# Measuring Political Positions from Legislative Debate Texts on Heterogenous Topics

Benjamin E. Lauderdale
London School of Economics
and Political Science
b.e.lauderdale@lse.ac.uk

Alexander Herzog
Clemson University
aherzog@clemson.edu

September 28, 2014

ABSTRACT

In this paper, we demonstrate a new strategy for scaling legislative speeches on heterogenous topics. Existing approaches to scaling political disagreement from texts perform poorly except when applied to narrowly selected texts discussing the same issue and written in the same style. Our approach exploits the fact that legislative speech is usually organized into a set of statements or speeches by legislators, sometimes preceding a vote or set of votes, i.e. a debate. We use this debate-level structure by scaling preferences from within-debate variation in word usage, thus holding constant the topic and other features of the debate. Then we combine these debate-specific scales into a general scale, recovering measures of spoken disagreement in legislatures, using an unsupervised model. We demonstrate this approach with applications to the Irish Dáil and the US Senate. In Ireland, we show that the dominant dimension of speech variation is government-opposition, that ministers are more extreme on this dimension than backbenchers, and that speeches are less polarized along these lines in later readings of bills. In the US, we show that partisan polarization in speeches varies in response to major political events and across the electoral cycle, as well as demonstrating strong gender differences in speech positioning that are not evident in roll-call voting.

# 1. INTRODUCTION

What can we learn from legislative speeches? Few citizens observe what their representatives say in legislative sessions, so the direct influence on electoral politics is usually limited. Few legislators are amenable to changing their votes in response to what their colleagues, so the direct influence of debate on policy outcomes is also limited. Yet legislators do speak: sometimes emphasizing agreement and sometimes emphasizing disagreement. A central difficulty in drawing inferences about what motivates legislators when they speak is the difficulty of summarizing the many speeches a legislator might give over a legislative session. It is possible to read one speech and characterize the position that a legislator takes in that instance; it is general not possible to read all speeches and to characterize the positions that all legislators take, in either relative or absolute terms. Thus, to systematically describe speech behavior, and to better understand the motivations behind it, it would be useful to be able to quantitatively summarize major features of what legislators say.

Treating political texts as quantitative data presents substantial difficulties in analysis (Grimmer and Stewart 2013). In the legislative context, topic models have proved useful for understanding how the different issues that legislators choose to address reflect their electoral strategies (Grimmer 2013). However topic is only one of the types of variation that can potentially be recovered from political texts. In addition to using textual data to discover the *issues* discussed, political scientists have sometimes been able to recover credible measures of the *positions* revealed across texts and authors (Slapin and Proksch 2008). However, there is generally an identification problem between these two sources of variation in texts: how can an unsupervised model for political text know whether word use variation is a function of the topic of a text, or what that text is saying about that topic? The difficulty of distinguishing between topical variation and spatial variation in word use has been recognized, with the standard approach being to fit a model and then attempt to figure

out through validation what the model has actually measured (Grimmer and Stewart 2013, 294).

In this paper, we propose a strategy for addressing this identification problem in the context of legislative debates, which are generally organized into a set of statements or speeches by legislators preceding a vote or set of votes. Our innovation is to confine the assumption that we can recover information about positions from word use variation to within the set of texts from a single legislative debate. We argue that just as the natural unit for legislative voting data is the roll-call, the natural unit for legislative speech is the debate. When researchers estimate scales from roll-call data, they assume that all legislators are motivated by the same latent factors on a given roll-call, but they do not assume that this is true across different roll-calls. Similarly, we can be far more confident that word use will depend in a consistent way on the latent positions we want to measure if we focus on the speeches during a single debate. The scaled variation in what legislators say in that debate, may then correlate more or less strongly with the general dimension(s) of legislative disagreement. Thus, just as roll-call analysis relies on aggregation across roll-calls of within-roll-call variation in voting behavior, the approach taken here relies on aggregation across debates of within-debate variation in speech behavior. Since most legislators do not speak in a given debate, each debate reveals information about the relative positions of only a subset of all legislators. Fortunately, there is sufficient speaker overlap across legislative debates to recover a common scale, given assumptions about the decision to speak in a given debate.

In this paper, we focus on a two-stage strategy for aggregating debate-specific scaling of political speeches, which we refer to as *Wordshoal*.[1] The first stage uses the existing text scaling model Wordfish (Slapin and Proksch 2008) to scale word use variation in each debate separately. The second stage uses a normal factor analysis model to aggregate these debate-specific scales into a general scale. As we note in our discussion of possible extensions, it

[1] A "shoal" is a group of fish, not necessarily traveling in the same direction.

is possible to integrate these analyses into a single hierarchical model, or to estimate both topics and preferences from the same texts. Because we only use within-debate variation in word usage to estimate preference differences across legislators, we could then estimate the issues at stake in each debate using across-debate variation in word usage. While this paper focuses on unidimensional aggregation of debate-specific dimensions, the prospect of both estimating the topics of legislative debates and the relative preferences of legislators within those debates is an important advance in the quantitative analysis of political speech.

We present two applications to validate our approach and to demonstrate how legislative speech-making varies in different political contexts. In the first application, we use speech data from the Irish Dáil to show that by aggregating debate-specific preference scales rather than aggregating debates before scaling, our model substantially outperforms Wordfish in terms of the validity of point estimates and the credibility of the uncertainty estimates. Irish legislative speech strongly reflect government-opposition dynamics, with members of parties that switch affiliation systematically changing their speech-making behavior, and government ministers taking pro-government positions more consistently than government backbenchers. We also show that government-opposition polarization in Ireland spiked during the debates over the emergency budgets at the onset of the financial crisis in 2008, but subsided as government unity subsequently disintegrated.

In the second application, we use US Senate voting and speech data to compare roll-call based position estimates to our speech-based position estimates. We show a substantial increase in party polarization of Senate speech behavior over the course of 1995–2014, increases that are larger than those in consistently highly polarized roll call behavior. This increase is due to replacement of moderate speakers with more extreme speakers as well as due to increasingly polarized debates from the same senators. We show that the partisan polarization of speeches varies in response to events, dipping significantly in the aftermath of 11 September 2001, and spiking upward in the period during late 2009 and 2010 that included

debates over the Affordable Care Act ("Obamacare"). We also see variation within the two-year Congressional cycle, with the lowest partisan polarization of speeches occurring in the month immediately preceding each election. We find that whereas female senators do not vote to the left of their same-party, same-state male colleagues (Simon and Palmer 2010), they do speak to the left of them by a large amount. We conclude by discussing several straightforward extensions of our estimation strategy that would further exploit the richness of legislative speech data.

## 2. MEASURING PREFERENCE VARIATION FROM TEXT DATA

The fundamental difficulty in trying to estimate political preferences from variation in the words used in political texts is that there are several more predictive sources of variation in word use. In descending order of importance in shaping the relative frequency of words, these are: (1) language, (2) style, (3) topic, and (4) preference or sentiment. Sources of variation higher on the list tend to overwhelm those lower on the list. If you have a text in German and a text in English, the variation in the frequency of different character strings is driven almost entirely by language. Style (or dialect) is nearly as important: the words used in legal documents, in political speeches, and in tweets vary enormously. Variation in word use due to topic is sometimes as important as differences due to dialect and style. The relative ordering of these is not important for present purposes, as the variation of interest here is that due to differences in the arguments being offered or the sentiments expressed towards a proposal, which we will refer to as *expressed preferences* or *stated positions*. This variation tends to be subtle in terms of relative word counts, and therefore difficult to detect unless the higher level sources of variation are held constant.[2] Arguably it is surprising that

[2]Analogously, scaling models applied to roll-call voting data only recover plausible measure of legislator preferences when those preferences are the dominant influence on voting behavior. This is not the case in legislatures with strong party discipline, such as the UK House of Commons (e.g. Spirling and McLean 2007).

we can ever measure preferences from relative word counts, however it seems to be a general feature of political debates that political disagreement is often expressed through relative emphasis on different terms, ideas and arguments (Budge 2001; Lowe 2013).

While few political scientists would attempt to measure preferences directly from a corpus consisting of untranslated texts in different languages, or of a mixture of tweets and legal opinions, several have attempted to recover preferences from sets of texts on different topics. Diermeier, Godbout, Yu and Kaufmann (2012), for example, analyze speeches from the US Senate by combining each senator's speeches across all topics into a single document. Similarly, Proksch and Slapin (2010) scale speeches from the European Parliament by aggregating contributions across many topics by national parties. By pooling speeches across many topics, these authors have implicitly hoped that variation in topic for different speakers or parties would each discuss a similar mixture of topics, and therefore topical variation would cancel out. While this might work in some cases, we think this is unlikely to work very well in general, and therefore a more attractive approach is to use only *within-topic* variation to generate text-based preference estimates.

The question, then, is how to condition on topic. One option is to try to design a hybrid topic-scaling model that simultaneously clusters texts into topics and then describes residual word-use variation within those topics as preference variation. Unfortunately, this kind of model would rely entirely on functional form to distinguish between the word-use variation that is attributable to topic and that which is attributable to expressed preferences. Fortunately, the structure of legislative debates offers an opportunity to hold constant topic-driven word-use variation so that preference variation can be measured more credibly. If 15 speakers make statements about a legislative proposal, the relative word counts across these texts are much more likely to vary as a function of preference variation than would be the

This is also not the case in legislatures where non-ideological government-opposition dynamics dominate voting behavior, such as the Brazilian Chamber of Deputies (e.g. Zucco and Lauderdale 2011).

case if one sampled 15 speeches from across all debates. Speakers may still not all talk about exactly the same aspects of that bill, some may wander off topic, or use metaphors that introduce nuisance word use variation. But using the debate structure is still a powerful form of conditioning: probably the most powerful form available in the legislative speech context.[3]

Having exploited this opportunity to condition on debate, and thus on topic, we must then determine how to aggregate debate-specific dimensions that involve only a small subset of legislators up to smaller number of dimensions that include all legislators. This needs to be done in a way that is robust to the possibility that some of the debate-specific dimensions of word use variation will have no relationship with one another, either due to higher order sources of word use variation or due to idiosyncratic features of the debates. In many legislatures, only a subset of 'debates' are really debates in the sense that they reveal political disagreement. For example, as Quinn, Monroe, Colaresi, Crespin and Radev (2010) document, a non-trivial fraction of speech in the US Senate consists of procedural statements or symbolic statements about notable constituents, the military, and sports. To extract the politically relevant variation, we propose scaling the debate-specific scales: to take these debate-specific dimensions as noisy manifestations of one (or more) underlying latent dimension(s).

Because this approach does not rely on word use variation in any single debate to perfectly reflect the general dimensions we are interested in estimating, it gains additional robustness against other sources of variation in word usage. All we need to discover a latent dimension of disagreement is for that dimension to have general predictive power for word use variation across the set of observed debates. The exact nature of that word use variation can be

[3]Thomas, Pang and Lee (2006) exploit the debate structure of speeches in the US Congress to estimate individual legislator' support or opposition to a bill. Our approach differs from their method in that we scale legislators on a latent dimension across multiple debates, while Thomas, Pang and Lee (2006) focus on document classification within debates.

different in different debates: a word that implies a left position in one debate may imply a right position in another debate, or may imply no particular position at all. If certain debates have speech variation that seems unrelated to other debates, the model will simply estimate that those debates fail to load strongly on the recovered common dimension.[4]

This second-level scaling will tend to extract variation that is common across multiple debates, reducing biases from idiosyncratic features of particular debates. If we imagine a debate about legislation with two components, and a legislature with left and right legislators, one threat to inferring positions from speech texts is that half of the legislators on both sides of the political divide choose to talk about each of the two components of the legislation. If they do this, we will tend to recover a dimension that reflects who talked about which component of the legislation, rather than their political disagreement. This is where the second-level aggregation is helpful: unless those same individuals consistently talk about different topics across all debates, we will still recover a sensible ideological dimension from the debates where that dimension more strongly predicts word usage.

Like all identification strategies, ours has no guarantees that the assumptions will hold, and so sanity checks and other forms of validation are still needed. But this is just as true in roll-call analysis, where estimated ideal points may variously reflect legislator preferences, party inducements, government-opposition incentives, and other factors. Our methodological argument is fundamentally based on an empirical assumption: that political disagreement is more clearly and consistently reflected in within-debate variations in word use than it is in across-debate variation in word use, and therefore that the scaling methods we use for legislative speech should reflect this regularity. We think this is a better assumption than

[4]This is analogous to the reason why the Bayesian IRT model for roll call votes introduced by Jackman (2001) follows the structure of the "2PL" (two parameter logistic) rather the Rasch model (1PL) from educational testing. Under the Rasch model, it is assumed that all test items (votes) are equally responsive to the latent dimension, they differ only in their difficulty (cutpoint). Under the 2PL model, test items can vary in their correlation with ability (position), and the model will effectively ignore items that turn out to be uncorrelated with the latent dimension. This is exactly the kind of property we need to deal with the fact that word use variation is often biased by topical and other more significant variation in word use.

the ones explicitly or implicitly used in previous studies, and so it is on this basis that we proceed to specifying an estimation procedure.

## 3. SCALING TEXTS FROM SETS OF POLITICAL DEBATES

### 3.1. *Scaling Individual Debates*

At the core of preference scaling of political texts is the idea of projecting highly multidimensional variation in word usage rates onto one (or more) continuous latent dimension(s). We begin by considering the unidimensional Poisson scaling model "Wordfish" (Slapin and Proksch 2008), as applied to a set of texts within a single political debate.

For all the following discussion, we index individuals $i \in 1, 2, \ldots, N$, index debates $j \in 1, 2, \ldots, M$, and index words $k \in 1, 2, \ldots, K$.

$$w_{ijk} \sim \mathcal{P}(\mu_{ijk}) \tag{1}$$

$$\mu_{ijk} = exp(\nu_{ij} + \lambda_{jk} + \kappa_{jk}\psi_{ij}) \tag{2}$$

That is, the frequency that legislator $i$ will use word $k$ in debate $j$ depends on a general rate parameter $\nu_{ij}$ for individual $i$'s word usage in debate $j$, word-debate usage parameters $\lambda_{jk}, \kappa_{jk}$ and the individual's debate-specific position $\psi_{ij}$. The $\nu_{ij}$ parameters captures the baseline rate of word usage in a given speech, which is simply a function of the length of the speech. The $\lambda_{jk}$ capture variation in the rate at which certain words are used. The $\kappa_{jk}$ capture how word usage is correlated with the individual's debate-specific position $\psi_{ij}$. This describes a standard text-scaling model, which could be applied to (1) all speeches given in a legislative session, (2) the aggregated speeches of each legislator, or (3) applied to the speeches in a specific debate. Lowe (2008) shows that correspondence analysis provides an

approximation to a Poisson ideal point model for text data. Lowe (2013) argues that in most applications it does not make much difference which model is used; however we have found that the Poisson scaling model is more robust when a single legislator gives a speech that is very different than his/her colleagues, which happens not infrequently in the legislatures we examine. Therefore, in the analysis that follows, we use the Poisson scaling model as our debate-level scaling model.

Our identification and estimation strategies are slightly different than those used by Slapin and Proksch (2008) or by Lowe in the R package "austin". We place normal priors with mean zero on all of the sets of the parameters in the model, with standard deviation 1 for the $\psi_{ij}$ and 5 for the other model parameters. We use a similar EM estimation procedure, however we implement the Newtonian optimization steps based on the derived gradient and hessian of the log-posterior in C++ to speed estimation (Eddelbuettel and François 2011). This implementation is 20-40 times faster than the R code in "austin", enabling us to quickly estimate the model on the word frequency matrices for each of the hundreds or thousands of debates that occur in a legislative term.

### 3.2. *Aggregating Debate-Level Scales*

Having fit a Poisson scaling model (Wordfish) on individual debates to scale the legislators' relative positions within each debate, we can use normal factor analysis on the resulting debate-specific positions $\psi_{ij}$ of legislator $i$ on debate $j$ to aggregate them into a single latent position $\theta_i$ for each legislator. The fastest factor analysis methods do not apply because of the large number of "missing observations" arising from the fact that not all legislators speak in each debate. However, since the computation of the $\psi_{ij}$ is most of the computational burden, we can still achieve a very fast estimation if we adopt a fully Bayesian treatment of the factor analysis model to recover $\theta_i$, treating the $\psi_{ij}$ as data and the missing $\psi_{ij}$ as

missing at random. This assumption about the missing $\psi_{ij}$ is a consequential one because it says that the positions that legislators would take are unrelated to their decisions about whether they actually speak in order to express those positions. We discuss this issue in Section 7, both in reference to interpreting the estimates from our present approach and as an opportunity to develop richer models that jointly model positions taken and the decision to speak.

These assumptions imply a model for the debate-specific dimensions of political disagreement $\psi_{ij}$ that is linear as a function of a single latent dimension $\theta_i$, with a normally distributed error.

$$\begin{aligned}
\psi_{ij} &\sim \mathcal{N}\left(\alpha_j + \beta_j \theta_i, \tau_i\right) && (3)\\
\theta_i &\sim \mathcal{N}(0,1) && (4)\\
\alpha_j, \beta_j &\sim \mathcal{N}(0, \left(\frac{1}{2}\right)^2) && (5)\\
\tau_i &\sim \mathcal{G}(1,1) && (6)
\end{aligned}$$

This specification means that the primary dimension of word-usage variation in individual debates $\psi$ can be more or less strongly associated with the aggregate latent dimension $\theta$ being estimated across all debates, with either positive or negative polarity for any particular debate. Essentially, this allows the model to select out those debate-specific dimensions that reflect a common dimension (large estimated values of $\beta_j$), while down-weighting the contribution of debates where the word-usage variation across individuals seems to be idiosyncratic ($\beta_j \approx 0$). The priors on $\theta_i$ and $\beta_j$ allow the model to remain agnostic about the relative polarity of individual debate dimensions, while constraining the common latent dimension of interest to a standard normal scale. Following a similar idea in roll-call analysis (Lauderdale 2010), $\tau_i$ varies by speaker to capture the possibility that some speakers are more inconsistent across debates than others.

This two stage procedure is very fast. We are able to estimate the model on recent sittings of the Irish Dáil, with about 1000 debates and 10,000 speeches, in a few minutes. These estimations recover full Bayesian posteriors over $\theta_i$ at the second stage. Since Poisson scaling models do not generally give meaningful uncertainty estimates (see Section 5), little is lost by failing to propagate uncertainty from the debate-specific scales. More fundamentally, it is the second-stage uncertainty that is of interest. The intervals we report indicate whether there are a sufficient number of debates with sufficiently overlapping sets of speakers in order to be confident that particular legislators are consistently speaking in different ways. We discuss costs and benefits of combining the two estimation stages in Section 7.

## 4. DATA

We use data from the Irish Dáil (the lower house in Ireland) and the US Senate to demonstrate our approach. The Irish data includes the two latest complete legislative sessions, the 29th Dáil (2002–2007) and the 30th Dáil (2007–2011). Data for the US Senate includes all speeches from the 104th to the 113th Senate, which covers almost 20 years of legislative debates (January 1995–June 2014). We collected all speeches from either existing databases of legislative debates or from official parliamentary records.[5] Before we scaled speeches and debates, we removed contributions from the person officially presiding over the chamber. In Ireland, this is either the Ceann Comhairle (speaker) or Leas-Cheann Comhairle (deputy speaker). In the US Senate, we removed speeches from the Presiding Officer and the President pro tempore. We further removed procedural debates, such as the discussion of the

[5]For Ireland, we retrieved speeches from "DPSI: Database of Parliamentary Speeches in Ireland" (Herzog and Mikhaylov 2013), which includes all speeches from the Irish Dáil from 1919 to 2013. Information in this database was collected from the Houses of the Oireachtas (the Irish national parliament) and is distributed under the *Public Sector Information (PSI) Licence for Re-Use of Information*, No. 2005/08/01. Speeches from the US Senate were collected from the digital version of the Congressional Record (http://www.gpo.gov/fdsys/browse/collection.action?collectionCode=CREC, last accessed June 5, 2014) using a web scraper and parser written in Python. We thank Justin Grimmer for sharing this script with us.

meeting agenda, prayers, tributes, elections of the speaker, points of order, and any other discussions concerning the rules of parliamentary procedure. Finally, we removed punctuation, numbers, and stop words and reduced words to their stem.

A key step in organizing the data was to identify speeches that belong to the same debate. We defined a debate as a set of speeches with the same title (as reported in the official parliamentary records) and that were held on the same day and included at least five speakers. One could argue against this definition that legislative debate on a single question can sometimes span multiple days or even weeks. Even setting aside the relative difficulty of operationalizing this kind of broader definition, we nevertheless think it is preferable to limit the definition of a debate to a single day because the content and context of a debate can change from one day to the next. Within each debate, we combined all contributions of a legislator into a single composite speech, excluding contributions with less than 50 words because they are usually interruptions.

Table 1 provides an overview of the speeches and debates included in our analysis. On average, there were 8,690 speeches across 770 debates in the Irish Dáil and 7,330 speeches across 574 debates in the US Senate. The 112th Senate (2011–2013) sticks out at as the least productive Senate in terms of number of speeches and debates (ignoring the 113th Senate, which is still in session). This is most likely the result of legislative gridlock after the Republican party took over the House in the 2010 election.

The average debate in our data set consists of 13 speakers, ranging from a minimum of 5 speakers (our lower threshold) to as many as 73 speakers. At the level of individual speeches, we find an average of 560 to 729 words per speech and legislature. The longest individual speech in our data set is 41,851 words long, which is senator Ted Cruz's (TX-R) 21-hour filibuster speech in September 2013.

Table 1: Summary statistics for speeches and debates

*Number of observations by legislature*

| | No. of speakers* | No. of speeches | No. of debates | No. of unique words |
|---|---|---|---|---|
| *Irish Dáil* | | | | |
| 29th (2002–2007) | 165 | 10,043 | 933 | 38,413 |
| 30th (2007–2011) | 165 | 7,211 | 608 | 31,156 |
| *US Senate* | | | | |
| 104th (1995–1997) | 101 | 9,449 | 667 | 60,872 |
| 105th (1997–1999) | 100 | 8,043 | 606 | 55,662 |
| 106th (1999–2001) | 100 | 7,874 | 625 | 56,300 |
| 107th (2001–2003) | 100 | 8,412 | 678 | 53,590 |
| 108th (2003–2005) | 99 | 7,934 | 630 | 55,725 |
| 109th (2005–2007) | 101 | 7,985 | 585 | 55,278 |
| 110th (2007–2009) | 101 | 7,633 | 615 | 55,294 |
| 111th (2009–2011) | 107 | 7,211 | 584 | 57,394 |
| 112th (2011–2013) | 101 | 5,188 | 442 | 45,915 |
| 113th (2013–2015)† | 104 | 3,572 | 306 | 42,253 |

*Number of speakers by debate and length of speeches*

| | Number of speakers by debate | | | Length of speeches by debate ($N$ words) | | |
|---|---|---|---|---|---|---|
| | mean | min | max | mean | min | max |
| *Irish Dáil* | | | | | | |
| 29th (2002–2007) | 11 | 5 | 46 | 726 | 21 | 7,579 |
| 30th (2007–2011) | 12 | 5 | 47 | 729 | 21 | 8,547 |
| *US Senate* | | | | | | |
| 104th (1995–1997) | 14 | 5 | 70 | 607 | 16 | 9,818 |
| 105th (1997–1999) | 13 | 5 | 65 | 595 | 19 | 9,144 |
| 106th (1999–2001) | 13 | 5 | 72 | 585 | 18 | 16,086 |
| 107th (2001–2003) | 12 | 5 | 73 | 560 | 19 | 13,392 |
| 108th (2003–2005) | 13 | 5 | 57 | 604 | 19 | 22,403 |
| 109th (2005–2007) | 14 | 5 | 58 | 595 | 21 | 11,302 |
| 110th (2007–2009) | 12 | 5 | 58 | 611 | 21 | 7,299 |
| 111th (2009–2011) | 12 | 5 | 57 | 624 | 21 | 8,099 |
| 112th (2011–2013) | 12 | 5 | 43 | 565 | 20 | 8,940 |
| 113th (2013–2015)† | 12 | 5 | 67 | 671 | 20 | 41,851 |

*Notes:*

* Speakers only include members with a seat in the legislature.

† Data includes all speeches and debates until June 5, 2014.

# 5. IRISH DÁIL

## 5.1. *Comparison to Wordfish*

In this section, we use speech data from the 29th and 30th Irish Dáil to demonstrate that our approach outperforms Wordfish when applied to speeches from a full legislative session. There were 165 legislators who spoke in the 29th and 30th Dáil. The 29th Dáil had a governing coalition of Fianna Fáil and the Progressive Democrats. The latter was a small center-right/liberal party that formed in 1985 and dissolved in 2009, with its remaining members joining Fianna Fáil. The 30th Dáil added the Green Party to that coalition. The largest opposition party in both parliaments was Fine Gael, the second largest party after Fianna Fáil at that time. Both parties are centrist parties with similar policy positions that have historically been divided over the question of Ireland's relationship with Great Britain. The other main opposition parties were the Labour Party and Sinn Féin.

In Figure 1, we show the mean party positions from Wordfish and from our approach by comparison to two benchmarks: whether the parties are in the governing coalition, and the left-right location of the parties as estimated using expert surveys conducted by Benoit and Laver (2006). Based on these estimates, it appears that in the Irish data our approach is primarily recovering government versus opposition conflict, rather than left-right ideology. There are two ways to see this. First, while the largest parties Fianna Fáil and Fine Gael are generally viewed to be ideologically moderate in left-right terms, we estimate them at or near the extremes of our dimensions. Note in particular the fact that the Labour Party is estimated to be more centrist than Fine Gael, which only makes sense if we think of this as government-opposition. Second, when the Green Party joins the coalition in the 30th Dáil, it moves from having a similar average position to Fine Gail to having nearly the same position as Fianna Fáil.

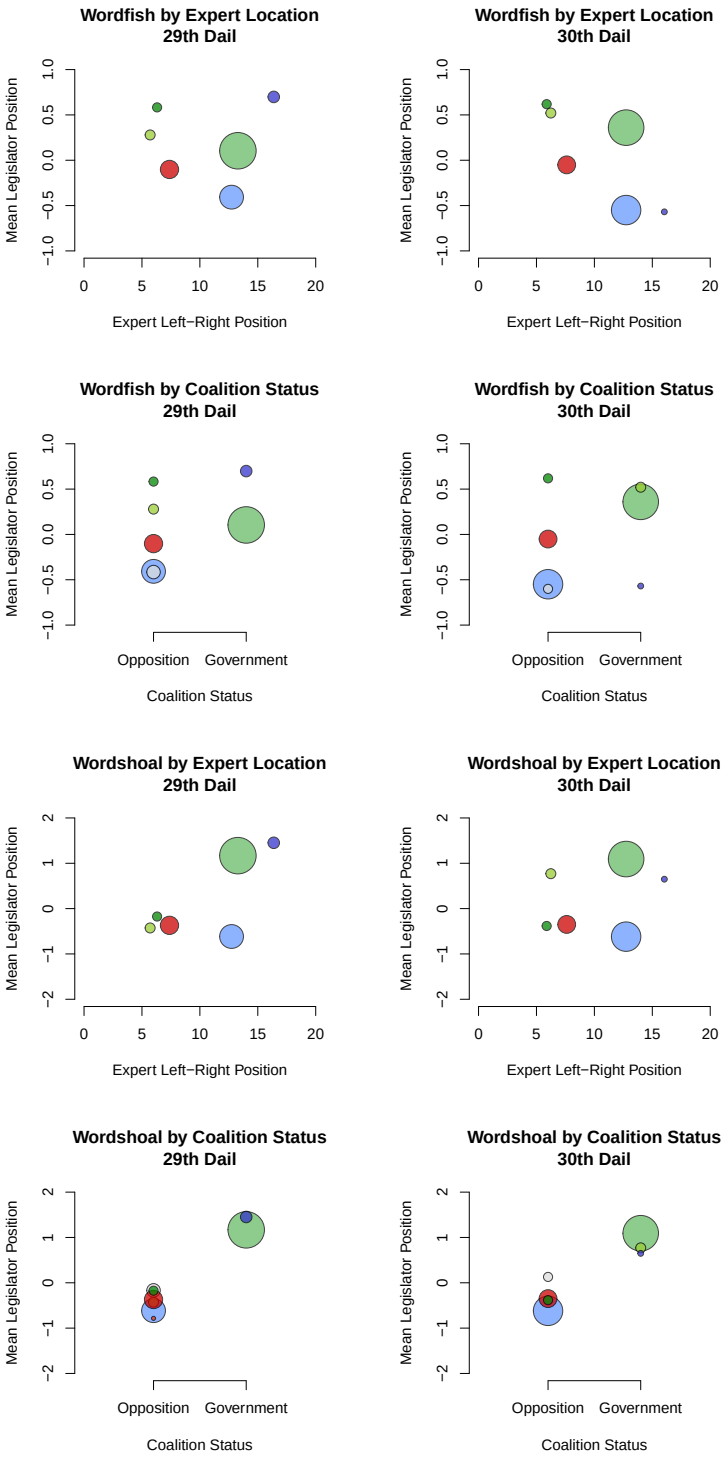


Figure 1: The top two rows show the association between party average Wordfish scores and expert assessed left-right position (1st row) and coalition status (2nd row). The bottom two rows show the corresponding relationship for Wordshoal scores. The left column shows the 29th Dáil, the right column the 30th Dáil. Point size is proportional to the number of legislators in each party.

In contrast, Wordfish estimates do not seem to consistently reflect the coalition structure of the Dáil. The Green Party has a similar estimated position to Fianna Fáil, both when they are in coalition and when they are not. The Progressive Democrats are at one extreme of the dimension in the 29th Dáil and the other in the 30th, despite no change in coalition status. Neither do these estimates seem to reflect the ideological cleavages of the Dáil as assessed by expert surveys. In particular, experts do not place the Labour Party between Fine Gail and Fianna Fáil, but Wordfish does in both the 29th and 30th Dáil. In general, the association between the party locations from Wordfish and from the expert surveys are very weak.

When we look at individual legislators, rather than the party means, we can see the association between our estimates and coalition status even more clearly. Figure 2 shows the relationship between the estimated legislator positions and the coalitions under both Wordfish and our estimates. In the 29th Dáil, the correlations are 0.89 for our estimates and 0.16 for Wordfish. In the 30th Dáil, the correlation between being in the coalition government and mean posterior position rises to 0.93, versus 0.36 using Wordfish.

Figure 3 shows that Wordfish gives implausibly narrow uncertainty intervals. The uncertainty estimates for legislators from Wordfish reflect relative fit of different positions in predicting words across all texts given the Poisson functional form assumption of that model. Wordfish, like LDA and other multinomial and poisson text models, is overconfident in its estimates for similar reasons to why Poisson regression coefficient estimates are overconfident when data are overdispersed. In contrast, the uncertainty intervals for the Wordshoal model more plausibly reflect uncertainty about the relative positions of speakers. These uncertainty intervals reflect the number of debates each legislator speaks in, the extent of overlap between speakers in different debates, and the extent to which legislators are consistently ordered (by debate-level Wordfish) across the debates they speak in. This is the relevant kind of uncertainty for assessing if we have enough data to say that a particular legisla-

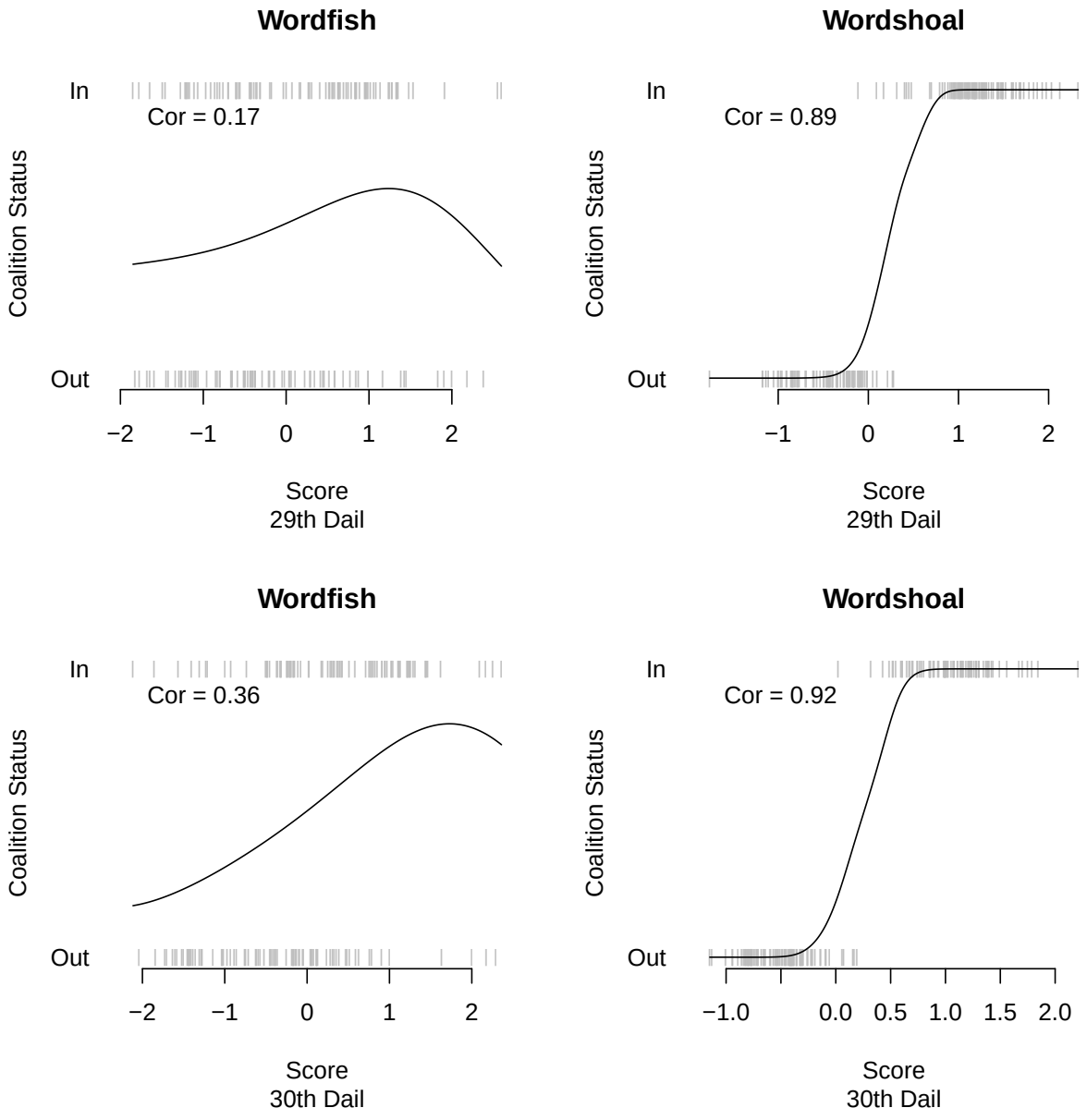


Figure 2: The association between the estimated positions of each legislator and their status as members of the coalition versus opposition, with correlation and local linear smooth, under Wordfish (left) and our approach (right), for the 29th (top) and 30th (bottom) Dáil.

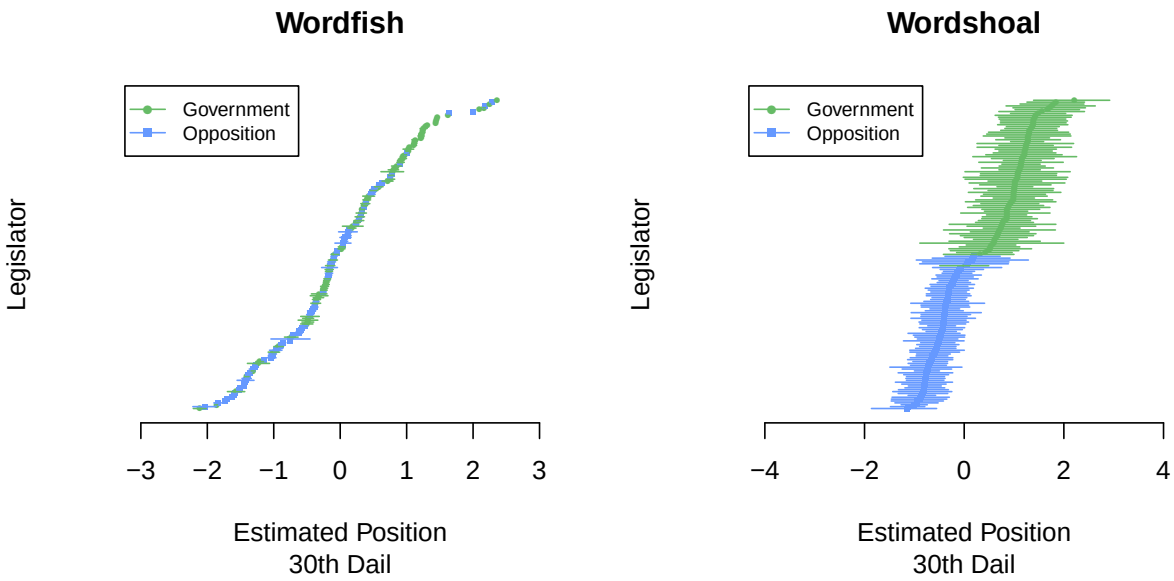


Figure 3: The 95% intervals associated with the estimates for each legislator under Wordfish (left) and Wordshoal (right), for the 30th Dáil.

tor takes different positions from another legislators across their many speeches in partially overlapping debates.

In sum, in these data Wordshoal recovers point estimates that measure a meaningful quantity and provide uncertainty intervals that reflect realistic uncertainty about that quantity. Applied in the manner of previous studies, Wordfish neither recovers plausible measures of policy preferences nor plausible measures of government-opposition disagreement. Wordshoal very clearly recovers the government-opposition dimension of disagreement in Ireland. Recalling the identification strategy underlying Wordshoal, and thinking about the Irish context, this is hardly surprising. Wordshoal aims to recover the dimension that best explains disagreement—as measured by variation in word use—across all debates. In a Westminster parliamentary system with strong party discipline like Ireland's, it is hardly surprising that the single factor that most consistently shapes speech behavior across every debate, is whether a legislator's party is in government or opposition.

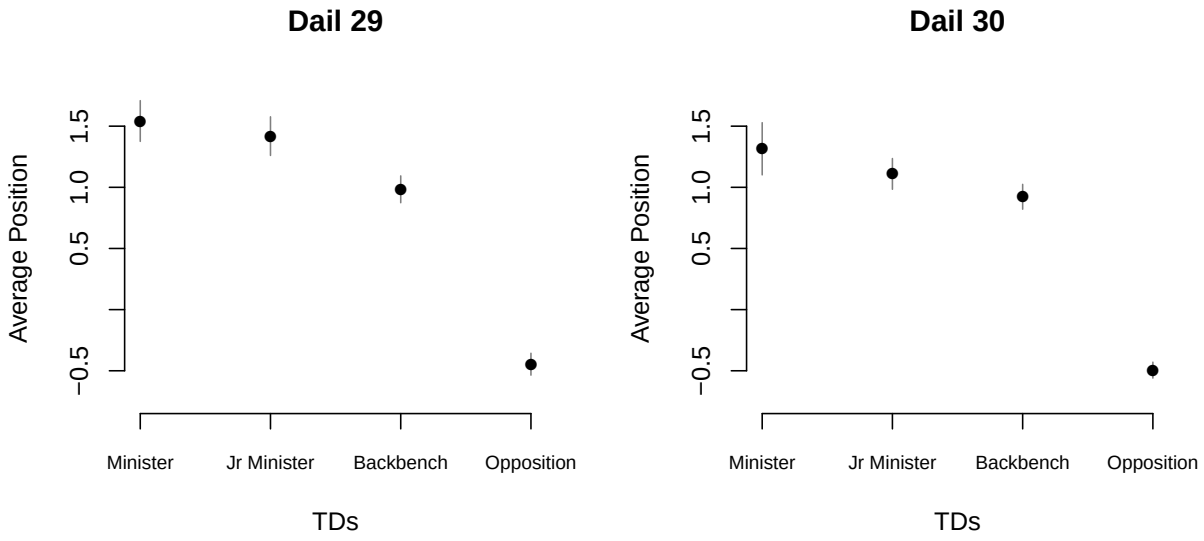


Figure 4: Mean positions of opposition speakers, government backbench TDs, junior ministers, and ministers for the 29th and 30th Dáil. Confidence intervals are constructed by non-parametric bootstrap, resampling over TDs.

### 5.2. *Cabinet Members versus the Backbench*

Having validated the estimates as reflecting a government-opposition dimension in speech, we can begin to explore how TDs vary in position along this dimension. A core feature of a parliamentary system like Ireland is the dominant role of the cabinet. Its members are bound by the doctrine of collective cabinet responsibility, which requires that cabinet members publicly support decisions made by the cabinet even if they privately disagree. We hence expect ministers to more reliably defend the government position than government backbenchers. We can assess whether this is the case in our data by comparing the average locations of TDs inside and outside the cabinet.

Figure 4 shows the average Wordshoal positions for cabinet ministers, junior ministers, government backbench TDs and opposition members.[6] As expected, we find that cabinet

[6]If a member had multiple positions or transferred from one position to another during the legislative term, we counted the position with the longest duration.

members are the most pro-government speakers. In the 29th Dáil, the average cabinet minister position is 1.52, which is significantly different from the position of backbench TDs at 0.98 ($t = 5.67$, $p = 3 \times 10^{-6}$). In the 30th Dáil, the difference is slightly smaller with positions at 1.27 and 0.92, respectively, but still significantly different ($t = 3.55$, $p = 0.001$). The average position of junior ministers is not significantly different from the average minister position, which indicates that cabinet members are equally bound by collective cabinet responsibility regardless of their rank.

The measured difference between ministers and government backbench TDs' speeches is in line with our expectations for a parliamentary system like Ireland. As a general principle, government backbenchers are less bound in what they say than cabinet members, and we see this clearly in our estimates. The next step in analyzing these data would be to explore whether backbench deviations from the government line seem to reflect strategic considerations, such as promoting the particularistic interests of constituencies. Regardless of one's interpretation of these results, observations of this type could not be made from voting records, which show almost perfect unity because of strong party discipline (Hansen 2009). Our estimation strategy therefore provide avenues for future research on legislative behavior in parliamentary systems that otherwise would not be possible.

### 5.3. *Variation in Disagreement*

In addition to this kind of comparison of estimated positions, the Wordshoal model facilitates assessments of the relative extent to which different kinds of debates align with the estimated common dimension. The greater the magnitude of the $\beta_j$, the greater the extent to which the common dimension predicts speech variation in a given debate. Thus, in the context of the Irish Dáil, if we compare the average magnitudes of the $\beta_j$ across debates on different parliamentary matters or over time, we can assess which debates see greater differences in

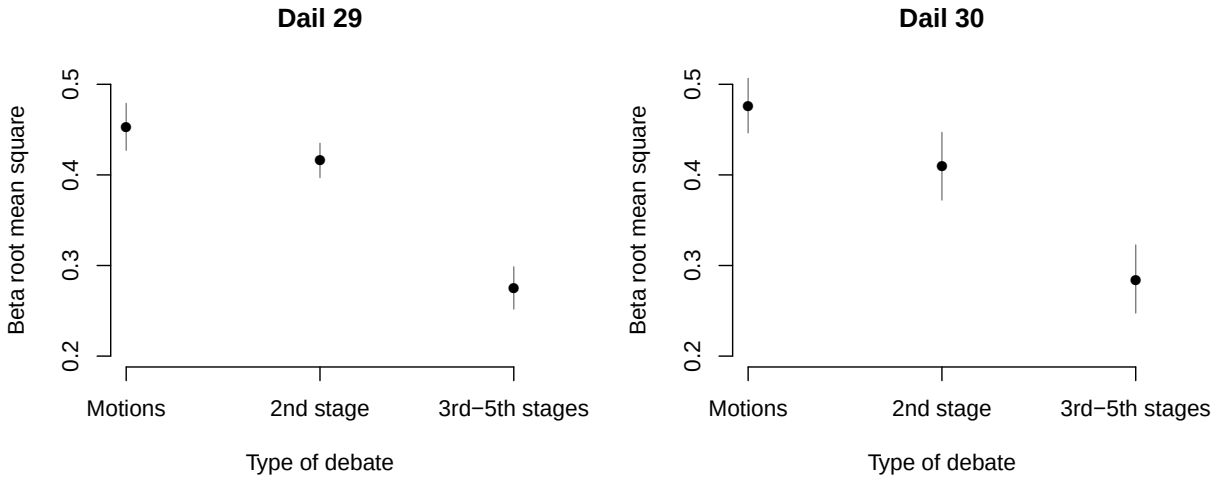


Figure 5: Weighted root mean square of $\beta_j$ by type of debate for the 29th and 30th Dáil.

the language used between the government and opposition. As our summary statistic for $\beta_j$, we use the root mean square of the $\beta_j$, weighted by the number of speeches in each debate $\text{speeches}_j$.

$$\sqrt{\frac{\sum_j \text{speeches}_j \cdot \beta_j^2}{\sum_j \text{speeches}_j}} \tag{7}$$

We first compare this statistic across different types of parliamentary debates. The majority of debates in our data take place during the second reading of a bill, which is the most important legislative stage after which a bill is formally accepted or rejected. The second most common type are motions, which are an instrument of parliament to scrutinize the work of the government. This includes ad hoc motions on topical issues, seasonal adjournment debates, and (less frequently) motions of confidence in the government or in individual cabinet members (Gallagher 2010).

Figure 5 shows the weighted root mean square of $\beta_j$ for motions, second-stage bills, and for debates during the remaining legislative stages, which we group together because of the small number of observations at each stage. We find high government-opposition division during debates on motions, which are mostly used by the opposition parties to

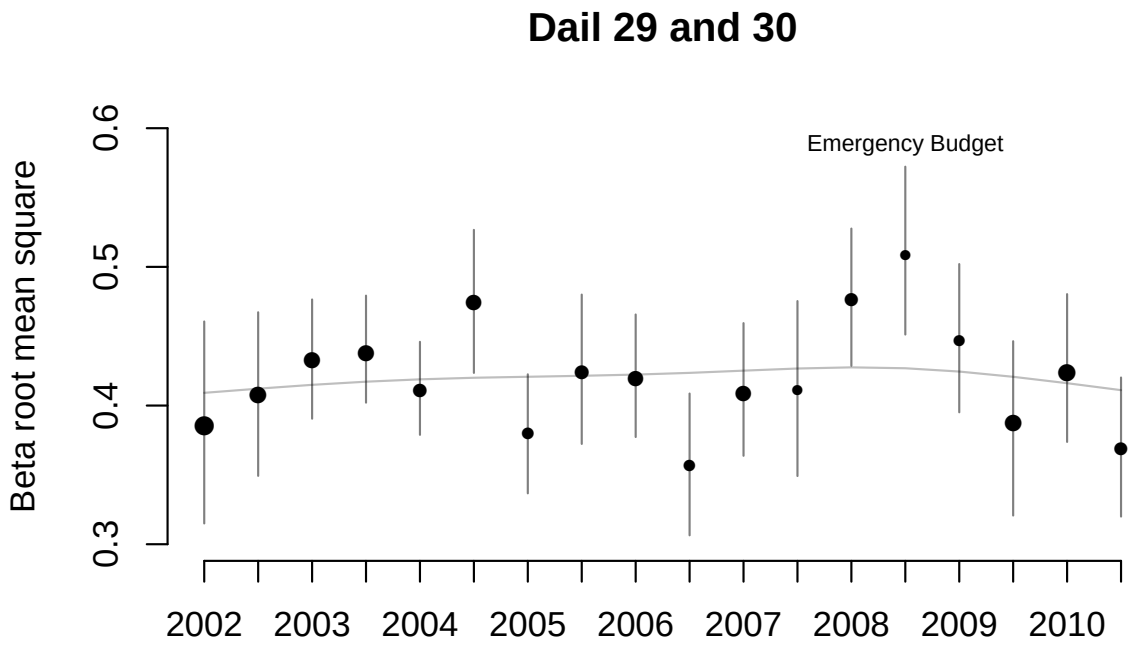


Figure 6: Weighted root mean square of $\beta_j$ over six-months periods. The size of each point is proportional to the relative number of second-stage debates in each period.

express grievance over government decisions, and hence elicit a strong divide between the opposition and government members. For similar reasons, we find nearly as high polarization in second-stage debates. In a parliamentary system with a majority government, legislation is almost exclusively initiated by the cabinet. The main debate of these bills then provide the opposition with another opportunity to criticize the government for its work. Once a bill has passed the second stage and is all but guaranteed to enter into law, debates become less strongly associated with the government-opposition dimension.

Finally, we turn to a comparison of debates over time. During the period covered by our data (2002–2011), Ireland went from boom to bust, with the end of rapid economic growth ("Celtic Tiger") in early 2000, followed by a period of average growth until the collapse of

the financial market and banking system in 2008/09. In Figure 6, we plot the weighted root mean square of $\beta_j$ for six-months period for both the 29th and 30th Dáil. Because we have found above that government-opposition polarization is higher during the second reading of a bill, we plot each point in Figure 6 proportional to the relative number of second-stage debates in each period.

The plot shows an increase in the root mean square of $\beta_j$ during 2008 despite a relative small number of second-stage debates. This increase coincides with the onset of the crisis and the first emergency budget of the government in October 2008. Our analysis therefore provides evidence for an increase in government-opposition polarization when legislators started to debate solutions to the financial crisis. After 2008, polarization returns to normal levels and drops below the trend line in the second half of 2009 and even further so in 2010, shortly before the Fianna Fáil-Green coalition collapsed over internal divisions. It seems that the increasing disagreement within the governing coalition decreased the observable government-opposition divide, which is in line with findings in Herzog and Benoit (2013) based on Irish budget debates.

We also find significant deviations from the trend line in 2004/2005 and in the second half of 2006. Both time periods coincide with the electoral cycle: local and European elections (held on the same day) in June 2004 and a national election in early 2007. While more data would be needed to assess whether these are consistent patterns, there is some suggestion here that government-opposition polarization increase during local/European elections, but decrease before a national election, which also coincide with a decrease in second-stage debates.

## 6. US SENATE

In our analysis of the US Senate, we pool all debates from January 1995 to early June 2014, covering the 104th to the 113th Senate. We fit a model where senators are assumed to have constant positions. While it is not difficult to model dynamic positions, an analysis using the constant position assumption enables a comparison of the degree to which polarization over this period has occurred due to senator replacement versus the same senators having more partisan debates. For purposes of comparison, we also fit a model with the same structure to all roll-call votes over the same period.[7]

Figure 7 shows the Wordshoal scores and 95% intervals of the senators serving in the 104th Senate (1995-1996) and the 113rd Senate (2013-2014). The partisan polarization of senators due to replacement is visually apparent from the increased degree to which the scores correlate with party. In the 104th, Democratic senators Ford (KY), Exon (NE), Hollings (SC), Bumpers (AR), Breaux (LA), Conrad (ND), Reid (NV), Dorgan (ND), Biden (DE), Baucus (MT), Pryor (AR), Nunn (GA) and Johnston (LA) spoke like Republicans. This list includes nearly all of the Democrats from the South as well as several from states like Montana and North Dakota that typically voted Republican in Presidential elections and Democratic in Congressional election in the preceding decades. The Republicans interspersed among the Democrats on the left side of the estimated dimension—Packwood (OR), Snowe (ME), Jeffords (VT), Kassebaum (KS), DeWine (OH), Brown (CO), D'Amato (NY), Roth (DE), and Chafee (RI) mostly come from the Northeast or were known as moderates during their careers. In contrast, in the 113th, there is much cleaner separation between the parties, with all of the overlapping senators either having served so briefly that there is substantial estimation uncertainty in their position or coming from electorally marginal states.

[7]Like the debate-score aggregation model, this is a heteroskedastic-by-legislator scaling model (Lauderdale 2010).

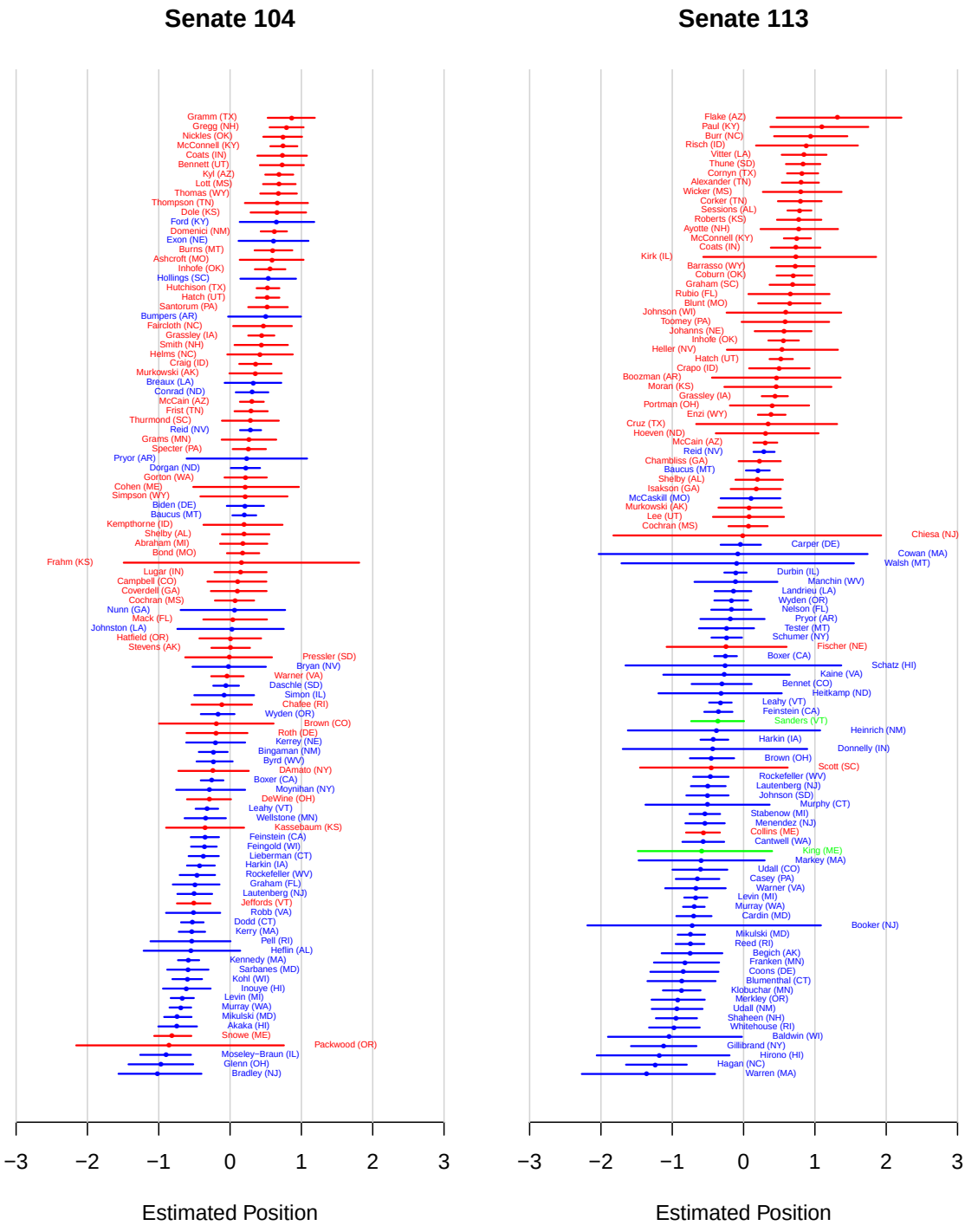


Figure 7: Wordshoal estimates for the 103rd and 113th US Senates.

### 6.1. *Speech Positions versus Roll Call Positions*

Figure 8 shows the increasing polarization from 1995 to 2014 of senators' Wordshoal speech positions (top panel) as well as the equivalent analysis for roll-call scores (bottom panel). A striking difference is that while roll-call scores have barely polarized over these two decades, the gap between the average party positions in speeches has doubled. New Republicans and Democrats have voted similarly on average to those same-party senators that they have replaced; however, newly elected senators speak in more partisan ways than those they replace. Whereas Republican and Democratic senators used to substantially overlap in how they spoke on the floor, that overlap has mostly disappeared over the last two decades due to turnover.[8]

### 6.2. *Speech Polarization versus Roll Call Polarization*

Turnover is not the sole cause of increasing polarization of speeches though. Following a similar logic to our analysis of the different readings of bills in the Dáil, we compare the speech-weighted root mean square $\beta_j$ over time for the US Senate. The top panel of Figure 9 shows this quantity within 8 month periods over the same Congresses. The 8 month periods run from January to August in the year after an election, from that September to April of the following year, and from May through December in the year of the next election. In general, polarization of debates, holding fixed the composition of the Senate, held steady during the Clinton administration, rose during the Bush administration, and has held steady at a higher level under Obama. As we saw earlier, increasing polarization due to senator

[8]Journalists have been writing articles for decades about the decline of inter-party socializing and collaboration in the Senate. For a recent example, "On Senate Menu, Bean Soup and a Serving of 'Hyperpartisanship'", http://www.nytimes.com/2014/08/20/us/politics/senate-dining-room-is-one-more-casualty-of-partisanship.html.

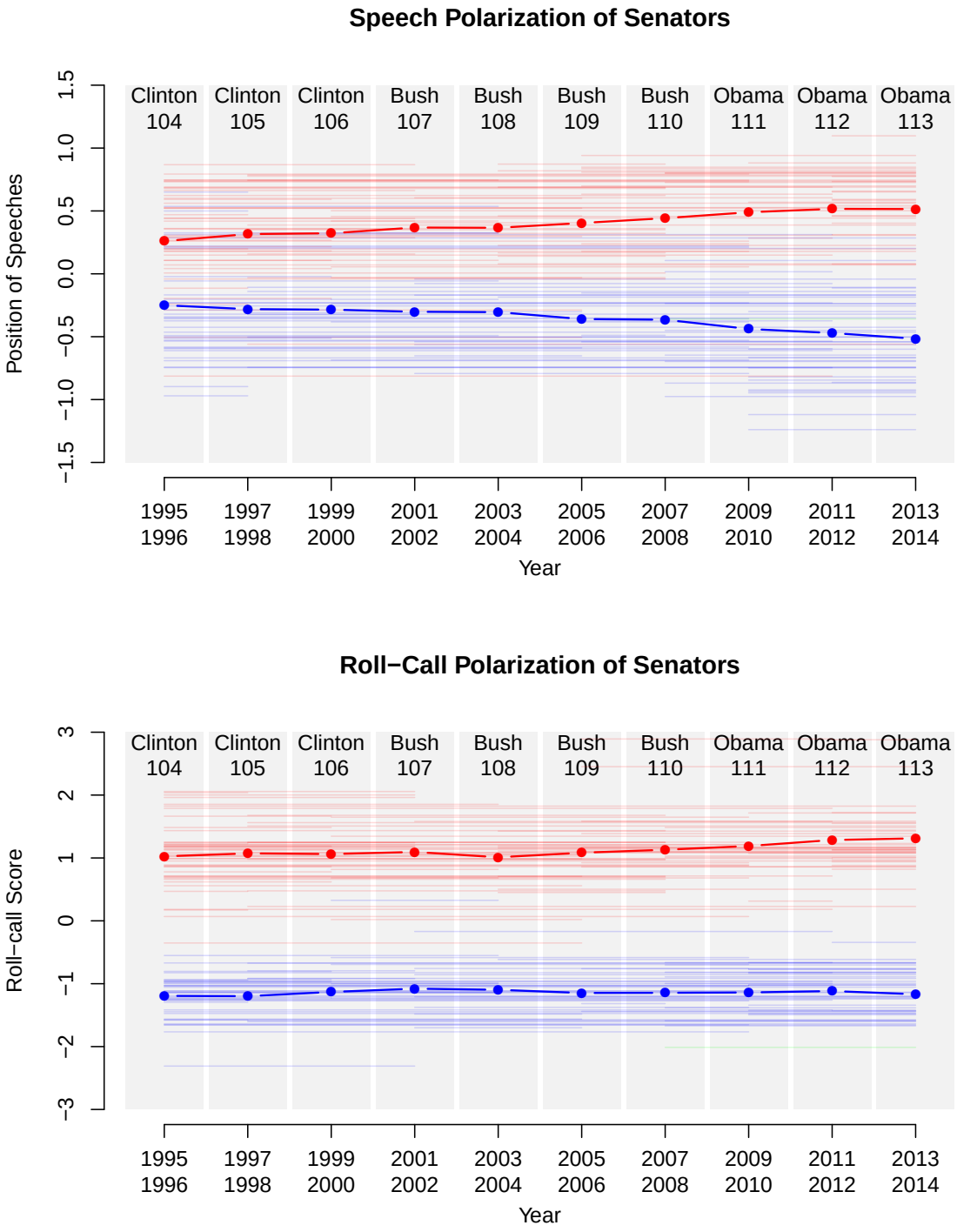


Figure 8: Average party positions in speeches (top panel) and in roll-call votes (bottom) from the 104th Senate (1995-1996) to the 113th Senate (2013-2014)

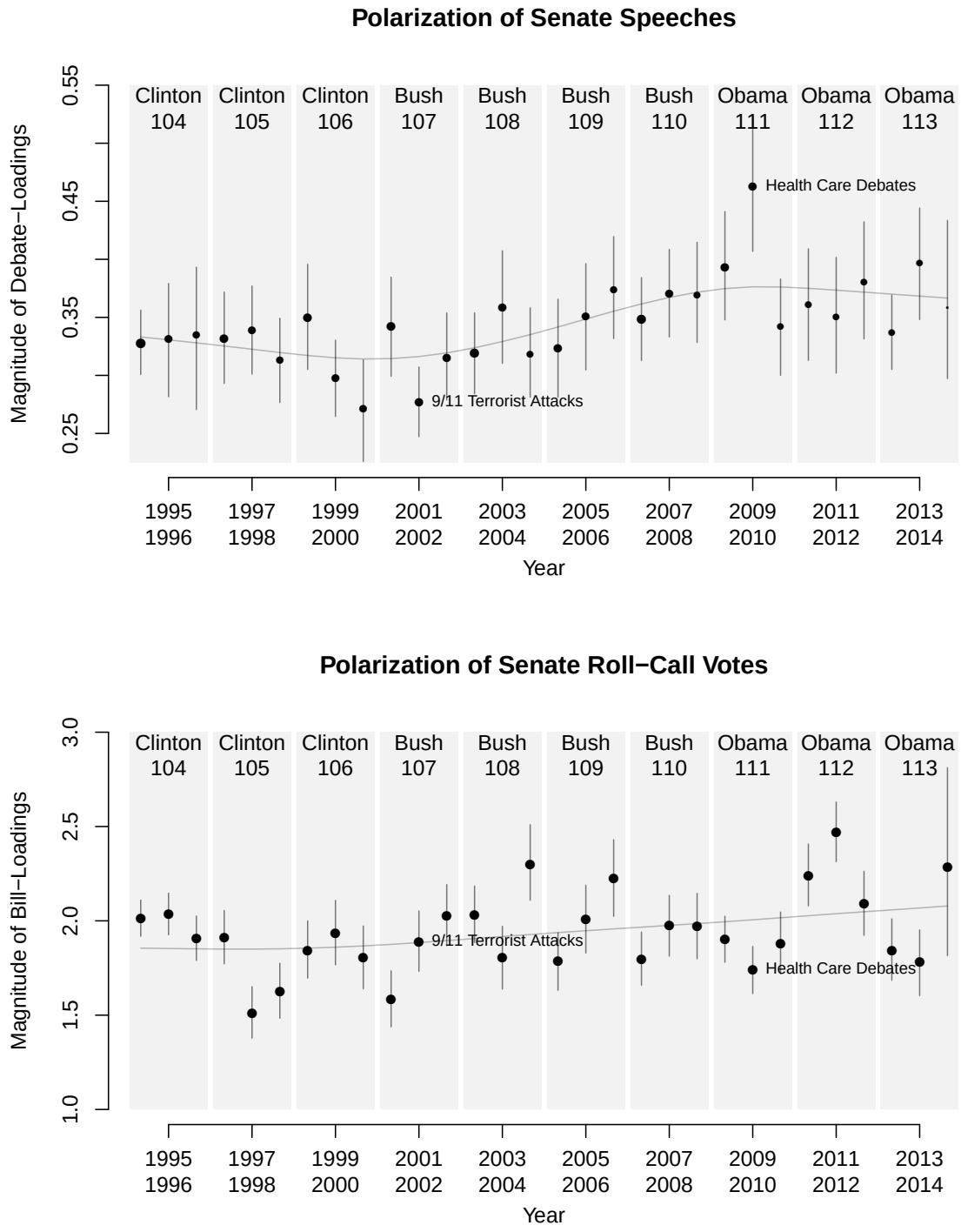


Figure 9: Average debate loadings (top panel) and average roll-call vote loadings (bottom) from the 104th Senate (1995-1996) to the 113th Senate (2013-2014)

turnover was occurring during all three periods.

Two 8-month periods deviate from this overall trend in the sense of being statistically significant outliers from the spline regression fit depicted in the figure. First, the period from September 2001 to April 2002, which began with the terrorist attacks on September 11 and included the US invasion of Afghanistan that started one month later.[9] The period with the highest polarization of debates, by far, is the period from September 2009 to April 2010 that included the Senate debates on health care legislation introduced by President Obama.[10]

The bottom panel of the same figure shows the equivalent trajectory for roll-calls, which shows different patterns. The health care debates, which occupied so much Senate time in 2009-2010, fail to register as they involved relatively few roll-call votes. The highest point in roll-call polarization instead comes from September 2011-April 2012, a period generating little major legislation, but occurring immediately after the debt-ceiling crisis of July-August 2011. There is some general upward trend over the period, but there is substantial variation within Congresses.

Does this variation within Congresses follow a pattern, either for the roll-call data or the speech data? The top panel of Figure 10 shows that variation within the congressional calendar is not large, however there is some suggestion that speech polarization is generally highest in the middle of a Congress. The most distinctive feature of the speech polarization series is that it has its lowest level in the month (October) immediately preceding a congressional elections. This suggests some strategic tendency towards moderation, either in the scheduling of debates or in individual speech behavior (or both). The pattern for roll-calls

[9]The only period with a lower point estimate for this measure of debate polarization is the final three months of the Clinton administration, which included the final month of the 2000 Bush v. Gore election and the period of the recount and Supreme Court case. However, very little of this partisanship reached the Senate floor, as it did not implicate Senate business. The estimate is relatively uncertain due to the smaller number of debates in the period, so it is not significantly below the trend.

[10]The second highest level of polarization occurs in the period from September 2013 to April 2014, which included a government shutdown triggered by Republican efforts to defund the health care legislation, although this is not significantly above the long-run trend.

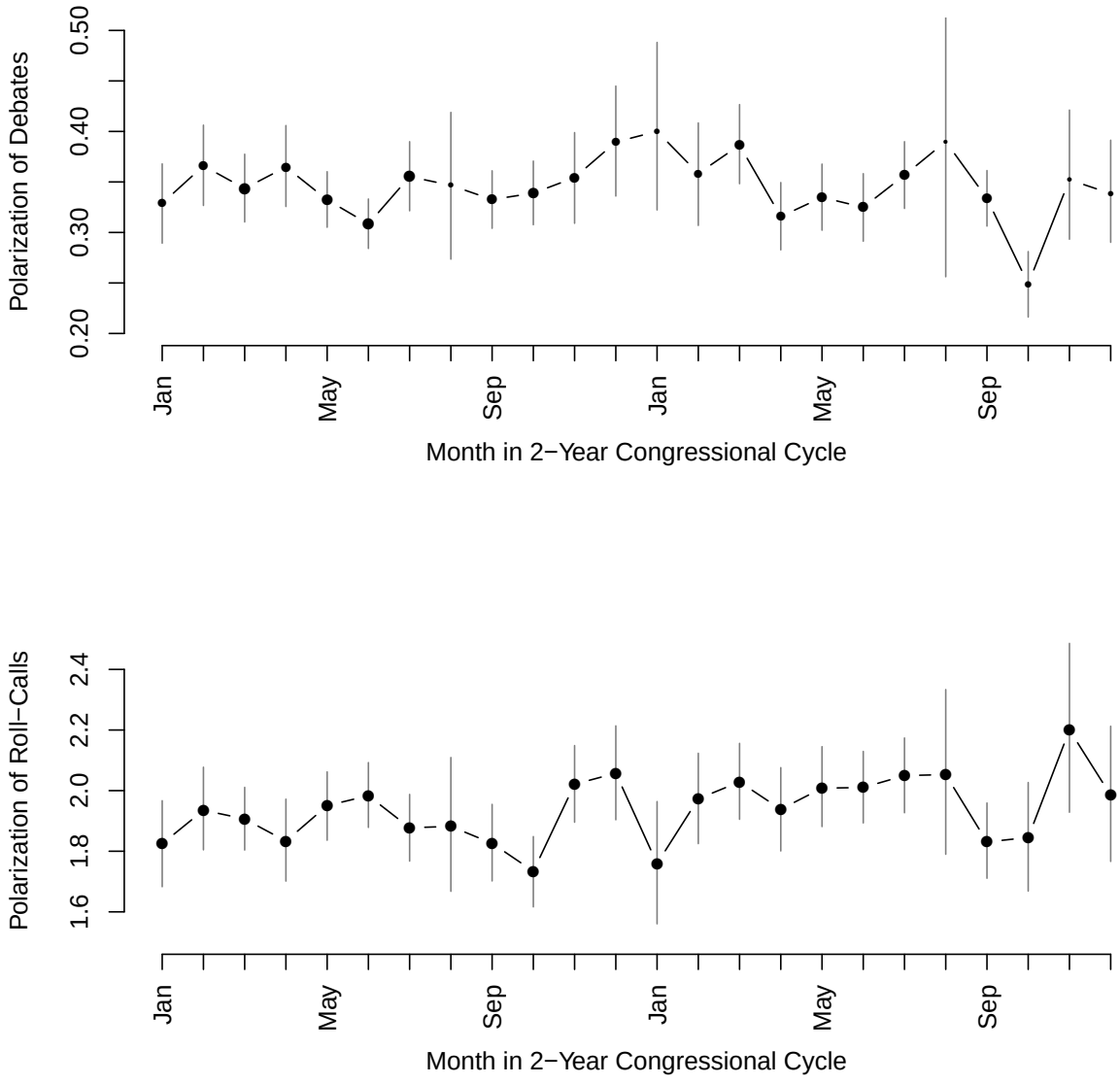


Figure 10: Average magnitude of debate loadings over the 24 months of a Congress, across the 104th to 113th Congress. The size of each point is proportional to the number of debates.

(bottom panel) shows some indication of an upward trend over the congressional cycle and some evidence of lower polarization in the two months before an election relative to that trend.

Across these analyses, we find evidence that the trend in speech polarization has some of the same temporal features as the trend in polarization in voting behavior, but what is unambiguously clear is that speech polarization is far more variable over the period of time we examine. While voting behavior has been consistently highly polarized by party since 1995, the polarization of speech behavior by senators has increased substantially due to both replacement and increasingly polarized debates, as well as increasing or decreasing in response to external events, the legislative agenda, and the electoral calendar.

These patterns highlight one of the reasons that speeches are worth studying in their own right. Rhetoric can be dialed up or down. Speeches are at once more visible and less costly opportunities for senators to emphasize or deemphasize partisan differences, as the political climate dictates. Our results suggest that senators use these opportunities in response to political conditions, even as aggregate voting behavior changes relatively little.

### 6.3. *Gender, Speeches and Votes*

There is a long-running debate among scholars about whether and how female legislators represent their constituents differently than male legislators. Some research on roll-call scores has shown that the average female representative in the US votes to the left of the average male representative from the same party. However, more recently, a study by Simon and Palmer (2010) suggests that this may be an artifact of the states from which women are most likely to be elected. They construct comparisons of female members of the US House, not to all other same-party representatives, but rather to male same-party representatives who immediately preceded or succeeded them. In this analysis, there is no difference between

same-party, same-seat men and women in roll-call behavior.

But roll-calls are not the only way that legislators represent their constituents, and speeches give a different lens on the positions that legislators are taking. We apply a variant on the Palmer and Simon identification strategy here, comparing the 9 female Republican and 20 female Democratic senators in our data set to same-party male senators who preceded them, succeeded them, or served alongside them in the other seat from the same state. At least one such male senator exists for 22 of the 29 females, and so we restrict our analysis to those senators.[11] We compare the female senator's speech scores and roll-call scores to the average scores of the 1-3 same-party males identified through this matching procedure.

On average, the female senators' speech scores are 0.52 to the left of their same-party, same-state male colleagues. This is a strongly significant difference: $t = -6.8$, $p = 8 \times 10^{-7}$, and the 95% interval runs from -0.68 to -0.36. It is also a substantively large difference: the within-party standard deviation of the Wordshoal scores is only 0.40. Consistent with the findings of Simon and Palmer (2010), we find no such difference in roll-call voting behavior. The difference in means for roll-call scores is just 0.02, with $t = -0.3$, $p = 0.74$, and the 95% interval runs from -0.19 to -0.14.[12]. To check that this is not the result of a small number of outliers, we also perform a sign test of whether the fraction of women who are to the left/right of their same-party, same-state colleagues is different from 0.5. Out of the 22 female senators, only 2 have speech scores to the right of their same-state male colleagues,[13] providing very strong evidence $p = 0.0001$ against the null hypothesis that there is no general difference between men and women in speech and that these differences arose by chance from individual-level variation in speech behavior. No such pattern is observed for roll-call votes,

[11] Within the time period we consider, there are no same-state male Democratic senators to match to Boxer (CA), Feinstein (CA), Cantwell (WA), Murray (WA), Carnahan (MO), McCaskill (MO), or Shaheen (NH).

[12] The within-party standard deviation of the roll-call scores is 0.46.

[13] Both of these are very small differences: Ayotte (NH) is 0.04 to the right of Gregg (NH) and Stabenow (MI) is 0.12 to the right of Levin (MI).

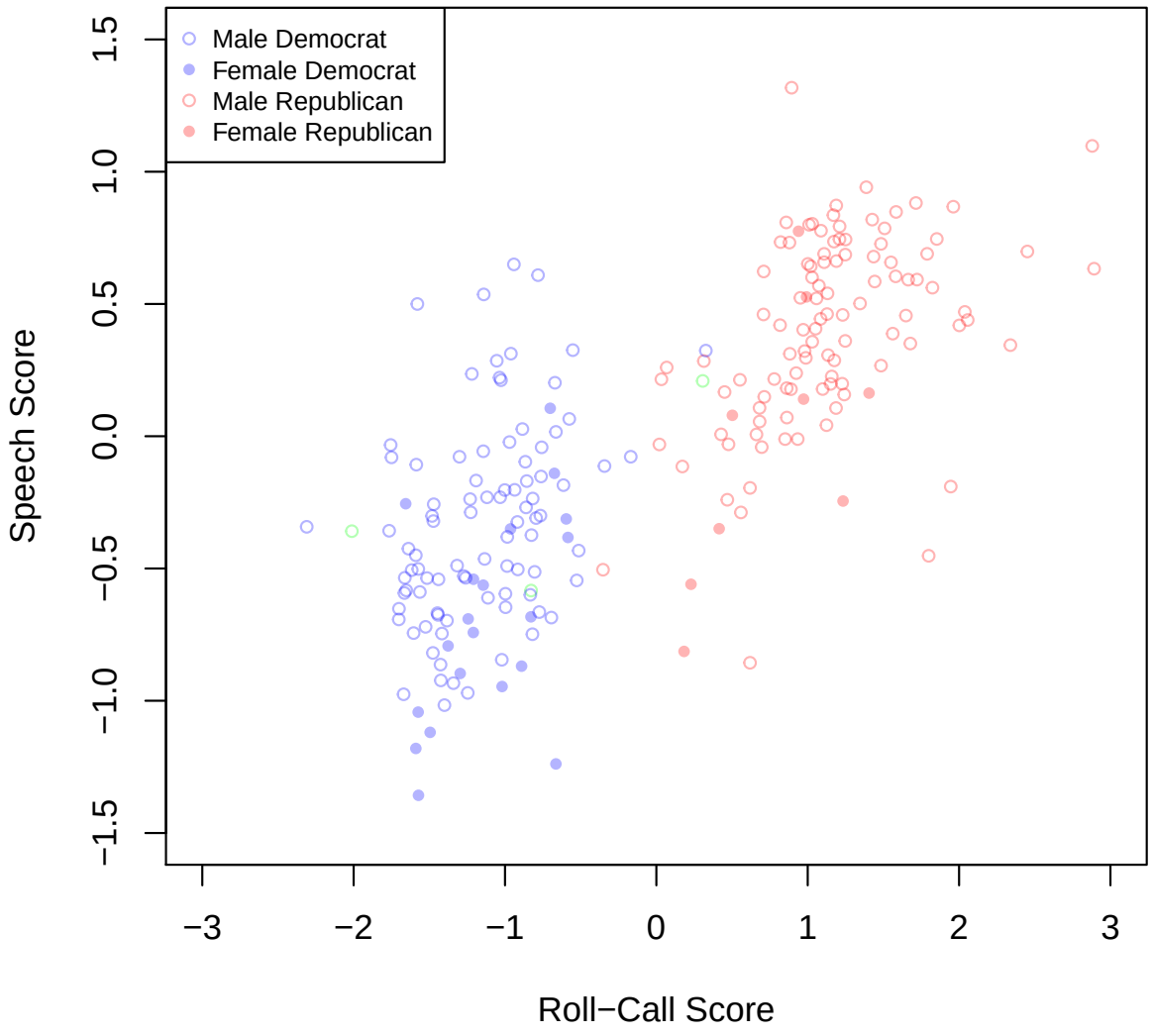


Figure 11: Wordshoal speech scores as a function of roll-call scores, for male versus female senators. Across all senators, the within-party correlation between the two scores is 0.41.

where 13 of 22 women have roll-call scores to the right of the same matched same-party, same-state men.

At its most basic level, these results indicate that across all Senate debates, female senators more frequently use the words that Democrats use more frequently, whatever those words happen to be in a given debate. A skeptical interpretation of this result is that this dimension could just be a mixture of linguistic features of female speech and left speech, and the fact that those attributes are correlated in the population (there are more female Democrats than Republicans) is the only reason we recover a dimension that is a mix of the two. It is reasonable to ask: have we have simply measured a dimension that is a mixture of left-right and male-female?

One way to address this concern is to fit a 2D factor model at the second level instead of the 1D factor model we have considered thus far. If the differences between male and female senators were due to a 1D model recovering a mixture of partisan and gender differences in speech, allowing two dimensions could enable the model to distinguish political variation in speech from gender variation in speech. Fitting a 2D model tells us whether the male-female differences show up in the same debates as the left-right differences (indicating women speak more to the left) or whether they arise in distinct sets of debates (indicating mixing of distinct dimensions). We do not show the details here, but adding a second dimension does nothing to diminish the result that female senators speak to the left of male senators from the same states. Under the 2D model, gender differences remain large along the axis that separates the parties, and are not significant along the orthogonal dimension.

## 7. LIMITATIONS AND EXTENSIONS

Extending the second-level model to have multiple dimensions is just one of many ways that our model can be extended.

### 7.1. *Selection into Speaking*

Some assumption about the process by which legislators choose to speak in a given debate is necessary to do any text scaling at all. All existing text-scaling methods implicitly assume that when a legislator does not speak, that fact is not informative about their position in a given debate. We make this assumption as well, but it is important to note that this might not be correct, particularly if legislators speak strategically or if party leaders strategically select members to represent the party in a debate (Proksch and Slapin 2012). The consequences of such non-random missingness depends on the mechanism: legislators might speak more when their positions are extreme because they care more about such issues, or they might speak more when they are moderate for electoral reasons. Similar assumptions are made in the analysis of roll-call voting, but missingness is far rarer in those data than in speeches. One could impose an alternative model of missingness in the aggregation model given measurable variables that predict the decision to speak. Such a model could be useful in assessing what incentives are driving speaking behavior.

Nonetheless, it is important to recognize that even without such a selection model, the simpler approach followed in this paper still yields valuable summaries of behavior. What we recover is a summary of the speeches that were actually given. For example, the fact that senator Snowe is estimated far from her co-partisans, among the left-wing Democrats, does not mean she is "really" a left-wing Democrat. What it does mean is that when she chooses speaks in the Senate, she uses similar language to the left-wing Democrats who speak in the same debates. Even if she is choosing those debates strategically, this is still an important fact about the speeches she actually gives.

### 7.2. *Hierarchical Estimation*

This paper argues that the political valence of particular words must be conditioned on the debate that those words were used in rather than treated as constant across all debates. Conditioning on debate allows us to control for topic to a far greater extent than is otherwise possible. However, the two stage procedure followed in this paper might take this logic too far. Some words are used similarly to denote position across different debates, and the presented approach ignores such information and the efficiencies it could provide. Fully hierarchical estimation would enable regularization of the text scaling parameters across debates to better exploit the information in the data. This compromise between the assumption that word usage in each debate is uninformative about other debates, and the assumption that word usage is identical across every debate, would potentially provide better estimates of preference than is possible at either extreme. The most immediate obstacle to a fully hierarchical specification is computational, and the payoffs of solving the estimation problems associated with this approach are uncertain.

### 7.3. *Disaggregation By Topic*

In the above specification, we estimate only a single dimension across all debates, and we fail to share information about word use parameters across different debates. However, in practice we expect that debates on similar topics will have similar word usage patterns, even though they will not be the same due to debate-specific features. Moreover, we expect that legislators preferences might vary by topic, and it might be useful to recover multidimensional preference estimates that were informed by the textual data rather than simply using multiple factors without substantive identification of their meaning. Since we have only used the within-debate variation in word usage to estimate preference variation, we could use the

across-debate variation in word usage to estimate which debates given us information about preferences on which latent dimensions.

In the two-stage estimation approach we use in this paper, this logic could be implemented by applying any of a variety of tools for topic-modeling to the textual data aggregated up to the level of the debates (i.e. across speakers). If we do this using a single-membership topic model, we can then separately estimate the second stage for each topic using the debate-specific preferences in the debates assigned to that topic. If we instead do this using a multiple-membership topic model like LDA (Blei, Ng and Jordan 2003), we can use the resulting matrix of mixture weights to define the mixture of dimension-specific $\theta_{id}$ as in Lauderdale and Clark (2014). The two-stage approach then becomes a three stage approach. First, estimate debate-specific preferences using correspondence analysis on the speeches of different legislators in each debate. Second, estimate the assignment of debates to topics using a topic model on the aggregated speeches of all legislators in each debate. Third, estimate topic-specific preferences for legislators using factor analysis on the debate-specific preference recovered from the correspondence analysis with the loadings of debates onto dimensions determined by the topic-model estimates.

Under the hierarchical estimation approach, we could implement the logic of topics using a hierarchical mixture model. This mixture model naturally applies to both the word parameters and the preference parameters in the model. For $\lambda_{jk}, \kappa_{jk}$, we expect variation in these parameters to depend on their membership $\eta_j$ in each of the components $t$ of a normal mixture distribution. That is, we expect the word-parameters to cluster by topic: the way that a word is related to preference is more similar within debates on the same topic than across debates on different topics. At the same time, we want to estimate separate underlying preference dimensions $\theta_{it}$ for each of these topics $t$. Estimating this model increases the computational burden only moderately over the model described above, because not only is the text scaling component of the estimation the same, all the parameters that are directly

relevant to the text scaling ($\lambda_{jk}, \kappa_{jk}, \psi_{ij}$) still have the same priors as in the model described above, once one conditions on topic membership. Thus the modifications to the model are all at higher levels of the hierarchy: replacing linear models for these parameters with mixtures of linear models for these parameters.

### 7.4. *Dynamic Positions*

As with disaggregating by topic, estimation of dynamic preferences can also be achieved from a closely related model that does not change the lower-level model for the texts. To model dynamics, one could apply one of the standard techniques for modeling dynamics (Poole and Rosenthal 1997; Martin and Quinn 2002) to $\theta_i$, allowing those parameters to vary over time. The same issues with inter-temporal identification would arise in our model. Combining this with the topic-model hierarchy is possible as well, but may not be feasible even for the most verbose legislatures.

## 8. CONCLUSION

It is appropriate to be skeptical about unsupervised estimators that purport to turn word counts into estimates of "expressed preferences" or "stated positions". This is partly because of the black box process by which such models sometimes assume—rather than demonstrate—that preferences are a major source of variation in word usage. But it is also because the longer experience of scaling roll-call voting data in political science has given us a sense of the various ways that measurement models can fail to measure what we want. Demonstrating that text scaling works well enough for purpose requires validation of the types provided in this paper.

The validation we have done suggests that *political disagreement*, rather than necessarily

*policy preferences*, is measured by our estimates in both Ireland and the US. This makes sense given the way that our estimation procedure is constructed. The many debate-specific scales will reflect various features of particular debates and the idiosyncratic preferences of particular legislators on those debates. However, the scaling of these scales will select out the common dimension of variation that most consistently shapes word usage across all spoken debates. It is perhaps not surprising that this tends to be the government-opposition cleavage in the Westminster-style system of the Irish Dáil . In the US, where a different constitutional structure and a two party system make the incentives for legislative speech different, we see patterns of speech behavior that, while different from roll-call based measures of preference, have a strong association with those measures both across and within parties. But there are important differences which reflect the different processes that generate speech and voting behavior. Party polarization of speeches is responsive to political events in the aggregate to a far greater extent than roll-call behavior. We also see that speech behavior can differ from roll-call behavior at the individual-level, as in the comparison of same-state, same-party female and male senators.

Whereas both computer scientists and political scientists have made enormous progress in recent years at developing and refining tools for recovering topics from texts, progress on the problem of recovering continuous measures of disagreement has advanced more slowly and is more peculiar to political science. Recovering positions—descriptions of relative disagreement—is a more difficult problem because of the nature of word usage, and there are fewer researchers working on it actively. As this paper indicates, there is much more work to be done in terms of method development and validation. But legislative speech is interesting in its own right, and is important for understanding the strategies adopted by legislators in response to the political and electoral environments that they face. These wide-ranging potential payoffs justify continued efforts to improve and extend the methods we use to measure features of legislative debates.

REFERENCES


Benoit, Kenneth and Michael Laver. 2006. *Party Policy in Modern Democracies*. London: Routledge.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Budge, Ian. 2001. "Validating party policy placements." *British Journal of Political Science* 31(1):210–223.

Diermeier, Daniel, Jean-François Godbout, Bei Yu and Stefan Kaufmann. 2012. "Language and Ideology in Congress." *British Journal of Political Science* 42(1):31–55.

Eddelbuettel, Dirk and Romain François. 2011. "Rcpp: Seamless R and C++ Integration." *Journal of Statistical Software* 40(8):1–18.
**URL:** *http://www.jstatsoft.org/v40/i08/*

Gallagher, Michael. 2010. The Oireachtas: President and parliament. In *Politics in the Republic of Ireland*, ed. John Coakley and Michael Gallagher. Routledge chapter 7, pp. 198–229.

Grimmer, Justin. 2013. *Representational Style in Congress: What Legislators Say and Why It Matters*. Cambridge University Press.

Grimmer, Justin and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267–297.

Hansen, Martin Ejnar. 2009. "The positions of Irish parliamentary actors 1937–2006." *Irish Political Studies* 24(1):29–44.

Herzog, Alexander and Kenneth Benoit. 2013. "The most unkindest cuts: government cohesion and economic crisis.".

Herzog, Alexander and Slava Mikhaylov. 2013. "DPSI: Database of Parliamentary Speeches in Ireland." Manuscript.

Jackman, Simon. 2001. "Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference, and Model Checking." *Political Analysis* 9(3):227–41.

Lauderdale, Benjamin E. 2010. "Unpredictable voters in ideal point estimation." *Political Analysis* 18(2):151–171.

Lauderdale, Benjamin E. and Tom S. Clark. 2014. "Scaling Politically Meaningful Dimensions Using Texts and Votes." *American Journal of Political Science* in press.

Lowe, Will. 2013. "Theres (basically) only one way to do: Some unifying theory for text scaling models." Paper presented at the American Political Science Association meeting September 2013, Chicago.

Lowe, William. 2008. "Understanding Wordscores." *Political Analysis* 16(4):356–371.

Martin, Andrew D. and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10:134–53.

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

Proksch, Sven-Oliver and Jonathan B. Slapin. 2010. "Position Taking in European Parliament Speeches." *British Journal of Political Science* 40(3):587–611.

Proksch, Sven-Oliver and Jonathan B. Slapin. 2012. "Institutional Foundations of Legislative Speech." *American Journal of Political Science* 56(3):520–37.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin and Dragomir R. Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Simon, Dennis M. and Barbara Palmer. 2010. "The Roll Call Behavior of Men and Women in the U.S. House of Representatives, 1937–2008." *Politics and Gender* 6:225–246.

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.

Spirling, Arthur and Iain McLean. 2007. "UK OC OK? Interpreting Optimal Classification Scores for the U.K. House of Commons." *Political Analysis* 15(1):85–96.

Thomas, Matt, Bo Pang and Lillian Lee. 2006. "Get Out the Vote: Determining Support or Opposition from Congressional Floor-debate transcripts." *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)* pp. 327–35.

Zucco, Cesar Jr. and Benjamin E. Lauderdale. 2011. "Distinguishing Between Influences on Brazilian Legislative Behavior." *Legislative Studies Quarterly* 36(3):363–96.