes	No
Category nesting, if any, is known	No	Yes	Yes	Yes	No
Relevant text features are known	No	No	Yes	Yes	Yes
Mapping is known	No	No	Yes	No	No
Coding can be automated	No	No	Yes	Yes	Yes
<b>B. Costs</b>					
Preanalysis Costs					
Person-hours spent conceptualizing	Low	High	High	High	Low
Level of substantive knowledge	Moderate/High	High	High	High	Low
Analysis Costs					
Person hours spent per text	High	High	Low	Low	Low
Level of substantive knowledge	Moderate/High	Moderate	Low	Low	Low
Postanalysis Costs					
Person-hours spent interpreting	High	Low	Low	Low	Moderate
Level of substantive knowledge	High	High	High	High	High

the categorization process can be performed algorithmically by a machine. We are at pains, in particular, to describe how five ways of categorizing texts—reading, human coding, automated dictionaries, supervised learning, and the topic model we describe here—fill distinctive niches as tools for political science.

Each of these five methods comes with unique costs and benefits. We find it useful to think of these costs along two main dimensions: (1) the extent to which the method requires detailed substantive knowledge and (2) the length of time it would take a single person to complete the analysis for a fixed body of text. Each of these two types of costs can be incurred at three stages of the analysis: the preanalysis phase where issues of conceptualization and operationalization are dealt with (perhaps in one or more pilot studies), the analysis phase where the texts of interest are categorized, and the postanalysis phase where the results from the analysis phase are interpreted and assessed for reliability and validity. Tables 1A and 1B depict how five major methods of text categorization compare in terms of their underlying assumptions and costs, respectively. The cell entries in Table 1A represent the minimal assumptions required by each method.

In the most general sense, the fundamental “method” for inferring meaning from text is *reading*. For exam-

ple, one reader of a specific journal article might attempt to place that article into one of a set of substantive categories (e.g., legislative studies/agenda setting/methodology/text analysis), while another reader might categorize the text in terms of its relevance (cite/request more information/ignore). Not only might the relevant categories change by reader, but a given reader will create new categories as more information about the text becomes apparent.

For some target sets of categories, we could delineate specific features of the text that make particular categories more likely. We can imagine that words like *Congress* or *legislature* make it more likely that we place an article under “legislative studies,” that typesetting in *LATEX* or multiple equations makes it more likely that we place it under “methodology,” and so on. For other target concepts, the relevant features are more abstract. To place it in the “cite” bin, we might require that the text display features like importance and relevance. Different readers may disagree on the salient features and their presence or absence in any particular text. This is important for the promise of automation via algorithm. We all use search engines that are useful at helping us find articles that are topically relevant (Google Scholar, JSTOR) or influential (Social Science Citation Index), but we would be

more skeptical of an algorithm that attempted to tell us whether a given article should be cited in our own work or not.

As one might expect—since all automated methods require at least some human reading—the act of reading a text rests on fewer assumptions than other methods of text categorization. The number of topics is not necessarily fixed in advance, the relationships between categories are not assumed *a priori*, texts can be viewed holistically and placed in categories on a case-by-case basis, and there is no attempt to algorithmically specify the categorization process. This allows maximum flexibility. However, the flexibility comes with nontrivial costs, especially when one attempts to read large, politically relevant texts such as the British *Hansard* or the U.S. *Congressional Record*. More specifically, human reading of text requires moderate-to-high levels of substantive knowledge (the language of the text and some contextual knowledge are minimal but nontrivial requirements) and a great deal of time in person-hours per text.<sup>3</sup> Finally, condensing the information in a large text requires a great deal of thought, expertise, and good sense. Even in the best of situations, purely qualitative summaries of a text are often open to debate and highly contested.

*Human coding* (see, for instance, Ansolabehere, Snowberg, and Snyder 2003; Budge et al. 2001; Ho and Quinn 2008; Jones, Wilkerson, and Baumgartner n.d.; and Klingemann et al. 2006) is the standard methodology for content analysis, and for coding in general, in social science. For such manual coding, the target categories of interest are assumed to be known and fixed. Coders read units of text and attempt to assign one of a finite set of codes to each unit. If the target categories have any relationship to each other (e.g., nesting), it is assumed to be known. There is typically no requirement that the readers use any particular feature in identifying the target category and the exact mapping from texts to categories is assumed unknown and never made explicit. One can tell, through reliability checking, whether two independent coders reach the same conclusion, but one cannot tell how they reached it. Manual coding is most useful when there are abundant human resources available, the target concepts are clearly defined *a priori*, but the mapping from texts to categories is highly complex and unknown (“I know it when I see it”).

By using clearly defined, mutually exclusive, and exhaustive categories to structure the coding phase, human coding methods require less substantive knowledge than would be necessary in a deep reading of the texts. Nev-

ertheless, the texts do still need to be read by a human (typically a research assistant) who is a competent reader of the language used in the texts. Further, some moderate contextual knowledge is required during this phase so that texts are interpreted in the proper context. While human coding is less costly than deep reading during the analysis phase, it has higher initial costs. In particular, arriving at a workable categorization scheme typically requires expert subject-matter knowledge and substantial human time.

The first steps toward automation can be found in *dictionary-based coding*, which easily carries the most assumptions of all methods here. Examples include Gerner et al. (1994), Cary (1977), and Holsti, Brody, and North (1964). In dictionary-based coding, the analyst develops a list (a dictionary) of words and phrases that are likely to indicate membership in a particular category. A computer is used to tally up use of these dictionary entries in texts and determine the most likely category.<sup>4</sup> So, as with manual coding, target categories are known and fixed. Moreover, the relevant features—generally the words or phrases that comprise the dictionary lists—are known and fixed, as is the mapping from those features into the target categories. When these assumption are met, dictionary-based coding can be fast and efficient.

As with human coding, dictionary methods have very high startup costs. Building an appropriate dictionary is typically an application-specific task that requires a great deal of deep application-specific knowledge and (often-times) a fair amount of trial and error. That said, once a good dictionary is built, the analysis costs are as low or lower than any competing method. A large number of texts can be processed quickly and descriptive numerical summaries can be easily generated that make interpretation and validity assessment relatively straightforward.

A more recent approach to automation in this type of problem is *supervised learning* (Hillard, Purpura, and Wilkerson 2007, 2008; Kwon, Hovy, and Shulman 2007; Purpura and Hillard 2006). Hand coding is done to a subset of texts that will serve as training data and to another subset of texts that serve as evaluation data (sometimes called “test data”). Machine-learning algorithms are then used to attempt to infer the mapping from text features to hand-coded categories in the training set. Success is evaluated by applying the inferred mapping to the test data and calculating summaries of out-of-sample predictive accuracy. Gains of automation are then realized by application to the remaining texts that have not been hand coded. There are a wide variety of possible algorithms and the field is growing. Again, note that target categories are assumed to be known and fixed. Some set

<sup>3</sup>The *Congressional Record* currently contains over four billion words and produces another half million—about the length of *War and Peace*—a day.

<sup>4</sup>One of the important early dictionary systems is the General Enquirer (Stone et al. 1966).

of possibly relevant features must be identified, but the algorithm determines which of those are relevant and how they map into the target categories. Some algorithms restrict the mapping from text features to categories to take a parametric form while others are nonparametric.<sup>5</sup>

Since supervised learning methods require some human coding of documents to construct training and test sets, these methods have high startup costs that are roughly the same as human-coding methods. Where they fare much better than human-coding methods is in the processing of the bulk of the texts. Here, because the process is completely automated, a very large number of texts can be assigned to categories quite quickly.

In the same way that supervised learning attempts to use statistical techniques to automate the process of hand coding, our topic model attempts to automate the topic-categorization process of reading. The key assumption shared with reading, and not shared with hand coding, dictionary-based coding, or supervised learning, is that the target categories and their relationships with each other are unknown. The target categories—here, the topics that might be the subject of a particular legislative speech—are an object of inference. We assume that words are a relevant feature for revealing the topical content of a speech, and we assume that the mapping from words to topics takes a particular parametric form, described below. The topic model seeks to identify, rather than assume, the topical categories, the parameters that describe the mapping from words to topic, and the topical category for any given speech.

The topic-modeling approach used in this article has a very different cost structure than all methods mentioned so far. Whereas other methods typically require a large investment in the initial preanalysis stage (human coding, dictionary methods, supervised learning) and/or analysis stage (reading, human coding), our topic model requires very little time or substantive knowledge in these stages of the analysis. Where things are reversed is in the postanalysis phase where methods other than deep reading are *relatively* costless but where our topic model requires more time and effort (but no more substantive knowledge) than other methods. The nature of the costs incurred by the topic model become more apparent below.

## A Model for Dynamic Multitopic Speech

The data-generating process that motivates our model is the following. On each day that Congress is in session a

<sup>5</sup>In Table 1A, we code the assumptions for the least stringent supervised learning techniques.

legislator can make speeches. These speeches will be on one of a finite number  $K$  of topics. The probability that a randomly chosen speech from a particular day will be on a particular topic is assumed to vary smoothly over time. At a very coarse level, a speech can be thought of as a vector containing the frequencies of words in some vocabulary. These vectors of word frequencies can be stacked together in a matrix whose number of rows is equal to the number of words in the vocabulary and whose number of columns is equal to the number of speeches. This matrix is our outcome variable. Our goal is to use the information in this matrix to make inferences about the topic membership of individual speeches.<sup>6</sup>

We begin by laying out the necessary notation. Let  $t = 1, \dots, T$  index time (in days);  $d = 1, \dots, D$  index speech documents;  $k = 1, \dots, K$  index possible topics that a document can be on; and  $w = 1, \dots, W$  index words in the vocabulary. For reasons that will be clearer later, we also introduce the function  $s : \{1, \dots, D\} \rightarrow \{1, \dots, T\}$ .  $s(d)$  tells us the time period in which document  $d$  was put into the *Congressional Record*. In addition, let  $\Delta^N$  denote the  $N$ -dimensional simplex.

## The Sampling Density

The  $d$ th document  $\mathbf{y}_d$  is a  $W$ -vector of nonnegative integers. The  $w$ th element of  $\mathbf{y}_d$ , denoted  $y_{dw}$ , gives the number of times word  $w$  was used in document  $d$ . We condition on the total number  $n_d$  of words in document  $d$  and assume that if  $\mathbf{y}_d$  is from topic  $k$

$$\mathbf{y}_d \sim \text{Multinomial}(n_d, \boldsymbol{\theta}_k).$$

Here  $\boldsymbol{\theta}_k \in \Delta^{W-1}$  is the vector of multinomial probabilities with typical element  $\theta_{kw}$ . One can think of  $\boldsymbol{\theta}_k$  as serving as a “prototype speech” on topic  $k$  in the sense that it is the most likely word-usage profile within a speech on this topic. This model will thus allow one to think about all the speeches in a dataset as being a mixture of  $K$  prototypes plus random error. We note in passing that a Poisson data-generating process also gives rise to the same multinomial model conditional on  $n_d$ . For purposes of interpretation, we will at some points below make use of

<sup>6</sup>The model we describe below differs from the most similar topic models in the computational linguistics literature (Blei and Lafferty 2006; Blei, Ng, and Jordan 2003; Wang and McCallum 2006) in several particulars. Among these are the dynamic model, the estimation procedure, and, most notably, the nature of the mixture model. In other models, documents have a mixture of topical content. This is perhaps appropriate for complex documents, like scientific articles. In ours, documents have a single topic, but we are uncertain which topic. This is appropriate for political speeches. Ultimately, our assumption allows us to distinguish between, for example, a speech on defense policy that invokes oil, and a speech on energy policy that invokes Iraq.

the transformation

$$\boldsymbol{\beta}_k = \left( \left[ \log\left(\frac{\theta_{k1}}{\theta_{k1}}\right) - c \right], \left[ \log\left(\frac{\theta_{k2}}{\theta_{k1}}\right) - c \right], \dots, \left[ \log\left(\frac{\theta_{kW}}{\theta_{k1}}\right) - c \right] \right)'$$

where  $c = W^{-1} \sum_{w=1}^W \log\left(\frac{\theta_{kw}}{\theta_{k1}}\right)$ .

If we let  $\pi_{tk}$  denote the marginal probabilities that a randomly chosen document is generated from topic  $k$  in time period  $t$ , we can write the sampling density for all of the observed documents as

$$p(\mathbf{Y} | \boldsymbol{\pi}, \boldsymbol{\theta}) \propto \prod_{d=1}^D \sum_{k=1}^K \pi_{s(d)k} \prod_{w=1}^W \theta_{kw}^{y_{dw}}.$$

As will become apparent later, it will be useful to write this sampling density in terms of latent data  $\mathbf{z}_1, \dots, \mathbf{z}_D$ . Here  $\mathbf{z}_d$  is a  $K$ -vector with element  $z_{dk}$  equal to 1 if document  $d$  was generated from topic  $k$  and 0 otherwise. If we could observe  $\mathbf{z}_1, \dots, \mathbf{z}_D$  we could write the sampling density above as

$$p(\mathbf{Y}, \mathbf{Z} | \boldsymbol{\pi}, \boldsymbol{\theta}) \propto \prod_{d=1}^D \prod_{k=1}^K \left( \pi_{s(d)k} \prod_{w=1}^W \theta_{kw}^{y_{dw}} \right)^{z_{dk}}.$$

### The Prior Specification

To complete a Bayesian specification of this model we need to determine prior distributions for  $\boldsymbol{\theta}$  and  $\boldsymbol{\pi}$ . We assume a semiconjugate Dirichlet prior for  $\boldsymbol{\theta}$ . More specifically, we assume

$$\boldsymbol{\theta}_k \sim \text{Dirichlet}(\boldsymbol{\lambda}_k) \quad k = 1, \dots, K.$$

For the data analysis below we assume that  $\lambda_{kw} = 1.01$  for all  $k$  and  $w$ . This corresponds to a nearly flat prior over  $\boldsymbol{\theta}_k$ . This prior was chosen before looking at the data.

The prior for  $\boldsymbol{\pi}$  is more complicated. Let  $\boldsymbol{\pi}_t \in \Delta^{K-1}$  denote the vector of topic probabilities at time  $t$ . The model assumes that a priori

$$\mathbf{z}_d \sim \text{Multinomial}(1, \boldsymbol{\pi}_{s(d)}).$$

We reparameterize to work with the unconstrained

$$\boldsymbol{\omega}_t = \left( \log \left[ \frac{\pi_{t1}}{\pi_{tK}} \right], \dots, \log \left[ \frac{\pi_{t(K-1)}}{\pi_{tK}} \right] \right)'.$$

In order to capture dynamics in  $\boldsymbol{\pi}_t$  and to borrow strength from neighboring time periods, we assume that  $\boldsymbol{\omega}_t$  follows a Dynamic Linear Model (DLM; Cargnoni, Müller, and West 1997; West and Harrison 1997). Specifically,

$$\boldsymbol{\omega}_t = \mathbf{F}'_t \boldsymbol{\eta}_t + \boldsymbol{\epsilon}_t \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_t) \quad t = 1, \dots, T \quad (1)$$

$$\boldsymbol{\eta}_t = \mathbf{G}_t \boldsymbol{\eta}_{t-1} + \boldsymbol{\delta}_t \quad \boldsymbol{\delta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t) \quad t = 1, \dots, T \quad (2)$$

Here equation (1) acts as the observation equation and equation (2) acts as the evolution equation. We finish this prior off by assuming prior distributions for  $\mathbf{V}_t$ ,  $\mathbf{W}_t$ , and  $\boldsymbol{\eta}_0$ . Specifically, we assume  $\mathbf{W}_t = \mathbf{W}$  for all  $t$  and  $\mathbf{V}_t = \mathbf{V}$  for all  $t$  in which Congress was in session with  $\mathbf{V}$  and  $\mathbf{W}$  both diagonal and

$$V_{ii} \sim \text{InvGamma}(a_0/2, b_0/2) \quad \forall i$$

$$W_{ii} \sim \text{InvGamma}(c_0/2, d_0/2) \quad \forall i$$

We assume

$$\boldsymbol{\eta}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathbf{C}_0).$$

In what follows, we assume  $a_0 = 5$ ,  $b_0 = 5$ ,  $c_0 = 1$ ,  $d_0 = 1$ ,  $\mathbf{m}_0 = \mathbf{0}$ , and  $\mathbf{C}_0 = 25\mathbf{I}$ . For days in which Congress was not in session we assume that  $V_t = 10\mathbf{I}$ . We have found that this helps prevent oversmoothing. We note that our substantive results are not terribly sensitive to other, more diffuse, priors for  $V_{ii}$  and  $W_{ii}$ . In a web appendix we detail how models fit with  $a_0 = b_0 = c_0 = d_0 = 1$  and  $a_0 = c_0 = 1$ ,  $b_0 = d_0 = 10$  produce extremely similar results.

In what follows we specify  $\mathbf{F}_t$  and  $\mathbf{G}_t$  as a local linear trend for  $\boldsymbol{\omega}_t$ :

$$\mathbf{F}_t = \begin{pmatrix} \mathbf{I}_{K-1} \\ \mathbf{0}_{K-1} \end{pmatrix} \quad t = 1, \dots, T$$

$$\mathbf{G}_t = \begin{pmatrix} \mathbf{I}_{K-1} & \mathbf{I}_{K-1} \\ \mathbf{0}_{K-1} & \mathbf{I}_{K-1} \end{pmatrix} \quad t = 1, \dots, T.$$

While we adopt a fairly simple model for the dynamics in the Senate data, the DLM framework that we make use of is extremely general. Details of the Expectation Conditional Maximization (ECM) algorithm used to fit this model are provided in the web appendix. Model fitting takes between 20 minutes and three hours depending on the quality of the starting values and the speed of the computer. No specialized hardware is required.

Viewed as a clustering/classification procedure, the model above is designed for “unsupervised” clustering. At no point does the user pretag documents as belonging to certain topics. As we will demonstrate below in the context of Senate speech data, our model, despite not using user-supplied information about the nature of the topics, produces topic labelings that adhere closely to generally recognized issue areas. While perhaps the greatest strength of our method is the fact that it can be used without any manual coding of documents, it can also be easily adapted for use in semisupervised fashion by constraining some elements of  $\mathbf{Z}$  to be 0 and 1. It is also possible to use the model to classify documents that were not in the original dataset used to fit the model.

## Applying the Topic Model to U.S. Senate Speech, 1995–2004

We present here an analysis of speech in the U.S. Senate, as recorded in the *Congressional Record*, from 1995 to 2004 (the 105th to the 108th Congresses). In this section, we briefly describe how we process the textual data to serve as input for the topic model and then discuss the specification of the model for this analysis.

### Senate Speech Data

The textual data are drawn from the United States Congressional Speech Corpus<sup>7</sup> (Monroe et al. 2006) developed under the Dynamics of Political Rhetoric and Political Representation Project (<http://www.legislativespeech.org>). The original source of the data is the html files that comprise the electronic version of the (public domain) *United States Congressional Record*, served by the Library of Congress on its THOMAS system (Library of Congress n.d.) and generated by the Government Printing Office (United States Government Printing Office n.d.).

These html files correspond (nearly) to separately headed sections of the *Record*. We identify all utterances by an individual within any one of these sections, even if interrupted by other speakers, as a “speech” and it is these speeches that constitute the document set we model. For the eight-year period under study, there are 118,065 speeches ( $D$ ) so defined.

The speeches are processed to remove (most) punctuation and capitalization and then all words are *stemmed*.<sup>8</sup> There are over 150,000 unique stems in the vocabulary of the Senate over this eight-year period, most of which are unique or infrequent enough to contain little information. For the analysis we present here, we filter out all stems that appear in less than one-half of 1% of speeches, leaving a vocabulary of 3,807 ( $W$ ) stems for this analysis.

This produces a  $118,065 \times 3,807$  input matrix of stem counts, which serves as the input to the topic model. This matrix contains observations of just under 73 million words.<sup>9</sup>

<sup>7</sup>Corpus (plural *corpora*) is a linguistic term meaning a textual database.

<sup>8</sup>A word’s *stem* is its root, to which affixes can be added for inflection (*vote* to *voted*) or derivation (*vote* to *voter*). Stemming provides considerable efficiency gains, allowing us to leverage the shared topical meaning of words like *abort*, *aborts*, *aborted*, *aborting*, *abortion*, *abortions*, *abortionist*, and *abortionists* instead of treating the words as unrelated. An algorithm that attempts to reduce words to stems is a *stemmer*. We use the Porter Snowball II stemmer (for English), widely used in many natural language processing applications (Porter 1980, n.d.).

<sup>9</sup>Details of the process are provided in the web appendix.

### Model Output

The model contains millions of parameters and latent variables. We can focus on two subsets of these as defining the quantities of substantive interest, the  $\beta$ ’s and the  $z$ ’s.

The  $\beta$  matrix contains  $K \times W (\approx 160,000)$  parameters. Each element  $\beta_{kw}$  of this matrix describes the log-odds of word  $w$  being used to speak about topic  $k$ . If  $\beta_{kw} > \beta_{kw'}$  it is the case that word  $w$  is used more often on topic  $k$  than word  $w'$ . This is the source of the semantic content, the meaning, in our model. That is, we use this to learn what each topic is *about* and how topics are related to one another.  $\beta$  describes the intratopic data-generating process, so it can be used to generate new “speeches” (with words in random order) on any topic. It can also be used, in conjunction with the other model parameters, to classify other documents. This is useful either for sensitivity analysis, as noted below, or for connecting the documents from some other setting (newspaper articles, open-ended survey responses) to the topical frame defined by this model.

$Z$  is a  $D \times K$  matrix with typical element  $z_{dk}$ . Each of the approximately 5,000,000  $z_{dk}$  values is a 0/1 indicator of whether document  $d$  was generated from topic  $k$ . The model-fitting algorithm used in this article returns the expected value of  $Z$  which we label  $\hat{Z}$ . Because of the 0/1 nature of each  $z_{dk}$ , we can interpret  $\hat{z}_{dk}$  (the expected value of  $z_{dk}$ ) as the probability that document  $d$  was generated from topic  $k$ .

We find that approximately 94% of documents are more than 95% likely to be from a single topic. Thus, we lose very little information by treating the maximum  $z_{dk}$  in each row as an indicator of “the topic” into which speech  $d$  should be classified, reducing this to  $D$  (118,000) parameters of direct interest. Since we know when and by whom each speech was delivered, we can generate from this measures of attention (word count, speech count) to each topic at time scales as small as by day, and for aggregations of the speakers (parties, state delegations, etc.). It is also possible to treat  $\hat{z}_d$  as a vector of topic probabilities for document  $d$  and to then probabilistically assign documents to topics.

### Model Specification and Sensitivity Analysis

We fit numerous specifications of the model outlined in the third section to the 105th–108th Senate data. In particular, we allowed the number of topics  $K$  to vary from 3 to 60. For each specification of  $K$  we fit several models using different starting values. Mixture models, such as that used here, typically exhibit a likelihood surface that is multimodal. Since the ECM algorithm used to fit the

model is only guaranteed to converge to a local mode, it is typically a good idea to use several starting values in order to increase one's chances of finding the global optimum.

We applied several criteria to the selection of  $K$ , which must be large enough to generate interpretable categories that have not been overaggregated and small enough to be usable at all. Our primary criteria were substantive and conceptual. We set a goal of identifying topical categories that correspond roughly to the areas of governmental competence typically used to define distinct government departments/ministries or legislative committees, such as "Education," "Health," and "Defense." This is roughly comparable to the level of abstraction in the 19 major topic codes of the Policy Agendas Project, while being a bit more fine-grained than the 10 major categories in Rohde's roll-call dataset (Rohde 2004) and more coarse than the 56 categories in the Comparative Manifestos Project. Conceptually, for us, a genuine topic sustains discussion over time (otherwise it is something else, like a proposal, an issue, or an event) and across parties (otherwise it is something else, like a frame). With  $K$  very small, we find amorphous categories along the lines of "Domestic Politics," rather than "Education"; as  $K$  increases, we tend to get divisions into overly fine subcategories ("Elementary Education"), particular features ("Education Spending"), or specific time-bound debates ("No Child Left Behind"). Results matching our criteria, and similar to each other, occur at  $K$  in the neighborhood of 40–45. We present here results for the  $K = 42$  model with the highest maximized log posterior. A series of sensitivity analyses are available in the web appendix.

## Reliability, Validity, Interpretation, and Application

This is a measurement model. The evaluation of any measurement is generally based on its reliability (can it be repeated?) and validity (is it right?). Embedded within the complex notion of validity are interpretation (what does it mean?) and application (does it "work"?).

Complicating matters, we are here developing multiple measures simultaneously: the assignment of speeches to topics, the topic categories themselves, and derived measures of substantive concepts, like attention. Our model has one immediate reliability advantage relative to human and human-assisted supervised learning methods. The primary feature of such methods that can be assessed is the human-human or computer-human intercoder reliability in the assignment of documents to the given topic frame, and generally 70–90% (depending on index and application) is taken as a standard. Our

approach is 100% reliable, completely replicable, in this regard.

More important are notions of validity. There are several concepts of measurement validity that can be considered in any content analysis.<sup>10</sup> We focus here on the five basic types of *external* or *criterion-based* concepts of validity. First, the measures of the topics themselves and their relationships can be evaluated for *semantic validity* (the extent to which each category or document has a coherent meaning and the extent to which the categories are related to one another in a meaningful way). This speaks directly to how the  $\beta$  matrix can be *interpreted*. Then, the derived measures of attention can be evaluated for *convergent construct validity* (the extent to which the measure matches existing measures that it should match), *discriminant construct validity* (the extent to which the measure departs from existing measures where it should depart), *predictive validity* (the extent to which the measure corresponds correctly to external events), and *hypothesis validity* (the extent to which the measure can be used effectively to test substantive hypotheses). The last of these speaks directly to the issue of how the  $z$  matrix can be *applied*.

### Topic Interpretation and Intratopic Semantic Validity

Table 2 provides our substantive labels for each of the 42 clusters, as well as descriptive statistics on relative frequency in the entire dataset. We decided on these labels after examining  $\hat{\beta}_k$  and also reading a modest number of randomly chosen documents that were assigned a high probability of being on topic  $k$  for  $k = 1, \dots, K$ . This process also informs the semantic validity of each cluster. Krippendorff (2004) considers this the most relevant form of validity for evaluating a content analysis measure. We discuss these procedures in turn.

In order to get a sense of what words tended to distinguish documents on a given topic  $k$  from documents on other topics we examined both the magnitude of  $\hat{\beta}_{kw}$  for each word  $w$  as well as the weighted distance of  $\hat{\beta}_{kw}$  from the center of the  $\hat{\beta}$  vectors other than  $\hat{\beta}_{kw}$  (denoted  $\hat{\beta}_{-kw}$ ). The former provides a measure of how often word  $w$  was used in topic  $k$  documents relative to other words in topic  $k$  documents. A large positive value of  $\hat{\beta}_{kw}$  means that word  $w$  appeared quite often in topic  $k$  documents. The weighted distance of  $\hat{\beta}_{kw}$  from the center of the  $\hat{\beta}_{-kw}$ ,

<sup>10</sup>The most common is, of course, face validity. Face validity is inherently subjective, generally viewed as self-evident by authors and with practiced skepticism by readers. We believe the results from the model as applied to the *Congressional Record* (see below) demonstrate significant face validity. But, by definition, there are no external criteria one can bring to bear on the issue of face validity and thus we focus on several other types of validity.

**TABLE 2 Topic Labels and Descriptive Statistics for 42-Topic Model**

Topic Labels	% <sup>a</sup>	Clarifying Notes
1. Judicial Nominations	1.0/2.4	
2. Supreme Court / Constitutional	1.1/3.0	incl. impeachment, DOJ, marriage, flag-burning
3. Campaign Finance	0.9/2.4	
4. Abortion	0.5/1.1	
5. Law & Crime 1 [Violence/Drugs]	1.3/1.8	violence, drug trafficking, police, prison
6. Child Protection	0.9/2.6	tobacco, alcohol, drug abuse, school violence, abuse
7. Health 1 [Medical]	1.5/2.4	emph. disease, prevention, research, regulation
8. Social Welfare	2.0/2.8	
9. Education	1.8/4.6	
10. Armed Forces 1 [Manpower]	1.0/1.5	incl. veterans' issues
11. Armed Forces 2 [Infrastructure]	2.3/3.0	incl. bases and civil defense
12. Intelligence	1.4/3.9	incl. terrorism and homeland security
13. Law & Crime 2 [Federal]	1.8/2.7	incl. the FBI, immigration, white-collar crime
14. Environment 1 [Public Lands]	2.2/2.5	incl. water management, resources, Native Americans
15. Commercial Infrastructure	2.0/2.9	incl. transportation and telecom
16. Banking and Finance	1.1/3.1	incl. corporations, small business, torts, bankruptcy
17. Labor 1 [Workers, esp. Retirement]	1.0/1.5	emph. conditions and benefits, esp. pensions
18. Debt / Deficit / Social Security	1.7/4.6	
19. Labor 2 [Employment]	1.4/4.5	incl. jobs, wages, general state of the economy
20. Taxes	1.1/2.7	emph. individual taxation, incl. income and estate
21. Energy	1.4/3.3	incl. energy supply and prices, environmental effects
22. Environment 2 [Regulation]	1.1/2.8	incl. pollution, wildlife protection
23. Agriculture	1.2/2.5	
24. Foreign Trade	1.1/2.4	
25. Procedural 3 [Legislation 1]	2.0/2.8	
26. Procedural 4 [Legislation 2]	3.0/3.5	
27. Health 2 [Economics—Seniors]	1.0/2.6	incl. Medicare and prescription drug coverage
28. Health 3 [Economics—General]	0.8/2.3	incl. provision, access, costs
29. Defense [Use of Force]	1.4/3.7	incl. wars/interventions, Iraq, Bosnia, etc.
30. International Affairs [Diplomacy]	1.9/3.0	incl. human rights, organizations, China, Israel, etc.
31. International Affairs [Arms Control]	0.9/2.3	incl. treaties, nonproliferation, WMDs
32. Symbolic [Tribute—Living]	1.9/1.3	
33. Symbolic [Tribute—Constituent]	3.2/1.9	
34. Symbolic [Remembrance—Military]	2.3/1.9	incl. tributes to other public servants, WWII Memorial
35. Symbolic [Remembrance—Nonmilitary]	2.4/2.3	
36. Symbolic [Congratulations—Sports]	0.6/0.4	
37. Jesse Helms re: Debt	0.5/0.1	almost daily deficit/debt 'boxscore' speeches
38. Gordon Smith re: Hate Crime	0.4/0.1	almost daily speeches on hate crime
39. Procedural 1 [Housekeeping 1]	20.4/1.5	
40. Procedural 5 [Housekeeping 3]	15.5/1.0	
41. Procedural 6 [Housekeeping 4]	6.5/1.6	
42. Procedural 2 [Housekeeping 2]	2.4/0.8	

<sup>a</sup>Percentage of documents (left of slash) and percentage of word stems (right of slash).

which we operationalize as

$$r_{kw} = \frac{\hat{\beta}_{kw} - \text{median}_{j \neq k}(\hat{\beta}_{jw})}{\text{MAD}_{\ell \neq k}(\hat{\beta}_{\ell w})},$$

where MAD represents the median absolute deviation, provides a measure of how distinctive the usage of word  $w$  is on topic  $k$  documents compared to other documents. To take an example, the word *the* always has a very high

$\beta$  value, as it is very frequently used. However, it is used roughly similarly across all of the topics, so its value of  $r$  is generally quite close to 0. We combine these measures by ranking the elements of  $\hat{\beta}_k$  and  $r_k$  and adding the ranks for each word  $w$ . This combined index gives us one measure of how distinctive word  $w$  is for identifying documents on topic  $k$ . Table 3 provides the top keys for each topic.<sup>11</sup>

Inspection of these tables produced rough descriptive labels for all of the clusters. After arriving at these rough labels we went on to read a number of randomly chosen speech documents that were assigned to each cluster. In general we found that, with the exception of the procedural categories, the information in the keywords (Table 3, extended) did an excellent job describing the documents assigned to each (substantive) topic. However, by reading the documents we were able to discover some nuances that may not have been apparent in the tables of  $\beta$  values, and those are reflected in the topic labels and clarifying notes of Table 2.

In general, the clusters appear to be homogeneous and well defined. Our approach is particularly good at extracting the primary meaning of a speech, without being overwhelmed by secondary mentions of extraneous topics. For example, since 9/11, the term *terrorism* can appear in speeches on virtually any topic from education to environmental protection, a fact that undermines information retrieval through keyword search.<sup>12</sup> It is worth noting that this technique will extract information about the centroid of a cluster's meaning and lexical use. There will be speeches that do not fall comfortably into any category, but which are rare enough not to demand their own cluster.<sup>13</sup>

Reading some of the raw documents also revealed some additional meaning behind the clusters. For instance, two of the clusters with superficially uninformative keywords turn out to be composed exclusively of pro forma "hobby horse" statements by Senator Jesse Helms about the current level of national debt and by Senator Gordon Smith about the need for hate crime legislation.

The  $\beta$  parameters identify words that, if present, most distinguish a document of this topic from all others, for the

<sup>11</sup> Longer lists of keywords and index values are provided in the web appendix.

<sup>12</sup> The reader can confirm this by searching on the word *terrorism* in THOMAS.

<sup>13</sup> As noted above, about 94% of all documents have a better than 95% chance of being generated from a single topic; over 97% of documents have a better than 75% chance of being generated from a single topic, and over 99% have a better than 50% chance of being generated from a single topic. The bulk of the documents that were not clearly on a single topic have high probabilities of being from two or more "procedural" categories and are thus clearly on some procedural topic.

time period under study and for the Senate as a whole. Our approach does *not* demand that all legislators talk about all topics in the same way. To the contrary, there is typically both a common set of terms that identifies a topic at hand (as shown in Table 3) and a set of terms that identifies particular political (perhaps partisan) positions, points of view, frames, and so on, within that topic.

For example, Table 3 lists the top 10 keys for Judicial Nominations (*nomine, confirm, nomin, circuit, hear, court, judg, judici, case, vacanc*), all of which are politically neutral references to the topic that would be used by speakers of both parties. Within these topically defined speeches, we can define keys that are at any given time (here the 108th) the most Democratic (which include *republican, white, hous, presid, bush, administr, lifetim, appoint, pack, controversi, divis*) or the most Republican (which include *filibust, unfair, up-or-down, demand, vote, qualifi, experi, distinguish*), clearly reflecting the partisan split over Bush appointees and Democratic use of the filibuster to block them.<sup>14</sup>

## Relationships between Topics and Metatopic Semantic Validity

An important feature of the topic model, another sharp contrast with other approaches, is that the  $\beta$  matrix is an estimate of the relationship between *each* word in the vocabulary and *each* topical cluster. As a result, we can examine the semantic relationships within and across groups of topics. Given the more than 150,000 parameters in the  $\beta$  matrix, there are many such relationships one might investigate. Here we focus on how the topics relate to each other as subtopics of larger metatopics, how they aggregate. The coherent meaning of the metatopics we find is further evidence of the semantic validity of the topic model as applied to the *Congressional Record*. This type of validation has not been possible with other approaches to issue coding.

One approach to discovering relationships among the 42 topics is agglomerative clustering of the  $\beta$  vectors,  $\hat{\beta}_1, \dots, \hat{\beta}_{42}$ , by topic. Agglomerative clustering begins by assigning each of the 42 vectors to its own unique cluster. The two vectors that are closest to each other (by Euclidean distance) are then merged to form a new cluster. This process is repeated until all vectors are merged into

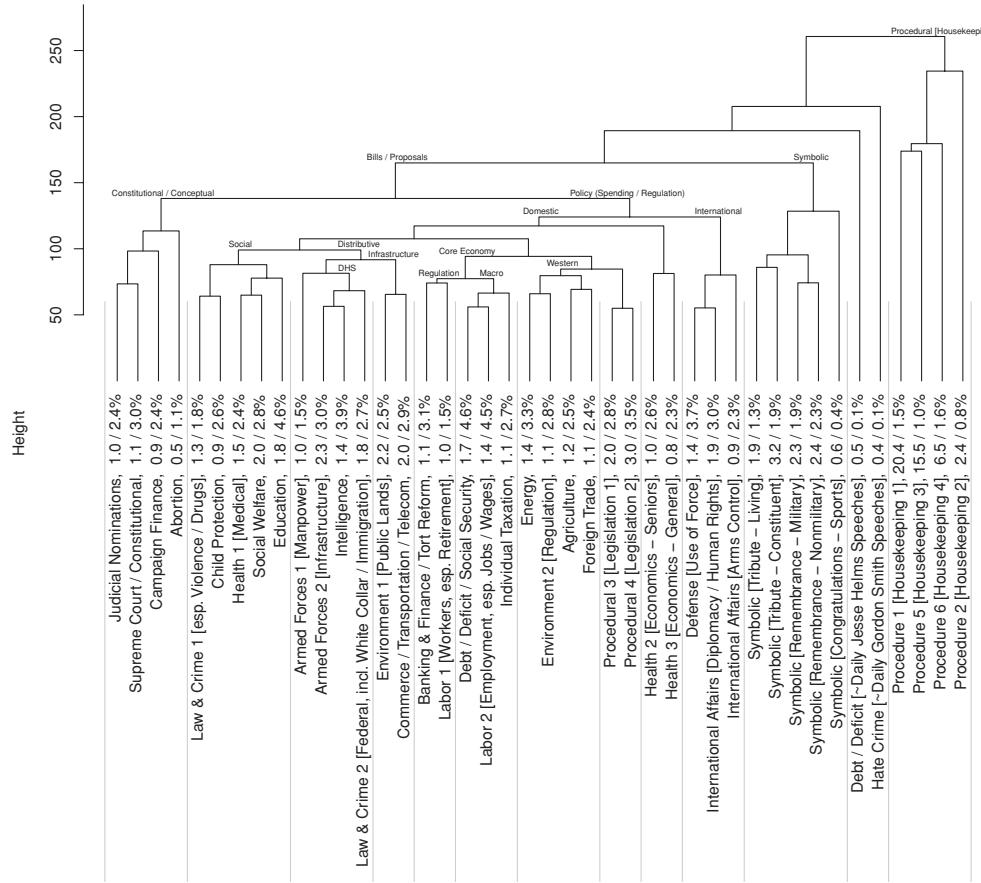
<sup>14</sup> These words all appear among the top keys using any of the variance-respecting feature selection techniques described in Monroe, Colaresi, and Quinn (2008). This includes the simplest method, roughly equivalent to ranking words by z-scores in a multinomial model of word choice with party as the only covariate, and a more computationally complex method based on regularization (a technique designed to reduce noise in such feature selection problems).

**TABLE 3 Topic Keywords for 42-Topic Model**

Topic (Short Label)	Keys
1. Judicial Nominations	<i>nomine, confirm, nomin, circuit, hear, court, judg, judici, case, vacanc</i>
2. Constitutional	<i>case, court, attornei, supreme, justic, nomin, judg, m, decis, constitut</i>
3. Campaign Finance	<i>campaign, candid, elect, monei, contribut, polit, soft, ad, parti, limit</i>
4. Abortion	<i>procedur, abort, babi, thi, life, doctor, human, ban, decis, or</i>
5. Crime 1 [Violent]	<i>enforc, act, crime, gun, law, victim, violenc, abus, prevent, juvenil</i>
6. Child Protection	<i>gun, tobacco, smoke, kid, show, firearm, crime, kill, law, school</i>
7. Health 1 [Medical]	<i>diseas, cancer, research, health, prevent, patient, treatment, devic, food</i>
8. Social Welfare	<i>care, health, act, home, hospit, support, children, educ, student, nurs</i>
9. Education	<i>school, teacher, educ, student, children, test, local, learn, district, class</i>
10. Military 1 [Manpower]	<i>veteran, va, forc, militari, care, reserv, serv, men, guard, member</i>
11. Military 2 [Infrastructure]	<i>appropri, defens, forc, report, request, confer, guard, depart, fund, project</i>
12. Intelligence	<i>intellig, homeland, commiss, depart, agenc, director, secur, base, defens</i>
13. Crime 2 [Federal]	<i>act, inform, enforc, record, law, court, section, crimin, internet, investig</i>
14. Environment 1 [Public Lands]	<i>land, water, park, act, river, natur, wildlif, area, conserv, forest</i>
15. Commercial Infrastructure	<i>small, busi, act, highhwai, transport, internet, loan, credit, local, capit</i>
16. Banking / Finance	<i>bankruptci, bank, credit, case, ir, compani, file, card, financi, lawyer</i>
17. Labor 1 [Workers]	<i>worker, social, retir, benefit, plan, act, employ, pension, small, employe</i>
18. Debt / Social Security	<i>social, year, cut, budget, debt, spend, balanc, deficit, over, trust</i>
19. Labor 2 [Employment]	<i>job, worker, pai, wage, economi, hour, compani, minimum, overtime</i>
20. Taxes	<i>tax, cut, incom, pai, estat, over, relief, marriag, than, penalti</i>
21. Energy	<i>energi, fuel, ga, oil, price, produce, electr, renew, natur, suppli</i>
22. Environment 2 [Regulation]	<i>wast, land, water, site, forest, nuclear, fire, mine, environment, road</i>
23. Agriculture	<i>farmer, price, produc, farm, crop, agricultur, disast, compact, food, market</i>
24. Trade	<i>trade, agreement, china, negoti, import, countri, worker, unit, world, free</i>
25. Procedural 3	<i>mr, consent, unanim, order, move, senat, ask, amend, presid, quorum</i>
26. Procedural 4	<i>leader, major, am, senat, move, issu, hope, week, done, to</i>
27. Health 2 [Seniors]	<i>senior, drug, prescript, medicar, coverag, benefit, plan, price, beneficiari</i>
28. Health 3 [Economics]	<i>patient, care, doctor, health, insur, medic, plan, coverag, decis, right</i>
29. Defense [Use of Force]	<i>iraq, forc, resolut, unit, saddam, troop, war, world, threat, hussein</i>
30. International [Diplomacy]	<i>unit, human, peac, nato, china, forc, intern, democraci, resolut, europ</i>
31. International [Arms]	<i>test, treati, weapon, russia, nuclear, defens, unit, missil, chemic</i>
32. Symbolic [Living]	<i>serv, hi, career, dedic, john, posit, honor, nomin, dure, miss</i>
33. Symbolic [Constituent]	<i>recogn, dedic, honor, serv, insert, contribut, celebr, congratul, career</i>
34. Symbolic [Military]	<i>honor, men, sacrific, memor, dedic, freedom, di, kill, serve, soldier</i>
35. Symbolic [Nonmilitary]	<i>great, hi, paul, john, alwai, reagan, him, serv, love</i>
36. Symbolic [Sports]	<i>team, game, plai, player, win, fan, basebal, congratul, record, victori</i>
37. J. Helms re: Debt	<i>hundr, at, four, three, ago, of, year, five, two, the</i>
38. G. Smith re: Hate Crime	<i>of, and, in, chang, by, to, a, act, with, the, hate</i>
39. Procedural 1	<i>order, without, the, from, object, recogn, so, second, call, clerk</i>
40. Procedural 5	<i>consent, unanim, the, of, mr, to, order, further, and, consider</i>
41. Procedural 6	<i>mr, consent, unanim, of, to, at, order, the, consider, follow</i>
42. Procedural 2	<i>of, mr, consent, unanim, and, at, meet, on, the, am</i>

*Notes:* For each topic, the top 10 (or so) key stems that best distinguish the topic from all others. Keywords have been sorted here by  $\text{rank}(\beta_{kw}) + \text{rank}(r_{kw})$ , as defined in the text. Lists of the top 40 keywords for each topic and related information are provided in the web appendix. Note the order of the topics is the same as in Table 2 but the topic names have been shortened.

FIGURE 1 Agglomerative Clustering of 42-Topic Model



*Notes:* Hierarchical agglomerative clustering of  $\hat{\beta}_1, \dots, \hat{\beta}_K$ . Clustering based on minimizing the maximum euclidean distance between cluster members. Each cluster is labeled with a topic name, followed by the percentage of documents and words, respectively, in that cluster.

a single cluster. The results of this process are displayed in the dendrogram of Figure 1.<sup>15</sup> Roughly speaking, the lower the height at which any two topics, or groupings of topics, are connected, the more similar are their word use patterns in Senate debate.<sup>16</sup>

Reading Figure 1 from the bottom up provides information about which clusters were merged first (those

<sup>15</sup>The order of topics given in Tables 2 and 3 is as determined here; the labels were determined prior to the agglomerative clustering.

<sup>16</sup>Further specifics, with code to reproduce this analysis and figure, are provided in the replication archive. Please note that the agglomerative clustering is not part of the model, but rather a tool (analogous to a regression table) for compactly displaying several important features of the estimates.

merged at the lowest height). We see that topics that share a penultimate node share a substantive or stylistic link. Some of these are obvious topical connections, such as between the two health economics clusters or between energy and environmental regulation. Some are more subtle. For example, the “Environment 1 [Public Lands]” category, which is dominated by issues related to management and conservation of public lands and water, and the “Commercial Infrastructure” category are related through the common reference to distributive public works spending. Both contain the words *project* and *area* in their top 25 keys, for example. The “Banking / Finance” category and the “Labor 1 [Workers]” category discuss different aspects of economic regulation and intervention,

the former with corporations and consumers, the latter with labor markets. Other connections are stylistic, rather than necessarily substantive. The symbolic categories, for example, all have *great*, *proud*, and *his* as keywords.

We can also read Figure 1 from the top down to get a sense of whether there are recognizable rhetorical metaclusters of topics. Reading from the top down, we see clear clusters separating the housekeeping procedural, hobby horse, and symbolic speech from the substantive policy areas. The more substantive branch then divides a cluster of conceptual and Constitutional issues from the more concrete policy areas that require Congress to appropriate funds, enact regulations, and so on. Within the concrete policy areas, we see further clear breakdowns into domestic and international policy. Domestic policy is further divided into clusters we can identify with social policy, public goods and infrastructure, economics, and “regional.” Note that what binds metaclusters is language. The language of the Constitutional grouping is abstract, ideological, and partisan. The social policy grouping is tied together by reference to societal problems, suffering, and need. The public goods / infrastructure grouping is tied together both by the language of projects and budgets, as well as that of state versus state particularism. The most interesting metacluster is the substantively odd “regional” grouping of energy, environment, agriculture, and trade. Exploration of the language used here shows that these are topics that divide rural and/or western senators from the rest—distributive politics at a different level of aggregation.

This approach has the potential to inform ongoing debates about how to characterize the underlying political structure of public policy. Whether such characterization efforts are of interest in and of themselves—we would argue they are—is not of as much relevance as the fact that they are necessary for understanding dimensions of political conflict (Clausen 1973; Poole and Rosenthal 1997), the dynamics of the political agenda (Baumgartner and Jones 2002; Lee 2006), the nature of political representation (Jones and Baumgartner 2005), or policy outcomes (Heitschusen and Young 2006; Katzenbach and Lapinski 2006; Lowi 1964). Katzenbach and Lapinski (2006) provide an eloquent defense of the exercise and a review of alternative approaches.

### **Speeches, Roll Calls, Hearings, and Construct Validity**

The construct validity of a measure is established via its relationships with other measures. A measure shows evidence of *convergent construct validity* if it correlates with other measures of the same construct. A measure shows

*discriminant construct validity* when it is uncorrelated with measures of dissimilar constructs (Weber 1990).

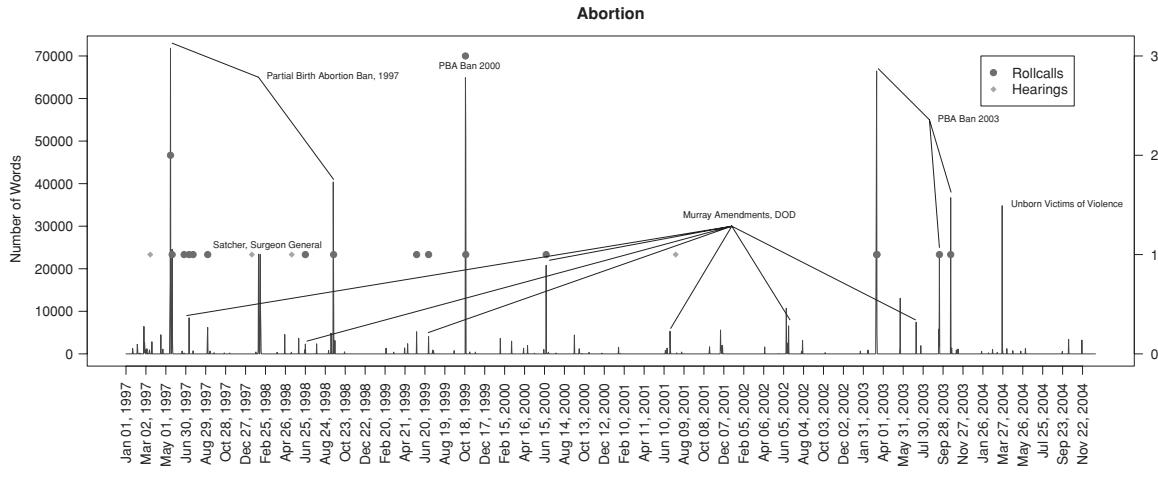
Construct validity has a double edge to it. If a new measure differs from an established one, it is generally viewed with skepticism. If a new measure captures what the old one did, it is probably unnecessary. In our case, the model produces measures we expect to converge with others in particular ways and to diverge in others. Consider a specific policy-oriented topic, like abortion. We expect that, typically, a roll call on abortion policy should be surrounded by a debate on the topic of abortion. This convergent relationship should appear in our measure of attention to abortion in speech and in indicators of roll calls on abortion policy.

Figure 2 displays the number of words given in speeches categorized by our model as “Abortion” over time. We also display the roll-call votes in which the official description contains the word *abortion*. We see the basic convergence expected, with number of roll calls and number of words correlated at +0.70. But note also that we expect divergence in the indicators as well. Attention is often given to abortion outside the context of an abortion policy vote, the abortion policy nature of a vote might be unclear from its description, and a particular roll call might receive very little debate attention.

Consider first, the spikes of debate attention that do not have accompanying roll-call votes. The first such spike is in February of 1998, when no vote was nominally on abortion. The occasion was the Senate confirmation of Clinton’s nominee for Surgeon General, David Satcher, and debate centered around Satcher’s positions on abortion. “Abortion” appears nowhere in the description of the vote. Hand-coding exercises would also not code the vote as abortion. For example, Rohde’s roll-call data (Rohde 2004) cover the House, but if extended to the Senate would clearly characterize the accompanying vote on February 10 as a confirmation vote, within a larger procedural category. None of Clausen (1973), Peltzman (1985), or Poole and Rosenthal (1997) extends forward to 1998, but all code previous Surgeon General confirmations at similar high levels of aggregation. For example, the C. Everett Koop confirmation vote, in 1981, is coded under the Clausen system as “Government Management,” under Peltzman as “Government Organization” (primarily) and “Domestic Social Policy” (secondarily), and under Poole and Rosenthal as “Public Health”.<sup>17</sup> Satcher would have been coded identically in each case. But it

<sup>17</sup>These codes are all listed in the D-NOMINATE dataset used for Poole and Rosenthal (1997) and archived on Poole’s website, <http://www.voteview.com>.

**FIGURE 2** The Number of Words Spoken on the ‘Abortion’ Topic Per Day



is clear from reading the transcript that the debate was about, and that attention was being paid to, abortion.

Another such spike is in March of 2004, when the Unborn Victims of Violence Act establishing penalties for violence against pregnant women was debated. The House vote on this identical bill is coded in the Rohde data under “Crime / Criminal Procedure” (Rohde 2004). Much of the debate attention, however, centered around the implications of the bill and possible amendments for abortion rights. In both cases, the spike in attention to abortion is real—captured by the speech measure and uncaptured by roll-call measures.

Similarly, the speech measure captures subtleties that the roll-call count does not. For example, on or around July 1 in every year from 1997 to 2003, Senator Murray offered an amendment to the Department of Defense Appropriations bill, attempting to restore access to abortions for overseas military personnel. The roll-call measure captures these through 2000, but misses them later. This is because the word *abortion* was removed from the description, replaced by a more opaque phrase: “to restore a previous policy regarding restrictions on use of Department of Defense medical facilities.” But with speech, these minor spikes in attention can be seen. Moreover, the speech measure captures when the amendment receives only cursory attention (a few hundred words in 1998) and when it is central to the discussion (2000, 2002).

Note also the relationship between speech and hearing data. The hearings data are sparse and generally examined at an annual level. At this level of aggregation, the two measures converge as expected—both show more

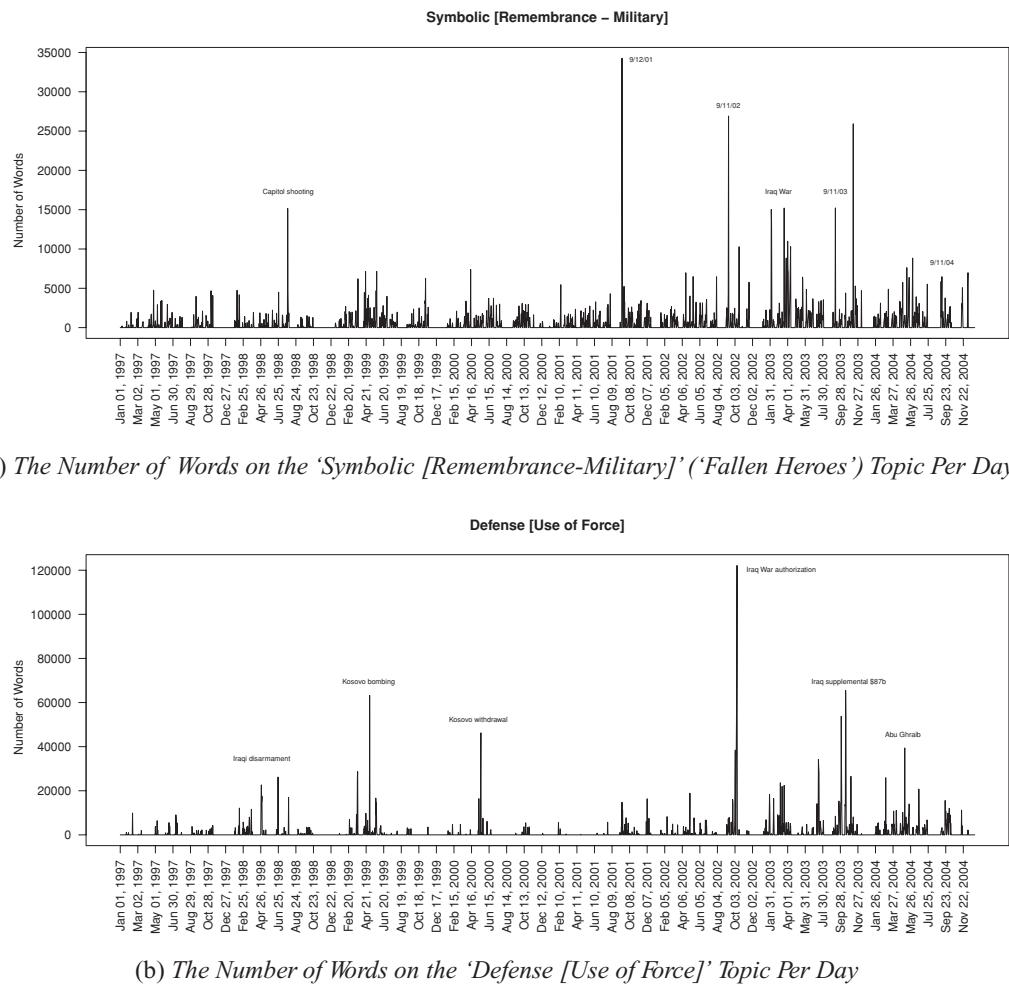
attention to abortion by the Senate during the Clinton presidency (1997–2000) than during the Bush presidency (2001–4). But at a daily level, the measures are clearly capturing different conceptual aspects of political attention. Higher cost hearings are more likely to capture attention that is well along toward being formulated as policy-relevant legislation. Speech is lower cost, so more dynamic and responsive at the daily level, more reflective of minority interests that may not work into policy, and potentially more ephemeral.

### Exogenous Events and Predictive Validity

*Predictive validity* refers to an expected correspondence between a measure and exogenous events uninvolved in the measurement process. The term is perhaps a confusing misnomer, as the direction of the relationship is not relevant. This means that the correspondence need not be a pure forecast of events from measures, but can be concurrent or postdictive, and causality can run from events to measures (Weber 1990). Of the limitless possibilities, it suffices to examine two of the most impactful political events in this time period: 9/11 and the Iraq War.

Figure 3a plots the number of words on the topic that corresponds to symbolic speech in support of the military and other public servants. Here we see a large increase in such symbolic speech immediately after 9/11 (the largest spike on the plot is exactly on September 12). There is another large spike on the first anniversary of 9/11 and then a number of consecutive days in March 2003 that feature moderate-to-large amounts of this type

**FIGURE 3 The Attention to ‘Symbolic [Remembrance—Military]’ and ‘Defense [Use of Force]’ Topics over Time**



of symbolic speech. This corresponds to the beginning of the Iraq War.

The number of words on the topic dealing with the use of military force is displayed in Figure 3b. The small intermittent upswings in 1998 track with discussions of Iraqi disarmament in the Senate. The bombing of Kosovo is represented as large spikes in spring 1999. Discussion within this topic increased again in May 2000 surrounding a vote to withdraw U.S. troops from the Kosovo peacekeeping operation. Post 9/11, the Afghanistan invasion brings a small wave of military discussion, while the largest spike in the graph (in October 2002) occurred during the debate to authorize military action in Iraq. This was followed, as one would expect,

by other rounds of discussion in fall 2003 concerning the emergency supplemental appropriations bill for Iraq and Afghanistan, and in the spring of 2004 surrounding events related to the increasing violence in Iraq, the Abu Ghraib scandal, and the John Negroponte confirmation.

### Hypothesis Validity and Application to the Study of Floor Participation

Hypothesis validity—the usefulness of a measure for the evaluation of theoretical and substantive hypotheses of interest—is ultimately the most important sort of validity. In this section we offer one example of the sort of

analysis to which attention measures can be applied directly. We return to a discussion of further applications in the concluding discussion.

One direct use of these data is to investigate floor participation itself to answer questions about Congressional institutions, the electoral connection, and policy representation. Prior empirical work has had severe data limitations, depending on low frequency events (e.g., floor amendments; Sinclair 1989; Smith 1989), very small samples (e.g., six bills; Hall 1996), or moderately sized, but expensive, samples (e.g., 2,204 speeches manually coded to three categories; Hill and Hurley 2002). Our data increase this leverage dramatically and cheaply.

Figure 4 summarizes the results from 50-count models (negative binomial) of the speech counts on all non-procedural topics and selected metatopical aggregations, for the 106th Senate, for all 98 senators who served the full session. Selected hypotheses, discussed below, are represented by shaded backgrounds.<sup>18</sup>

Congressional behavior is of core relevance to questions about the existence and possible decline of “norms” of committee deference, specialization, and apprenticeship (Hall 1996; Matthews 1960; Rohde, Ornstein, and Peabody 1985; Shepsle and Weingast 1987; Sinclair 1989; Smith 1989). As noted by Hall, this is a difficult empirical question as the primary leverage has come from floor amendment behavior, a relatively rare occurrence (1996, 180–81). Figure 4 shows that committee membership, but not necessarily service as chair or ranking member, continues to have a substantial impact on the tendency to participate in debate across policy topics. The apprenticeship norm, as indicated by a negative impact of freshman status, also seems to be present in more technical policy areas, but notably not in common electoral issues like abortion or the size of government. Examination of the data over time could further inform the question of decline (Rohde, Ornstein, and Peabody 1985; Sinclair 1989; Smith 1989) and, with the cross-topic variation provided here, the role of expertise costs (Hall 1996) versus norms (Matthews 1960) in both deference and apprenticeship.

Since at least Mayhew, Congressional scholars have also been interested in how career considerations affect the electoral connection (Fenno 1978, 1996; Hill and Hurley 2002; Maltzman and Sigelman 1996; Mayhew 1974). The sixth and seventh rows of Figure 4 identify two career cycle effects in the electoral connection and symbolic/empathy speech. A senator approaching election is more likely to give speeches in the symbolic (“I am proud

<sup>18</sup>These are graphical tables (Gelman, Pasarica, and Dodhia 2002; Kastellec and Leoni 2007). Alternative specifications, for both standardized and unstandardized coefficients, and for equivalent models of word count, all show substantively similar results.

to be one of you”) and social (“I care about you”) categories than is one whose next election is further in the future. Conversely, senators who subsequently retired gave many fewer such speeches, adding further evidence to the literature on *participatory shirking* (Poole and Rosenthal 1997; Rothenberg and Sanders 2000).

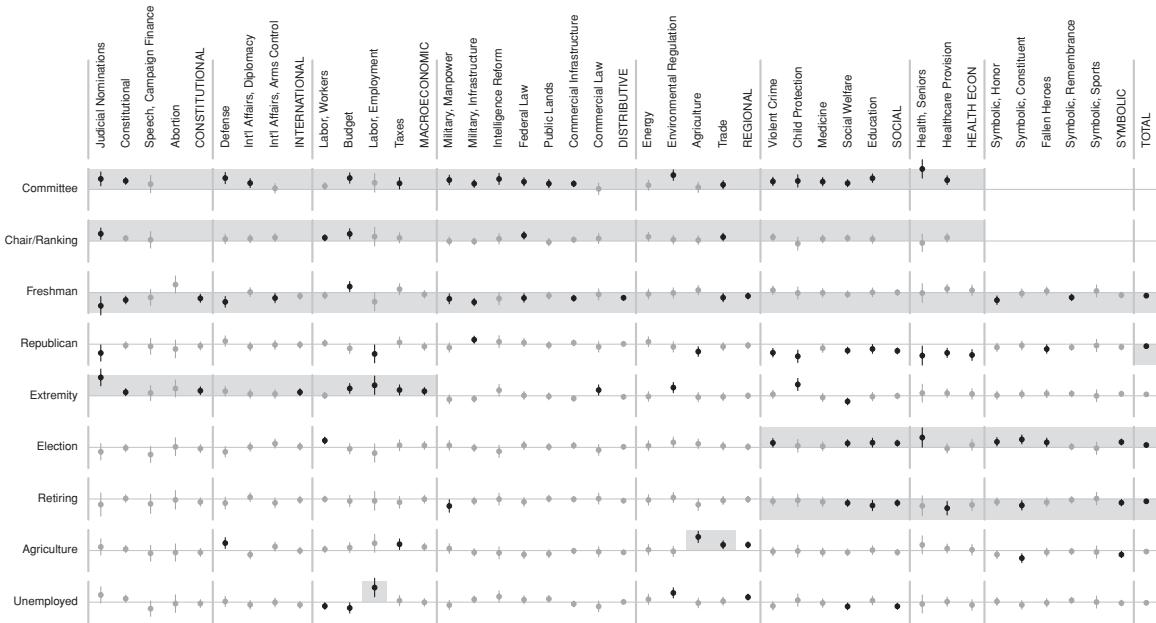
The last two rows of Figure 4 provide evidence of two (arbitrary) examples of policy representation, unemployment and agriculture. This reflects the notion of representation as congruence between constituency and representation, a subject of considerable scholarly attention (Anscombe, Snyder, and Stewart 2001 is a prominent example, in a literature that traces at least to Miller and Stokes 1963). Previous studies of congruence have generally been limited, on the legislator side, to measures of position based on elite surveys or roll calls. Jones and Baumgartner (2005) examine the year-by-year congruence of relative attention to topic (via hearings) with aggregate (not constituency-level) demand measured by Gallup “most important problem” data.

The party and ideology results (rows four and five) also contain interesting insights for our broader interests in how speech can inform our understanding of the landscape of political competition. Democrats are more likely to speak on social issues and more likely to speak in general. Given that Democrats were in the minority in the 106th Senate, this does lend some support to the assertion that speech is better than more constrained legislative behaviors at revealing thwarted minority party preferences and strategies.

Extremity (absolute DW-NOMINATE score) is associated with more speeches on constitutional, international, and economics topics, but not generally on social issues or geographically driven topics. This could be taken as evidence that the former set of topics is of greater interest to ideological extremists. Or—our view—it could be taken as evidence that these are the topics that *define* the current content of the primary dimension captured by roll-call-based ideal point estimation procedures. The lack of association between “extremism” and attention to other topics is suggestive that those other topics define higher dimensions of the political space.

## Discussion

In this article we have presented a method for inferring the relative amount of legislative attention paid to various topics at a daily level of aggregation. Unlike other commonly used methods, our method has minimal startup costs, allows the user to infer category labels (as well as the mapping from text features to categories), and can be

**FIGURE 4** Speech Count Models, 106th Senate

*Notes:* Each column represents a negative binomial model of speeches delivered on a given topic, or group of topics, in the 106th Senate, with one observation per senator who served the entire two years (98 in total). Each row of the table represents a covariate: “Committee” (binary indicating whether the senator is on a topic-relevant committee); “Chair/Ranking” (binary indicating the senator is the chair or ranking member of a topic-relevant committee); “Freshman” (binary); “Republican” (binary); “Extremity” (absolute value of Dimension 1 Poole-Rosenthal DW-NOMINATE scores); “Agriculture” (log of state agricultural income per capita, 1997); “Election” (dummy, up for election in next cycle); “Retiring” (dummy, retired before next election); “Unemployment” (state unemployment rate, 1999). Plotted are standardized betas and 95% confidence intervals, darker where this interval excludes zero. Shaded areas represent hypotheses discussed in the text.

applied to very large corpora in reasonable time. While other methods have one or more of these features, no other general method possesses all of these desirable properties.

While our method has several advantages over other common approaches to content analysis, it is not without its own unique costs. In particular, the topic model discussed in this article requires more user input *after* the initial quantitative analysis is completed. Since no substantive information is built directly into the model, the user must spend more time interpreting and validating the results *ex post*.

This article presents several ways that such interpretation and validation can be performed. Specifically, we demonstrate how (a) keywords can be constructed and their substantive content assessed, (b) agglomerative clustering can be used to investigate the semantic relationships *across* topics, (c) construct validity of our daily measures of topic attention can be evaluated by looking at their covariation with roll calls and hearings on the topic

of interest, and (d) predictive validity of our measures can be assessed by examining their relationships with exogenous events (such as 9/11 or the Iraq War) that are widely perceived to have shifted the focus of attention in particular ways. In each case, we find strong support for the validity of our measures.

While our method is useful, it will not (and should not) replace other methods. Instead, our data and method supplement and extend prior understandings of the political agenda in ways that have been to date prohibitively expensive or near impossible. Our method is particularly attractive when used as an exploratory tool applied to very large corpora. Here it quickly allows new insights to emerge about topic attention measured at very fine temporal intervals (in our example days). In some applications this will be enough; in others more detailed (and expensive) confirmatory analysis will be in order.

There are many potential applications beyond those we have given here for such measures of attention as

this. The dynamic richness of our data allows topic-specific examination of policy-agenda dynamics, and questions of incrementalism or punctuated equilibrium (Baumgartner and Jones 1993). The dynamic richness also allows us to move beyond static notions of congruence into dynamic notions of responsiveness, illuminating the topics and conditions under which legislators lead or follow public opinion (Jacobs and Shapiro 2000; Stimson, MacKuen, and Erikson 1995).

Moving another step, there are many possible indirect applications of the topic model. Once speeches are separated by topic, we can examine the substantive content—the values and frames—that underlie partisan and ideological competition. We can, for example, track in detail the dynamics by which issues and frames are adopted by parties, absorbed into existing ideologies, or disrupt the nature of party competition (Carmines and Stimson 1989; Monroe, Colaresi, and Quinn 2008; Poole and Rosenthal 1997; Riker 1986).

Further, once we know the content of party competition, we can evaluate the positioning of individual legislators. That is, as hinted above, the topic model is a valuable first step toward using speech to estimate ideal points from legislative speech. This allows dynamically rich, topic-by-topic ideal point estimation, and insights into the content and dimensionality of the underlying political landscape (Lowe 2007; Monroe and Maeda 2004; Monroe et al. 2007).

Perhaps most exciting, our method travels beyond English and beyond the Congressional setting, where conventional methods and measures can be prohibitively expensive or difficult to apply. We hope this might provide an important new window into the nature of democratic politics.

## References

- Adler, E. Scott, and John Wilkerson. 2006. "Congressional Bills Project." Technical report, University of Washington, Seattle. NSF 00880066 and 00880061.
- Ansolabehere, Stephen, Erik C. Snowberg, and James M. Snyder. 2003. "Statistical Bias in Newspaper Reporting: The Case of Campaign Finance." MIT Department of Political Science Working Paper.
- Ansolabehere, Stephen, James M. Snyder, and Charles Stewart. 2001. "Candidate Positioning in U.S. House Elections." *American Journal of Political Science* 45(1): 136–59.
- Baumgartner, Frank R., Christoffer Green-Pedersen, and Bryan D. Jones. 2006. "Comparative Studies of Policy Agendas." *Journal of European Public Policy* 13(7): 959–74.
- Baumgartner, Frank R., and Bryan D. Jones. 1993. *Agendas and Instability in American Politics*. Chicago: University of Chicago Press.
- Baumgartner, Frank R., and Bryan D. Jones, eds. 2002. *Policy Dynamics*. Chicago: University of Chicago Press.
- Blei, David M., and John D. Lafferty. 2006. "Dynamic Topic Models." 23rd International Conference on Machine Learning, Pittsburgh, PA.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.
- Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tannenbaum. 2001. *Mapping Policy Preferences: Parties, Electors and Governments, 1945–1998*. Oxford: Oxford University Press.
- Cargnoni, Claudia, Peter Müller, and Mike West. 1997. "Bayesian Forecasting of Multinomial Time Series Through Conditional Gaussian Dynamic Models." *Journal of the American Statistical Association* 92(438): 640–47.
- Carmines, Edward G., and James A. Stimson. 1989. *Issue Evolution: Race and the Transformation of American Politics*. Princeton, NJ: Princeton University Press.
- Cary, Charles D. 1977. "A Technique of Computer Content Analysis of Transliterated Russian Language Textual Materials: A Research Note." *American Political Science Review* 71(1): 245–51.
- Clausen, Aage R. 1973. *How Congressmen Decide: A Policy Focus*. New York: St. Martin's Press.
- Fenno, Richard F. 1978. *Home Style: House Members in Their Districts*. Boston: Little, Brown.
- Fenno, Richard F. 1996. *Senators on the Campaign Trail*. Norman: University of Oklahoma Press.
- Gelman, Andrew, Cristian Pasarica, and Rahul Dodhia. 2002. "Let's Practice What We Preach: Turning Tables into Graphs." *The American Statistician* 56(2): 121–30.
- Gerner, Deborah J., Philip A. Schrodt, Ronald A. Francisco, and Judith L. Weddle. 1994. "Machine Coding of Event Data Using Regional and International Sources." *International Studies Quarterly* 38(1): 91–119.
- Hall, Richard L. 1996. *Participation in Congress*. New Haven, CT: Yale University Press.
- Heitschusen, Valerie, and Garry Young. 2006. "Macropolitics and Changes in the U.S. Code: Testing Competing Theories of Policy Production, 1874–1946." In *The Macropolitics of Congress*, ed. E. Scott Adler and John S. Lapinski. Princeton, NJ: Princeton University Press, 129–50.
- Hill, Kim Quaile, and Patricia A. Hurley. 2002. "Symbolic Speeches in the U.S. Senate and Their Representational Implications." *Journal of Politics* 64(1): 219–31.
- Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2007. "An Active Learning Framework for Classifying Political Text." Presented at the annual meeting of the Midwest Political Science Association.
- Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2008. "Computer Assisted Topic Classification for Mixed Methods Social Science Research." *Journal of Information Technology and Politics* 4(4): 31–46.
- Ho, Daniel E., and Kevin M. Quinn. 2008. "Measuring Explicit Political Positions of Media." Harvard Department of Government Working Paper.
- Holsti, Ole R., Richard A. Brody, and Robert C. North. 1964. "Measuring Affect and Action in International Reaction

- Models: Empirical Materials from the 1962 Cuban Crisis." *Journal of Peace Research* 1(3/4): 170–90.
- Jacobs, Lawrence R., and Robert Y. Shapiro. 2000. *Politicians Don't Pander: Political Manipulation and the Loss of Democratic Responsiveness*. Chicago: University of Chicago Press.
- Jones, Bryan D., and Frank R. Baumgartner. 2005. *The Politics of Attention: How Government Prioritizes Problems*. Chicago: University of Chicago Press.
- Jones, Bryan D., John Wilkerson, and Frank R. Baumgartner. n.d. "The Policy Agendas Project." <http://www.policyagendas.org>.
- Kastellec, Jonathan P., and Eduardo L. Leoni. 2007. "Using Graphs Instead of Tables in Political Science." *Perspectives on Politics* 5(4): 755–71.
- Katzenbach, Ira, and John S. Lapinski. 2006. "The Substance of Representation: Studying Policy Content and Legislative Behavior." In *The Macropolitics of Congress*, ed. E. Scott Adler and John S. Lapinski. Princeton, NJ: Princeton University Press, 96–126.
- King, Gary, and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57(3): 617–42.
- Kingdon, John W. 1995. *Agendas, Alternatives, and Public Policies*. Boston: Little, Brown.
- Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge, and Michael McDonald. 2006. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990–2003*. Oxford: Oxford University Press.
- Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*. 2nd ed. New York: Sage.
- Kwon, Namhee, Eduard Hovy, and Stuart Shulman. 2007. "Identifying and Classifying Subjective Claims." Eighth National Conference on Digital Government Research, Digital Government Research Center.
- Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97: 311–31.
- Lee, Frances. 2006. "Agenda Content and Senate Party Polarization, 1981–2004." Presented at the annual meeting of the Midwest Political Science Association.
- Library of Congress. n.d. "THOMAS." <http://thomas.loc.gov>.
- Lowe, William. 2007. "Factors, Ideal Points, and Words: Connecting Legislators' Preferences to Legislative Speech." Measures of Legislators' Policy Preferences and the Dimensionality of Policy Spaces, Washington University, St. Louis.
- Lowe, William. 2008. "Understanding Wordscores." *Political Analysis* 16(4): 356–71.
- Lowi, Theodore J. 1964. "American Business, Public Policy, Case-Studies, and Political Theory." *World Politics* 16: 677–715.
- Maltzman, Forrest, and Lee Sigelman. 1996. "The Politics of Talk: Unconstrained Floor Time in the U.S. House of Representatives." *Journal of Politics* 58(3): 819–30.
- Matthews, Donald. 1960. *U.S. Senators and Their World*. Chapel Hill: University of North Carolina Press.
- Mayhew, David R. 1974. *Congress: The Electoral Connection*. New Haven, CT: Yale University Press.
- Miller, Warren E., and Donald Stokes. 1963. "Constituency Influence in Congress." *American Political Science Review* 57: 45–56.
- Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Partisan Conflict." *Political Analysis* 16(4): 372–403.
- Monroe, Burt L., and Ko Maeda. 2004. "Rhetorical Ideal Point Estimation: Mapping Legislative Speech." Society for Political Methodology, Stanford University.
- Monroe, Burt L., Cheryl L. Monroe, Kevin M. Quinn, Dragomir Radev, Michael H. Crespin, Michael P. Colaresi, Jacob Bazar, and Steven P. Abney. 2006. "United States Congressional Speech Corpus." <http://www.legislativespeech.org>.
- Monroe, Burt L., Kevin M. Quinn, Michael P. Colaresi, and Ko Maeda. 2007. "Estimating Legislator Positions from Speech." Measures of Legislators' Policy Preferences and the Dimensionality of Policy Spaces, Washington University, St. Louis.
- Peltzman, Sam. 1985. "An Economic Interpretation of the History of Congressional Voting in the Twentieth Century." *American Economic Review* 75: 656–75.
- Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll-Call Voting*. Oxford: Oxford University Press.
- Porter, Martin F. 1980. "An Algorithm for Suffix Stripping." *Program* 14(3): 130–37.
- Porter, Martin F. n.d. <http://snowball.tartarus.org/algorithms/english/stemmer.html>.
- Purpura, Stephen, and Dustin Hillard. 2006. "Automated Classification of Congressional Legislation." Technical report, John F. Kennedy School of Government.
- Riker, William H. 1986. *The Art of Political Manipulation*. New Haven, CT: Yale University Press.
- Rohde, David W. 2004. "Roll-Call Voting Data for the United States House of Representatives, 1953–2004." Technical report, Political Institutions and Public Choice Program, Michigan State University.
- Rohde, David, Norman Ornstein, and Robert Peabody. 1985. "Political Change and Legislative Norms in the U.S. Senate, 1957–74." In *Studies of Congress*, ed. Glenn Parker. Washington, DC: Congressional Quarterly Press, 147–88.
- Rothenberg, Lawrence S., and Mitchell S. Sanders. 2000. "Severing the Electoral Connection: Shirking in the Contemporary Congress." *American Journal of Political Science* 44: 316–25.
- Shepsle, Kenneth A., and Barry R. Weingast. 1987. "Why Are Committees Powerful?" *American Political Science Review* 81: 929–45.
- Sinclair, Barbara. 1989. *The Transformation of the U.S. Senate*. Baltimore: Johns Hopkins University Press.
- Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Positions from Texts." *American Journal of Political Science* 52(3): 705–22.

- Smith, Steven S. 1989. *Call to Order: Floor Politics in the House and Senate*. Washington, DC: Brookings Institution.
- Stimson, James A., Michael B. MacKuen, and Robert S. Erikson. 1995. "Dynamic Representation." *American Political Science Review* 89(3): 543–65.
- Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Enquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press.
- United States Government Printing Office. n.d. "The Congressional Record, GPO Access." <http://www.gpoaccess.gov/crecord>.
- Wang, Xuerui, and Andrew McCallum. 2006. "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends." 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia.
- Weber, Robert Phillip. 1990. *Basic Content Analysis*. New York: Sage.
- West, Mike, and Jeff Harrison. 1997. *Bayesian Forecasting and Dynamic Models*. New York: Springer.
- Wolbrecht, Christina. 2000. *The Politics of Women's Rights: Parties, Positions, and Change*. Princeton, NJ: Princeton University Press.

# Structural Topic Models for Open-Ended Survey Responses

**Margaret E. Roberts** University of California, San Diego  
**Brandon M. Stewart** Harvard University  
**Dustin Tingley** Harvard University  
**Christopher Lucas** Harvard University  
**Jetson Leder-Luis** California Institute of Technology  
**Shana Kushner Gadarian** Syracuse University  
**Bethany Albertson** University of Texas at Austin  
**David G. Rand** Yale University

*Collection and especially analysis of open-ended survey responses are relatively rare in the discipline and when conducted are almost exclusively done through human coding. We present an alternative, semiautomated approach, the structural topic model (STM) (Roberts, Stewart, and Airoldi 2013; Roberts et al. 2013), that draws on recent developments in machine learning based analysis of textual data. A crucial contribution of the method is that it incorporates information about the document, such as the author's gender, political affiliation, and treatment assignment (if an experimental study). This article focuses on how the STM is helpful for survey researchers and experimentalists. The STM makes analyzing open-ended responses easier, more revealing, and capable of being used to estimate treatment effects. We illustrate these innovations with analysis of text from surveys and experiments.*

Despite broad use of surveys and survey experiments within political science, the vast majority of analysis deals with responses to options along a scale or from preestablished categories. Yet, in most areas of life, individuals communicate either by writing or by speaking, a fact reflected in earlier debates about

open-and closed-ended survey questions. Collection and especially analysis of open-ended data are relatively rare in the discipline and when conducted are almost exclusively done through human coding. We present an alternative, semiautomated approach, the structural topic model (STM) (Roberts, Stewart, and Airoldi 2013; Roberts et al.

---

Margaret E. Roberts is Assistant Professor, Department of Political Science, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093 (molly.e.roberts@gmail.com). Brandon M. Stewart is a PhD Student, Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St., Cambridge, MA 02138 (bstewart@fas.harvard.edu). Dustin Tingley is Paul Sack Associate Professor of Political Economy, Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St., Cambridge, MA 02138 (dtingley@gov.harvard.edu). Christopher Lucas is a PhD Candidate, Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St., Cambridge, MA 02138 (clucas@fas.harvard.edu). Jetson Leder-Luis is an Undergraduate, Division of the Humanities and Social Sciences, California Institute of Technology, Caltech MSC 462, Pasadena, CA 91126 (jetson@caltech.edu). Shana Kushner Gadarian is Assistant Professor, Department of Political Science, Maxwell School of Citizenship and Public Affairs, Syracuse University, 100 Eggers Hall, Syracuse, NY 13244 (sgadaria@maxwell.syr.edu). Bethany Albertson is Assistant Professor, Department of Government, University of Texas at Austin, Mailcode A1800, Austin, TX 78712 (balberts@austin.utexas.edu). David G. Rand is Assistant Professor, Departments of Psychology and Economics, Yale University, 2 Hillhouse Rd, New Haven, CT 06511 (david.rand@yale.edu). Questions or comments can be sent to corresponding author at dtingley@gov.harvard.edu.

Our thanks to the Caltech SURF program, IQSS's Program on Text Analysis, and Dustin Tingley's dean support for supporting Jetson's initial participation during the summer of 2012. Brandon Stewart gratefully acknowledges funding from a National Science Foundation Graduate Research Fellowship. Alex Storer helped get computers to do their job. We thank the following for helpful comments and suggestions: Neal Beck, Justin Grimmer, Jennifer Jerit, Luke Keele, Gary King, Mik Laver, Rose McDermott, Helen Milner, Rich Nielsen, Brendan O'Connor, Mike Tomz, and participants in the Harvard Political Economy and Applied Statistics Workshops, UT Austin Government Department IR Seminar, Visions in Methodology 2013, and Stanford Methods Seminar. Replication files are available in the AJPS Data Archive on Dataverse (<http://dvn.iq.harvard.edu/dvn/dv/ajps>). The supplementary appendix is available at <http://scholar.harvard.edu/les/dtingley/les/ajpsappendix.pdf>.

*American Journal of Political Science*, Vol. 58, No. 4, October 2014, Pp. 1064–1082

©2014, Midwest Political Science Association

DOI: 10.1111/ajps.12103

1064

2013), that draws on recent developments in machine learning based analysis of textual data. A crucial contribution of the method is that it incorporates information about the document, such as the author's gender, political affiliation, and treatment assignment (if an experimental study). Elsewhere, we demonstrate its usefulness for analyzing other sources of text of interest across political science (Lucas et al. 2013). This article focuses on how the STM is helpful for survey researchers and experimentalists. The STM makes analyzing open-ended responses easier, more revealing, and capable of being used to estimate treatment effects. We illustrate these innovations with several experiments and an analysis of open-ended data in the American National Election Study (ANES).

In practice, we believe that many survey researchers and experimentalists avoid open-ended response data because they are costly to analyze in a systematic way. There are also debates about the desirability of using open- and closed-ended response formats. We provide relatively low-cost solutions that occupy a middle ground in these debates and innovate in two ways. First, we show how survey researchers and experimentalists can efficiently analyze open-ended data alongside a variety of common closed-ended data, such as a subject's party preferences or assignment to an experimental condition. Second, we provide a suite of tools that enable preprocessing of textual data, model selection, and visualization. We also discuss best practices and tools for human intervention in what otherwise is an unsupervised learning model, such as how a researcher could implement pre-analysis plans, as well as a discussion of limitations to the unsupervised learning model<sup>1</sup> at the foundation of our research strategy.

We proceed by first laying out the advantages and limitations of incorporating open-ended responses into research designs. Next, we present our estimation strategy and quantities of interest, as well as contrast our approach to existing methodologies. Having set up our research strategy, we analyze open-ended data from a survey experiment on immigration preferences and a laboratory experiment on public goods provision. We also analyze the "most important problem" data from the ANES. In each example, we showcase both our methodology for including covariates as well as the software tools we make available to researchers. Finally, we conclude with a discussion of future research possibilities.<sup>2</sup>

<sup>1</sup>As opposed to supervised models that require a hand-coded training set; see Grimmer and Stewart (2013) for details.

<sup>2</sup>All methods in this article are implemented in the R package *stm*, available at [www.structuraltopicmodel.com](http://www.structuraltopicmodel.com). The supplemental ap-

## Why Open-Ended Responses?

There was a point at which research on survey methodology actively debated whether questions should be open or closed in form (Geer 1991; Krosnick 1999; Lazarsfeld 1944). Today, the majority of survey analyses are composed predominately of closed-ended questions, and open-ended questions are rarely analyzed. This is despite the fact that prominent scholars writing on the topic identified advantages with each methodology (Krosnick 1999; Lazarsfeld 1944).

There are advantages and disadvantages to both closed- and open-ended data. One view of open-ended responses is that they provide a direct view into a respondent's own thinking. For example, RePass (1971, 391) argues that open-ended questions query attitudes that "are on the respondent's mind at the time of the interview," attitudes that were presumably salient before the question and remain so afterward. Similarly Iyengar (1996, 64) notes that open-ended questions have the advantage of "nonreactivity." That is, unlike closed-ended questions, "open-ended questions do not cue respondents to think of particular causes or treatments."<sup>3</sup>

A major concern about open-ended questions is that open-ended questions chiefly require that subjects "articulate a response, not their underlying attitudes" (Geer 1988, 365). Furthermore, nonresponses to open-ended questions may stem from ineloquence rather than indifference; subjects may not respond to open-ended questions because they lack the necessary rhetorical device (Geer 1988). A related concern is that open-ended questions may give respondents too little of a frame of reference in order to form a coherent response (Schuman 1966).

Open-ended responses have traditionally been considered more difficult to analyze than their closed counterparts (Schuman and Presser 1996), as human coding is almost always used. The use of human coders typically involves several steps. First, the researcher needs to define the dimensions on which open-ended data will be coded by humans and generate examples in order to guide the coders. This is typically guided by the researcher's own prior theoretical expectations and potentially reading of some examples. Next, human coders are unleashed on the data and numerical estimates for each document

pendix includes estimation details, a comparison to alternative models, a range of simulation studies, and additional tools for applied users.

<sup>3</sup>On this point, Kelley (1983, 10) notes that the opinions of the American electorate are so wide ranging that any closed list is bound to omit good opinions.

compared across coders (Artstein and Poesio 2008; Lombard, Snyder-Duch, and Bracken 2006).

Our view is that while such pragmatic concerns are reasonable, they ought not be our ultimate consideration, and instead what is crucial is whether open-ended questions give real insights (Geer 1991, 360). Rarely have survey researchers/experimentalists used automated text analysis procedures, and when they have, covariate information, either in the form of randomized treatment conditions or pretreatment covariates (e.g., gender or political ideology), is not used in the textual analysis (Simon and Xenos 2004). Researchers still might have good reason to use human coders, but we believe adoption of our methods at a minimum will assist them in using human coders more effectively.

## Our Contributions

The model below has a number of advantages over only using human coders. First, it allows the researcher to *discover* topics from the data, rather than assume them. These topics may or may not correspond to a researcher's theoretical expectations. When they do correspond, researchers can leverage the wide variety of quantities of interest that the STM generates. When they do not correspond, researchers may consider revising their theoretical model for future work or retain their model and turn to standard human coding procedures.

Second, it allows analysts to do this while studying how the prevalence and content of topics change with information that is particular to each respondent (e.g., whether the respondent received the treatment or background demographic data). We argue our model can fruitfully be used at either an exploratory stage prior to using human coders or as part of making credible inferences about the effect of treatments/frames/covariates on the content of open-ended responses. Thus, our approach can serve a variety of purposes. The next sections demonstrate the usefulness of text analysis tools for analyzing open-ended responses.

## Statistical Models of Text

The core innovation of the article is to bridge survey and experimental techniques, which include randomization of frames or encouragements to adopt a particular emotional status or way of looking at political issues, with new techniques in text analysis. Our approach also allows the analyst to incorporate covariates (e.g. attributes

of the respondent, treatment condition), with a model of the topics that are inferred directly from the written text. For experimental applications, this enables us to calculate treatment effects and uncertainty estimates on open-ended textual data. We believe that we are the first to do so in a way that builds in the structural information about the experiment, though we share similar motivations with Simon and Xenos (2004) and Hopkins (2010). In this section, we outline the notation and core ideas for statistical topic models; then we overview the STM, including quantities of interest, and conclude by discussing extensive material available in the supplemental appendix.

## A Heuristic Understanding of Statistical Topic Models

Statistical topic models allow for rich latent topics to be automatically inferred from text. Topic models are often referred to as "unsupervised" methods because they *infer* rather than *assume* the content of the topics under study, and they have been used across a variety of fields (Blei, Ng, and Jordan 2003; Grimmer 2010; Quinn et al. 2010; Wang and Blei 2011). We emphasize that this is conceptually different from "supervised" methods where the analyst defines the topics *ex ante*, usually by hand-coding a set of documents into preestablished categories (e.g., Laver, Benoit, and Garry 2003).

Within the class of unsupervised statistical topic models, topics are defined as distributions over a vocabulary of words that represent semantically interpretable "themes." Topic models come in two varieties: *single-membership* models and *mixed-membership* models. Previous work in political science has focused on single-membership models which have emphasized document meta-data (Grimmer 2010; Quinn et al. 2010, see also Grimmer and Stewart 2013 for a general review). In mixed-membership models, the most notable of which is latent Dirichlet allocation (LDA; Blei 2012; Blei, Ng, and Jordan 2003), a document is represented as a mixture of topics, with each word within a given document belonging to exactly one topic; thus, each document can be represented as a vector of proportions that denote what fraction of the words belong to each topic. In single-membership models, each document is restricted to only one topic, so all words within it are generated from the same distribution. We focus on mixed-membership models, highlighting the comparison to single-membership alternatives in the appendix.

In mixed-membership models, each document (indexed by  $d$ ) is assumed to be generated as follows. First, a

distribution over topics ( $\theta_d$ ) is drawn from a global prior distribution. Then, for each word in the document (indexed by  $n$ ), we draw a topic for that word from a multinomial distribution based on its distribution over topics ( $z_{d,n} \sim \text{Mult}(\theta_d)$ ). Conditional on the topic selected, the observed word  $w_{d,n}$  is drawn from a distribution over the vocabulary  $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$  where  $\beta_{k,v}$  is the probability of drawing the  $v$ -th word in the vocabulary for topic  $k$ . So, for example, our article (the one you are reading), which is just one article among all journal articles ever written, might be represented as a mixture over three topics that we might describe as survey analysis, text analysis, and experiments. Each of these topics is actually a distribution over words with high-frequency words associated with that topic (e.g., the experiment's topic might have "experiment, treatment, control, effect" as high-probability words). LDA, the model described above, is completed by assuming a Dirichlet prior for the topic proportions such that  $\theta_d \sim \text{Dirichlet}(\alpha)$ .<sup>4</sup>

The expressive power of statistical topics models to discover topics comes at a price. The resulting posterior distributions have many local modes, meaning that different initializations can produce different solutions. This can arise even in simple mixture models in very low dimensions (Anandkumar et al. 2012; Buot and Richards, 2006; Sontag and Roy 2009). Later in this section, we present a framework for model evaluation focused on semantic interpretability as well as robustness checks.

## Structural Topic Model

The STM innovates on the models just described by allowing for the inclusion of covariates of interest into the prior distributions for document-topic proportions and topic-word distributions. The result is a model where each open-ended response is a mixture of topics. Rather than assume that topical prevalence (i.e., the frequency with which a topic is discussed) and topical content (i.e., the words used to discuss a topic) are constant across all participants, the analyst can incorporate covariates over which we might expect to see variance.

We explain the core concept of the model here (complete details in the appendix. As in LDA, each document arises as a mixture over  $K$  topics. In the STM, topic

<sup>4</sup>Estimation for LDA in Blei, Ng, and Jordan (2003) proceeds by variational expectation-maximization (EM), where the local variables  $\theta_d$ ,  $\vec{z}_d$  are estimated for each document in the E-step, followed by maximization of global parameters  $\alpha$ ,  $\beta_{1:K}$ . Variational EM uses a tractable factorized approximation to the posterior. See Grimmer (2011).

proportions ( $\theta$ ) can be correlated, and the prevalence of those topics can be influenced by some set of covariates  $X$  through a standard regression model with covariates  $\theta \sim \text{LogisticNormal}(X\gamma, \Sigma)$ . For each word ( $w$ ) in the response, a topic ( $z$ ) is drawn from the response-specific distribution, and conditional on that topic, a word is chosen from a multinomial distribution over words parameterized by  $\beta$ , which is formed by deviations from the baseline word frequencies ( $m$ ) in log space ( $\beta_k \propto \exp(m + \kappa_k)$ ). This distribution can include a second set of covariates  $U$  (allowing, for example, Democrats to use the word "estate" more frequently than Republicans while discussing taxation). We discuss the difference between the two sets of covariates in more detail in the next subsection.

Thus, there are three critical differences in the STM as compared to the LDA model described above: (1) topics can be correlated; (2) each document has its own prior distribution over topics, defined by covariate  $X$  rather than sharing a global mean; and (3) word use within a topic can vary by covariate  $U$ . These additional covariates provide a way of "structuring" the prior distributions in the topic model, injecting valuable information into the inference procedure.<sup>5</sup>

The STM provides fast, transparent, replicable analyses that require few a priori assumptions about the texts under study. Yet it is a computer-assisted method, and the researcher is still a vital part of understanding the texts, as we describe in the examples section. The analyst's interpretive efforts are guided by the model and the texts themselves. But as we show, the STM can relieve the analyst of the burden of trying to develop a categorization scheme from scratch (Grimmer and King 2011) and perform the often mundane work of associating the documents with those categories.

## Estimating Quantities of Interest

A central advantage to our framework for open-ended survey response analysis is the variety of interpretable quantities of interest beyond what is available from LDA. In all topic models, the analyst estimates for each document the proportion of words attributable to each topic, providing a measure of topic *prevalence*. The model also calculates the words most likely to be generated by each

<sup>5</sup>We estimate the model using semi-collapsed variational EM. In the E-step, we solve for the joint optimum of the document's topic proportions ( $\theta$ ) and the token-level assignments ( $z$ ). Then in the M-step, we infer the global parameters  $\kappa$ ,  $\gamma$ ,  $\Sigma$ , which control the priors on topical prevalence and content. The STM prior is not conjugate to the likelihood and thus does not enjoy some of the theoretical guarantees associated with mean-field variational inference in the conjugate exponential family.

topic, which provides a measure of topical *content*. However, in standard LDA, the document collection is assumed to be unstructured; that is, each document is assumed to arise from the same data-generating process irrespective of additional information the analyst might possess. By contrast, our framework is designed to incorporate additional information about the document or its author into the estimation process. This allows us to measure systematic changes in topical prevalence and topical content over the conditions in our experiment, as measured by the  $X$  covariates for prevalence and the  $U$  covariates for content. Thus, for example, we can easily obtain measures of how our treatment condition affects both how often a topic is discussed (prevalence) and the language used to discuss the topic (content). Using our variational approximation to the posterior distribution, we can propagate our uncertainty in the estimation of the topic proportions through our analysis.<sup>6</sup>

The inference on the STM quantities of interest is best understood by reference to the familiar regression framework. For example, consider topical prevalence; if we observed the topics for each survey response, we could generate a regression where the topic is the outcome variable, and the treatment condition or other respondent controls (e.g., gender, income, party affiliation), along with any interactions, are the explanatory variables. This regression would give us insight into whether our treatment condition caused respondents to spend a larger portion of their written response discussing a particular topic. In our framework for analysis, we conduct this same regression, while simultaneously estimating the topics. This framework builds on recent work in political science on single-membership models, specifically Quinn et al. (2010) and Grimmer (2010), which allow topical prevalence to vary over time and author, respectively. Our model extends this framework by allowing topical prevalence to vary with *any* user-specified covariate. We also extend the framework to topical content. Word use within a particular topic comes from a regression, in this case a multinomial logistic regression, where the treatment condition and other covariates can change the rate of use for individual words within a topic.

In addition to these corpus-level changes, we also get an estimate of the proportion of words in each survey response attributable to a particular topic. Thus, we can retrieve the same types of quantities that would arise from human coding without the need to construct a coding scheme in advance. These document-level parameters

<sup>6</sup>We include uncertainty by integrating over the approximate posterior using the method of composition. See the appendix for more details.

can be used to construct useful summaries such as most representative documents for each topic, most representative documents for each treatment condition, or variation in topic use across other covariates not in the model.

We can also use the model to summarize the semantic meaning of a topic. Generally, these summaries are the highest probability words within a topic; however, this tends to prioritize words that have high frequency overall but may not be semantically interesting. Following the insights of Bischof and Airoldi (2012), who demonstrate the value of exclusivity in summary words for topics, we label topics using simplified frequency-exclusivity (FREX) scoring (Roberts, Stewart, and Airoldi 2013; Roberts et al. 2013). This summarizes words with the harmonic mean of the probability of appearance under a topic and the exclusivity to that topic. These words provide more semantically intuitive representations of topics.

In Figure 1, we list some of the quantities of interest with a simple interpretation. These quantities can be combined to create more complex aggregates, but we expect these summaries will suffice for most applications.

## Model Specification and Selection

Researchers must make important model specification and selection decisions. We briefly discuss the choice of covariates and the number of topics. We discuss theoretical implications of model specification choices, quantitative metrics, and methods for semiautomated model evaluation and selection.<sup>7</sup>

**Choices in Model Specification.** In the STM framework, the researcher has the option to choose covariates to incorporate in the model. These covariates inform either the topic prevalence or the topical content latent variables with observed information about the respondent. The analyst will want to include a covariate in the topical prevalence portion of the model ( $X$ ) when she believes that the observed covariate will affect *how much* the respondent is to discuss a particular topic. The analyst also has the option to include a covariate in the topical content portion of the model ( $U$ ) when she believes that the observed covariate will affect the words which a respondent uses to discuss a particular topic. These two sets of covariates can overlap, suggesting that the topic proportion and the way the topic is discussed change with

<sup>7</sup>We use standard text preprocessing conventions, such as stemming (Manning, Raghavan, and Schütze 2008). The appendix provides complete details along with software to help users manage and preprocess their collections of texts.

---

**FIGURE 1 Quantities of Interest from STM**


---

1. QOI: Topical Prevalence Covariate Effects
    - Level of Analysis: Corpus
    - Part of the Model:  $\theta, \gamma, X$
    - Description: Degree of association between a document covariate  $X$  and the average proportion of a document discussing each topic.
    - Example Finding: Subjects receiving the treatment on average devote twice as many words to Topic 2 as control subjects.
  
  2. QOI: Topical Content Covariate Effects
    - Level of Analysis: Corpus
    - Part of the Model:  $\kappa, U$
    - Description: Degree of association between a document covariate  $U$  and the rate of word use within a particular topic.
    - Example Finding: Subjects receiving the treatment are twice as likely to use the word “worry” when writing on the immigration topic as control subjects.
  
  3. QOI: Document-Topic Proportions
    - Level of Analysis: Document
    - Part of the Model:  $\theta$
    - Description: Proportion of words in a given document about each topic.
    - Example Use: Can be used to identify the documents that devote the highest or lowest proportion of words to a particular topic. Those with the highest proportion of words are often called “exemplar” documents and can be used to validate that the topic has the meaning the analyst assigns to it.
  
  4. QOI: Topic-Word Proportions
    - Level of Analysis: Corpus
    - Part of the Model:  $\kappa, \beta$
    - Description: Probability of observing each word in the vocabulary under a given topic. Alternatively, the analyst can use the FREX scoring method described above.
    - Example Use: The top 10 most probable words under a given topic are often used as a summary of the topic’s content and help inform the user-generated label.
- 

particular covariate values. The STM includes shrinkage priors or regularization, which draws the covariate effects toward zero. An analyst concerned about overfitting to the covariates can increase the degree of regularization.

The analyst must also choose the number of topics. There is no “right” answer to this choice. Varying the number of topics varies the level of granularity of the view into the data. Therefore, the choice will be dependent both on the nature of the documents under study and the goals of the analysis. While some corpora like academic journal articles might be analyzed with 50–100 topics (Blei 2012) due to the wide variety in their content, survey responses to focused questions may only consider a few topics. The appropriateness of particular levels of aggregation will vary with the research question.

**Model Selection Methods.** It would be useful if all of these choices could be evaluated using a simple diagnostic. It is tempting to compute an approximation to the marginal likelihood and calculate a model selection statistic, but we echo previous studies in emphasizing that this maximizes model fit and not substantive interpretation (Chang et al. 2009). Instead, we advocate quantitative evaluations of properties of the topic-word distributions. Specifically, we argue that a semantically interpretable topic has two qualities: (1) it is *cohesive* in the sense that high-probability words for the topic tend to co-occur within documents, and (2) it is *exclusive* in the sense that the top words for that topic are unlikely to appear within top words of other topics.

These two qualities are closely related to Gerring’s (2001) “consistency” and “differentiation” criteria for

concepts in empirical social science.<sup>8</sup> Semantic cohesion has previously been studied by Mimno et al. (2011) who develop a criterion based on co-occurrence of top topic words and show that it corresponds with human evaluation by subject matter experts.<sup>9</sup> While semantic coherence is a useful criterion, it only addresses whether a topic is internally consistent; it does not, for example, penalize topics that are alike. From the standpoint of social science inference, we want to be sure both that we are evaluating a well-defined concept and that our measure captures all incidence of the concept in the survey responses.

For this, we turn to the *exclusivity* of topic words, drawing on previous work on exclusivity and diversity in topic models (Bischof and Airoldi 2012; Eisenstein, Ahmed, and Xing 2011; Zou and Adams 2012). If words with high probability under topic  $i$  have low probabilities under other topics, then we say that topic  $i$  is exclusive. A topic that is both cohesive and exclusive is more likely to be semantically useful.

In order to select an appropriate model, we generate a set of candidate models (generated by differing initializations, tuning parameters, or processing of the texts) and then discard results that have the lowest value for the bound.<sup>10</sup> We then plot the exclusivity and semantic coherence of the remaining models and select a model on the semantic coherence-exclusivity “frontier,” that is, where no model strictly dominates another in terms of semantic coherence and exclusivity. We then either randomly select a model or manually examine the remainder and select the model most appropriate to our particular research question. We provide methods for calculating exclusivity and semantic cohesion with our estimation software.

While this simple measure is computationally efficient and interpretable, it cannot replace human judgment. The insight of the investigator is paramount here, and we strongly suggest careful reading of example texts. In these cases, the STM can direct the reader to the most useful documents to evaluate by providing a list of ex-

<sup>8</sup>These qualities also appear in the evaluation of single-membership clustering algorithms (Jain 2010). We speculate these qualities are implicitly central to many conceptual paradigms, both in quantitative as well as qualitative political science.

<sup>9</sup>Newman et al. (2010) first proposed the idea of using point-wise mutual information to evaluate topic quality. Mimno et al. (2011) then proposed a closely related measure, which they named *semantic coherence*, demonstrating that it corresponded with expert judgments of National Institutes of Health (NIH) officials on a corpus of NIH grants as well as human judgments gathered through Amazon’s Mechanical Turk.

<sup>10</sup>The number of models retained can be set by the researcher.

emplar texts for each topic. An intermediate step between automated diagnostics and judgement of the principal investigator is to use human evaluations on tasks for cluster quality. Chang et al. (2009) and Grimmer and King (2011) describe human evaluation protocols for testing topic quality that can easily be applied to our setting. Of course, researchers are free to not use these selection methods, or to create other methods. Researchers might also incorporate pre-analysis plans, which specify sets of words they expect to appear together in topics of interest and select based upon those criteria.

## Validating the Model: Simulations Tests and Examples

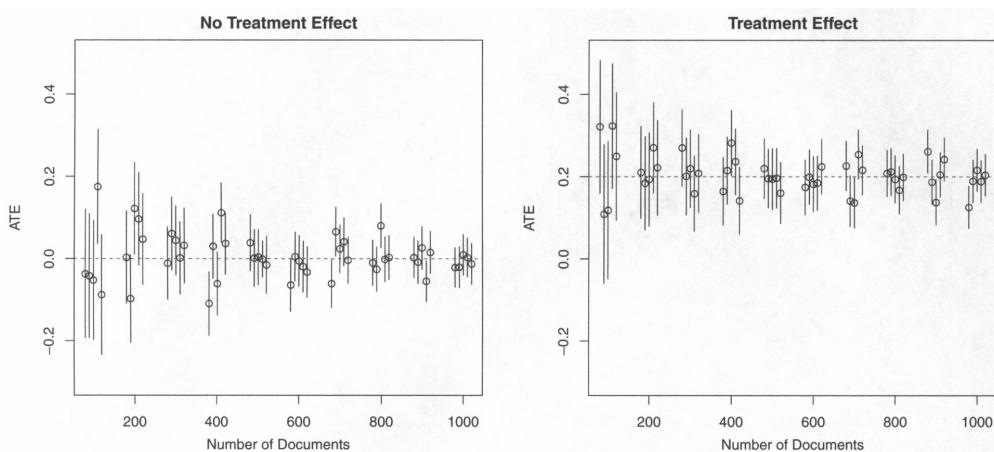
When introducing any new method, it is important to test the model in order to validate that it performs as expected. Specifically, we were driven to answer two critical questions about the performance of the structural topic model:

1. Does the model recover treatment effects correctly (i.e., low false positives and low false negatives)?
2. How does analysis compare to first estimating topics with LDA and then relating the topics to covariates?

In the supplemental appendix, we address both of these questions in turn using a battery of tests that range from Monte Carlo experiments on purely simulated data through applied comparisons of the examples presented in the next section. Here, we briefly address each question, providing an overview of our simulations and deferring the details to the appendix.

As shown in Figure 2, the model recovers the effect of interest when it exists, and it does not induce a spurious effect when the effect is actually zero (false positives). A separate, but related, concern is the effects of multiple testing. While our simulation results demonstrate that STM does not systematically overestimate treatment effects, it does not address concerns of accurate p-values in the presence of multiple testing. In the appendix, we discuss how false discovery rate methods and preexperiment plan approaches can be incorporated into the topic model framework to address these concerns. In the appendix, we also show results of a permutation test on one of our applied examples. In this test, we randomly permute the treatment variable across documents and refit the model, showing that we do not find spurious treatment effects.

FIGURE 2



*Note:* Estimated average treatment effect (ATE) with 95% confidence intervals, holding expected number of words per document fixed at 40 and the concentration parameter fixed at 1/3. The STM is able to recover the true ATE both in cases where there is no treatment effect (left) and cases with a sizable treatment effect (right). As expected, inferences improve as sample sizes increase.

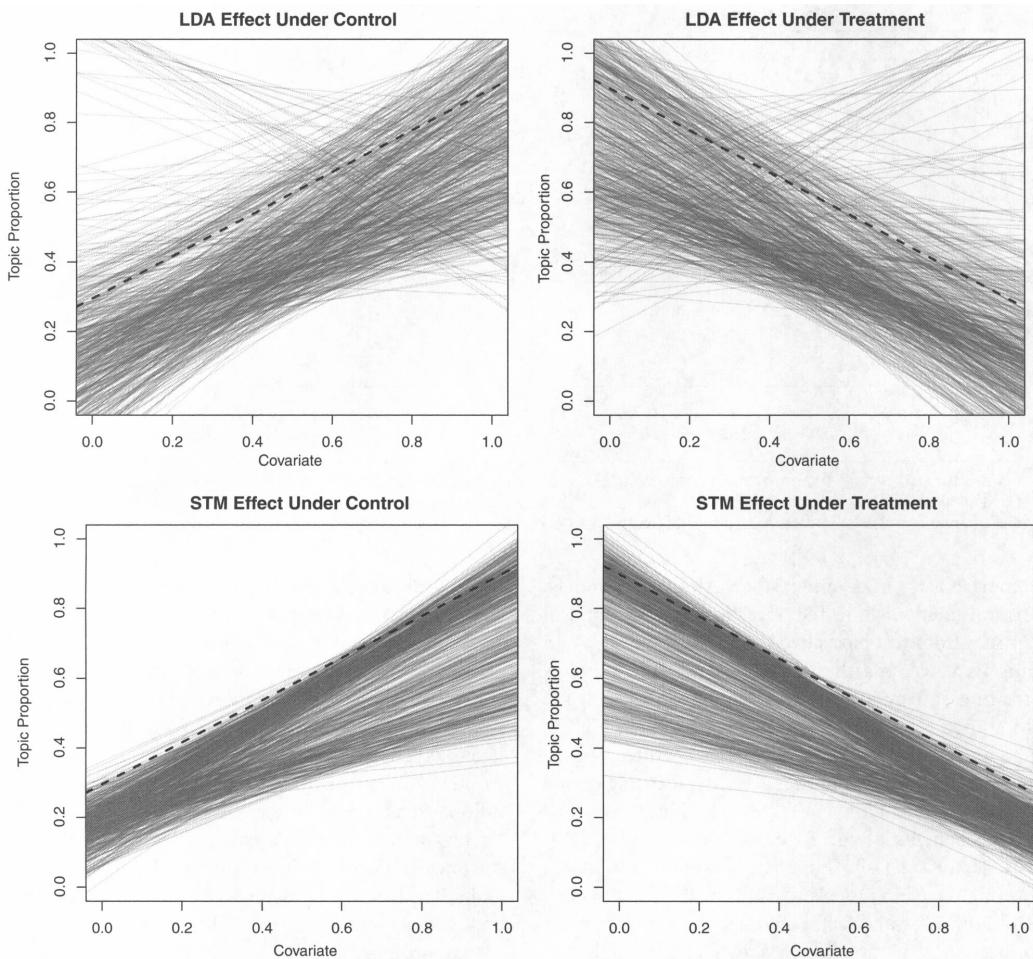
**Comparison to LDA and Other Alternate Models.** Statistical methods for the measurement of political quantities from text have already seen widespread use in political science, and the number of available methods is growing at a rapid rate. How does analysis with the STM compare to existing unsupervised models? In the appendix, we contrast our approach with three prominent alternative models in the literature, focusing on the advantages of including covariates. Specifically, we contrast the benefits of the STM with vanilla LDA (Blei 2012), factor analysis (Simon and Xenos 2004), and single-membership models (Grimmer 2010; Quinn et al. 2010). Both LDA and factor analysis provide the mixed-membership structure, which allows responses to discuss multiple topics, but cannot incorporate the rich covariate information we often have available, whereas the single-membership models developed in political science can incorporate a narrow set of covariate types and are limited to a single topic per document, which may be too restrictive for our application. Compared to other unsupervised techniques, we believe the STM provides the most versatility for survey researchers and experimentalists.

In the appendix, we provide an extensive comparison to LDA, which shows that the STM provides more accurate estimation of quantities of interest when compared to using LDA with covariates in a two-stage process. We show Monte Carlo simulations consistent with the

theoretical expectation that LDA will tend to attenuate continuous covariate relationships on topical prevalence. Figure 3 shows one such simulation for the case of a continuous covariate that operates differently under the treatment and control conditions. LDA is unable to capture the dynamics of the effect in many of the simulated data sets. The appendix also overviews some diagnostics for LDA models that indicate when the inclusion of additional structure as in the STM is useful for inference. Finally, we provide an analysis of actual documents from the immigration experiment discussed below using LDA and characterize the differences between those solutions and the ones attained by the STM.

In a companion paper, we provide a thorough contrast of our method to supervised learning techniques (Lucas et al. 2013). Supervised methods provide a complement to unsupervised methods and can be used when an analyst is interested in a specific, known quantity in the text. Thus, supervised methods can be seen as occupying a place on the spectrum between closed-ended questions, which provide an *a priori*, analyst-specified assessment of the quantities of interest, and the unsupervised analysis of open-ended responses which provide a data-driven assessment of the quantities of interest with post hoc analyst assessment. We provide an implicit comparison to supervised approaches by comparing our unsupervised methods to human coders in the ANES data analysis section below and the appendix.

FIGURE 3 STM versus LDA Recovery of Treatment Effects.



*Note:* Each line represents the estimated effect from a separate simulation, with the bold line indicating the true data-generating process. While the two-stage LDA process often captures the approximate effect, it exhibits considerably higher variance.

**Additional Material.** The appendix provides a number of additional details which we split into three major sections:

1. Model Estimation gives details on the variational expectation maximization-based approach to optimization of the model parameters.
2. Model Validation tests includes simulations mentioned above as well myriad other validations.

3. Getting Started overviews two additional software tools that we provide (i.e., `txtorg`, a tool for preprocessing and handling *large* bodies of text with extensive non-English language support, and a topic visualization tool for helping users browse their documents and assess model results) and discussions of common questions that might arise and how they connect to the topic modeling framework, including multiple testing, mediation analysis, and pre-analysis plans.

## Data Analysis

The purpose of this section is to illustrate the application of the method to actual data. We show how to estimate the relationships between covariates and topics with corresponding uncertainty estimates, how to interpret model parameters, and how to automatically identify passages that are the best representations of certain topics. To illustrate these concepts, we rely on several recent studies that recorded open-ended text as well as recently released data from the ANES.

### Public Views of Immigration

Gadarian and Albertson (forthcoming) examine how negatively valanced emotions influence political behavior and attitudes. In one of their surveys, they focus on immigration preferences by using an experimental design that in the treatment encourages some subjects to become worried about immigration and in control to simply think about immigration. To categorize these open-ended responses, they turned to human coders who were instructed to code each response along the dimensions of enthusiasm, concern, fear, and anger, each along a 3-point scale.

**Topic Analysis.** To estimate the STM, we use an indicator variable for the treatment condition, a 7-point party identification self-report, and an interaction between party identification and treatment condition as covariates. The interaction term lets us examine whether individuals who are Republican respond to the treatment condition differently from those who are Democrats. In this particular application, the influence of these parameters was estimated on topic proportions ("prevalence") within responses. To address multi-modality, we estimated our model 50 times, with 50 different starting values, and applied the model selection procedure described in earlier. This left us with 10 models, from which we selected one based on exclusivity and semantic coherence criterion. However, a close examination of these ten models indicates that all have very similar results in terms of the topics discovered and differences in topic proportions across treatment conditions.

We estimated three topics in total in our analysis. The two topics most associated with the treatment and control groups, respectively, are presented in Figure 4. Topic 1 is the "crime" and "welfare" or "fear" topic, and Topic 2 emphasizes the human elements of immigrants, such as "worker" and "mexican." To get an intuitive sense of the

**FIGURE 4 Vocabulary Associated with Topics 1 and 2**

<b>Topic 1:</b> illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag, take, us, free
<b>Topic 2:</b> immigr, illeg, legal, border, need, worri, mexico, think, countri, law, mexican, make, america, worker, those, american, fine, concern, long, fenc

**FIGURE 5 A Representative Response from Topic 1**

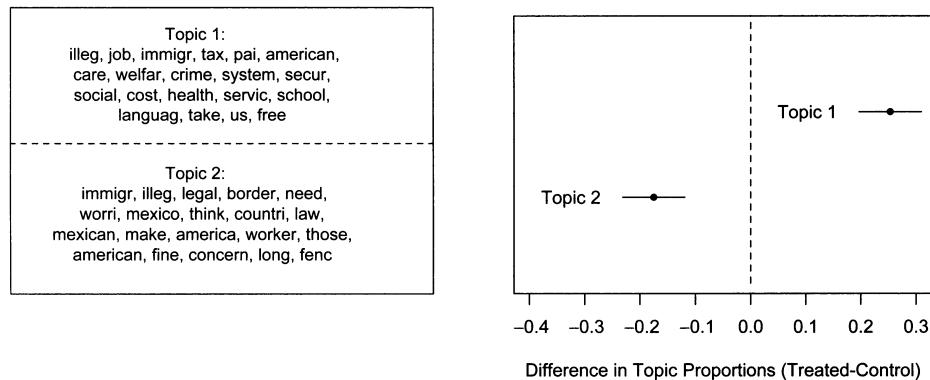
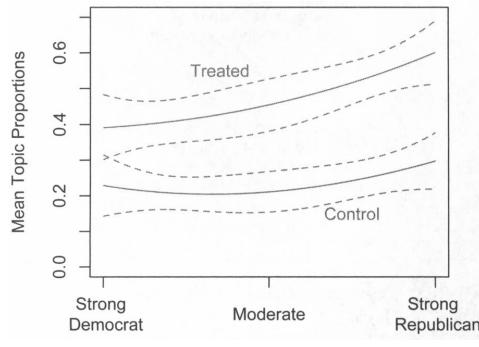
problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals.
crime lost jobs benefits paid to illegals health care and food....we cannot feed the world when we have americans starving, etc.

**FIGURE 6 A Representative Response from Topic 2**

i worry about the republican party doing something very stupid. this country was built on immigration, to deny anyone access to citizenship is unconstitutional. what happened to give me your poor, sick, and tired?
border control, certain illegal immigrants tolerated, and others immediately deported.

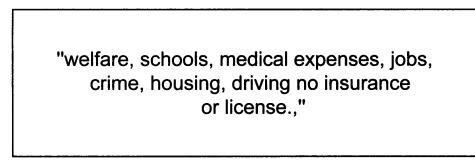
topics, Figures 5 and 6 plot representative responses for topic 1 and 2.<sup>11</sup>

<sup>11</sup>The predicted probability of that response being in the given topic is high relative to other responses within the corpus for Topics 1 and 2.

**FIGURE 7 Words and Treatment Effect Associated with Topic 1****FIGURE 8 Party Identification, Treatment, and the Predicted Proportion in Topic 1**

**Covariate Analysis.** Next, we move to differences across the treatment groups. On average, the difference between the proportion of a treated response that discusses Topic 1 and the proportion of an untreated response that discusses Topic 1 is .28 (.23, .33).<sup>12</sup> This shows that the study's encouragement to express worries about immigration was effective. In addition, on average over both treatment and control, Republicans talk about fear and anger toward immigrants much more than Democrats do: By our estimates, the difference between the proportion of a Republican response that talked about Topic 1 and the proportion of a Democrat response that talked about Topic 1 was .09 (.04, .14).

<sup>12</sup>Estimates within the parentheses represent a 90% confidence interval.

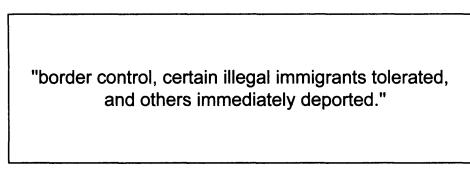
**FIGURE 9 Fearful Response with High Topic 1**

The ability to estimate moderating effects on the treatment/control differences is a key contribution of our technique. The interaction between party identification and treatment also heavily influences topics. The difference between the proportion of a treated Republican response that talked about Topic 1 and the proportion of an untreated Democrat response that talked about Topic 1 is very large: .33 (.28, .39). What does this mean? An untreated Democrat will talk about Topic 1 20% of the time and Topic 2 40% of the time. A treated Republican will talk about Topic 1 54% of the time and Topic 2 20% of the time.<sup>13</sup>

Words associated with the topic and topic proportions by treatment are displayed graphically in Figure 7. The second plot in Figure 7 shows a treatment effect of response proportions in Topics 1 and 2, comparing treated to untreated. Figure 8 shows a Loess-smoothed line of the proportion of each response in Topic 1 on party identification, where 7 is a strong Republican and 0 is a strong Democrat. In general, these accord with our expectations

<sup>13</sup>Words that do not fall into Topic 1 or Topic 2 fall into Topic 3, a topic we do not discuss here because it is least associated with treatment.

**FIGURE 10** Fearful Response with Low Topic 1



about how the treatment and party identification should be associated with the responses.

**Aggregate Comparison with Human Coders.** The traditional way text has been analyzed in survey or experimental settings is to have human coders code each response based on a set of coding instructions. Fortunately, in this example, Gadarian and Albertson (forthcoming) did just this, using two research assistants. How do our results compare with those of the coders? The comparison between the hand coding and the results from our algorithm is described in detail in the appendix. In summary, the results from the STM and the hand coding are similar, and both methods find a treatment effect. In addition, there is significant correlation between the hand coding of individual responses and the predicted topic proportions from our unsupervised learning model.

Of course, since our model is unsupervised, the topics discovered by our model does not perfectly match the topics the coders were instructed to use. The coders categorized the vast majority of responses into fear and anger, but because the topic model by design tries to distinguish between documents, its definition of topics does not align directly with fear and anger, and some documents with a low proportion of Topic 1 from our analysis are also hand-coded with fear and anger. We would expect that the documents with low predicted proportion of Topic 1 but hand-coded as fear and anger would have fewer characteristics associated with Topic 1; for example, they might talk about crime and Social Security less relative to other reasons for being fearful of or angry at immigrants. Figure 9 presents a document that has a high predicted proportion of Topic 1 that the coders both agree includes fear or anger, whereas Figure 10 presents a response with low relation to Topic 1 but still coded to be of high fear.

It is clear from these responses that both are somewhat fearful of illegal immigrants, but the reasoning behind their emotion is different. In addition, these two may have a very different view of legal immigration in general. One advantage of the topic model is that even if the overwhelming majority of people are either fearful of

or angry at illegal immigrants, it will refine the topics in order to distinguish between documents, so even if a category predetermined by the researcher applies to almost all responses, the topic model can find a finer distinction between them.

## Intuition versus Reflection in Public Goods Games

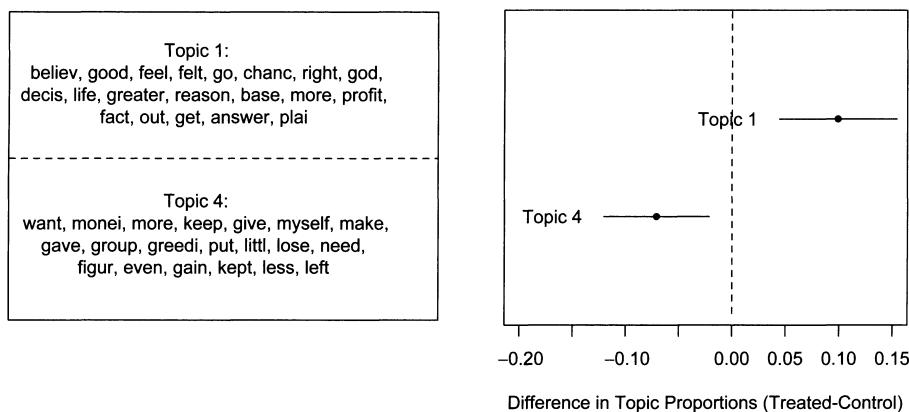
Rand, Greene, and Nowak (2012) study how intuitive versus reflective reasoning influences decision making in public goods games using a number of experimental conditions. In the “free-write” experimental contrast, subjects were primed to consider a time when they have acted out of intuition in a situation where their action worked out well or a time when they reflected and carefully reasoned in a situation where their action worked out well. After this encouragement, everyone played a single incentivized, one-shot public goods game. In the “time” experimental contrast, subjects were either forced to make a decision quickly or encouraged to take their time, after which all players participated in the same public goods game. Rand, Greene, and Nowak (2012) find that subjects contribute more under the treatments where subjects are primed for intuition or are under time pressure, concluding that cooperation is intuitive. After both the free-write and time experiments, subjects were asked to write about the strategy they used while playing the public goods game. We analyze the players’ descriptions of their strategies and their relationship to game contributions.

**Decision Explanations across Treatment Conditions.** We contrast the topics present in the strategy descriptions across the different treatment conditions.<sup>14</sup> The topic model reflects how the experimental conditions influence strategy descriptions. In the free-write experimental contrast, respondents primed to think intuitively talk about their strategy very differently from those who received the reflection priming. Listed in Figure 11, Topic 1 is associated with the intuitive priming, and Topic 4 is associated with the reflection priming. Topic 1 has words that reflect intuition, for example, “good,” “feel,” “chance,” “felt,” and “believ.” Topic 4, on the other hand, includes words such as “want,” “money,” “more,” and “myself.” The estimated topical difference between the two treatments is shown in the second graph of Figure 11.

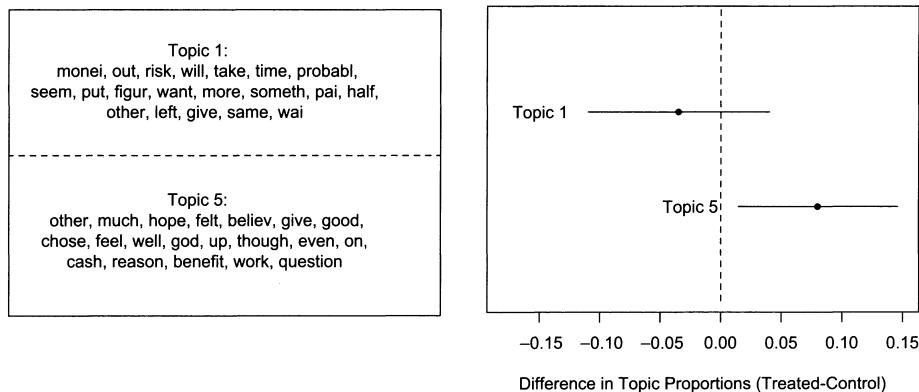
In the time experimental contrast, people who are given less time to think about their decision in the public goods game use more feeling and trusting words to

<sup>14</sup>We estimated a five-topic model with intuition or time pressure as the treatment.

**FIGURE 11 Topics from Intuition vs. Reflection Priming and Intuition Treatment Effect**



**FIGURE 12 Topics from Time Pressure Experiment and Time Pressure Treatment Effect**

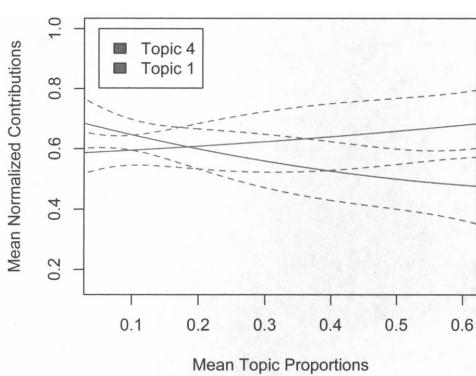
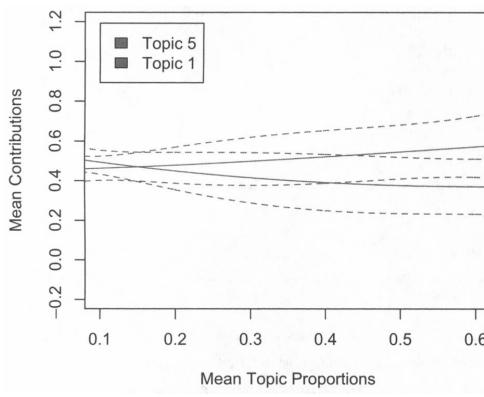


describe their decisions, as shown in Topic 5 on the left-hand side of Figure 12. Words in this topic reflect concern over morality and feeling, with reference to words like “believe,” “felt,” “hope,” and “god.” In contrast, people who are given more time to think about their decision of whether or not to contribute use a more calculating vocabulary to describe their decision, with words like “money,” “risk,” and “figure.” This topic is shown below in Topic 1 of Figure 12. The treatment effect for both of these topics is shown on the right-hand side of Figure 12.

The topics in the intuition priming experiment and the time pressure experiment show some similarities. Topic 1 and 5, respectively, use words associated with feeling and trusting. Topic 2 and Topic 1, respectively, are

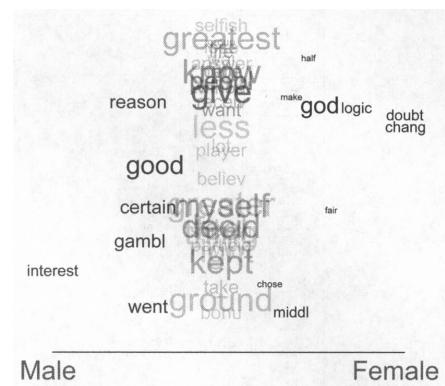
more related to thinking, maximizing payoff, and making choices. These results show a nice coherence in the experimental design and shows how topic models can directly connect to the theoretical model, where the intuitive-cooperation theory expects these exact distinctions to be important.

**Respondents Who Talk about Their Intuition Cooperate More.** We also examine the relationship between references to a topic and contributions in the game. As Rand, Greene, and Nowak (2012) find, forcing people to think quickly increases contributions, and priming people to think intuitively also increases contributions. We expect, therefore, that people who talk about intuition, or whose

**FIGURE 13 Intuition Topics and Contributions****FIGURE 14 Time Pressure Versus Delay Topics and Contributions**

responses are more in line with Topic 1 in the intuition priming and Topic 5 in the time pressure experiment, will also contribute more, but that people who talk about strategy and maximizing their profits, more in line with Topic 2 in the intuition priming and Topic 1 in the time pressure experiment, will contribute less.

In both cases, respondents with a higher predicted proportion of the intuitive topic in their response are more likely to contribute. Figure 13 and Figure 14 plot a Loess, smoothed line of contributions plotted with 90% confidence intervals on the predicted topic proportions for each document. For each of the experiments, responses with a higher predicted proportion of the intuitive topic have overall higher contribution. However, as the predicted proportion of the reasoning topic in-

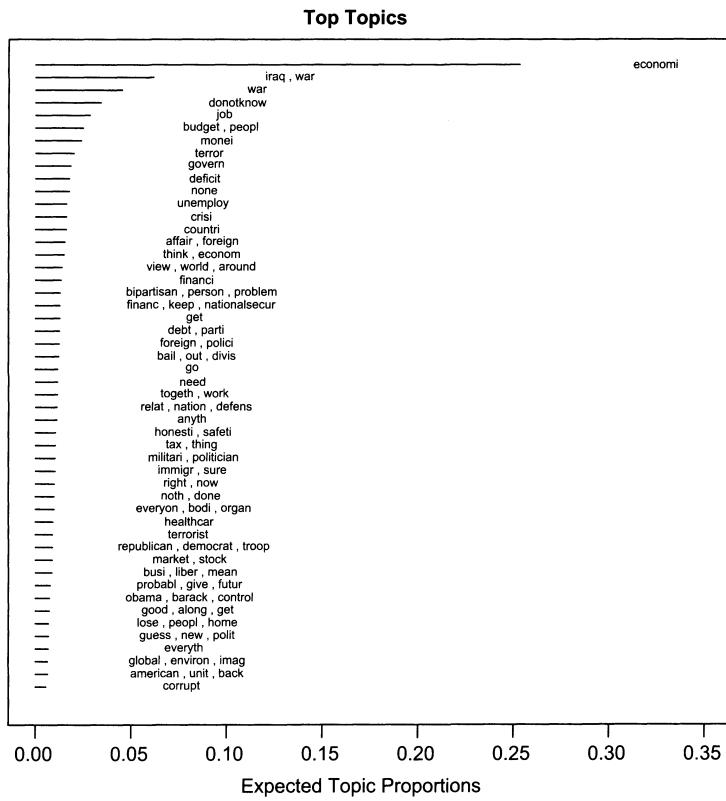
**FIGURE 15 Intuitive Topic Allowing for Different Vocabularies Based on Gender****FIGURE 16 Comparison of Women and Men's Vocabulary After Intuition Treatment**

Female	Male
Topic 3: keep, peopl, give, know, myself, good, decid, god, thing, feel, go, more, right, well, logic, greater, want, monei, someon, less, kept, cent, other, awai, sure, believ, amount, benefit, doubt, chang, hope, best, make, greatest, worri, much, chose, figur, same, thought	Topic 3: good, keep, peopl, give, know, myself, decid, thing, feel, go, more, right, well, greater, want, monei, less, someone, kept, cent, other, awai, sure, amount, believ, benefit, hope, best, god, greatest, worri, much, reason, figur, thought, same, ground, gambl, guess, wai

creases, the overall level of contributions falls. Therefore, people who talk more about intuition, trust, and their feelings, are more likely to contribute. People who talk more about strategy and maximizing profits are less likely to contribute.

**Using Covariates for the Vocabulary: Gender.** The topic model is not only able to assess the influence of covariates on the topic proportions, but it is also able to use covariates to show how different types of respondents use different vocabulary to talk about the same topic. A researcher might be interested in how women

FIGURE 17 STM Topics from ANES Most Important Problem



talk about their strategy when primed with an intuition treatment, compared to how men talk about their strategy when primed with intuition. Figure 15 shows a word cloud of the intuition topic where the word size represents the frequency with which a word is used. We find that men talk about their intuition with certainty, whereas women describe their intuition with doubt and in terms of their morality. On the left side of the plot are words that men use more frequently within the intuition topic, including “interest,” “gamble,” and “certain.” On the right side of the plot are words that women use more frequently within the intuition topic, including “god”, “middle”, and “doubt”, words associated with morality and uncertainty. Figure 16 also shows the most frequent words for the intuition topic specific to men and women.

By allowing topics to vary in vocabulary by gender, we expect that we will also better be able to measure treatment effects. If women and men simply explain their intuition differently, we should not treat these two slightly

different vocabularies as completely different topics. By allowing flexibility in the topics, we expect that we can more precisely measure the prevalence of topics within documents. Future research will explore the relationship between better estimated vocabularies and more precise treatment effects on topic proportions.

## ANES

Our model has more general applications to open-ended survey questions even when no treatment condition is included in the study. In this section, we apply the structural topic model to analyze open-ended responses from the American National Election Survey (ANES). A sample of 2,323 respondents was interviewed after the 2008 presidential election. Each respondent was asked to identify the most important and second most important political problem facing the United States, as well as the most important and second most important personal issue in the

**TABLE 1 Comparison of STM to Hand Coding**

STM Topic	STM Count	ANES Topic	Hand-Coding Count
Economy	891	The Economy	653
War or Iraq War	151	War, or Iraq War	189
Don't Know	53	Don't Know	72
Unemployment and Job	61	Employment	83

election. The original data were recently recoded by the ANES into a set of stable categories using human coding. We show that the STM is consistent with the human coding of the open-ended responses, while also uncovering new categories that are specific to the 2008 election.

We analyze open-ended responses that identify the most important political problem for each individual and use party identification, education, age, and an interaction between party identification and education as the covariates. Figure 17 displays the top topics from a 60-topic model and the frequency of these topics within our data.<sup>15</sup> The topics correspond closely to general topics we would expect: High-frequency topics include "econ," "iraq, war," "war," "donotknow," and "job."

The ANES hired human coders to code each of the open-ended responses into one of 69 categories.<sup>16</sup> Table 1 compares the aggregate categorization of the top categories between the STM and hand coding. The aggregate numbers of responses coded into each category are very similar across the STM and the ANES hand-coded data, even though the topic categories are not perfectly aligned between the STM and the pre-determined human categories. The major difference in the aggregate numbers is between the "Economy" categories. This difference is because the ANES has many categories related to the economy, including a catch-all "Economics" category, separate from "The Economy," "Budget," and "Unemployment" categories.

Not only are the aggregate numbers similar, but also many of the individual responses coded by the STM are similarly coded by the hand-coding scheme. For example, we compare responses that the STM estimated to have more than 20% of the topic "terror" with responses that were hand-coded to be at least partially in the topic "Terrorism." Of the responses that were coded by the STM into "terror," only three of them were not also hand-coded into

"Terrorism." These three had to do with prices of natural resources, oil in particular, which is often associated with terrorism and the Middle East. Of the responses that were hand-coded into "Terror," 20% of them were not coded by the STM into "terror." These responses usually included the word "terrorist," which the STM found to be a separate category from "terror."

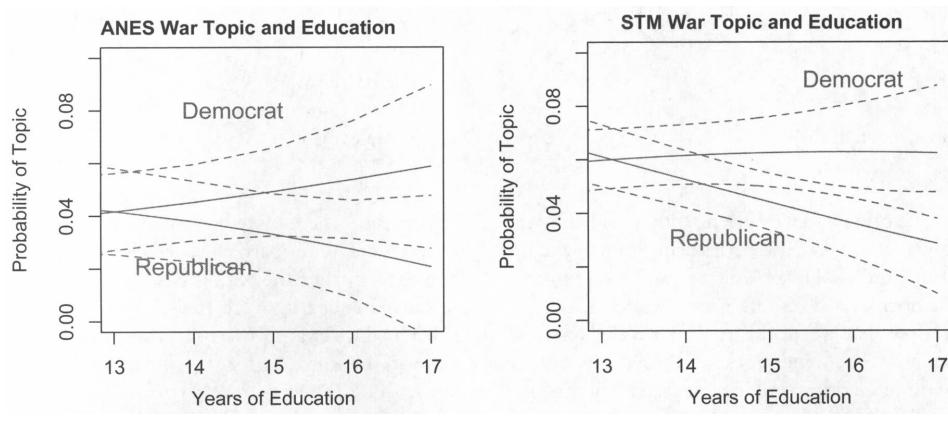
While most of the responses in the ANES were placed into one category, the human coding did allow an individual response to fall into multiple categories. Overall, about 19% of responses were hand-coded into multiple categories. Whereas Table 1 shows that the STM is consistent with the hand coding in putting responses into categories, a comparison of multiple categories provides yet another dimension along which we can compare hand coding to the STM because the STM also allows responses to be a mixture over topics. We would expect, for example, that responses that the ANES coded into a single category would also be heavily centered over one topic in the results from the STM.

For each response, we use the STM to calculate the number of topics that contribute to at least 20% of the individual's response. Using this method of comparison, we find high correlation between the number of topics the STM assigns to a response and the hand coding assigned to a response. Of the responses the hand coding coded into only one category, 80% were also coded by the STM into only one category. Of the responses that the STM coded into one category, 81% were also hand-coded into one category. Overall, 94% of the responses were coded either in the same number of categories between the STM and hand coding or only had one category difference. To consider a specific example, of the responses that were hand-coded into the "Economy" and "Unemployment" categories, the two most prevalent topics from the STM were also "econ" and "unemployment."

The STM recovers covariate relationships very similar to those discovered by the ANES hand coders. Figure 18 shows the relationship between party identification, education, and the "Iraq War" topic in each case. The relationship between the covariates and the "Iraq War" topic look very similar between the two different models; Democrats write more about the Iraq War than

<sup>15</sup>Words used to label the topics are the most likely words within that topic. The number of words printed is determined by an algorithm that calculates the gap between the probability of the last word printed and the probability of the next most likely word.

<sup>16</sup>Some responses were placed into multiple categories if the coders determined that the response included multiple topics.

**FIGURE 18 Comparison of Covariate Relationships**

Republicans, especially at higher levels of education. One benefit of the STM is that it produces a continuous measure of topics within each document, whereas the hand-coded data produces only a categorization. The continuous measure is much easier to work with when looking at covariate relationships, which is one advantage of using the topic model in place of, or in addition to, human coding.

An additional advantage of using the unsupervised approach to categorize open-ended responses in the ANES data is that categories are specific to the time period in which the survey was administered. Since this survey was administered directly after the 2008 election, many respondents indicated that their most important problem was “Obama,” the “democrats,” or “republicans,” none of which are categories within the ANES hand-coding scheme.<sup>17</sup> Categorizations can change quickly with time period, and researchers might be interested in these changes. Thus, with respect to categorization, the STM allows for flexibility and adaptability that is difficult with human coding.

Of course, there are disadvantages to using the unsupervised approach to the categorization of these open-ended responses. In particular, predetermined categories that have a low incidence rate within the open-ended responses are unlikely to show up within the STM. For example, human coders assigned one response to the category “China.” Since very few responses mention China, the STM does not discover a topic related to China, and therefore this categorization would be lost using the unsupervised approach. Due to low-frequency responses like

<sup>17</sup>The ANES has a general category for “Politicians,” but it does not include categories for specific politicians.

the one response to China, there are also some topics recovered by the STM that are not particularly meaningful, with most frequent words like “get” or “go,” which represent a hodgepodge of low-frequency responses. These topics would become better defined with more data.

However, the costs of human coding thousands of responses may balance out downsides associated with the unsupervised approach. Indeed, as the number of responses increases, the STM becomes more accurate and human coding becomes more unwieldy. At minimum, the ANES could use this unsupervised method in conjunction with human coding. This approach would save significant time and money, would allow for the discovery of topics specific to the survey time period, and would point human coders to ambiguous responses where human classification is more important.

## Conclusion

Spurred by recent efforts in digitalization of text, researchers are rapidly reinventing the evidence base of the social sciences by introducing new measures of political and social behavior constructed from large text archives (Grimmer and Stewart 2013; Hopkins and King 2012; King 2009; Lazer et al. 2009). At the same time, political science has shown increasing interest in survey experiments, merging the discipline’s long-standing use of surveys with the inferential strengths of the experimental framework (Druckman et al. 2006). Yet, interestingly, these two trends have been remarkably

distinct, and analyses of textual data from surveys and experiments are rare. We show how the structural topic model recently introduced by Roberts and colleagues (Roberts, Stewart, and Airolidi 2013; Roberts et al. 2013) provides a unified approach that enables scholars to take advantage of recent advances in automated text analysis. Crucially, it enables the analysis to investigate a range of interesting quantities of interest while incorporating information like a respondent's experimental condition or other covariates like partisan affiliation or gender. This in turn enables the estimation of standard quantities of interest while incorporating measurement uncertainty. Furthermore, this approach is suitable for other sources of text, not just open-ended responses in surveys (e.g., Lucas et al. 2013).

While the model can dramatically simplify the analysis of open-ended responses, the proposed methodology is not without limitations. In some cases, the flexibility of the model comes at the cost of complicating the survey analysis process. We highlight two such areas and detail the tools we provide to help alleviate these difficulties.

**Model Selection.** Analysts may be unaccustomed to the model selection issues that arise from the multi-modal objective function. We address these concerns through a principled, semiautomated model selection procedure that helps to guide analysts towards automated solutions. In the supplemental appendix, we engage with a range of potential criticisms that arise as a consequence of the subjective model choice, such as concerns about false positives arising from the use of treatment assignment for both measurement and effect estimation. As with any research design, best practices will depend on the inferential goals and criteria within a research field.

**Power Calculations.** In planning a survey experiment, researchers often turn to power calculations for anticipating the number of respondents necessary to recover an effect of a given size. These calculations quickly become quite complicated in the case of the STM due to the introduction of the uncertainty in the measurement model. We note that traditional power calculations provide a lower bound on the number of respondents necessary to identify an effect. The looseness of the bound is determined by how well differentiated the language is within the outcome categories. The data for both survey experiments discussed here are included within the **stm** package, which should provide some intuition for the necessary sample sizes.

The STM can be used for exploration of a corpus about which little is known *ex ante*, or along with rig-

orous preanalysis plans that define clear prior predictions of expected topics. When the researcher has a body of text with metadata where the documents take on a mixture of documents, the STM can be useful for exploration, description, and prediction. Opportunities for future research are immense, including new substantive applications, incorporation of respondent ranking in text (e.g., denote one problem the "first" most important problem and another the "second"), and technical extensions like techniques for nonparametric selection of the number of topics (Paisley, Wang, and Blei 2012).

## References

- Anandkumar, Anima, Dean Foster, Daniel Hsu, Sham Kakade, and Yi-Kai Liu. 2012. "A Spectral Algorithm for Latent Dirichlet Allocation." *Advances in Neural Information Processing Systems* 25: 926–34.
- Artstein, R., and M. Poesio. 2008. "Inter-Coder Agreement for Computational Linguistics." *Computational Linguistics* 34(4): 555–96.
- Bischof, Jonathan, and Edoardo Airolidi. 2012. "Summarizing topical content with word frequency and exclusivity." In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, eds John Langford and Joelle Pineau. New York, NY: Omnipress, 201–208.
- Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55(4): 77–84.
- Blei, David M., Andrew Ng, and Michael Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(Jan): 993–1022, 2003.
- Buot, M. L. G., and D. S. P. Richards. 2006. "Counting and Locating the Solutions of Polynomial Systems of Maximum Likelihood Equations, I." *Journal of Symbolic Computation* 41(2): 234–44.
- Chang, Jonathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." *Advances in Neural Information Processing Systems*, 288–296.
- Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100(4): 627–35.
- Eisenstein, Jacob, Amr Ahmed, and Eric P. Xing. 2011. "Sparse Additive Generative Models of Text." *Proceedings of the 28th International Conference on Machine Learning* 1041–48.
- Gadarian, Shana, and Bethany Albertson. forthcoming. "Anxiety, Immigration, and the Search for Information." *Political Psychology*.
- Geer, John G. 1988. "What Do Open-Ended Questions Measure?" *Public Opinion Quarterly* 52(3): 365–71.
- Geer, John G. 1991. "Do Open-Ended Questions Measure 'Salient' Issues?" *Public Opinion Quarterly* 55(3): 360–70.

- Gerring, John. 2001. *Social Science Methodology: A Unified Framework*. Cambridge: Cambridge University Press.
- Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18(1): 1–35.
- Grimmer, Justin. 2011. "An Introduction to Bayesian Inference via Variational Approximations." *Political Analysis* 19(1): 32–47.
- Grimmer, Justin, and Gary King. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization." *Proceedings of the National Academy of Sciences* 108(7): 2643–50.
- Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3): 267–97.
- Hopkins, Daniel, and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1): 229–47.
- Hopkins, Daniel. 2013. "The Exaggerated Life of Death Panels: The Limits of Framing Effects in the 2009–2012 Health Care Debate." Working Paper. Georgetown University.
- Iyengar, Shanto. 1996. "Framing Responsibility for Political Issues." *Annals of the American Academy of Political and Social Science* 546(1): 59–70.
- Jain, A. K. 2010. "Data Clustering: 50 Years beyond K-Means." *Pattern Recognition Letters* 31(8): 651–66.
- Kelley, Stanley. 1983. *Interpreting Elections*. Princeton, NJ: Princeton University Press.
- King, Gary. 2009. "The Changing Evidence Base of Social Science Research." In *The Future of Political Science, 100 Perspectives*, eds. Gary King, Kay Schlozman, and Norman Nie, New York: Routledge Press, 91–93.
- Krosnick, Jon A. 1999. "Survey Research." *Annual Review of Psychology* 50: 537–67.
- Laver, M., K. Benoit, and J. Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(2): 311–31.
- Lazarsfeld, Paul F. 1944. "The Controversy over Detailed Interviews—An Offer for Negotiation." *Public Opinion Quarterly* 8(1): 38–60.
- Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, et al. 2009. "Computational Social Science." *Science* 323(5915): 721–23.
- Lombard, M., J. Snyder-Duch, and C. C. Bracken. 2006. "Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability." *Human Communication Research* 28(4): 587–604.
- Lucas, Christopher, Richard Nielsen, Margaret Roberts, Brandon Stewart, Alex Storer, and Dustin Tingley. 2013. "Computer Assisted Text Analysis for Comparative Politics." Working paper.
- Manning, Christopher D., Prahalak Raghavan, and Hinrich Schütze. 2008. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing Semantic Coherence in Topic Models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, 262–72.
- Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. "Automatic Evaluation of Topic Coherence." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 100–108.
- Paisley, John, Chong Wang, and Blei, David. M. 2012. "The discrete infinite logistic normal distribution." *Bayesian Analysis*, 7(4): 997–1034.
- Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54(1): 209–28.
- Rand, David G., Joshua D. Greene, and Martin A. Nowak. 2012. "Spontaneous Giving and Calculated Greed." *Nature* 489(7416): 427–30.
- RePass, David E. 1971. "Issue Salience and Party Choice." *American Political Science Review* 65(2): 389–400.
- Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. 2013. "Structural Topic Models." Working paper.
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. 2013. "The Structural Topic Model and Applied Social Science." *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.
- Schuman, Howard. 1966. "The Random Probe: A Technique for Evaluating the Validity of Closed Questions." *American Sociological Review* 31(2): 218–22.
- Schuman, Howard, and Stanley Presser. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Thousand Oaks, CA: Sage.
- Simon, A. F., and M. Xenos. 2004. "Dimensional Reduction of Word-Frequency Data as a Substitute for Intersubjective Content Analysis." *Political Analysis* 12(1): 63–75.
- Sontag, D., and D. M. Roy. 2009. "Complexity of Inference in Topic Models." *Advances in Neural Information Processing: Workshop on Applications for Topic Models: Text and Beyond*.
- Wang, Chong, and David M. Blei. 2011. "Collaborative Topic Modeling for Recommending Scientific Articles." In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 448–56.
- Zou, James, and Ryan Adams. 2012. "Priors for Diversity in Generative Latent Variable Models." *Advances in Neural Information Processing Systems* 25: 3005–13.

# Conclusion



# We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together

Justin Grimmer, *Stanford University*

**I**nformation is being produced and stored at an unprecedented rate. It might come from recording the public's daily life: people express their emotions on Facebook accounts, tweet opinions, call friends on cell phones, make statements on Weibo, post photographs on Instagram, and log locations with GPS on phones. Other information comes from aggregating media. News outlets disseminate news stories through online sources, and blogs and websites post content and receive comments from their readers. Politicians and political elites contribute their own messages to the public with advertising during campaigns. The federal government disseminates information about where it spends money, and local governments aggregate information about how they serve their citizens.

The promise of the “big data” revolution is that in these data are the answers to fundamental questions of businesses, governments, and social sciences. Many of the most boisterous claims come from computational fields, which have little experience with the difficulty of social scientific inquiry. As social scientists, we may reassure ourselves that we know better. Our extensive experience with observational data means that we know that large datasets alone are insufficient for solving the most pressing of society’s problems. We even may have taught courses on how selection, measurement error, and other sources of bias should make us skeptical of a wide range of problems.

This statement is true; “big data” alone is insufficient for solving society’s most pressing problems—but it certainly can help. This paper argues that big data provides the opportunity to learn about quantities that were infeasible only a few years ago. The opportunity for descriptive inference creates the chance for political scientists to ask causal questions and create new theories that previously would have been impossible (Monroe et al. 2015). Furthermore, when paired with experiments or robust research designs, “big data” can provide data-driven answers to vexing questions. Moreover, combining the social scientific research designs makes the utility of large datasets even more potent.

The analysis of big data, then, is not only a matter of solving computational problems—even if those working on big data in industry primarily come from the natural sciences or computational fields. Rather, expertly analyzing big data also requires thoughtful measurement (Patty and Penn 2015), careful research design, and the creative deployment

of statistical techniques. For the analysis of big data to truly yield answers to society’s biggest problems, we must recognize that it is as much about social science as it is about computer science.

## THE VITAL ROLE OF DESCRIPTION

Political scientists prioritize causal inference and theory building, often pejoratively dismissing measurement—*inferences characterizing and measuring conditions as they are in the world*—as “mere description” or “induction.” Gerring (2012) showed, for example, that 80% of articles published in *American Political Science Review* focus on causal inference. The dismissal of description is ironic because much of the empirical work of political scientists and theories that they construct are a direct product of description. Indeed, political scientists have developed a wide range of strategies for carefully measuring quantities of interest from data, validating those measures, and distributing them for subsequent articles. Therefore, although descriptive inference often is denigrated in political science, our field’s expertise in measurement can make better and more useful causal inferences from big data.

The VoteView project is perhaps the best example of political science’s expertise with measurement and why purely descriptive projects affect the theories we construct and the causal-inference questions we ask (McCarty, Poole, and Rosenthal 2006; Poole and Rosenthal 1997).<sup>1</sup> VoteView is best known for providing NOMINATE scores—that is, measures of where every representative to serve in the US House and Senate falls on an ideological spectrum. The authors are emphatic that NOMINATE measures only low-dimensional summaries of roll-call voting behavior. Like other measurement techniques, these summaries are a consequence of both the observed data and the assumptions used to make the summary (Clinton and Jackman 2009; Patty and Penn 2015). Extensive validations suggest, however, that the measures are capturing variation in legislators’ expressed ideology (Clinton, Jackman, and Rivers 2004; Poole 1984; Poole and Rosenthal 1985; 1997).

The impact of the VoteView project is broad and substantial. NOMINATE measures appear in almost every paper about the US Congress and in much of the work of other scholars related to US politics. These findings have fueled numerous debates. Perhaps one of the most famous findings

is that polarization in Congress—that is, the ideological distance between the two parties—has grown substantially in the past 40 years (McCarty, Poole, and Rosenthal 2006; Poole and Rosenthal 1984). This basic descriptive insight, which characterizes the state of the world rather than explaining why, has led to a large literature on the origins of polarization (McCarty, Poole, and Rosenthal 2006; 2009; Theriault 2008) and its consequence for governance (e.g., Krehbiel 1998). The findings on polarization also have reached the media, providing evidence for claims about the historic distance between the two parties. They even have been extended to include all candidates and donors across all levels of government (Bonica 2014) as well as all users of massive social networking websites (Bond and Messing 2014).

*Social scientists know that large amounts of data will not overcome the selection problems that make causal inference so difficult.*

The opportunities for important descriptive inferences abound in big data. For example, census data and social media posts can contribute to an important developing literature about how some of the fastest growing demographic groups (e.g., biracial Americans) reconcile their competing social and political identities (Davenport 2014). Aggregated newspaper articles can provide unprecedented accounts of the media's agenda (Boydston 2013). Online discussions can answer broad questions about how often the public talks about politics during daily life. Each descriptive inference is important on its own and, if linked to broader populations (Nagler and Tucker 2015), would facilitate causal inferences and theoretical advances.

Each example also demonstrates the distinctive way that social scientists use machine-learning algorithms. Social scientists typically use machine-learning techniques to measure a certain characteristic or latent quantity in the world—a qualitatively different goal than computer scientists, who use the measures for prediction (Chang et al. 2009; Grimmer and Stewart 2013; Quinn et al. 2010). To measure latent quantities, social scientists must make consequential and untestable assumptions to compress data into some measure, similar to the assumptions necessary for causal inference. To assess how those assumptions affect the inferences made, social scientists developed a suite of methods for validating latent measures. These tools are invaluable in making descriptive inferences from big data that are useful for the most vexing problems—which provides our first example of how the analysis of big data is best viewed as a subfield of the social sciences.

#### **RESEARCH DESIGN IN LARGE DATASETS**

Descriptive inferences tell us about the world as it is. Big data proponents, however, argue that it also can tell us about the world as it could be. Big data, we often are told, will facilitate “data-driven” decision making. Companies and policy makers are told that they can use the large collections of information to be aware of the consequences of their actions before they are taken. Academics are told they can use the massive

datasets to test causal theories that would be impractical in smaller datasets.

Of course, social scientists know that large amounts of data will not overcome the selection problems that make causal inference so difficult. Instead, a large literature has emerged to argue that causal inferences require a rigorous research design, along with a clear statement of the assumptions necessary for that design to yield accurate causal estimates (Imai, King, and Stuart 2008; Sekhon 2009). The best studies then will provide an argument about why those assumptions are satisfied and an analysis of what happens if they are violated.

Big data alone is insufficient to make valid causal inferences; however, having more data certainly can improve causal

inferences in large-scale datasets. Consider, for example, using matching methods and the characteristics of observations to make treatment and control units comparable (Ho et al. 2007; Rosenbaum and Rubin 1983). A challenge in matching methods is that there may be few units similar on a wide range of characteristics; therefore, there may be potential discrepancies on observable characteristics, let alone differences on unobserved traits. However, massive datasets may provide ideal settings for matching, wherein the multitude of units ensures that the matches are close or that the treatment and control units are similar (Monroe et al. 2015).

Other research designs used to estimate causal effects also could benefit from a massive number of observations. For example, numerous papers use regression-discontinuity designs to estimate a valid local estimate of an intervention's effect (Lee 2008; Lee, Moretti, and Butler 2004). One limitation of the design is that there often are too few units very close to the discontinuity; therefore, units farther away must be used to obtain precise estimates. If there is a discontinuity in a large dataset, however, it is necessary to borrow information from units that are far from the discontinuity.

Massive datasets and social networking sites provide opportunities to design experiments on a scale that was previously impossible in the social sciences. Subtle experiments on a large number of people provide the opportunity to test social theories in ecologically valid settings. The massive scale of the experiments also provides the chance to move away from coarse treatments estimated at the population level to more granular treatments in more specific populations. The result will be a deeper understanding of the social world. Designing experiments and developing robust observational research designs requires more than computational tools. Social science is necessary, then, for big data to provide data-driven decision making.

#### **COMBINING MACHINE LEARNING AND CAUSAL INFERENCE**

Large collections of data not only improve the causal inferences we make. The computational tools that often are associated with

the analysis of big data also can help scholars who are designing experiments or making causal inferences from observational data. This is because many problems in causal inference have a close analogue in machine learning. Indeed, scholars who recognize this connection already have improved how experiments are designed and analyzed.

Consider, for example, blocking observations in an experiment—that is, grouping together observations before

social scientists—measuring quantities of interest from noisy data and inferring causal effects—are abundant. Therefore, for big data to be useful, we must draw on the substantial knowledge base that social scientists have amassed about how to most effectively use quantitative tools to solve social scientific problems. Recognizing the value of social science will lead to fruitful collaboration. Although social scientists have little experience with massive datasets, we have extensive

*Computational advances have led to monumental changes in the tools that everyday people use to live their life, immense progress in how the data are stored, and unprecedented tools to analyze large collections.*

random assignment to improve the precision of estimated effects. Higgins and Sekhon (2014) leveraged insights from graph theory to provide a blocking algorithm with guarantees about the similarity of observations assigned to the same block. Moore and Moore (2013) used tools to provide a blocking algorithm for experiments that arrive sequentially. Machine-learning tasks also are helpful for the closely related task of matching. Hazlett (2014) used a kernel method to create a flexible matching method to reduce imbalances between treatment and control units.

Machine-learning methods also can improve what we learn from experiments and the types of experiments that are conducted. Not only are effect estimates interesting for the entire population of units in the experiment; we also might be interested in how the treatment effects vary across units. Furthermore, machine-learning methods are effective at identifying actual differences in response. For example, Imai and Ratkovic (2013) extended variable selection methods to estimate treatment-effect heterogeneity, whereas Green and Kern (2012) used Bayesian additive regression trees to capture systematic heterogeneity in treatment effects.

Indeed, combining machine learning to make causal inferences is one of the fastest growing and most open fields in political methodology. There is much work to be done in estimating causal effects in texts (Roberts et al. 2014) and political networks (Fowler et al. 2011). There also are numerous opportunities to combine experimental design with machine-learning algorithms to learn how high-dimensional treatments affect response. This area presents an opportunity for leveraging the insights from social science, the computational tools from machine learning, and the big data sources that now are abundant.

#### WE ARE ALL SOCIAL SCIENTISTS NOW

The big data revolution has been hailed as a triumph of computation and, indeed, it is. Computational advances have led to monumental changes in the tools that everyday people use to live their life, immense progress in how the data are stored, and unprecedented tools to analyze large collections. The results are the largest and most detailed datasets in the history of the world. However, the big data revolution also is a recognition that the problems addressed by quantitative

experience with causal inference. Data scientists have significantly more experience with large datasets but they tend to have little training in how to infer causal effects in the face of substantial selection.

Social scientists must have an integral role in this collaboration; merely being able to apply statistical techniques to massive datasets is insufficient. Rather, the expertise from a field that has handled observational data for many years is required. For “big data” to actually be revolutionary, we must recognize that we are all social scientists now—regardless of in which field our degree is. ■

---

#### NOTE

1. Of course, there are other important data collections in the study of the US Congress that have many of the same characteristics, including the Policy Agendas Project (Jones, Wilkerson, and Baumgartner 2009) and the Congressional Bills Project (Adler and Wilkerson 2014).

---

#### REFERENCES

- Adler, E. Scott, and John Wilkerson. 2014. “Congressional Bills Project.” Available at [www.congressionalbills.org](http://www.congressionalbills.org). Accessed August 1, 2014.
- Bond, Robert, and Solomon Messing. “Quantifying Social Media’s Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook.” Stanford University Unpublished Manuscript.
- Bonica, Adam. 2014. “Mapping the Ideological Marketplace.” *American Journal of Political Science* 58 (2): 367–87.
- Boydston, Amber. 2013. *Making the News: Politics, the Media, and Agenda Setting*. Chicago: University of Chicago Press.
- Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. 2009. “Reading Tea Leaves: How Humans Interpret Topic Models.” In *Neural Information Processing Systems Proceedings*, 288–96.
- Clinton, Joshua D., and Simon Jackman. 2009. “To Simulate or NOMINATE?” *Legislative Studies Quarterly* 34 (4): 593–621.
- Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. “The Statistical Analysis of Roll Call Data.” *American Political Science Review* 98 (02): 355–70.
- Davenport, Lauren. 2014. “Politics between Black and White.” Redwood City, CA: Stanford University Unpublished Manuscript.
- Fowler, James H., Michael T. Heaney, David W. Nickerson, John F. Padgett, and Sinclair Betsy. 2011. “Causality in Political Networks.” *American Politics Research* 2: 437–80.
- Gerring, John. 2012. “Mere Description.” *British Journal of Political Science* 42 (4): 721–46.
- Green, Donald P., and Holger L. Kern. 2012. “Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees.” *Public Opinion Quarterly* 76 (3): 491–511.

- Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97.
- Hazlett, Chad. 2014. "Kernel Balancing (KBAL): A Balancing Method to Equalize Multivariate Distance Densities and Reduce Bias without a Specification Search." Cambridge, MA: MIT Unpublished Manuscript.
- Higgins, Michael J., and Jasjeet S. Sekhon. 2014. "Improving Experiments by Optimal Blocking: Minimizing the Maximum Within-Block Distance." Berkeley: University of California Unpublished Manuscript.
- Ho, Dan, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15 (3): 199–236.
- Imai, Kosuke, Gary King, and Elizabeth Stuart. 2008. "Misunderstandings between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (2): 481–502.
- Imai, Kosuke, and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *Annals of Applied Statistics* 7 (1): 443–70.
- Jones, Bryan, John Wilkerson, and Frank Baumgartner. 2009. "The Policy Agendas Project." Available at <http://www.policyagendas.org>. Accessed August 1, 2014.
- Krehbiel, Keith. 1998. *Pivotal Politics: A Theory of US Lawmaking*. Chicago: University of Chicago Press.
- Lee, David, Enrico Moretti, and Matthew Butler. 2004. "Do Voters Affect or Elect Policies? Evidence from the US House." *Quarterly Journal of Economics* 119 (3): 807–59.
- Lee, Frances. 2008. "Dividers, Not Uniters: Presidential Leadership and Senate Partisanship, 1981–2004." *Journal of Politics* 70 (4): 914–28.
- McCarty, Nolan, Keith Poole, and Howard Rosenthal. 2006. *Polarized America: The Dance of Inequality and Unequal Riches*. Cambridge, MA: MIT Press.
- . 2009. "Does Gerrymandering Cause Polarization?" *American Journal of Political Science* 53 (3): 666–80.
- Monroe, Burt L., Jennifer Pan, Margaret E. Roberts, Maya Sen, and Betsy Sinclair. 2015. "No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science." *PS: Political Science and Politics* 48 (1): this issue.
- Moore, Ryan T., and Sally A. Moore. 2013. "Blocking for Sequential Political Experiments." *Political Analysis* 21 (4): 507–23.
- Nagler, Jonathan, and Joshua Tucker. 2015. "Drawing Inferences and Testing Theories with Big Data." *PS: Political Science and Politics* 48 (1): this issue.
- Patty, John, and Elizabeth Maggie Penn. 2015. "Analyzing Big Data: Social Choice and Measurement." *PS: Political Science and Politics* 48 (1): this issue.
- Poole, Keith. 1984. "Least Squares Metric, Unidimensional Unfolding." *Psychometrika* 49 (3): 311–23.
- Poole, Keith, and Howard Rosenthal. 1984. "The Polarization of American Politics." *Journal of Politics* 46 (4): 1061–79.
- . 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29 (2): 357–84.
- . 1997. *Congress: A Political-Economic History of Roll Call Voting*. Oxford: Oxford University Press.
- Quinn, Kevin et al. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209–27.
- Roberts, Margaret E. et al. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science*. DOI 10.1111/ajps.12103.
- Rosenbaum, Paul R., and Donald R. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.
- Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12: 487–508.
- Theriault, Sean M. 2008. *Party Polarization in Congress*. Cambridge: Cambridge University Press.

*Annual Review of Sociology*

# Data ex Machina: Introduction to Big Data

David Lazer<sup>1,2</sup> and Jason Radford<sup>1,3</sup>

<sup>1</sup>Department of Political Science and College of Computer and Information Science, Northeastern University, Boston, Massachusetts 02115; email: d.lazer@neu.edu, j.radford@neu.edu

<sup>2</sup>Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts 02138

<sup>3</sup>Department of Sociology, University of Chicago, Chicago, Illinois 60637

Annu. Rev. Sociol. 2017. 43:19–39

First published as a Review in Advance on May 17, 2017

The *Annual Review of Sociology* is online at soc.annualreviews.org

<https://doi.org/10.1146/annurev-soc-060116-053457>

Copyright © 2017 by Annual Reviews.  
All rights reserved



ANNUAL  
REVIEWS **Further**

Click here to view this article's  
online features:

- Download figures as PPT slides
- Navigate linked references
- Download citations
- Explore related articles
- Search keywords

## Keywords

big data, computational social science, technology, research methodology, quantitative sociology, research ethics, research design, social media, CDR data, data linkage, networks, mobility

## Abstract

Social life increasingly occurs in digital environments and continues to be mediated by digital systems. Big data represents the data being generated by the digitization of social life, which we break down into three domains: digital life, digital traces, and digitalized life. We argue that there is enormous potential in using big data to study a variety of phenomena that remain difficult to observe. However, there are some recurring vulnerabilities that should be addressed. We also outline the role institutions must play in clarifying the ethical rules of the road. Finally, we conclude by pointing to a number of nascent but important trends in the use of big data.

## INTRODUCTION

Archives of human activity go back millennia; however, the increasingly comprehensive digital archives of human behavior, combined with the exponential growth of computational power, create the potential for a transformation of fields such as sociology. Core constructs of sociology, such as interaction, collective action, expression, and diffusion of behavior lurk in these archives. It is possible to study the connectivity of entire human societies, including who communicates with whom about what, how people move through space, who says what, and who buys what, all with a temporal granularity of seconds. The coming generation will witness a transformation of sociological theory through these improvements in our ability to observe dynamic social systems (boyd & Crawford 2012, Golder & Macy 2014).

Big data does pose distinctive challenges for scholars. These digital archives are not the product of scientific design. The information captured in these archives is not what a social scientist would choose. Further, what is captured is constantly, and sometimes abruptly, changing. The signal in big data is vulnerable to manipulation, sometimes purposive, sometimes incidental. Further, relevant behaviors are split into many archives, with no practical way of conjoining them. For example, there may not be a strong conceptual or empirical distinction or boundary to be drawn between behaviors captured by different cell phone carriers, but most research on cell phone data is based on data from a single carrier.

Big data also presents enormous institutional challenges to sociology. The large majority of sociologically relevant analysis of big data is done by computer scientists, and there is relatively little reflection of the big data revolution in top sociology journals. Only 6 of 182 articles published between 2012 and 2016 in the *American Journal of Sociology* (*AJS*) and 9 of 240 in *American Sociological Review* (*ASR*) involve the use of big data.<sup>1</sup>

The objective of this review, therefore, is to critically assess the potential of big data for the collective intellectual endeavor that is sociology. In what follows, we provide a brief overview of what big data is and then review recent big data projects. We break these projects down into three types of big data, enumerate the promises and pitfalls common across them, and offer guidance to sociology for moving forward. We conclude by drawing together the challenges that remain across all three areas and outline the work that needs to be done to put the field on more solid ethical, methodological, and epistemic ground. Our hope is to demonstrate the value of big data research while articulating its most pressing problems and offering reasonable strategies for advancing the field.

## WHAT IS BIG DATA?

The term “big data” has proliferated across society and owes its recent popularization to a report by McKinsey (Manyika et al. 2011). The issue of scale of data is relevant across the academy, from astronomy to the humanities. The manifestation of big data, of course, varies. In astronomy, big data takes the form of images that take petabytes of storage; in the humanities, it can take the form of millions of digitized books.

The dominant data paradigm for the quantitative social sciences has been tabular—variables in columns, cases in rows. Big sociological data, in contrast, comes in an infinite array of other

<sup>1</sup>For our review, we looked for any studies using big data sources or existing data with big data methods such as automated text analysis. We especially looked for social media data, data from online platforms like crowdfunding and online dating websites, data from smartphones including CDR data, and large scale administrative data sets. At *AJS*, these works include Burt (2012), Knigge et al. (2014a), Legewie (2016), Legewie & Schaeffer (2016), Lin & Lundquist (2013), and de Vaan et al. (2015). For *ASR*, the relevant literature includes Bail (2012), Curington et al. (2015), Diekmann et al. (2014), Goldberg et al. (2016), Hall et al. (2015), Knigge et al. (2014b), Leung (2014), van de Rijt et al. (2013), and Vasi et al. (2015).

forms, such as pictures, video, tweets, text, space, networks, streaming, etc. In early formulations, its “bigness” meant that the data could not be processed with extant software (Gartner 2011, Manovich 2012). Big data was big to the extent that new technologies had to be created by specialists to collect, store, and analyze it. The term big data thus refers to data that are so large (volume), complex (variety), and/or variable (velocity) that the tools required to understand them must first be invented (Laney 2001, Monroe 2013). Big data thus requires a computational social science—a mash-up of computer science for inventing new tools to deal with complex data, and social science, because the substantive substrate of the data is the collective behavior of humans (Lazer et al. 2009).

The big data challenges of today have analogues in history. Two generations ago, running multivariate linear models required punch cards and a computer the size of a basement. The challenge was to build tools that leveraged ever-increasing processing power. Today, the technical and theoretical challenges confronting big data research are different, stemming from the complex and heterogeneous form of contemporary big data. These challenges, too, can be met. Sociologists must learn how to cull through large amounts of unstructured text data, program mobile phones, and build data pipelines that scrape, process, store, and make available large amounts of data of wide varieties and types.

## Big Data Sources

Big data is vast and heterogeneous, encompassing everything from YouTube to digital archives of books. A comprehensive review of all big data sources is beyond the scope of this review. There are many discrete literatures around different big data sources, and even a complete list of those literatures would soon be obsolete. Instead, we offer a fuzzy typology of big data sources, based on the loci of data collection. We begin with a discussion of digital life—the capturing of digitally mediated social behaviors—that likely accounts for the majority of big data research. We then discuss digital trace data, the archival exhaust of the modern bureaucratic organization, and conclude with digitalized life data, the movement of intrinsically analog behavior into digital form.

**Digital life.** An increasing fraction of life is intrinsically digitally mediated. Twitter, Facebook, and Wikipedia are all platforms where behaviors are all online. These behaviors may relate to offline occurrences, but behaviors like tweeting are inherently digital. Behaviors on such platforms are typically substantially captured by platform owners, because their business models rely on inferences (such as ad targeting) that can be made from these data streams. Further, it is possible for third parties to harvest data from these platforms. Facebook allows users to download portions of their data; the entire edit history of Wikipedia can be downloaded for analysis; Google allows some access to search query volume. Twitter data are the most used by scholars because of their accessibility. Third parties can negotiate and pay for access for large samples of Twitter data or harvest sizable targeted subsamples.

Digital platform data may be viewed in two ways. The first is to view these platforms as generalizable microcosms of society (Tufekci 2014). Kossinets & Watts (2006) examine email to study the role of different social foci in emergent network structure. Barberá et al. (2015) use Twitter to study political mobilization. The dissemination of news and rumors has been studied on Wikipedia (Keegan et al. 2013, Keegan & Brubaker 2015), Twitter (Bakshy et al. 2011, Romero et al. 2011, Yang & Counts 2010), and across platforms (Goel et al. 2012, Kim et al. 2014). Online markets such as Airbnb and Kickstarter have been used to study patterns of social inequality (Edelman & Luca 2014, Greenberg & Mollick 2017). The Billion Prices Project uses price data scraped from online retailers to track inflation (Cavallo & Rigobon 2016). In this vein too is

Google Flu Trends, which sought to track cases of flu using Google search data (Ginsberg et al. 2009).

The second way to view these platforms is as distinctive realms in which much of the human experience now resides. Exemplary in this regard is the research on whether Facebook creates or accentuates an informational filter around individuals such that people only see ideologically compatible content (Bakshy et al. 2015; also see Lazer 2015). The objective of this article is not to resolve the filter bubble question more generally—the study's finding of only moderate filtering effects tells us nothing about Google or Twitter, for example. And its entire scientific relevance rests on the (correct) assumption that Facebook is, by itself, general enough to be worthy of study.

We note that not all behavior on a platform may be captured. Indeed, the entire Internet may be viewed as an enormous platform for human expression and, although there are efforts to capture records of the Internet, most notably by the Internet Archive, generally these capture only snapshots of particular moments for particular URLs.

Further, as we discuss below, the mapping between what happens on these platforms and the phenomena of interest to a sociologist may be somewhat weak. All friends are not Facebook friends, and not all Facebook friends are friends.

**Digital traces.** The modern complex organization creates a steady output of records that chronicle actions taken (sometimes labeled metadata). Call detail records (CDRs) from phone calls are illustrative (Onnela et al. 2007, Toole et al. 2015). CDRs from cell phones, for example, typically offer time stamps and duration of calls, identifiers for initiator and recipient of calls, and identifiers for the cell towers accessed during the call. Such information has been used to critically examine the strength of weak ties (Onnela et al. 2007), predict individual and collective level unemployment (Toole et al. 2015), and model the spread of malaria (Wesolowski et al. 2012). Governmental data, such as voter records, political contribution data (Bonica 2014), and tax data (Chetty et al. 2016) are other examples. What distinguishes trace data from digital life is that the trace is only a record of the action, not the action itself.

**Digitalized life.** Lastly, digitalized life represents the capture of nonintrinsically digital life (i.e., most of life) in digital form. Thus, for example, phones can be programmed to continuously identify nearby Bluetooth devices, thus capturing the proximity of individuals (Eagle et al. 2009). The constant video recording of major modern cities creates ongoing records of human interaction. Informational objects that predate computers can be easily scanned into manipulable digital form. Exemplary in this regard are Google Books and related endeavors (Michel et al. 2011), which involved scanning millions of books, as well as the use of newspaper data to study the dynamics of fame (van de Rijt et al. 2013).

**Instrumentation of human behavior.** Crosscutting these three types of big data is the possibility for the proactive and purposive instrumentation of human behavior. It is possible to identify and collect data on subsets of behaviors: to monitor a stream of certain types of tweets, track web browsing behaviors from particular locations, or scrape select data from select websites, such as with the Billion Prices Project (Cavallo & Rigobon 2016). Smartphones can be instrumented to collect a wide array of ambient data (e.g., Eagle et al. 2009). Specially designed hardware, so-called sociometers, can track minute details regarding face-to-face social interactions, and have been used, for example, to study the role of gender in collaboration (Onnela et al. 2014).

These examples highlight the variety of uses for these new data. These projects rely heavily on a new breed of tools and techniques, largely from computer science.

## Summary

Big data represents both the new kinds of digital data available to sociologists as well as the tools and technologies required to access these data. The explosion of applications of big data to long-standing social questions reveals a variety of important opportunities for sociologists to extend our knowledge. In the following, we outline these opportunities and provide suggestions for taking advantage of them. These applications, too, both in their successes and their failures, reveal new challenges for big data research, which must be addressed if we are to advance in a responsible, equitable, and scientifically sound way.

## OPPORTUNITIES

Our review of the literature reveals a set of distinct opportunities present in a world replete with big data. Much of these data are what is generally called massive, passive data: data generated in the process of meaningful social behavior rather than data reported for research. The prevalence of such systems and digital devices means that whole systems are captured by these data. And because these systems are always running, they act as controls for experiments offline, online, and even in the lab. One final opportunity in big data comes from making big data small by finding the special populations within it (Foucault Welles 2014).

### Massive, Passive: Behavioral Data at Scale

In principle, big data archives offer measures of actual behaviors, as compared with self-reports of behaviors. The literature is rife with evidence of the problems with self-reported behavior. Generally, self-reported behavior is noisy, with a variety of systemic biases. For example, people systematically lie about everything from whether they voted, to what their weight and height are. Certain types of behavior are entirely inaccessible via self-reports.

To focus on one example, social network researchers have long struggled to accurately measure social ties (Bernard et al. 1984, Marsden 1990). Respondents are biased in the ties they remember and how they remember them. Respondents provide very different networks when asked who they go to for advice, who they would ask for a loan, who their friends are, and who they spend the most time with. And the same question may be interpreted differently by different people. These differences present a range of problems for survey-based social network research, including ambiguity in defining what a tie means, difficulty interpreting concepts such as social influence, and uncertainty about omitted ties.

Behavioral data enable us to observe social networks through interactions. Eagle et al. (2009) compare self-report data to behavioral data. They handed out 94 smartphones to students for nine months. The phones were programmed with applications that “recorded and sent the researcher data about call logs, Bluetooth devices in proximity of approximately five meters, cell tower IDs, application usage, and phone status” (2009, p. 15274). The researchers constructed a behavioral measure of participants’ social networks using the call log data and spatial proximity captured via Bluetooth. Participants were behaviorally related if they talked to one another or were in physical proximity. In the middle of the study, participants filled out a questionnaire about their physical and social proximity to one another. The survey constituted the self-report measure of students’ networks.

Comparing reported physical proximity to behavioral measures of proximity, Eagle et al. find recency and salience biases in subjects’ recall. People tended to overreport physical proximity in general. Additionally, friends were much more accurate in recalling their physical proximity to

one another than those who did not consider one another friends. The researchers also find that subjects who reported being friends behaved in stable and substantially different ways from those who reported not being friends. Friends were much more likely to be spatially proximate to one another at night or on the weekends even if they were at work or somewhere else. Moreover, self-reported friends in January exemplified this pattern five months later in May.

This study helps validate passive network inference by showing that behavioral measures can capture self-reported friendship. At the same time, it reveals a substantial shortcoming of self-reported measures in capturing our weak ties—the colleagues and acquaintances who we see on a regular basis but who we do not consider close friends.

The Copenhagen Network Study (Stopczynski et al. 2014a,b) extends behavioral research on social networks by comparing different modes of behavioral inference. Researchers handed out 1,000 phones to students entering the Technical University of Denmark in 2012 and 2013. Researchers used the phones to infer Bluetooth proximity, geographical proximity via GPS, and interaction via calls and text messages. They combined this with students' Facebook data, their proximity to routers, and qualitative field observations from an anthropologist. They find that nearly the entire call network is captured by Bluetooth proximity, and 80% of students' Facebook friends could be captured by Bluetooth proximity. Yet only 20% of students' Facebook friend network was captured by call logs. Stopczynski et al. conclude that passive behavioral measures are not interchangeable and different digital systems can lead to different social networks with different properties. That is, there is not a single social network for anyone, but a series of shifting networks based on the organizations and technologies individuals use to form and sustain relationships. Thus, the choice of which systems to use to collect data from subjects and how to integrate these different data sources will likely affect study results.

### Nowcasting

Monitoring social phenomena is an essential part of social science research. Surveys such as the University of Michigan's Survey of Consumers or the Bureau of Labor Statistics' Current Employment Statistics survey are fielded regularly to monitor the economic health of the United States. The Centers for Disease Control and Prevention and World Health Organization use a network of testing labs, healthcare providers, and government agencies to generate regular estimates of flu prevalence and virulence.

These regularly updated, vital statistics monitor phenomena that are essential to the workings of contemporary institutions. However, these surveillance systems are very expensive to operate, time consuming to deploy, and inaccurate at high levels of temporal and geographic granularity. Through the digitization of social life, these phenomena are increasingly becoming visible in big data. Such digitization offers the potential to reduce the costs, improve the accuracy, and increase the scale of societal monitoring.

Researchers have used existing large-scale digital data such as CDR data and scraped web data as sensors for monitoring social phenomena. CDR data have been used to detect unemployment (Toole et al. 2015). Google Flu Trends sought to track cases of flu using Google search data (Ginsberg et al. 2009; although see Lazer et al. 2014). Finally, Beauchamp (2016) and Hopkins & King (2010) use Twitter data to generate estimates of public opinion. Each of these studies represents an attempt to validate the use of big data to generate estimates of socially relevant phenomena. This area of research, called nowcasting, seeks to generate descriptions of the world as they happen.

The Billion Prices Project is one such ambitious project. Using price data scraped from online retailers, Cavallo, Rigobon, and colleagues produce daily estimates of prices that are essential to

everything from exchange rates to inflation to the real value of wages (see Cavallo & Rigobon 2016 for an overview). Much like traditional price indexing, researchers at the Billion Prices Project scrape data for a preselected “bag” of goods from a curated set of online retailers from around the world. They compare the prices among the same goods within retailers within countries every day to compute a daily consumer price index for over 70 countries.

This approach has unearthed several new empirical insights. The standard data used to study price changes showed that price changes were normally distributed, with most changes being small. However, Cavallo (2017) finds that the distribution of price changes is actually bimodal, with large price increases and decreases, and very few small price changes. Cavallo argues that the price imputations made during the construction of standard data produce small but erroneous price changes. In essence, measurement error led to a decade’s worth of poor theory on how prices change.

A second result of the Billion Prices Project has been to document violations of the law of one price (Cavallo et al. 2014). This law asserts that a good should generally have the same price no matter where it is sold, controlling for differences in the value of currency. Cavallo et al. show that the law of one price only holds among countries that share a currency.

Finally, Cavallo (2013) demonstrates the social impact of nowcasting. For example, the government of Argentina began manipulating its estimates of inflation in 2007. Cavallo attempts to identify the extent of manipulation by estimating the divergence between Argentina’s official estimates and those derived from online prices. As a control, Cavallo also compares Consumer Price Indexes computed via online prices to those published by the governments of Brazil, Venezuela, Chile, and Colombia. The results show that although indexes computed using online prices track prices in the control countries, they diverge substantially in the case of Argentina. In 2015, Argentina stopped publishing inflation numbers, and the Billion Prices Project has been used to infer inflation in their place.

The Billion Prices Project demonstrates some of the implications of using big data to estimate social phenomena. First, it acts as an alternative method of estimating social phenomena, illuminating problems in traditional methods. Second, it can act as a measure of something that is otherwise unmeasured or whose measure may be disputed. Finally, nowcasting demonstrates big data’s potential to generate measures of more phenomena, with better geographic and temporal granularity, much more cost-effectively than traditional methods. It is worth noting that this granularity is most useful when fused with traditional methods rather than used as a replacement for them (Lazer et al. 2014).

## Data on Social Systems

Perhaps most exciting about big data is the opportunity to build a science of society, a science that would study society at scale, composed of subsystems and individuals that are dynamically connected in particular ways and locations. For example, State et al. (2015) examine networks based on interactions among millions of people on Twitter and between Yahoo! email users. Both data sets revealed strong intracultural correlations among individuals across the globe.

Analysis of the aggregate activity of Twitter reveals expected and unexpected patterns in global behavior. For example, there are strong diurnal patterns across the globe (Golder & Macy 2011). People are generally happier in the morning and earlier in the week, but these patterns vary substantially with work and season and differ systematically for so-called night owls. Dodds et al. (2011) show that big data can capture annual emotional cycles, particularly around holidays, as well as divergences during global events such as the financial crisis.

Data on systems have been used to answer long-standing questions about human mobility. Online data and data from mobile phones have been used to characterize human mobility (Brockmann

et al. 2006, González et al. 2008). This mobility information has been linked to geographical, cultural, and political information in order to study patterns of interaction at the national and regional levels (Blanford et al. 2015, Sevtsuk & Ratti 2010).

One such study is that of Toomet et al. (2015), who use cell phone data to examine the patterns of interaction among members of the Estonian majority and Russian-speaking minority members in the Estonian capital of Tallinn. They used CDR data from Estonia's largest mobile provider, Telia Eesti (formerly EMT), to capture individuals' locations to within several hundred meters. When individuals made calls or sent texts in a shared location at a particular time, Toomet et al. counted that as copresence. To distinguish Estonian and Russian-speaking users, they used the preferred language settings that cell phone providers link to a phone's SIM (subscriber identity module) card. Finally, Toomet and colleagues inferred individuals' home and work locations based on their most frequent geographic location during work and evening hours.

Toomet et al. analyze the ethnic dissimilarity of copresent mobile users at work, home, and during free time. They find high similarity (i.e., substantial segregation) among individuals at home and at work. However, they find substantial dissimilarity among individuals during free time periods. This mixing during free time was most prominent in the city's central district, but also existed even in the less integrated suburbs. The findings suggest that although major social institutions engender segregation, people desegregate in less structured social life. The study raises questions for the long history of research on segregation and social inequality, which almost uniformly define segregation as residential segregation (Massey & Denton 1993, Wilson 1987). As Small (2004, ch. 5) suggests, segregation is as much a process of where you live and work as it is the cultural and physical boundaries that shape who can and cannot participate in the life of a neighborhood. This study shows that high-resolution social mobility data offer a more nuanced picture of segregation bridging residence, employment, and community.

### Natural and Field Experiments

Digitally mediated social systems such as Facebook and Twitter and newly digital institutions like the Internal Revenue Service (IRS) capture data irrespective of events in the environment. In addition, the administrators of these systems make innumerable changes to them over time. As such, it is plausible that all kinds of natural experiments may be hidden within large-scale data. Big data offers a milieu for studying the effect of external events on ongoing social processes. For example, Phan & Airolid (2015) use Facebook data to examine how social networks were affected by Hurricane Ike in 2008. Ayers et al. (2011) use search traffic data to determine whether tax increases on cigarettes lead to an increased interest in tobacco cessation.

Big data can capture the effects of experiments in the field through data linkage. Exemplary in this regard is a series of studies by Chetty and collaborators who merge IRS data with data from prior research (Chetty et al. 2014a,b, 2016). Chetty et al. (2016) linked participants in the Moving to Opportunity (MTO) field experiment to their tax filings with the IRS decades later. MTO sought to investigate neighborhood effects by examining the impact of receiving vouchers and moving to wealthier and safer neighborhoods on the economic, social, and psychological well-being of low-income families. By linking IRS data to the MTO data, Chetty et al. (2016) were able to measure the effect of the experimental treatment on children's future earnings. They found that the economic impact only occurred for children who were younger when they moved, indicating that neighborhood effects require long exposure periods to affect children's future earnings.

Big data systems themselves can create natural experiments by changing user behavior through subtle and not-so-subtle changes to their policies and practices. For example, Brown et al. (2010) use changes to eBay's interface to test the effect of suppressing price information, specifically shipping costs, on purchase decisions. Researchers can use changes to policies, designs, or algorithms

as experimental manipulations. The very malleability of digital worlds make them powerful vessels for conducting very large experiments.

Facebook has been the setting for several large-scale field experiments using randomized manipulations of what people could see about their peers to study social influence (Bond et al. 2012, Kramer et al. 2014). For example, Bond et al. conducted a field experiment using all Facebook users in the United States who were over the age of 18 and who logged into Facebook on the date of the US election in 2012. Six hundred thousand users were put into an informational condition and received a message encouraging them to vote, providing information on where to vote, and allowing them to click an “I Voted” button. Sixty million users were randomly put into the social condition and received the same information but with a list of all of their friends who clicked the “I Voted” button. Finally, a control group of six hundred thousand users saw no message at all.

They examined the extent to which study participants actually voted, engaged with the “I Voted” button, or clicked the link to view voter information. To measure actual voting they linked a subsample of the study population to public voting records. They found a direct effect: Those in the informational condition were more likely to vote than those in the control group, and those in the social condition were more likely than those in the informational condition to vote. They also found an indirect effect: Those who had friends in the social condition were 0.255% more likely to vote per friend in the condition and 0.012% more likely to seek out information per friend in the condition. These findings are important, proving that online social networks contribute to the diffusion of offline behavior. It presents another insight beyond this. The effects of the study are minuscule but, in aggregate, still represent hundreds of thousands of voters. Big data systems offer an unprecedented degree of precision in measuring small but meaningful effects.

### Making Big Data Small

The power of big data is often the small data contained within. Although many online platforms such as Reddit, Wikipedia, and Microsoft’s Xbox Live are dominated by young, white, Western men, within their millions of users, one can still find hundreds or thousands of people who are older, nonwhite, female, and non-Western. Studies that “make big data small” (Foucault Welles 2014) either use big data to observe traditionally hard to reach populations or utilize the vast array of very specific kinds of cases to generate robust estimates.

Big data provides access to data on traditionally underrepresented populations. Twitter data have been used to study people who suffer from PTSD, suicidal ideation, and depression (Coppersmith et al. 2014; De Choudhury et al. 2013, 2016). Jackson & Foucault Welles (2016) and Barberá et al. (2015) use Twitter to identify individuals at the center of emergent social movements including Black Lives Matter, the Gezi Park protest in Turkey, Occupy Wall Street, and the Spanish Indignados.

Barberá et al. use Twitter to compare the patterns of mobilization between a successful protest mobilization at Taksim Gezi Park in 2013 (one locus of the Arab Spring) and two unsuccessful mobilizations by Occupy Wall Street and the Indignados in the spring of 2012. They searched Twitter’s public application program interface (API) for keywords and hashtag keywords to collect samples of tweets from the three movements. They also collected two other samples of tweets from widespread, nonprotest activity to act as a control (the 2014 Academy Awards and a year’s worth of tweets related to raising the minimum wage in the United States). They constructed mobilization networks among users who posted messages containing the keywords and hashtags and users who reposted those messages (i.e., retweets).

They find peripheral members of the Gezi Park protest mobilized more people than the core Occupy and Indignados protesters did. They find no such core-periphery patterns in the non-political mobilization cases. They argue that successful mobilization involves getting peripheral

members to recruit still more peripheral people to protest: The mobilized must mobilize others themselves. This adds to contemporary social movement theory, showing that resources and organizational capacity are not enough to mobilize protest. Network diffusion can separate success from failure.

As studies of online protest show, using big data for research takes advantage of the increasing digitization of social life. Social movement studies have largely depended on newspapers and organizational archives because they are the few sources regularly logging social movement activities (Earl et al. 2004). Social media have become an essential mode for people at the margins of society to connect with one another and express thoughts and sentiments that otherwise go unsaid.

A second use for the small data within big data is in the robust estimates they can generate from the many narrow samples contained within. One example of this is the use of big data to generate population estimates. Paramount in this vein is Wang et al.'s (2015) use of surveys from a panel of Xbox users to predict national election polls in the United States. Xbox users skew heavily toward young, white, Western men. However, even though men made up 93% of the panel, the poll of roughly 340,000 still contained roughly 24,000 women. Using multilevel regression and poststratification, Wang et al. generate estimates of state and national polls for "demographic cells" (Wang et al. 2015, p. 981) such as college educated women over 55 in Florida who identify as Republican and voted for McCain in 2008. They aggregate these cell estimates to predict polls at the state and national level. They find that this reweighting of highly skewed big data accurately predicts national and most state-level polls for the presidential race and Obama's eventual election victory.

In addition, large samples contain enough unusual cases to robustly estimate heterogeneous effects. Small data sets are blunt tools able only to detect large average effects. However, many associations of interest in sociology are contingent on individual and contextual factors. The study of intersectionality is premised on the belief that the main forces in society differ by the particular combinations of race, class, gender, sexuality, and other identities (Collins 1998). Big data allows for robustly estimating effects at the intersection. Others are developing algorithms that seek to inductively identify groups for whom effects are especially large (Athey & Imbens 2016, Green & Kern 2012, Imai & Ratkovic 2013, Taddy et al. 2016).

As survey researchers have long known, a well-defined sample of even a small size can tell us more than millions of poorly defined cases (Squire 1988). These studies by Wang et al., Athey & Imbens, and others show that we can recover the power of well-defined, small data from big data. The difference is that big data and big populations like Xbox users already exist, making them substantially more cost efficient to use than traditional sample building methods.

## VULNERABILITIES

The core issue with any data is who and what get represented. With surveys, one might ask, for example, which respondents are accessible, and what they can accurately reveal. The scale of big data sets creates the illusion that they contain all relevant information on all relevant people. However, the difference between big and everything is still infinite, and the core issues of social science research around validity and generalizability still apply. Further, certain big data can be quite brittle, vulnerable to changes in the data generation process and to attacks motivated by the fact that they are materially consequential.

## Generalizability

Big data are almost always convenience samples offering a distinct set of advantages and disadvantages. However, the data now easily available are unlike most convenience samples hitherto

common in the social sciences. Many are often convenience censuses: a complete record of a certain set of individuals or behaviors that match certain criteria. Scale and seeming comprehensiveness of data often obscure major issues regarding inclusion and selection and therefore representativeness and generalizability.

Many big data census efforts aspire to capture all possible data but do so without a systematic sampling frame. For example, projects like EventRegistry (Leban et al. 2014) and GDELT (Leetaru & Schrodt 2013) identify global events from major media sources such as the *New York Times* and Associated Press, news aggregators such as Google News and LexisNexis, and regional news sources, which are added continuously. They crawl as many sources as possible with the goal of creating as near a census of world events as possible. Yet these projects lack principles for sampling on critical features such as geography, publication frequency, journalistic practices, and type of news. This means that different kinds of events like sporting events or weather may be covered in geographically and temporally uneven ways. Beyond this, more data cannot solve the long-standing problems of using news sources to study events (Earl et al. 2004).

“Big data hubris” (Lazer et al. 2014, p. 1203) is the belief that volume can solve all problems. With near-census projects, trying to create a census without a sampling frame causes errors associated with selection, missing data, and thin coverage to be inestimable (Japec et al. 2015). To use an analogy, these near-census projects are like sending millions of people out into the streets to count the population of the United States. You would count a large number of people, but you could not know the kinds of people who are counted twice or not counted at all, and therefore you could not know what kinds of people are over- or underrepresented and to what extent.

However, big data often is a census of a particular, conveniently accessible social world. All of Twitter is a census of Twitter. Data from Kickstarter are a census of Kickstarter. However, even census data have limits when they come from a single platform. Tufekci (2014) notes that Twitter has become to social media scholars what the fruit fly is to biologists—a model organism. Tufekci argues that relying on a single platform produces issues for generalizability unique to model organisms. For example, different social media platforms have different rules for following or friending others and posting or sending messages. Thus, the patterns of friendship and diffusion differ across platforms (Tufekci 2014, p. 507). Different model organisms behave differently and therefore can give divergent results. In addition, platforms differ in the age, gender, race, and class backgrounds of their users (Perrin 2015). Even phone call data can exclude or overrepresent certain populations (Pestre et al. 2016). Not all social processes and forces are well-represented in a given model organism.

Generalizability is always a question of reference: To what do we want to generalize? With big data, the widespread availability of convenient census and near-census data leads many to overstate the reach of their findings. As research using multiple social media platforms shows, the results from one population of users do not necessarily apply to another. And convenient near-census data may play to our big data hubris that volume will trump sampling. The solution is to use data from multiple sources to validate the findings from any one of them. This will take a commitment from scientists to create institutional structures that can provide access to these data sources. In addition, near-census projects need to take sampling more seriously in order to estimate error in the data produced (Japec et al. 2015).

## Too Many Big Data

Being tied to individual platforms presents another problem: when the scientifically relevant behavior spans these platforms. For example, consider trying to measure a simple construct like “who an individual regularly talks with.” People use different modes to interact with different people,

such as text messaging, face-to-face, or by phone. And people can use functionally equivalent tools within a mode, such as cell phones, land lines, Skype, and Google Hangouts to reach different people. Thus, the data on who someone interacts with exist in a variety of different data sets. As society continues to develop new ways for people to interact with one another, our data on social interaction will continue to fragment. We call this the problem of too many big data.

This issue is even worse for many important sociocultural constructs, such as friendship, which are arguably more cognitive and normative than behavioral. How does one observe love, affection, or deceit from cell phone data?

These issues with too many big data are potentially surmountable. They point to the need to create a fusion of emerging computational methods with existing social science methods. For example, one study has shown promise in tracking interaction across digital and physical modes. As we discussed earlier, the Copenhagen Network Study has proven we can track interaction across digital and physical modes. The results provide a rare glimpse into the ways in which these data can overlap or diverge from one another. However, this multimode data collection is expensive and intrusive.

Other work could compare use of a particular platform relative to population averages to understand the potential limits of the platform. It is also plausible that nonbehavioral constructs like friendship, love, and trust are observable in the sense that there may be certain behaviors strongly correlated with certain cognitive constructs. To identify the best signals, we need a Rosetta stone connecting behavioral big data constructs to theoretically motivated social constructs (Margolin et al. 2013).

### Artifacts and Reactivity

The caveats to big data extend beyond the problems of only studying a specific data set. Big data systems are themselves susceptible to various kinds of error and misappropriation. In the next section, we discuss how social forces can manipulate these data in unexpected and difficult-to-detect ways. In this section, we talk about artifacts, the errors and anomalies that systems produce, and reactivity, the changes in data resulting from technical changes rather than underlying changes in behavior.

Platforms do not merely represent data but generate them. In some cases, it can be difficult to distinguish observations resulting from errors in the system from those representing a real change in underlying behavior. For example, in the Google Ngram project, the word “fuck” is used with startling frequency in books published through 1800, and drops to near zero during the 1800s. Upon closer inspection, it is clear that this did not reflect some dramatic shift in social mores, but rather is an artifact of contemporary optical character recognition systematically misinterpreting an archaic version of “s” as an “f.”<sup>2</sup>

When platforms change how they operate, both behavior and the way behavior is recorded can change. Earlier we argued that these changes can act as natural experiments. Here, we show these changes can have negative consequences for science. Google Flu Trends was perhaps one of the most prominent uses of big data, estimating the prevalence of flu using Google search traffic (Ginsberg et al. 2009, Ortiz et al. 2011). However, it seems likely that Google changed its search to make it more useful for finding health-related information, leading people to perform more searches for the flu during the peak of the season. The result was that the number of

<sup>2</sup><http://searchengineland.com/when-ocr-goes-bad-googles-ngram-viewer-the-f-word-59181>.

searches increased relative to the number of flu cases, and Google Flu Trends started dramatically overreporting the flu (Lazer et al. 2014).

Behavior on any platform is part emergent, part mechanical. The hashtag convention on Twitter was not imagined or proposed by the company but instead originated from users who found keywords searchable when they began with the hash. Only later did Twitter make these hashtags into hyperlinks. There will always be changes in a platform, and artifacts inevitably slip in. Social scientists must be aware of and wary of the technical and social history of the platforms from which their data are generated.

### The Ideal User Assumption: Bots, Puppets, and Manipulation

In big data analysis, we often assume that the data have been generated by a specific type of user, most often single, unique people who express themselves honestly through their personal accounts. This ideal user assumption fails to hold under a wide variety of critical circumstances. Many accounts are not operated by human beings. Additionally, users can have multiple accounts, sometimes with the intent to conceal the user's true identity. Finally, people, organizations, and even nation states put platforms to unintended uses. Taken together, the characteristics of the ideal user need to be validated rather than assumed, and nonideal users generating big data need to be studied in their own right.

The first type of violation of the ideal user assumption is when we incorrectly assume all users are human. This is most often violated in the case of robots and organizations. It is likely that robots are present in all big data systems, whether bots on social media (Ferrara et al. 2016), nonplayer characters in online games (Lee & Ramler 2015), or even robocalls in CDR data (Gupta et al. 2015). Many organizations and institutions are present and active in these systems as well, including corporations, governments, and terrorist organizations (McCorriston et al. 2015). Robots and organizations engage on these platforms for a variety of prosocial and antisocial purposes, making them difficult to detect and theorize (Lee et al. 2011, Ferrara 2015).

The ideal-user assumption is also violated when we assume everyone is who they claim to be. Humans misrepresent themselves in a variety of ways outlined by Wang et al. (2006). Individuals can create multiple accounts for different purposes, including alters to express different identities or so-called sock puppets to fabricate support for themselves or their cause (Bu et al. 2013, Zheng et al. 2006). Individuals can also be deceptive about critical aspects of their identity, as in the case of catfishing. Finally, people can use a fraudulent identity, presenting themselves to be another particular person.

The final violation of the ideal user assumption is when users manipulate the platform in unintended ways to achieve surreptitious goals. In analyzing big data, we often assume that the data are being generated by people behaving in good faith. Yet, many users attempt to game the system or create rigged systems (Ferrara 2015). News media have reported cases in which countries and election campaigns have used robots and real people on social media to artificially inflate their own positions and influence (Durgin 2016, Matthew 2015). In 2016, the Chinese peer-to-peer lending platform Ezubao was revealed to be a Ponzi scheme in which 95% of the proposals were fake (Xinhua 2016). Finally, actors, particularly nation states, can directly control these platforms. As King et al. (2014) find, the Chinese government conducts mass censorship of online media to suppress news of any events with the potential to yield extragovernmental political mobilizations such as protests.

Social media platforms that are highly permeable and very cheap to manipulate are easy targets for robots, deception, and manipulation. However, even high cost, impermeable platforms like CDR and newswire data can be subject to this activity. The challenge is that each of these violations

of the ideal user assumption typically occurs under different conditions. Sock puppetry and political manipulation are probably more likely to occur in contentious arenas and topics, whereas fraud and market manipulation are probably likely to occur at the intersection of the financially vulnerable and the financially influential. In analyzing big data, sample control is critical for making inferences about human beings acting in good faith. However, it is important to remember that these forms of nonideal-user behavior are worthy of study on their own.

## RESEARCH ETHICS

There are major ethical issues regarding the acquisition and use of big data for researchers, institutions, and society at large. Some issues are new, but many are new versions of long-standing issues. The problem, however, is that there is no consensus on what the rules should be, and the policies and recommendations set forth by scientific associations vary substantially, often contradicting one another. Rules will eventually become clear, but the risks to researchers, universities, and the public remain high until they do. Further, these ethical issues in turn suggest interesting researchable questions, ranging from issues around reidentification (Sweeney 2002) to the meaning and management of consent by subjects (Stopczynski et al. 2014a,b).

The National Research Council proposed to amend the definition of human subjects research to “a systematic investigation designed to develop or contribute to generalizable knowledge by obtaining data about a living individual directly through interaction or intervention, or by obtaining identifiable private information about an individual” (NRC 2014, recommendation 2.1, p. 40). However, rules regarding obtaining subject consent are typically obviated when the data have been collected by a third party and have been anonymized. Is this an acceptable standard given the prevalence of collaboration between companies and academics? Furthermore, given the ease of deidentifying many kinds of data, what standards can be set for anonymization?

Regulating informed consent, the heart of human subjects research, is only one of the central issues that have yet to be settled. Other open issues include the rights of secondary subjects, measuring harms from the loss of privacy, and regulating the status of leaked data like the Panama Papers, which revealed international tax evasion by the world’s elite, or the data dump of the US adultery website Ashley Madison. The role of the university becomes critical here because it is part of the regulatory apparatus enforcing rules and protecting scholars, but also because it provides the training infrastructure that empowers compliance.

## FUTURE TRENDS

Researchers have used big data to answer old questions in new ways and new questions never before answerable. Their successes and failures have helped us identify the promises and pitfalls of research in this area and the kinds of investments institutions need to make for the future of this research. In this final section, we point to six trends that will likely affect the big data landscape in hopes of helping sociology get ahead of the curve.

### More Data Are Coming

Big data will continue to grow into more domains. For example, EventRegistry provides data on events, and it also acts as a repository providing real-time access to more than 100,000 news publishers (<http://www.EventRegistry.org>). Big data will also continue to reach back into the past as libraries digitize their collections, newspapers digitize their archives, and initiatives such as

Google Books and Project Gutenberg digitize books. The question for data coverage will continue to be less about whether the data exist, but more about what can be studied with the available data.

More linkages between different big data will become more common. The work of Chetty and colleagues (2014a,b, 2016) with IRS data is only the beginning. Another opportunity is connecting online data to offline data, such as linking social media accounts to voter records and data collected by brokers such as Axiom. One revolutionary form of big data linkage is wikification, which involves linking words and phrases in text to entities in Wikipedia (Mihalcea & Csoma 2007). Entity linkage allows researchers to use the structured and unstructured data Wikipedia maintains on entities to enhance the contextual information associated with texts.

### Different Data Are Coming

The majority of data being created on big data platforms are still unusable for social scientific research. These are the images, audio, and video being created, discussed, and shared. The tools for providing meaningful structure to these data at scale have lagged behind those of other types of data, such as text. They are quickly catching up. For example, image processing using convolutional neural nets and other deep learning methods has become as easy to use in Python and R as methods for text analysis. Furthermore, the tools to analyze these data are increasingly being made available through publicly accessible interfaces like Google Cloud Vision API. With publicly accessible models, researchers upload their files to the service, which uses pretrained models to make inferences about the file and then sends these inferences back as metadata.

### Models Will Become More Generic

Google Cloud Vision API is a prime example of another critical trend in machine learning: creating generic models and making them available to the public. There is a long precedent of creating and sharing models for processing unstructured data. The Linguistic Inquiry and Word Count dictionary is perhaps the best known in the social sciences (Tausczik & Pennebaker 2010). However, such models are now being published in a variety of methods. For example, Jozefowicz et al. (2016) trained a deep learning model on the One Billion Word Benchmark and are publishing the model itself as an alternative to the model released in 2014 in Stanford's GloVe (Pennington et al. 2014). Like the Vision API, researchers can use these models on their own texts to generate word embeddings.

Generic models allow researchers to use pretrained machine learning models on their own data, rather than having to deal with the issues of data processing and model specification. These out-of-the-box machine learning projects aspire to use big data to create the most effective models and then make those models the standard for processing unstructured data. However, generic models are not necessarily better at applied tasks than specialized models. And, in the absence of social theory, such generic models may miss obvious social patterns in data, potentially reinforcing long-standing social biases (Caliskan et al. 2017).

### Data from Multiple Platforms Will Become Standard

As big data systems proliferate and multiple systems offer similar services, it will become increasingly possible and easier for researchers to perform studies on different platforms. The CrowdBerkeley project is offering data for multiple crowdfunding websites such as Kickstarter, Indiegogo, and Kiva. Multiple forms of big data will also be used to create parallel measures. For example, new models of political partisanship are being generated by Federal Election

Commission data, Twitter, press releases, and floor speeches (Barberá 2015, Bonica 2014, Gentzkow et al. 2016, Tsur et al. 2015). Interestingly, each of these measures provides different narratives about the emergence of partisanship and demonstrates the importance of using different data to approach the same phenomena.

### Qualitative Approaches to Big Data

While big data is typically viewed as quantitative social science on steroids, we anticipate innovative approaches weaving together qualitative methods and computational approaches to large-scale data. There is a long history of archival research in the social sciences. Digital archives present the challenge of vast amounts of information beyond the capacity of armies of grad students to read. Searching and sorting of archives becomes essential to qualitative understanding. At the simplest level, this might simply require keyword searches, but certainly more complex, computationally enabled approaches will emerge. For example, consider hypothetical research examining the Internet Archive's version of <http://www.congress.gov>. The snapshots of the members' home pages present too large a data set to comprehensively read, but it is feasible to query the data for policy statements on health care by every member for targeted reading and hand-coding.

### Methodological Integration

The prior point highlights a more general lesson: Big data will increasingly be integrated with existing research methods in sociology. Big data offers strengths and weaknesses that are quite different than existing data sources (Lazer et al. 2014). The most compelling sociological research in the twenty-first century will not be big data but a fusion of data sources related to important questions. Survey data will be linked to a tiny portion of archival data, providing inferential power to the entire archive that it otherwise would not have. Interesting or typical cases in big data can be identified for qualitative exploration. The scientific payoff should, in turn, be insight into phenomena that heretofore have been neglected, related to the connectivity and dynamics of entire societies.

The future of big data is as bright and fraught as its past. While sociology has generally lagged behind in using big data, there are many opportunities for the field to take advantage of and many challenges and debates to confront. Further, the increasing presence of digitally mediated social activity and the increasingly digital social life mean that the need to integrate big data approaches into sociology will increase for the foreseeable future, with the corresponding need for sociologists to contribute to our understanding of an increasingly digital and digitalized world.

### DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

### ACKNOWLEDGMENTS

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF0920053 (the ARL Network Science CTA) and in part by a grant from the US Army Research Office W911NF1210556. Any opinions expressed are the authors' alone.

## LITERATURE CITED

- Athey S, Imbens G. 2016. Recursive partitioning for heterogeneous causal effects. *PNAS* 113(27):53–60
- Ayers JW, Ribisl K, Brownstein JS. 2011. Using search query surveillance to monitor tax avoidance and smoking cessation following the United States' 2009 “SCHIP” cigarette tax increase. *PLOS ONE* 6(3):e16777
- Bail CA. 2012. The fringe effect. *Am. Sociol. Rev.* 77(6):55–79
- Bakshy E, Hofman JM, Mason WA, Watts DJ. 2011. Everyone's an influencer: quantifying influence on twitter. *Proc. 4th ACM Conf. Web Search Data Mining*, pp. 65–74. New York: ACM
- Bakshy E, Messing S, Adamic LA. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348(6239):1130–32
- Barberá P. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Polit. Anal.* 23(1):76–91
- Barberá P, Wang N, Bonneau R, Jost JT, Nagler J, et al. 2015. The critical periphery in the growth of social protests. *PLOS ONE* 10(11):1–15
- Beauchamp N. 2016. Predicting and interpolating state-level polls using Twitter textual data. *Am. J. Political Sci.* 61:490–503
- Bernard HR, Killworth P, Kronenfeld D, Sailer L. 1984. The problem of informant accuracy: the validity of retrospective data. *Annu. Rev. Anthropol.* 13:495–517
- Blanford JI, Huang Z, Savelyev A, MacEachren AM. 2015. Geo-located tweets. Enhancing mobility maps and capturing cross-border movement. *PLOS ONE* 10(6):e0129202
- Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, et al. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–98
- Bonica A. 2014. Mapping the ideological marketplace. *Am. J. Polit. Sci.* 58(2):367–86
- boyd d, Crawford K. 2012. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inf. Commun. Soc.* 15(5):662–79
- Brockmann D, Hufnagel L, Geisel T. 2006. The scaling laws of human travel. *Nature* 439(7075):462–65
- Brown J, Hossain T, Morgan J. 2010. Shrouded attributes and information suppression: evidence from the field. *Q. J. Econ.* 125(2):859–76
- Bu Z, Xia Z, Wang J. 2013. A sock puppet detection algorithm on virtual spaces. *Knowl.-Based Syst.* 37:366–77
- Burt RS. 2012. Network-related personality and the agency question: multirole evidence from a virtual world. *Am. J. Sociol.* 118(3):543–91
- Caliskan A, Bryson JJ, Narayanan A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–86
- Cavallo A. 2013. Online and official price indexes: measuring Argentina's inflation. *J. Monet. Econ.* 60(2):152–165
- Cavallo A. 2017. Scrapped data and sticky prices. *Rev. Econ. Stat.* In press. [http://dx.doi.org/10.1162/REST\\_a\\_00652](http://dx.doi.org/10.1162/REST_a_00652)
- Cavallo A, Neiman B, Rigobon R. 2014. Currency unions, product introductions, and the real exchange rate. *Q. J. Econ.* 129(2):529–95
- Cavallo A, Rigobon R. 2016. The Billion Prices Project: using online prices for measurement and research. *J. Econ. Perspect.* 30(2):151–78
- Chetty R, Friedman JN, Rockoff JE. 2014a. Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates. *Am. Econ. Rev.* 104(9):2593–632
- Chetty R, Friedman JN, Rockoff JE. 2014b. Measuring the impacts of teachers II: teacher value-added and student outcomes in adulthood. *Am. Econ. Rev.* 104(9):2633–79
- Chetty R, Hendren N, Katz LF. 2016. The effects of exposure to better neighborhoods on children: new evidence from the Moving to Opportunity experiment. *Am. Econ. Rev.* 106(4):855–902
- Collins PH. 1998. It's all in the family: intersections of gender, race, and nation. *Hypatia* 13(3):62–82
- Coppersmith G, Harman C, Dredze M. 2014. Measuring post traumatic stress disorder in Twitter. *Proc. 8th Int. AAAI Conf. Weblogs Soc. Media*, pp. 579–82. <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8079>
- Curington CV, Lin K-H, Lundquist JH. 2015. Positioning multiraciality in cyberspace. *Am. Sociol. Rev.* 80(4):764–88

- De Choudhury M, Gamon M, Counts S, Horvitz E. 2013. Predicting depression via social media. *Proc. 7th Int. AAAI Conf. Weblog Soc. Media*, pp. 128–37. [http://course.duruofei.com/wp-content/uploads/2015/05/Choudhury\\_Predicting-Depression-via-Social-Media\\_ICWSM13.pdf](http://course.duruofei.com/wp-content/uploads/2015/05/Choudhury_Predicting-Depression-via-Social-Media_ICWSM13.pdf)
- De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *Proc. 2016 CHI Conf. Hum. Factors Comput. Syst.*, pp. 2098–2110. New York: ACM Press
- De Vaan M, Vedres B, Stark D. 2015. Game changer: the topology of creativity. *Am. J. Sociol.* 120(4):1144–94
- Diekmann A, Jann B, Przepiorka W, Wehrli S. 2014. Reputation formation and the evolution of cooperation in anonymous online markets. *Am. Sociol. Rev.* 79(1):65–85
- Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM. 2011. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. *PLOS ONE* 6(12):e26752
- Durgin C. 2016. Inside Donald Trump's Potemkin Twitter army. *National Review*, Apr. 8. <http://www.nationalreview.com/article/433870/donald-trumps-twitter-supporters-might-be-fake>
- Eagle N, Pentland AS, Lazer D. 2009. Inferring friendship network structure by using mobile phone data. *PNAS* 106(36):15274–78
- Earl J, Martin A, McCarthy JD, Soule SA. 2004. The use of newspaper data in the study of collective action. *Annu. Rev. Sociol.* 30(1):65–80
- Edelman BG, Luca M. 2014. *Digital discrimination: the case of Airbnb.com*. Working Pap. 14–054, NOM Unit, Harvard Bus. Sch.
- Ferrara E. 2015. “Manipulation and abuse on social media” by Emilio Ferrara with Ching-man Au Yeung as coordinator. *ACM SIGWEB Newsletter*, Spring 4:1–9
- Ferrara E, Varol O, Davis C, Menczer F, Flammini A. 2016. The rise of social bots. *Commun. ACM* 59(7):96–104
- Foucault Welles B. 2014. On minorities and outliers: The case for making big data small. *Big Data Soc.* 1(1):1–2
- Gartner. 2011. *Gartner says solving “big data” challenge involves more than just managing volumes of data*. News Release, June 27. <http://www.gartner.com/newsroom/id/1731916>
- Gentzkow M, Shapiro JM, Taddy M. 2016. Measuring polarization in high-dimensional data: method and application to congressional speech. NBER Work. Pap. 22423, Natl. Bur. Econ. Res., Cambridge, MA
- Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–14
- Goel S, Watts DJ, Goldstein DG. 2012. The structure of online diffusion networks. *Proc. 13th ACM Conf. Electron. Commer.*, pp. 623–38. New York: ACM
- Goldberg A, Srivastava SB, Manian VG, Monroe W, Potts C. 2016. Fitting in or standing out? The tradeoffs of structural and cultural embeddedness. *Am. Sociol. Rev.* 81(6):1190–222
- Golder SA, Macy MW. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–81
- Golder SA, Macy MW. 2014. Digital footprints: opportunities and challenges for online social research. *Annu. Rev. Sociol.* 40:129–52
- González MC, Hidalgo CA, Barabási A-L. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–82
- Green DP, Kern HL. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin. Q.* 76(3):491–511
- Greenberg J, Mollick E. 2017. Activist choice homophily and the crowdfunding of female founders. *Adm. Sci. Q.* 62:341–74
- Gupta P, Srinivasan B, Balasubramaniyan V, Ahamad M. 2015. *Phoneypot: data-driven understanding of telephony threats*. Brief. Pap., NDSS Symp. 2015, San Diego, CA
- Hall M, Crowder K, Spring A. 2015. Neighborhood foreclosures, racial/ethnic transitions, and residential segregation. *Am. Social. Rev.* 80(3):526–49
- Hopkins DJ, King G. 2010. A method of automated nonparametric content analysis for social science. *Am. J. Political Sci.* 54(1):229–47
- Imai K, Ratkovic M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7(1):443–70

- Jackson SJ, Foucault Welles B. 2016. #Ferguson is everywhere: initiators in emerging counterpublic networks. *Inf. Commun. Soc.* 19(3):397–418
- Japec L, Kreuter F, Berg M, Biemer P, Decker P, et al. 2015. *AAPOR report on big data*. Am. Assoc. Public Opin. Res., Oakbrook Terrace, IL
- Jozefowicz R, Vinyals O, Schuster M, Shazeer N, Wu Y. 2016. Exploring the limits of language modeling. arXiv:1602.02410 [cs.CL]
- Keegan BC, Brubaker JR. 2015. “Is” to “was”: coordination and commemoration in posthumous activity on Wikipedia biographies. *Proc. 18th ACM Conf. Comput. Support. Coop. Work Soc. Comput.*, pp. 533–46. New York: ACM
- Keegan BC, Gergle D, Contractor N. 2013. Hot off the wiki: structures and dynamics of Wikipedia’s coverage of breaking news events. *Am. Behav. Sci.* 57(5):595–622
- Kim M, Newth D, Christen P. 2014. Trends of news diffusion in social media based on crowd phenomena. *Proc. 23rd Int. Conf. World Wide Web*, pp. 753–58. New York: ACM
- King G, Pan J, Roberts ME. 2014. Reverse-engineering censorship in China: randomized experimentation and participant observation. *Science* 345(6199):1–10
- Knigge A, Maas I, van Leeuwen MHD. 2014a. Sources of sibling (dis)similarity: total family impact on status variation in the Netherlands in the nineteenth century. *Am. J. Sociol.* 120(3):908–48
- Knigge A, Maas I, van Leeuwen MHD, Mandemakers K. 2014b. Status attainment of siblings during modernization. *Am. Sociol. Rev.* 79(3):549–74
- Kossinets G, Watts DJ. 2006. Empirical analysis of an evolving social network. *Science* 311(5757):88–90
- Kramer ADI, Guillory JE, Hancock JT. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *PNAS* 111(29):8788–90
- Laney D. 2001. *3D data management: controlling data volume, velocity and variety*. Res. Note, META Group, Stamford, CT. <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>
- Lazer D. 2015. The rise of the social algorithm. *Science* 348(6239):1090–91
- Lazer D, Kennedy R, King G, Vespignani A. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343(6176):1203–5
- Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, et al. 2009. Life in the network: the coming age of computational social science. *Science* 323(5915):721
- Leban G, Fortuna B, Brank J, Grobelnik M. 2014. Event Registry: learning about world events from news. *Proc. 23rd Int. Conf. World Wide Web*, pp. 107–10. New York: ACM
- Lee C-S, Ramler I. 2015. Rise of the bots: bot prevalence and its impact on match outcomes in League of Legends. *Int. Worksh. Netw. Syst. Support Games (NetGames)*, Zagreb, Dec. 3–4, pp. 1–6
- Lee K, Eoff BD, Caverlee J. 2011. Seven months with the devils: a long-term study of content polluters on Twitter. *5th Int. AAAI Conf. Weblogs Soc. Media*. <https://pdfs.semanticscholar.org/1dd5/355e62b9fc37a355e135d5909ed28128d653.pdf>
- Leetaru K, Schrodt PA. 2013. GDELT: global data on events, location, and tone, 1979–2012. *Int. Stud. Assoc. Annu. Conf., San Diego*. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.686.6605&rep=rep1&type=pdf>
- Legewie J. 2016. Racial profiling and use of force in police stops: how local events trigger periods of increased discrimination. *Am. J. Sociol.* 122(2):379–424
- Legewie J, Schaeffer M. 2016. Contested boundaries: explaining where ethnoracial diversity provokes neighborhood conflict. *Am. J. Sociol.* 122(1):125–61
- Leung MD. 2014. Dilettante or Renaissance person? How the order of job experiences affects hiring in an external labor market. *Am. Sociol. Rev.* 79(1):136–58
- Lin K-H, Lundquist J. 2013. Mate selection in cyberspace: the intersection of race, gender, and education. *Am. J. Sociol.* 119(1):183–215
- Manovich L. 2012. Trending: the promises and the challenges of big social data. In *Debates in the Digital Humanities*, Vol. 2, ed. MK Gold, pp. 460–75. Minneapolis, MN: Univ. Minn. Press
- Manyika J, Chui M, Brown B, Bughin J, Dobbs R, et al. 2011. *Big data: the next frontier for innovation, competition, and productivity*. Rep., McKinsey Global Inst. <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>

- Margolin D, Lin Y-R, Brewer D, Lazer D. 2013. Matching data and interpretation: towards a Rosetta stone joining behavioral and survey data. *7th Int. AAAI Conf. Weblogs Soc. Media*, pp. 9–10. <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6267>
- Marsden PV. 1990. Network data and measurement. *Annu. Rev. Sociol.* 16:435–63
- Massey DS, Denton NA. 1993. *American Apartheid: Segregation and the Making of the Underclass*. Cambridge, MA: Harvard Univ. Press
- Matthew S. 2015. Revealed: how Russia’s “troll factory” runs thousands of fake Twitter and Facebook accounts to flood social media with pro-Putin propaganda. *The Daily Mail*, March 28
- McCorriston J, Jurgens D, Ruths D. 2015. Organizations are users too: characterizing and detecting the presence of organizations on Twitter. *9th Int. AAAI Conf. Web Soc. Media*. [http://www-cs.stanford.edu/~jurgens/docs/mccorriston-jurgens-ruths\\_icwsm-2015.pdf](http://www-cs.stanford.edu/~jurgens/docs/mccorriston-jurgens-ruths_icwsm-2015.pdf)
- Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–82
- Mihalcea R, Csomai A. 2007. Wikify!: Linking documents to encyclopedic knowledge. *Proc. 16th ACM Conf. Inf. Knowl. Manag.*, pp. 233–42. New York: ACM
- Monroe BL. 2013. The five Vs of big data political science: introduction to the Virtual Issue on Big Data in Political Science. *Polit. Anal.* 19(5):66–86
- NRC (Natl. Res. Coun.). 2014. *Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences*. Washington, DC: Natl. Acad. Press
- Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, et al. 2007. Structure and tie strengths in mobile communication networks. *PNAS* 104(18):7332–36
- Onnela J-P, Waber BN, Pentland A, Schnorf S, Lazer D. 2014. Using sociometers to quantify social interaction patterns. *Sci. Rep.* 4:5604
- Ortiz JR, Zhou H, Shay DK, Neuzil KM, Fowlkes AL, Goss CH. 2011. Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. *PLOS ONE* 6(4):1–9
- Pennington J, Socher R, Manning CD. 2014. GloVe: global vectors for word representation. *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, pp. 1532–43
- Perrin A. 2015. *Social networking usage: 2005–2015*. Rep., Pew Res. Cent., Washington, DC
- Pestre G, Letouzé E, Zagheni E. 2016. *The ABCDE of big data: assessing biases in call-detail records for development estimates*. Presented at Annu. Bank Conf. Dev. Econ., June 20–21, Washington, DC. <http://pubdocs.worldbank.org/pubdocs/publicdoc/2016/6/551311466182785065/Pestre-Letouze-Zagheni-ABCDE-May-2016.pdf>
- Phan TQ, Airoltdi EM. 2015. A natural experiment of social network formation and dynamics. *PNAS* 112(21):6595–600
- Romero DM, Meeder B, Kleinberg J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. *Proc. 20th Int. Conf. World Wide Web*, pp. 695–704. New York: ACM
- Sevtsuk A, Ratti C. 2010. Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *J. Urban Technol.* 17(1):41–60
- Small ML. 2004. *Villa Victoria: The Transformation of Social Capital in a Boston Barrio*. Chicago: Univ. Chicago Press
- Squire P. 1988. Why the 1936 *Literary Digest* poll failed. *Public Opin. Q.* 52(1):125–33
- State B, Park P, Weber I, Macy M. 2015. The mesh of civilizations in the global network of digital communication. *PLOS ONE* 10(5):e0122543
- Stopczynski A, Pietri R, Pentland A, Lazer D, Lehmann S. 2014a. Privacy in sensor-driven human data collection: a guide for practitioners. arXiv:1403.5299 [cs.CY]
- Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, et al. 2014b. Measuring large-scale social networks with high resolution. *PLOS ONE* 9(4):e95978
- Sweeney L. 2002. K-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* 10(05):557–70
- Taddy M, Gardner M, Chen L, Draper D. 2016. A nonparametric Bayesian analysis of heterogenous treatment effects in digital experimentation. *J. Bus. Econ. Stat.* 34(4):661–72

- Tausczik YR, Pennebaker JW. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29(1):24–54
- Toole JL, Lin Y-R, Muehlegger E, Shoag D, González MC, Lazer D. 2015. Tracking employment shocks using mobile phone data. *J. R. Soc. Interface* 12(107):20150185
- Toomet O, Silm S, Saluveer E, Ahas R, Tammaru T. 2015. Where do ethno-linguistic groups meet? How copresence during free-time is related to copresence at home and at work. *PLOS ONE*. 10(5):e0126093
- Tsur O, Calacci D, Lazer D. 2015. A frame of mind: using statistical models for detection of framing and agenda setting campaigns. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Joint Conf. Nat. Lang. Process., Beijing, July 26–31*, pp. 1629–38. <https://pdfs.semanticscholar.org/f5c8/dbcea0112227486b7fc3bd20a73726ffea88.pdf>
- Tufekci Z. 2014. Big questions for social media big data: representativeness, validity and other methodological pitfalls. arXiv:1403.7400 [cs.SI]
- van de Rijt A, Shor E, Ward C, Skiena S. 2013. Only 15 minutes? The social stratification of fame in printed media. *Am. Sociol. Rev.* 78(2):266–89
- Vasi IB, Walker ET, Johnson JS, Tan HF. 2015. “No fracking way!” Documentary film, discursive opportunity, and local opposition against hydraulic fracturing in the United States, 2010 to 2013. *Am. Sociol. Rev.* 80(5):934–59
- Wang GA, Chen H, Xu JJ, Atabakhsh H. 2006. Automatically detecting criminal identity deception: an adaptive detection algorithm. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* 36(5):988–99
- Wang W, Rothschild D, Goel S, Gelman A. 2015. Forecasting elections with non-representative polls. *Int. J. Forecast.* 31(3):980–91
- Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, et al. 2012. Quantifying the impact of human mobility on malaria. *Science*. 338(6104):267–70
- Wilson WJ. 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Chicago: Univ. Chicago Press
- Xinhua. 2016. Online P2P lender suspected of \$US 7.6 billion fraud. *Xinhua*, Feb. 1. [http://news.xinhuanet.com/english/2016-02/01/c\\_135065022.htm](http://news.xinhuanet.com/english/2016-02/01/c_135065022.htm)
- Yang J, Counts S. 2010. Predicting the speed, scale, and range of information diffusion in Twitter. *ICWSM* 10:355–58
- Zheng R, Li J, Chen H, Huang Z. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *J. Am. Soc. Inf. Sci. Technol.* 57(3):378–93