How to Analyze Political Attention with Minimal Assumptions and Costs[*]

Kevin M. Quinn[†] Burt L. Monroe[‡] Michael Colaresi[§] Michael H. Crespin[¶]
Dragomir R. Radev[||]

July 14, 2008

Abstract

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 United States Senate from 1997-2004. 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 inter-related, and a richer view of democratic agenda dynamics than had previously been possible.

[*] An earlier version of this paper 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, Jake Bowers, Janet Box-Steffensmeier, Patrick Brandt, Barry Burden, Suzie Linn, John Freeman, Will Howell, Simon Jackman, Brad Jones, Bryan Jones, Kris Kanthak, Gary King, Glen Krutz, Frances Lee, Bob Luskin, Andrew Martin, Iain McLean, Nate Monroe, Becky Morton, Stephen Purpura, Phil Schrodt, Gisela Sin, Betsy Sinclair, Michael Ward, John Wilkerson, Dan Wood, Chris Zorn and seminar participants at UC Davis, 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 paper. 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. This paper 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 paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. In addition, Quinn thanks the Center for Advanced Study in the Behavioral Sciences for its hospitality and support.

[†] Associate Professor, Department of Government and Institute for Quantitative Social Science, Harvard University, kevin_quinn@harvard.edu.

[‡] Associate Professor, Department of Political Science, The Pennsylvania State University, burtmonroe@psu.edu.

[§] Associate Professor, Department of Political Science, University of Colorado, michael.colaresi@colorado.edu.

[¶] Assistant Professor, Department of Political Science, University of Georgia, crespin@uga.edu.

[||] Associate Professor, School of Information and Department of Electrical Engineering and Computer Science, University of Michigan, radev@umich.edu.

1 Introduction

What 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. Studies of legislative representation within and across issues (Lowi, 1964; Mayhew, 1974; Riker, 1986), policy agenda change (Baumgartner and Jones, 1993; Kingdon, 1995; Baumgartner et al., 2006), and issue evolution (Carmines and Stimson, 1989; Wolbrecht, 2000) all seek to answer some or all of the questions above.

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 (Jones et al., N.D.; Adler and Wilkerson, 2006), 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.[1] 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 et al. (2007, 2008). To the extent that automated approaches substitute computers for human coders, the costs of coding is reduced and

[1] 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.

the reliability is increased (King and Lowe, 2003) (although both still require substantial human effort in the creation and validation of dictionaries or manually coded training data). 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.

In this paper, 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 key words 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 divisions 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. Instead, 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 US 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 United States 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 key words *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.

The paper proceeds as follows. In Section 2, we place our approach in the context of other approaches, human and computer-assisted, for classifying texts. In Section 3, we describe the details of our model. In Section 4, we discuss the particulars of estimating the model for the Senate, including a brief discussion of the process by which the text of the *Congressional Record* was converted into numerical data. In Section 5 we discuss and interpret the results, with particular attention to evidence for five types of measurement validity. This section includes an application of the model to examine theoretical hypotheses about institutional, career, and electoral influences on legislator behavior.

2 Categorizing Texts: Methods, Assumptions and Costs

There are many methods for content analysis and inferring meaning from text. Each of these imposes its own particular set of assumptions and, as a result, has particular advantages and weaknesses for any given question or set of texts. For present purposes, we focus attention on the basic problem of categorizing texts – placing texts into discrete target categories or bins.[2] 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,

[2] An equally interesting problem is placing texts, or their authors, in a continuous space, the problem addressed by such techniques as WORDSCORES (Laver et al., 2003; Lowe, 2008) and rhetorical ideal point estimation (Monroe and Maeda, 2004; Monroe et al., 2007b).

and (5) whether 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 pre-analysis 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 post-analysis phase where the results from the analysis phase are interpreted and assessed for reliability and validity. Tables 1 and 2 depict how five major methods of text categorization compare in terms of their underlying assumptions and costs respectively. The cell entries in Table 1 represent the minimal assumptions required by each method.

In the most general sense, the fundamental "method" for inferring meaning from text is *reading.* For example, 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 make 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 non-trivial costs, especially when one attempts to read large, politically relevant texts such as the British *Hansard* or the US *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 non-trivial requirements) and a great deal of time in person-hours per text.[3] 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, Jones et al. (N.D.); Budge et al. (2001); Klingemann et al. (2006); Ansolabehere et al. (2003), and Ho and Quinn (2008)) 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 – such as a hierarchical nesting of categories[4] – 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

[3] The *Congressional Record* currently contains over four billion words and produces another half million – about the length of *War and Peace* – a day. Someone capable of reading a million words a day could catch up in 23 years.

[4] A prominent example would be the Policy Agendas Project, which has minor topic codes nested within major topics codes (Jones et al., N.D.).

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. Nevertheless, 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 et al. (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.[5] 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 (oftentimes) a fair amount of trial and error. That said, once

[5] One of the most important early dictionary based systems is the General Enquirer (Stone et al., 1966).

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* (Purpura and Hillard, 2006; Hillard et al., 2008, 2007; Kwon et al., 2007). 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 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 non-parametric.[6]

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

[6] In Table 1 we code the assumptions for the least stringent supervised learning techniques.

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 paper has a very different cost-structure than all methods mentioned so far. Whereas other methods typically require a large investment in the initial pre-analysis 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 post-analysis 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.

Such topic modeling, using "unsupervised learning" methods, is a fast-growing area in computational linguistics in areas of inquiry with similar structural needs: categorization of massive text collections by subject, with minimal assumptions about the underlying categories, and with minimal costs (Blei et al., 2003; Blei and Lafferty, 2006, 2007; Wang and McCallum, 2006; Gruber and Weiss, 2007).

3 A Model for Dynamic Multi-Topic Speech

The data generating process that motivates our model is the following. On each day that Congress is in session a 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.[7]

We begin by laying out the necessary notation. Let $t = 1, \ldots, T$ index time (in days); $d = 1, \ldots, D$ index speech documents; $k = 1, \ldots, K$ index possible topics that a document can be on; and $w = 1, \ldots, W$ index words in the vocabulary. For reasons that will be more clear later, we also introduce the function $s : \{1, \ldots, D\} \rightarrow \{1, \ldots, T\}$. $s(d)$ tells us the time period in which document d was put into the *Congressional Record*. In addition, let Δ^N denote the N-dimensional simplex.

3.1 The Sampling Density

The dth document $\mathbf{y}_d$ is a W-vector of non-negative integers. The wth 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

$$\mathbf{y}_d \sim \mathcal{M}ultinomial(n_d, \boldsymbol{\theta}_k).$$

Here $\boldsymbol{\theta}_k \in \Delta^{W-1}$ is the vector of multinomial probabilities with typical element θ_{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 the transformation

$$\boldsymbol{\beta}_k = \left(\left[\log\left(\frac{\theta_{k1}}{\theta_{k1}}\right) - c \right], \left[\log\left(\frac{\theta_{k2}}{\theta_{k1}}\right) - c \right], \ldots, \left[\log\left(\frac{\theta_{kW}}{\theta_{k1}}\right) - c \right] \right)'$$

[7] The model we describe below differs from the most similar topic models in the computational linguistics literature (Blei et al., 2003; Blei and Lafferty, 2006; 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.

where $c = W^{-1} \sum_{w=1}^{W} \log\left(\frac{\theta_{kw}}{\theta_{k1}}\right)$.

If we let π_{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 $\mathbf{z}_1, \ldots, \mathbf{z}_D$. 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 $\mathbf{z}_1, \ldots, \mathbf{z}_D$ we could write the sampling density above as:

$$p(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\theta}) \propto \prod_{d=1}^{D} \prod_{k=1}^{K} \left(\pi_{s(d)k} \prod_{w=1}^{W} \theta_{kw}^{y_{dw}}\right)^{z_{dk}}$$

3.2 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 semi-conjugate Dirichlet prior for $\boldsymbol{\theta}$. More specifically, we assume

$$\boldsymbol{\theta}_k \sim \mathcal{D}irichlet(\boldsymbol{\lambda}_k) \quad k = 1, \ldots, K$$

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*

$$\mathbf{z}_d \sim \mathcal{M}ultinomial(1, \boldsymbol{\pi}_{s(d)}).$$

We reparameterize to work with the unconstrained

$$\boldsymbol{\omega}_t = \left(\log\left[\frac{\pi_{t1}}{\pi_{tK}}\right], \ldots, \log\left[\frac{\pi_{t(K-1)}}{\pi_{tK}}\right]\right)'$$

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) (West and Harrison, 1997; Cargnoni et al., 1997). Specifically,

$$\boldsymbol{\omega}_t = \mathbf{F}_t' \boldsymbol{\eta}_t + \boldsymbol{\epsilon}_t \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_t) \quad t = 1, \ldots, T \tag{1}$$

$$\boldsymbol{\eta}_t = \mathbf{G}_t \boldsymbol{\eta}_{t-1} + \boldsymbol{\delta}_t \qquad \boldsymbol{\delta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t) \quad t = 1, \ldots, T \tag{2}$$

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

$$V_{ii} \sim \mathcal{I}nv\mathcal{G}amma(a_0/2, b_0/2) \quad \forall i$$

and

$$W_{ii} \sim \mathcal{I}nv\mathcal{G}amma(c_0/2, d_0/2) \quad \forall i$$

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:

$$\mathbf{F}_t = \begin{pmatrix} \mathbf{I}_{K-1} \\ \mathbf{0}_{K-1} \end{pmatrix} \quad t = 1, \ldots, T$$

$$\mathbf{G}_t = \begin{pmatrix} \mathbf{I}_{K-1} & \mathbf{I}_{K-1} \\ \mathbf{0}_{K-1} & \mathbf{I}_{K-1} \end{pmatrix} \quad t = 1, \ldots, T.$$

This produces a local linear trend model for $\boldsymbol{\omega}_t$.

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. By properly choosing $\mathbf{F}_t, \mathbf{G}_t$ and forms for $\mathbf{V}_t$ and $\mathbf{W}_t$ a very wide range of dynamics an be captured. 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 3 hours depending on the quality of the starting values and the speed of the computer. No specialized hardware is required.

It is worth reemphasizing that, viewed as a clustering / classification procedure, the model above is designed for "unsupervised" clustering. At no point does the user pre-tag 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 semi-supervised fashion, by constraining some elements of $\mathbf{Z}$ to be 0 and 1 in order to force particular documents into particular topic categories. This is not appropriate for our questions here, but could be useful in settings where one wanted results to be as maximally matched to some preexisting coding frame as possible. It is also possible to use the model to classify documents that were not in the original dataset used to fit the model.

4 Applying the Topic Model to U. S. Senate Speech, 1995-2004

We present here an analysis of speech in the United States Senate, as recorded in the *Congressional Record*, from 1995-2004 (the 105th to the 108th Congresses). In this section, we briefly desribe 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.

4.1 Senate Speech Data

The textual data are drawn from the United States Congressional Speech Corpus[8] (Monroe et al., 2006) developed under the Dynamics of Political Rhetoric and Political Representation Project (Monroe et al., 2005). The original source of the data are the html files that comprise the electronic version of the (public domain) *United States Congressional Record*, served by the

[8] *Corpus* (plural *corpora*) is a linguistic term meaning a textual database.

Library of Congress on its THOMAS system (The Library of Congress, N.D.) and generated by the Government Printing Office (The 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*[9] 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 one percent 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.[10]

4.2 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 β's and the z's.

The $\boldsymbol{\beta}$ matrix contains $K \times W$ ($\approx$ 160,000) parameters. Each element β_{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. $\boldsymbol{\beta}$ describes the intra-topic 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

[9] 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 leverage the shared topical meaning of words like *abort, aborts, aborted, aborting, abortion, abortions, abortionist, 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).

[10] The production of these data is a fairly elaborate process. Details of the process are provided in the web appendix.

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.

$\mathbf{Z}$ 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 paper returns the expected value of $\mathbf{Z}$ which we label $\hat{\mathbf{Z}}$. 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 $\hat{\mathbf{z}}_d$ as a vector of topic probabilities for document d and to then probabilistically assign documents to topics.

4.3 Model Specification and Sensitivity Analysis

We fit numerous specifications of the model outlined in Section 3 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.[11]

[11] Obviously, the optimally useful K lies between one ("Senate Speech") and 118,000 (speeches 1 ... 118,000).

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 ten major categories in Rohde’s rollcall data set (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 timebound debates (“No Child Left Behind”).

As noted above, we evaluated results for K between 3 and 60 along with some runs with K set equal to 100. Results matching our criteria, and similar to each other in broad strokes, occur at K in the neighborhood of 40-45. We present here results for the K=42 model with the highest maximized log posterior.

In order to ensure that our results are not dramatically affected by small-to-moderate changes in priors or the time period under study, we engaged in a series of sensitivity analyses. First, we fit models that employed more diffuse priors for V_{ii} and W_{ii} and then compared the document classification probabilities $\hat{\mathbf{z}}_d$ to those based on the prior described above. Here we found very high agreement between the results presented below and those based on more diffuse priors. Second, we dropped documents from the last 500 and 1000 days of the original 2920 day dataset and then compared the document classification probabilities $\hat{\mathbf{z}}_d$ derived from the subsetted data to those based on the the full data. This was done for documents that appeared in both the original and the subsetted data. Again, the estimates were very similar– even after dropping about 1/3 of the data. The estimates from the model fit to the subsetted data were also used to predict the topic membership of those documents that were dropped from the analysis. Here we found that the predicted topic memberships accorded well with those derived from a model fit to the full 2920 day data. Full results and a more detailed description of these sensitivity analyses are available in the web appendix.

5 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 inter-coder 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.

The reliability of category lists themselves is theoretically measurable in all types of systems. In theory, we could have hundreds of scholars, and our model, examine the same set of documents, determine categories that match some agreed conceptual definition, and compare them. Not only would this be impossibly expensive, but approximations of the exercise suggest there will be little agreement and no gold standard. Compare, for example, the category lists and typologies of the Jones et al. (N.D.); Clausen (1973); Lee (2006); Rohde (2004); Heitschusen and Young (2006), and Katznelson and Lapinski (2006) which all cover roughly the same domain of Congressional attention and have minimal overlap.[12]

[12] A related issue concerns the particular substantive labels we attach to the categories below. While the ones we have chosen would be unlikely to be reproduced exactly in repeated experiments, this is less important than

More important are notions of validity. There are several concepts of measurement validity that can be considered in any content analysis.[13] 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 documents 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 β 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 validty* (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.*

5.1 Topic Interpretation and Intra-Topic Semantic Validity

Table 3 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 $\hat{\boldsymbol{\beta}}_k$ 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 $\hat{\boldsymbol{\beta}}$ vectors other than $\hat{\beta}_{kw}$ (denoted $\hat{\boldsymbol{\beta}}_{-kw}$). The

the reliability of the objective summaries. We provide lists of topic-specific keywords, for example, which are 100% replicable. In fact, Mei et al. (2007) recommend an automated labeling scheme for such models, essentially listing these major keywords as the label. This would render the reliability of the labels a moot question, but decrease the usefulness of the labels for the reader.

[13] 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 is no external criteria one can bring to bear on the issue of face validity and thus we focus on several other types of validity.

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 $\hat{\boldsymbol{\beta}}_{-kw}$, which we operationalize as:

$$r_{kw} = \frac{\hat{\beta}_{kw} - \underset{j \neq k}{\text{median}} \left(\hat{\beta}_{jw} \right)}{\underset{\ell \neq k}{\text{MAD}} \left(\hat{\beta}_{\ell w} \right)},$$

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 β 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 $\hat{\boldsymbol{\beta}}_k$ and $\mathbf{r}_k$ 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 4 provides the top keys for each topic.[14]

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 4, 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 $\hat{\boldsymbol{\beta}}$ values, and those are reflected in the topic labels and clarifying notes of Table 3.

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.[15] It is worth noting that this technique will extract

[14] Longer lists of keywords and index values are provided in the web appendix .

[15] The reader can confirm this by searching on the word "terrorism" in THOMAS.

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.[16]

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 β parameters identify words that, if present, most distingush a document of this topic from all others, for the 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 4) *and* a set of terms that identify particular political (perhaps partisan) positions, points of view, frames, and so on, within that topic.

For example, Table 4 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.[17]

[16] 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 a some procedural topic.

[17] These words all appear among the top keys using any of the variance-respecting feature selection techniques described in (Monroe et al., 2007a). 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).

5.2 Relationships Between Topics and Meta-Topic Semantic Validity

An important feature of the topic model, another sharp contrast with other approaches, is that the β 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 β matrix, there are many such relationships one might investigate. Here we focus on how the topics relate to each other as subtopics of larger meta-topics, how they aggregate. The coherent meaning of the meta-topics 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 β vectors, $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_{42}$, 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 a single cluster. The results of this process are displayed in the dendrogram of Figure 1.[18] 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.[19]

Reading Figure 1 from the bottom up provides information about which clusters were merged first (those 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

[18] The order of topics given in Tables 3 and 4 is as determined here; the labels were determined prior to the agglomerative clustering.

[19] 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.

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 vs. 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 (Lowi, 1964;

Heitschusen and Young, 2006; Katznelson and Lapinski, 2006). Katznelson and Lapinski (2006) provide an eloquent defense of the exercise and a review of alternative approaches.

5.3 Speeches, Rollcalls, 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 rollcall 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 rollcalls 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 rollcall votes in which the official description contains the word “abortion”. We see the basic convergence expected, with number of rollcalls 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 rollcall might receive very little debate attention.

Consider first, the spikes of debate attention that do not have accompanying rollcall 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 rollcall 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 Peltzmann as "Government Organization" (primarily) and "Domestic Social Policy" (secondarily), and under Poole and Rosenthal as "Public Health".[20] Satcher would have been coded identically in each case. But it 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 rollcall measures.

Similarly, the speech measure captures subtleties that the rollcall count does not. For example, on or around July 1 in every year from 1997-2003, Senator Murray offered an amendment to the Department of Defense Appropriations bill, attempting to restore access to abortions for overseas military personnel. The rollcall 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

[20] 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`.

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 (2001-2004). 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.

5.4 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 3 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 12th). 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 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 4. 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 US 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.

5.5 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 utlimately 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 (Smith, 1989; Sinclair, 1989)), very small samples (e.g., six bills (Hall, 1996)), or moderately sized, but expensive, samples (e.g., 2204 speeches manually coded to three categories (Hill and Hurley, 2002)). Our data increase this leverage dramatically and cheaply.

Figure 5 summarizes the results from 50 count models (negative binomial) of the speech counts on all non-procedural topics and selected meta-topical aggregations, for the 106th Senate, for all 98 senators who served the full session. Selected hypotheses, discussed below, are represented by shaded backgrounds.[21]

Congressional behavior is of core relevance to questions about the existence and possible decline of "norms" of committee deference, specialization, and apprenticeship (Hall, 1996; Smith, 1989;

[21] These are graphical tables (Gelman et al., 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.

Sinclair, 1989; Shepsle and Weingast, 1987; Rohde et al., 1985; Matthews, 1960). As noted by Hall, this is a difficult empirical question as the primary leverage has come from floor amendment behavior, a relatively rare occurrence (pp.180-1). Figure 5 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 et al., 1985; Smith, 1989; Sinclair, 1989) and, with the cross-topic variation provided here, the role of expertise costs (Hall, 1996) vs. 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 (Hill and Hurley, 2002; Maltzman and Sigelman, 1996; Fenno, 1996, 1978; Mayhew, 1974). The sixth and seventh rows of Figure 5 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 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* (Rothenberg and Sanders, 2000; Poole and Rosenthal, 1997).

The last two rows of Figure 5 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 (Ansolabehere et al. (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 rollcall-based. 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 lanscape 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 are 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 rollcall-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.

6 Discussion

In this paper we have presented a method for inferring the 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 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 paper 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 paper presents several ways that the such interpretation and validation can be performed. Specifically, we demonstrate how (a) key words 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 reasons to believe our measures from the 105th to 108th U.S. Senate are valid.

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 measures such measures of attention as this. The dynamic richness of our data allow 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 et al., 1995). We can also examine issues of concern in particular policy areas, such as legislative oversight in foreign policy (Colaresi et al., 2007).

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 (Monroe et al., 2007a; Poole and Rosenthal, 1997; Carmines and Stimson, 1989; 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 (Monroe et al., 2007b; Lowe, 2007; Monroe and Maeda, 2004).

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).

Baumgartner, Frank R., and Bryan D. Jones. 1993. *Agendas and Instability in American Politics*. The University of Chicago Press.

Baumgartner, Frank R., and Bryan D. Jones, editors. 2002. *Policy Dynamics*. Chicago: The 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., and John D. Lafferty. 2007. "Latent Dirichlet Allocation." *Annals of Applied Statistics* 1(1):17–35.

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, 19451998*. 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):640–647.

Carmines, Edward G., and James A. Stimson. 1989. *Issue Evolution: Race and the Transformation of American Politics*. 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):245–251.

Clausen, Aage R. 1973. *How Congressmen Decide: A Policy Focus*. St. Marin's Press.

Colaresi, Michael P., Kevin M. Quinn, and Burt L. Monroe. 2007. "From Lapdog to Watchdog: Dynamic Transparency Costs, Legisaltive Oversight, and National Security Policy." American Political Science Association, Chicago.

Fenno, Richard F. 1978. *Home Style: House Members in Their Districts*. Little, Brown.

Fenno, Richard F. 1996. *Senators on the Campaign Trail*. 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):121–130.

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.

Gruber, Amit, and Yair Weiss. 2007. "Hidden Topic Markov Models." Artificial Intelligence and Statistics (AISTATS), San Juan, Puerto Rico.

Hall, Richard L. 1996. *Participation in Congress*. New Haven: 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* (E. Scott Adler, and John S. Lapinski, editors), Princeton, NJ: Princeton University Press, pp. 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):219–231.

Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2007. "An Active Learning Framework for Classifying Political Text." Midwest Political Science Association, Chicago.

Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2008. "Computer Assisted Topic Classification for Mixed Methods Social Science Research." *Journal of Information Technology and Politics* Forthcoming.

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):170–190.

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*. The 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.

Katznelson, Ira, and John S. Lapinski. 2006. "The Substance of Representation: Studying Policy Content and Legislative Behavior." In *The Macropolitics of Congress* (E. Scott Adler, and John S. Lapinski, editors), Princeton, NJ: Princeton University Press, pp. 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):617–642.

Kingdon, John W. 1995. *Agendas, Alternatives, and Public Policies*. 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*. New York: Sage, second edition.

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* 97:311–331.

Lee, Frances. 2006. "Agenda Content and Senate Party Polarization, 1981-2004." Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago.

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):Forthcoming.

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: Uncontrained 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*. University of North Carolina Press.

Mayhew, David R. 1974. *Congress: The Electoral Connection*. New Haven: Yale University Press.

Mei, Qiaozhu, Xuehua Shen, and ChengXiang Zhai. 2007. "Automatic Labeling of Multinomial Topic Models." 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, CA, pp. 490–9.

Miller, Warren E., and Donald Stokes. 1963. "Constituency Influence in Congress." *American Political Science Review* 57:45–56.

Monroe, Burt L., Steven P. Abney, Michael P. Colaresi, Kevin M. Quinn, and Dragomir Radev. 2005. "The Dynamics of Political Rhetoric and Political Representation." Center for Political Studies, University of Michigan.

Monroe, Burt L., and Ko Maeda. 2004. "Rhetorical Ideal Point Estimation: Mapping Legislative Speech." Society for Political Methodology, Stanford University, Palo Alto.

Monroe, Burt L., Cheryl L. Monroe, Kevin M. Quinn, Dragomir Radev, Michael H. Crespin, Michael P. Colaresi, Jacob Balazar, and Steven P. Abney. 2006. "United States Congressional Speech Corpus." Center for Political Studies, University of Michigan.

Monroe, Burt L., Kevin M. Quinn, and Michael P. Colaresi. 2007a. "Identifying the Content of Partisan Conflict: Methods for Lexical Feature Selection." Midwest Political Science Association, Chicago, IL.

Monroe, Burt L., Kevin M. Quinn, Michael P. Colaresi, and Ko Maeda. 2007b. "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* 75:656–675.

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):130–137.

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*. Yale University Press.

Rohde, David, Norman Ornstein, and Robert Peabody. 1985. "Political Change and Legislative Norms in the U.S. Senate, 1957–74." In *Studies of Congress* (Glenn Parker, editor), Washington, DC: Congressional Quarterly PRess, pp. 147–88.

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, East Lansing, MI.

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*. Johns Hopkins University Press.

Smith, Steven S. 1989. *Call to Order: Floor Politics in the House and Senate*. 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.

The Library of Congress. N.D. "THOMAS." `http://thomas.loc.gov`.

The 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, PA, pp. 424–33.

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 University Press.

	Method				
		Human		*Supervised*	*Topic*
Assumption	*Reading*	*Coding*	*Dictionaries*	*Learning*	*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

Table 1: A Summary of Some Common Assumptions Necessarily Made by Major Methods of Discrete Categorization of Text

	Method				
		Human		*Supervised*	*Topic*
Pre-Analysis Costs	*Reading*	*Coding*	*Dictionaries*	*Learning*	*Model*
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
Post-Analysis Costs					
Person-hours spent interpreting	High	Low	Low	Low	Moderate
Level of substantive knowledge	High	High	High	High	High

Table 2: A Summary of Some Relative Costs Associated with Major Methods of Discrete Categorization of Text

Topic Labels	%^{a}	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	

^{a}*Percentage of documents (left of slash) and percentage of word stems (right of slash).*

Table 3: Topic labels and descriptive statistics for 42-topic model.

Topic (Short Label)	Keys
1. Judicial Nominations	*nomine, confirm, nomin, circuit, hear, court, judg, judici, case, vacanc*
2. Constitutional	*case, court, attornei, supreme, justic, nomin, judg, m, decis, constitut*
3. Campaign Finance	*campaign, candid, elect, monei, contribut, polit, soft, ad, parti, limit*
4. Abortion	*procedur, abort, babi, thi, life, doctor, human, ban, decis, or*
5. Crime 1 [Violent]	*enforc, act, crime, gun, law, victim, violenc, abus, prevent, juvenil*
6. Child Protection	*gun, tobacco, smoke, kid, show, firearm, crime, kill, law, school*
7. Health 1 [Medical]	*diseas, cancer, research, health, prevent, patient, treatment, devic, food*
8. Social Welfare	*care, health, act, home, hospit, support, children, educ, student, nurs*
9. Education	*school, teacher, educ, student, children, test, local, learn, district, class*
10. Military 1 [Manpower]	*veteran, va, forc, militari, care, reserv, serv, men, guard, member*
11. Military 2 [Infrastructure]	*appropri, defens, forc, report, request, confer, guard, depart, fund, project*
12. Intelligence	*intellig, homeland, commiss, depart, agenc, director, secur, base, defens*
13. Crime 2 [Federal]	*act, inform, enforc, record, law, court, section, crimin, internet, investig*
14. Environment 1 [Public Lands]	*land, water, park, act, river, natur, wildlif, area, conserv, forest*
15. Commercial Infrastructure	*small, busi, act, highwai, transport, internet, loan, credit, local , capit*
16. Banking / Finance	*bankruptci, bank, credit, case, ir, compani, file, card, financi, lawyer*
17. Labor 1 [Workers]	*worker, social, retir, benefit, plan, act, employ, pension, small, employe*
18. Debt / Social Security	*social, year, cut, budget, debt, spend, balanc, deficit, over, trust*
19. Labor 2 [Employment]	*job, worker, pai, wage, economi, hour, compani, minimum, overtim*
20. Taxes	*tax, cut, incom, pai, estat, over, relief, marriag, than, penalti*
21. Energy	*energi, fuel, ga, oil, price, produce, electr, renew, natur, suppli*
22. Environment 2 [Regulation]	*wast, land, water, site, forest, nuclear, fire, mine, environment, road*
23. Agriculture	*farmer, price, produc, farm, crop, agricultur, disast, compact, food, market*
24. Trade	*trade, agreement, china, negoti, import, countri, worker, unit, world, free*
25. Procedural 3	*mr, consent, unanim, order, move, senat, ask, amend, presid, quorum*
26. Procedural 4	*leader, major, am, senat, move, issu, hope, week, done, to*
27. Health 2 [Seniors]	*senior, drug, prescript, medicar, coverag, benefit, plan, price, beneficiari*
28. Health 3 [Economics]	*patient, care, doctor, health, insur, medic, plan, coverag, decis, right*
29. Defense [Use of Force]	*iraq, forc, resolut, unit, saddam, troop, war, world, threat, hussein*
30. International [Diplomacy]	*unit, human, peac, nato, china, forc, intern, democraci, resolut, europ*
31. International [Arms]	*test, treati, weapon, russia, nuclear, defens, unit, missil, chemic*
32. Symbolic [Living]	*serv, hi, career, dedic, john, posit, honor, nomin, dure, miss*
33. Symbolic [Constituent]	*recogn, dedic, honor, serv, insert, contribut, celebr, congratul, career*
34. Symbolic [Military]	*honor, men, sacrific, memori, dedic, freedom, di, kill, serve, soldier*
35. Symbolic [Nonmilitary]	*great, hi, paul, john, alwai, reagan, him, serv, love*
36. Symbolic [Sports]	*team, game, plai, player, win, fan, basebal, congratul, record, victori*
37. J. Helms re Debt	*hundr, at, four, three, ago, of, year, five, two, the*
38. G. Smith re Hate Crime	*of, and, in, chang, by, to, a, act, with, the, hate*
39. Procedural 1	*order, without, the, from, object, recogn, so, second, call, clerk*
40. Procedural 5	*consent, unanim, the, of, mr, to, order, further, and, consider*
41. Procedural 6	*mr, consent, unanim, of, to, at, order, the, consider, follow*
42. Procedural 2	*of, mr, consent, unanim, and, at, meet, on, the, am*

Table 4: For each topic, the top ten (or so) key stems that best distinguish the topic from all others. Keywords have been sorted here by rank($\beta_k w$) + rank(r_{kw}), as defined in the text. Lists of the top forty 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 3 but the topic names have been shortened.

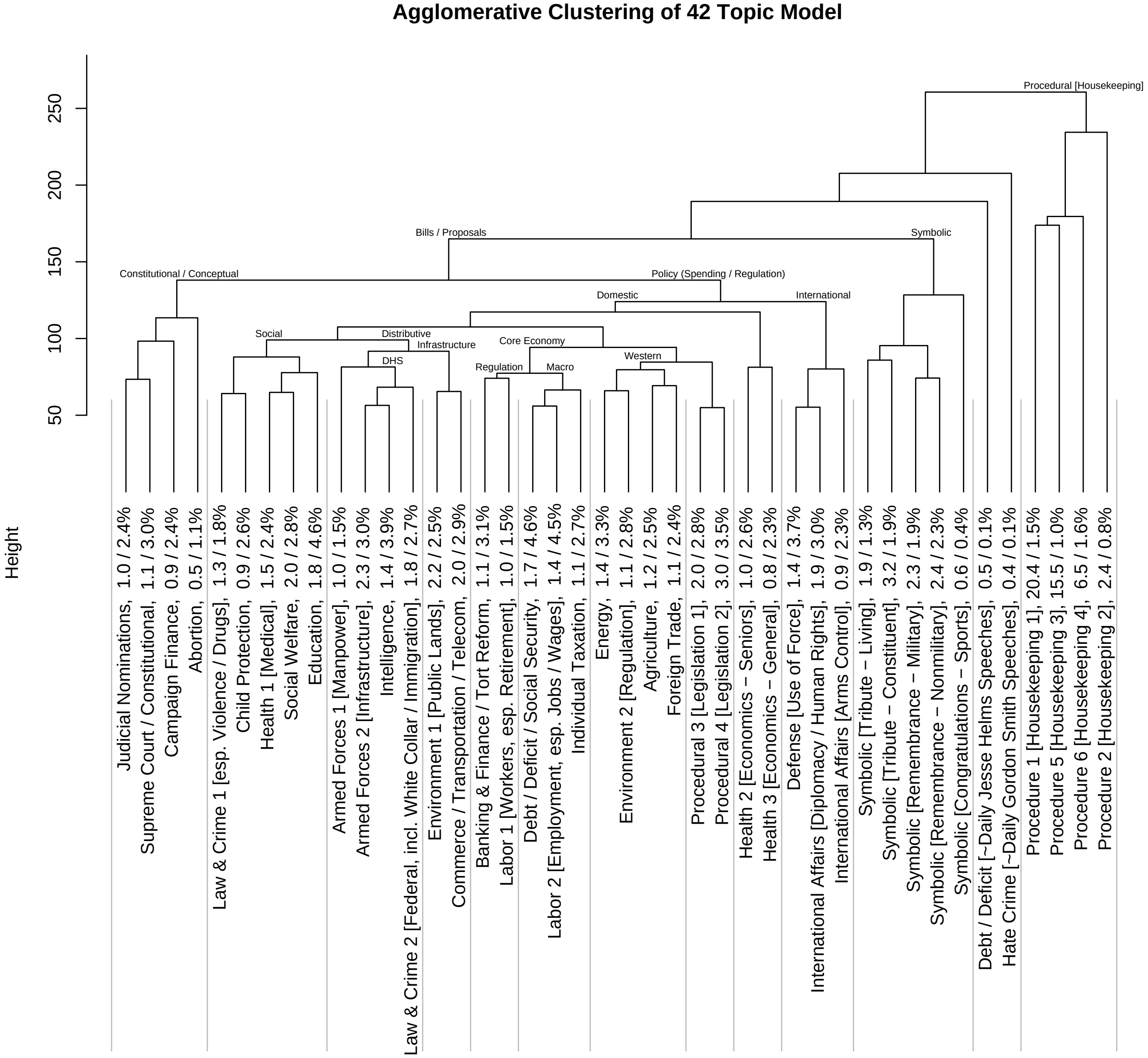

Figure 1: *Hierarchical Agglomerative Clustering of* $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_K$*.* Clustering based on minimizing the maximum Euclidean distance between cluster members. Each cluster is labelled with a topic name, followed by the percentage of documents and words, respectively, in that cluster.

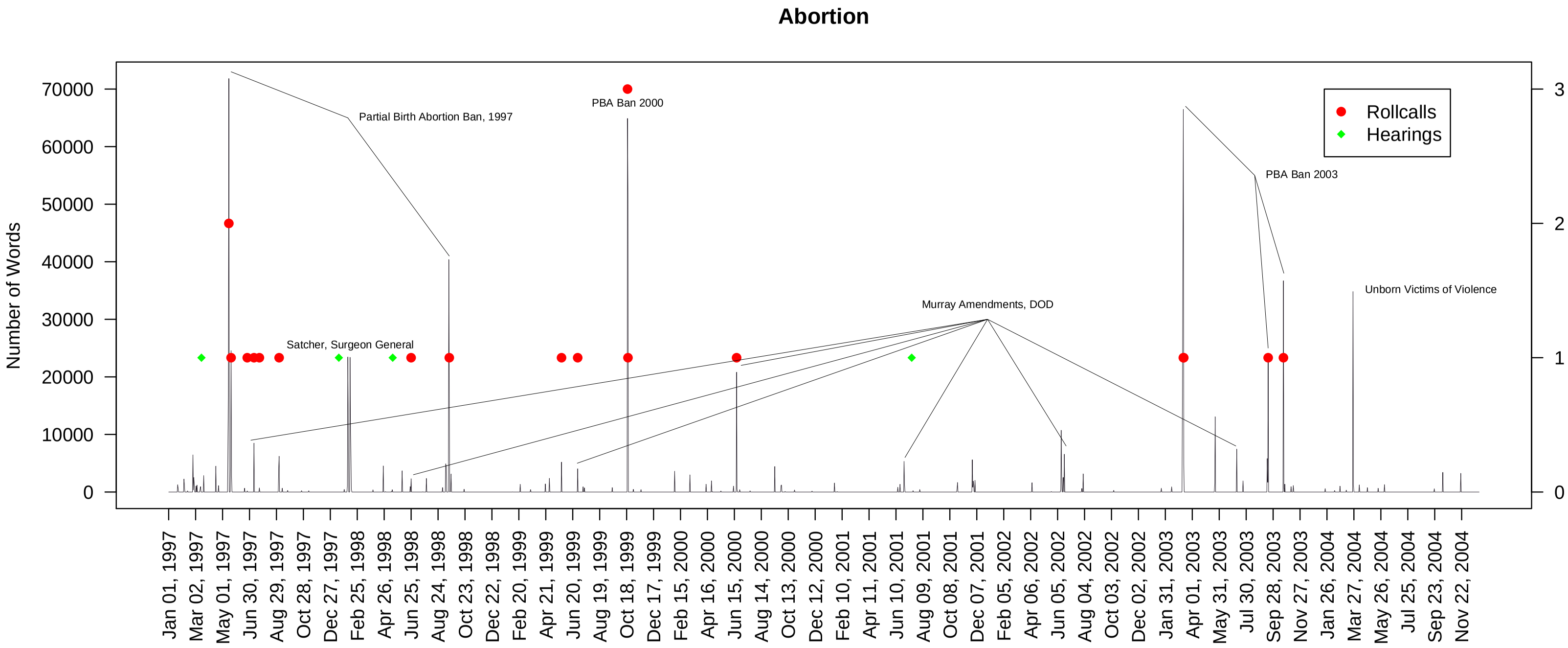

Figure 2: *The Number of Words Spoken on the 'Abortion' Topic Per Day.*

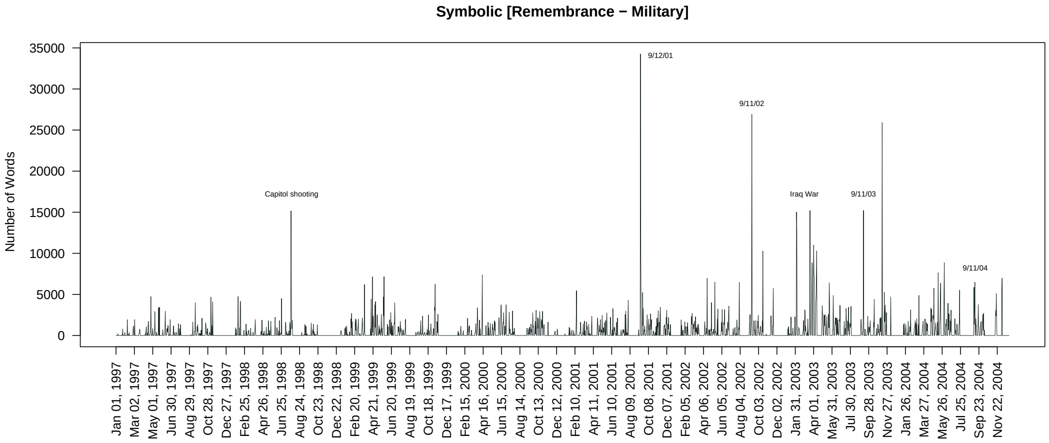

Figure 3: *The Number of Words on the 'Symbolic [Remembrance - Military]' ('Fallen Heroes') Topic Per Day.*

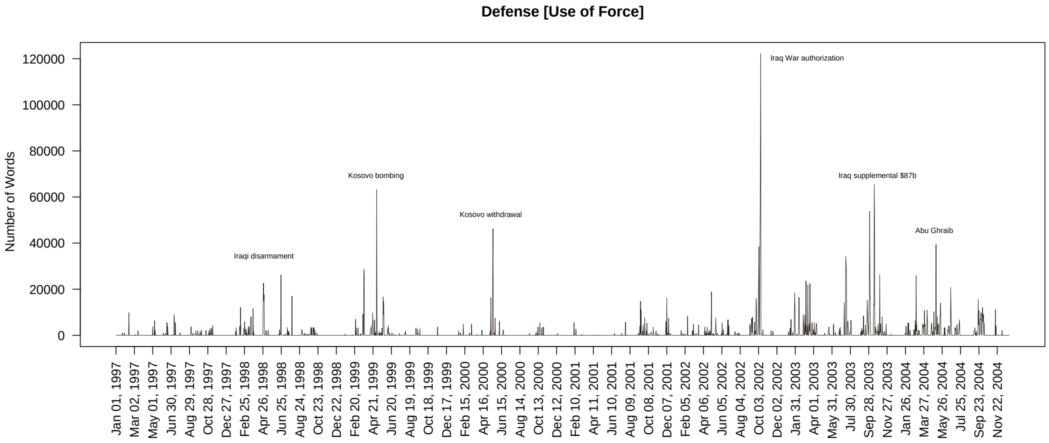

Figure 4: *The Number of Words on the 'Defense [Use of Force]' Topic Per Day.*

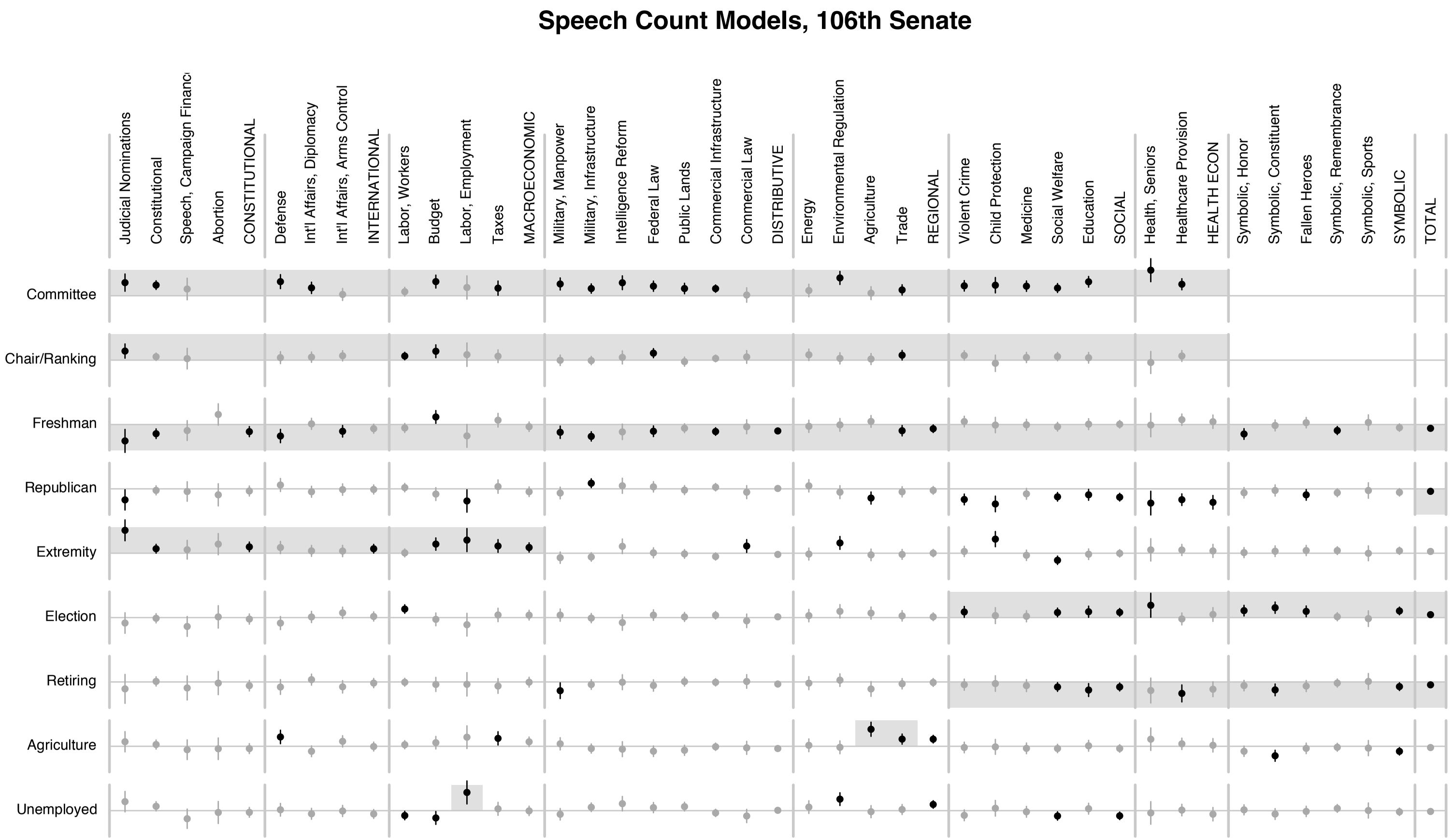

Figure 5: *Each column represents a negative binomial model of speeches delivered on a given topic, or group of topics, in the 106th Senate, with one observation per Senator who served the entire two years (98 in total). Each row of the table represents a covariate: "Committee" (binary indicating whether 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.*