Advance Access publication February 4, 2009

Political Analysis (2008) 16:351–355
doi:10.1093/pan/mpn017

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

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

Philip A. Schrodt
Department of Political Science, University of Kansas, Lawrence, KS 66045

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

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

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

[1] http://www.wjh.harvard.edu/~inquirer/; accessed August 4, 2008.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

References

Advanced Research Projects Agency (ARPA). 1993. *Proceedings of the Fifth Message Understanding Conference (MUC-5).* Los Altos, CA: Morgan Kaufmann.

Alexa, Melina, and Cornelia Zuell. 1999. *A review of software for text analysis.* Mannheim, Germany: Zentrum für Umfragen, Methoden und Analysen.

Beigman Klebanov, Beata. 2006. "Measuring semantic relatedness using people and WordNet." *Proceedings of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers,* pp 13–7. New York, NY: Association for Computational Linguistics.

Benoit, Kenneth, and M. Laver. 2003. Estimating Irish party positions using computer wordscoring: The 2002 elections. *Irish Political Studies* 17:97–107.

Bond, Doug, J. Craig Jenkins, Charles L. Taylor, and Kurt Schock. 1997. Mapping mass political conflict and civil society: The automated development of event data. *Journal of Conflict Resolution* 41:553–79.

Cardie, Claire, and John Wilkerson. 2008. Special issue: Text annotation for political science. *Journal of Information Technology and Politics* 5:1–6.

Goodrich, Melanie. 2008. "A coding methodology for open-ended survey questions. Poster presented to the Political Methodology Society, Ann Arbor, MI, July 10–12, 2008.

Grimmer, Justin. 2008. "Expanding the study of political representation: Measuring and explaining representatives' expressed agenda. Poster presented to the Political Methodology Society, Ann Arbor, MI, July 10–12, 2008.

Holsti, Ole R. 1969. *Content analysis for the social sciences and humanities.* Reading, MA: Addison-Wesley.

Janowitz, Morris. 1968. Harold D. Lasswell's contribution to content analysis. *Public Opinion Quarterly* 32:646–53.

Jurafsky, Daniel, and James H. Martin. 2008. *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics.* 2nd ed. Upper Saddle River, NJ: Prentice-Hall.

Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology.* 2nd ed. Thousand Oaks, CA: Sage.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review* 97:311–31.

Manning, Christopher D., and Hinrich Schütze. 2002. *Foundations of statistical natural language processing.* 5th ed. Cambridge, MA: MIT Press.

Miller, George. 1990. "WordNet: An on-line lexical database." *International Journal of Lexicography* 3:235–312.

Monroe, Burt L., and Ko Maeda. 2004. "Talk's cheap: Text-based ideal point estimation. Paper presented to the Political Methodology Society, Palo Alto, July 29–31, 2004.

Neuendorf, Kimberly A. 2001. *The content analysis guidebook.* New York: Sage.

Pemstein, Daniel. 2008. "Predicting strategic roll calls with legislative text. Poster presented to the Political Methodology Society, Ann Arbor, MI, July 10–12, 2008.

Popping, Roel. 2000. *Computer-assisted text analysis.* New York: Sage.

Quinn, Kevin M., Burt L. Monroe, Michael P. Colaresi, Michael H. Crespin, and Dragomir Radev. 2006. "An automated method of topic-coding legislative speech over time with application to the 105th–108th United States Senate. Paper presented to the Political Methodology Society, Davis, CA, July 20–22, 2006.

Roberts, Carl W. 1997. Text analysis for the social sciences: Methods for drawing inferences from texts and transcripts. Mahwah, NJ: Lawrence Erlbaum.

Sagarzazu, Inaki. 2008. *"Look who's talking: Analyzing the Dynamics of Political Discourse.* Poster presented to the Political Methodology Society, Ann Arbor, MI, July 10–12, 2008.

Salton, Gerard. 1989. *Automatic Text Processing.* Reading, MA: Addison-Wesley.

Schrenk, Michael. 2007. *Webbots, spiders, and screen scrapers.* San Francisco, CA: No Starch Press.

Schrodt, Philip. 2003. "Analyzing text using statistical methods. Workshop presented to the Political Methodology Society, Minneapolis, MN, July 17–19, 2003.

Simon, Adam F., and Michael Xenos. 2004. "Dimensional reduction of word-frequency data as a substitute for intersubjective content analysis." *Political Analysis* 12:63–75.

Smith, Peter D. 1990. *An introduction to text processing.* Cambridge, MA: MIT Press.

Stone, Philip J., Dexter C. Dunphry, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The general inquirer: A computer approach to content analysis.* Cambridge, MA: MIT Press.

Weber, Robert P. 1990. *Basic content analysis.* 2nd ed. Newbury Park, CA: Sage Publications.