COMMENT BY

ARTHUR SPIRLING At the time of this writing in late 2012, the United States risks falling off a "fiscal cliff." Absent a bipartisan agreement between a Democratic president and a Republican House of Representatives, taxes will rise and public spending will be cut automatically in a bid to decrease a large budget deficit, regardless of the (seemingly baleful) consequences. The received wisdom is that reaching a legislative deal to prevent this outcome is a difficult proposition: the parties bicker and are intransigent, and they operate in a Congress that is "the most polarized since the end of Reconstruction," according to Ezra Klein, a columnist and blogger for the FN1 *Washington Post.*[1]

For social scientists, at least three research questions arise from this purported nadir of American politics and the rancor and bitterness that supposedly characterize it: First, is it true? Second, does it matter? And third, how did this state of affairs come about? Broadly, it is these queries that this paper by Jacob Jensen and coauthors seek to answer. In so doing, they collect an extraordinary new and voluminous data set that incorporates a century of congressional speech, use innovative measures of political partisanship, and compare their results with a corpus of published phrases (Google Ngrams, drawn from Google Books) to look for possibly causal relationships between what politicians say and what is said by their publics. What emerges from their efforts is an impressive, data-driven look at political polarization and its development since the Reconstruction Era. Unsurprisingly, given its sheer scope, the analysis is not without flaws; commensurately, I will comment here on possible avenues for improvement and refinement, primarily on technical grounds. In addition, as it is clear that the paper and its data will inspire future research, my concluding section will attempt to point such efforts in fruitful directions, in terms of both methods and substance.

THE PAPER'S CONTRIBUTION AND ITS MOTIVATION It is important to emphasize the wealth of text data the authors have gathered: it is, to this discussant's knowledge, unprecedented in the study of U.S. politics. In sum, it is 130 years of information from the speech of the nation's representatives in Congress, and after much reduction it still includes some 690,000 phrase observations. The authors have matched these data to the party of the speaker, which no doubt required thorough cleaning and

[1] Ezra Klein, "14 Reasons Why This Is the Worst Congress Ever," *Wonkblog (Washington Post),* July 13, 2012. www.washingtonpost.com/blogs/wonkblog/wp/2012/07/13/13-reasons-why-this-is-the-worst-congress-ever/.

much careful effort. The authors' inferential task is then similarly ambitious: to track the polarization of the parties over time, and to see what effects this varying polarization might have on other important outcomes, such as political violence. They perform extremely computationally intensive operations on their text data from the *Congressional Record,* and then do the same for the Google Books corpus. All of this is as impressive as it is important, and the paper deserves to be well cited—quite apart from the fact that the data set itself will form the backbone of many future studies. The paper is candid, thoughtful, and circumspect, and it comes at a time when methods for "text-as-data" are coming to the fore in the toolkit of political science (see, for example, Quinn and others 2010, Grimmer and Stewart forthcoming), and when "polarization" is a buzzword both in popular media and in academia (McCarty, Poole, and Rosenthal 2008, Fiorina, Abrams, and Pope 2010).

However, no good deed goes unpunished, and no good paper goes uncriticized. This is most assuredly a good paper, and any harshness in the comments that follow should indicate the degree to which reading it provokes thought—approving or otherwise.

TWO PROBLEMS WITH TRIGRAMS The core of the authors' measurement strategy is the trigram, a three-word sequence. Because they both "stem" and "stop" the raw text, meaning that words are truncated to their "roots" and function words (articles, conjunctions, auxiliary verbs, prepositions, and pronouns) are removed, it is not necessarily the case that any given trigram appears as is in the speeches. For example, "capital gains tax" becomes "capit.gain.tax," and any parts of sentences containing nothing but function words, such as "what he did with them," will disappear from the counting process altogether. The authors can hardly be blamed for attempting to reduce the dimensions of the feature space: although operating at the "token" (in this case, single word) level would be more general, the estimation problem would become much more computationally difficult—perhaps prohibitively so. Since one imagines that political phrases are precisely about context and a particular relationship with the words around them, stemmed and stopped trigrams are a reasonable pragmatic choice, capturing subtleties of meaning and allowing relative ease of interpretation while retaining tractability on the statistical side. Nonetheless, social scientists might have a few concerns. First, the idea of trigrams in this context is to capture some notion of word order. That is, phrases like "capit.gain.tax" and "nation.debt.increas" relate to concerns specific to political economy in a way that these words uttered separately do not. For many classification exercises, working with such simplifications of spoken language present almost no cost. But life may be less rosy when, as here, issues of the speaker's sentiment are at stake. To cite a crude example, one imagines that the sentences "I do not support the New Deal" and "I do support the New Deal" would be spoken by legislators of quite different ideological stripes. But depending on the specific choice of stop words, both reduce to "support.new.deal" for the purposes of the present analysis, with potentially confusing consequences for interpretation.

A second, more subtle issue when working with stopped, stemmed trigrams arises from the fact that not all partisan phrases will be treated equally. As a running example, consider two very different three-word phrases: "Martin Luther King" and "By Almighty God." Notice that the second phrase includes a noun ("God") that has many synonyms: Creator, Lord, Heavenly Father, and so on. In principle, then, members of Congress could use any of these alternatives and communicate approximately the same meaning, and each would be counted separately under the authors' scheme. This is much less true of "Martin Luther King," a phrase that refers to an obvious individual and for which there are few close substitutes. As a result, speakers who wish to make a comment about that individual have little choice but to coordinate on "Martin Luther King" as a phrase. The consequence is that even if a particular concept—such as talking of God in whatever way—is highly discriminatory (and, one might hypothesize, indicative of Republicans), the diversity of options will reduce the chances that it appears as such. Matters are even worse in this particular case, because "By Almighty God" includes a stop word, which will be removed and some other word joined to the other two, further diversifying the nature of its appearance in the texts at hand.

What to do? One obvious robustness check would be to vary the stemming and stopping rules and verify that the results are similar. Another is to be more explicit about sentence structure and word order. Here the work of Huma Lodhi and others (2002) might prove profitable, and in particular their use of string kernels, which allow the researcher to break up documents into sets of *n*-contiguous characters and then base analysis on the relative frequency of these characters. There is no stemming or stopping with such procedures, and thus the statements regarding the New Deal above would be categorized as different. In addition, future work might consider identifying synonyms, perhaps with the help of a thesaurus or its equivalent, although this would involve more human coding a priori than the authors were perhaps willing to undertake for this study.

POLARIZED VIEWS, OR PARTISAN TOPICS? The metric used in the paper to calculate a phrase's (that is, trigram's) partisanship gives extra weight to word sequences that are used frequently by one party but infrequently by the other. That is, "partisan" words are those that discriminate between Democrats and Republicans. But as the authors themselves acknowledge, this difference in use may come from two very different sources: parties may talk about different things (guns versus immigration, for example), or about the same things in different ways ("illegal aliens" versus "undocumented workers"), or perhaps some combination of the two occurs. Depending on what the researcher wants to do with the results generated, this conflation is of varying concern. The broad goal of this paper is to measure "polarization," which is usually taken to mean a difference of opinion on the same topic, such as taxes, abortion, or immigration, because ideological distance decreases the ability of a given Congress and administration to deliver public policy efficiently. That is, we care about parties' positions rather than the valence they accord to different issues. If all the authors have captured is a difference in what subjects are "important" to the parties, then they have deviated some distance from the original goal. Notice here that validating the trigrams—in the sense that they predict party membership well in a holdout sample—cannot discriminate between ideological and topical division as an organizing principle for congressional speech.

Inspection of the trigrams identified as partisan does not help on this matter. As the authors note, the most recent examples do indeed appear to capture different views on the same issues, but in many years the selected trigrams appear entirely uninformative, "fiscal.year.end" (Republicans, 1897), "unit.state.oblig" (Republicans, 1919), and "unit.state.transmit" (Democrats, 1967) being easily found examples. One way to proceed may be that described by Burt Monroe, Michael Colaresi, and Kevin Quinn (2008), who limit attention to the difference on particular topics, thus getting immediately to the estimand of interest for the current authors, and in a model-based way. Note further that "topic" in this context could refer to some exogenously imposed issue that must be discussed, such as an OPEC oil shock, rather than one endogenously introduced for the specific purpose of partisan legislating.

UNSURE ABOUT UNCERTAINTY The paper's core measure, β_{pc}, is the correlation between the frequency of use of a phrase and the party of a speaker (coded 1 if the member is a Republican, −1 if a Democrat). Thus, if β_{pc} is negative, the phrase is associated with Democrats, and if positive, with Republicans. If the correlation is large in absolute terms (the authors do not say how large), the phrase is denoted as "polarizing." Unusually for an estimated quantity, there is no uncertainty around this metric. This is unfortunate for several reasons. First, when comparing words within a given Congress, it would presumably be helpful to know how different the use of phrases actually is. Suppose, for example, that "Franklin.Delano. Roosevelt" receives a score of –0.3, implying it is a Democratic phrase; suppose further, however, that the bounds on that correlation are (–0.7, 0.1). In that case it clearly includes zero—or perfect nonpartisanship—implying that one cannot claim it is "truly" a Democratic phrase. Second, the same logic applies over time, too: the fact that a phrase is used more in a later Congress should affect one's certainty about its status as a polarizing term, even if the relative proportion of times it is used by the different parties remains constant. This matters given that Congress says and does more and more today than in the past, and it is precisely the notion of "never been so bad" that the authors seek to tackle. How might the authors proceed? Obviously, correlations have sampling distributions, and one can place confidence intervals around them. If that is objectionable, one might proceed via a bootstrap approach, although as with the confidence interval approach, care is needed in demarcating exactly what is being sampled from and dealing with the fact that it is the normalized frequency that enters the correlation calculation (set to be zero, on average, for every Congress).

INFERENCE, TIME, AND INSTITUTIONS The authors look at several time series that they expect to be correlated with, if not causally related to, the polarization of Congress. They find, among other things, that polarization is related to political violence, but not to legislative efficiency. That is, the work of government still gets done even if the parties disagree. Of course, such claims raise obvious issues of simultaneity and endogeneity: for example, the more a party gets done (such as Obamacare), the more the other party may respond by acting in polarized fashion. The authors also find that polarizing phrases in the Google Books corpus diffuse into Congress over time, but that less polarized language diffuses from Congress into books. The authors are quite candid that making causal inferences about such time series is fraught with difficulty: to put it most crudely, it is hard to know which causes which, and getting at the mechanisms behind the causation is even harder. One interesting observation that might lead to more helpful theorizing about all these problems is given by the authors in their comments on House control: they note that partisanship of language tends to switch in the direction of the (new) minority party. The authors speculate that this may be due to a more vocal minority attempting to filibuster majority progress. An alternative possibility is that minority parties represent more of a draw from the core of their party, since moderates tend to come from more evenly bipartisan districts and are more vulnerable to electoral forces there, so that when a party loses power, it tends disproportionately to lose its most centrist voices (see Canes-Wrone, Brady, and Cogan 2002 for a discussion of this literature).

Thinking about parties in this way introduces a more general notion of institutions (of which parties are one type) and norms of behavior. Parties are known to "whip" their members—that is, to pressure them to vote in certain ways—and it seems plausible that they would cajole them to speak in certain ways as well. In addition, the rules that Congress uses to run itself vary over time, and future work in this area should note that such changes are likely to be reflected in the debates, and the debate structure, observed in practice.

TOWARD A STRUCTURAL MODEL? As noted above, the authors have not been shy about linking their data on speeches with the historical record in books. A further project might attempt to compare and contrast like with like, at least as it pertains to national legislatures. Recent times have seen the digitization of the British Hansard House of Commons records: every speech, every member, every session (www.hansard-archive.parliament.uk). Although it would certainly be interesting to look at polarization in comparative perspective, a more compelling target for analysis is the changing nature of language across the systems. Consider, for example, the term "liberal." In the United States this adjective is typically applied to those on the political left and connotes social permissiveness combined with notions of strict regulation of industry and a relatively generous welfare state. In Europe, in contrast, and particularly in the United Kingdom, "liberal" is less likely to be used to describe such views. Indeed, as traditionally considered, liberalism refers to free trade and a more laissez-faire method of economic production. Precisely where these European and American notions diverged is of profound interest in understanding both American "exceptionalism," in the Tocquevillian sense, and the general development of European political movements—including socialism—that are curiously absent in the United States (Hartz 1955). The authors' methods provide some clues as to how one might proceed in such an inquiry. One option is to take each trigram including the word (or appropriate stem) "liberal" and take account of the words preceding and following it. It may be that in some initial historical period, before the American Revolution, the words were used identically on both sides of the Atlantic (pertaining to trade, or speech, or meetings), but that "liberty" later took on a special ideological meaning in the United States that it did not in Britain. Furthermore, the Google Books corpus has separate data bases for British and

American English publications: whether parliaments follow presses, or presses follow parliaments, is a question for both countries. There is, of course, nothing unique about terms such as "liberal," and the best approach would be—as the authors are—agnostic about what divides groups both within and outside the parliaments of their countries.

The authors of this paper have shown how political science and economics can come together fruitfully to yield insights of value to both. Further collaborative work between the disciplines on methods of measurement is surely in order, too. Put most crudely, the social sciences do not yet have a generally accepted (or perhaps even a useful) structural model of text generation that would allow researchers to connect the language choices observed in the data with a model of rational human behavior, the parameters of which can be directly interpreted in terms of quantities we care about. In this respect the contrast between analysis of congressional speech and analysis of congressional votes is stark. For the latter, the last 15 years has seen an explosion in the application of item response models to roll-call data (see, for example, Poole and Rosenthal 1997, Clinton, Jackman, and Rivers 2004). The theoretical model underpinning the techniques typically used is that of "spatial voting" (in the sense of Davis, Hinich, and Ordeshook 1970), which is based on the proposition that elected representatives compare the status quo with the outcome promised by the new bill and choose the option that offers greater utility. Of course, not every feature of the structural model is identified (in particular, one cannot obtain outcome locations without additional assumptions), but the reduced-form estimates nonetheless correspond to some "helpful"—if somewhat idealized—world of human interaction and decisionmaking. Thus, one can readily ask, in a comparative statics fashion, what is expected to happen on seeing a bill become more attractive to a member of Congress along some dimension, how ideologically cohesive a party is, or (by imposing more structure) how representatives have moved through ideological space over time.

Matters are much less clear with text. In particular, we lack a satisfying theoretical model of human behavior that describes how and why different words, or different words in combination, are selected from some possible dictionary such that they communicate a political point or maximize utility in some way. In part, this lacuna is due to the fact that the strategy space—what agents can do given the situation they face—is extremely complicated: rather than simply vote aye or nay, politicians must decide which words (out of thousands) strike the right tone, quite apart from any selection of topic to discuss. Second, although some strategic and reactive voting certainly does occur in legislatures, ignoring this variation seems fairly harmless in the case of voting (but see Spirling and McLean 2007). It is much less innocuous in the case of speeches, which by their very nature are responses to one another: studying their words and phrases as independent observations seems a bold, and possibly disastrous, statistical choice. Although political scientists have given presumed data generating processes for documents, especially in the context of "topic models" (for example, Quinn and others 2010), they are generally vague in terms of the role of human decisionmaking. Thus, there is room for improvement in this part of political economy: writing down a (simple) structural model that could be fit to data in some reduced form should be the goal. We have plenty of data—the authors have shown us that; we now need to work together as social scientists to put this type of data to its best use.

REFERENCES FOR THE SPIRLING COMMENT

Canes-Wrone, Brandice, David Brady, and John Cogan. 2002. "Out of Step, Out of Office: Electoral Accountability and House Members Voting." *American Political Science Review* 96, no. 1: 127–40.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98, no. 2: 355–70.

Davis, Otto, Melvin Hinich, and Peter Ordeshook. 1970. "An Expository Development of a Mathematical Model of the Electoral Process." *American Political Science Review* 64: 426–48.

Fiorina, Morris, Samuel Abrams, and Jeremy Pope. 2010. *Culture War? The Myth of a Polarized America,* 3rd ed. London: Pearson.

Grimmer, Justin, and Brandon Stewart. Forthcoming. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents." *Political Analysis.*

Hartz, Louis. 1955. *The Liberal Tradition in America: An Interpretation of American Political Thought since the Revolution.* New York: Harcourt, Brace & World.

Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Christianini, and Chris Watkins. 2002. "Text Classification Using String Kernels." *Journal of Machine Learning Research* 2: 419–44.

McCarty, Nolan, Keith Poole, and Howard Rosenthal. 2008. *Polarized America: The Dance of Ideology and Unequal Riches.* MIT Press.

Monroe, Burt, Michael Colaresi, and Kevin Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16, no. 4: 372–403.

Poole, Keith, and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting.* Oxford University Press.

Quinn, Kevin, Burt Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54: 209–28.

Spirling, Arthur, and Ian McLean. 2007. "UK OC OK? Interpreting Optimal Classification Scores for the United Kingdom House of Commons." *Political Analysis* 15, no. 1: 85–96.

GENERAL DISCUSSION Bradford DeLong praised the authors for their contribution to documenting and explaining polarization in American politics. He thought it important to differentiate between ideological polarization and partisan polarization, with the latter being much more in evidence today. To illustrate the difference, DeLong noted that a century ago Theodore Roosevelt began his political career as an ideological firebrand, yet was also very willing not only to cut deals across partisan lines but even to wreck his own party's electoral chances to promote the policies he supported. That was an example of ideological but not partisan polarization. By contrast, the current Congress demonstrates so much partisan polarization—predominantly but not overwhelmingly on the Republican side—that it cannot even enact policies on which the two parties have historically agreed.

Steven Davis found it difficult to interpret the paper's results that drew on Google Books without knowing more about the composition of the Google Books database. In particular, he wondered whether that composition had shifted over time as economic factors—changes in the pricing of books, the emergence of new media—changed the relative supply and demand for different types of books. Such shifts, for example in the relative output of serious nonfiction books versus cheap romances or sci-fi novels, could call into question whether phrase counts from Google Books provided a valid and stable measure of political discourse. Davis also hypothesized that the more widely a book is circulated, the greater its impact on polarization, and so he asked whether data were available to allow weighting of books by their sales.

David Romer said that although he agreed with Arthur Spirling that a structural model of speech would be ideal, at the very least the paper would benefit from some simple statistical baselines. For example, measuring polarization by counting trigrams might automatically lead to finding the most frequently discussed topics to be the most polarized, even when there is broad agreement on the topic. If instead the trigrams could be compared against a null data set, like that which might be generated by