# Using Text to Scale Legislatures with Uninformative Voting

Nick Beauchamp*
NYU Department of Politics

August 8, 2012

Abstract

This paper shows how legislators' written and spoken text can be used to ideologically scale individuals even in the absence of informative votes, by positioning members according to their similarity to two reference texts constructed from the aggregated speech of every member of each of two major parties. The paper develops a new Bayesian scaling that is more theoretically sound that the related Wordscores approach, and a new vector-based scaling that works better than either at matching the vote-based scaling DW-Nominate in the US Senate. Unsupervised methods such as Wordfish or principal component analysis are found to do less well. Once validated in the US context, this approach is then tested in a legislature without informative voting, the UK House of Commons. There the scalings successfully separate members of different parties, order parties correctly, match expert and rebellion-based scalings reasonably well, and work across different years and even changes in leadership. The text-based scaling developed here both matches existent scalings fairly well, and may be a much more accurate window into the true ideological positions of political actors in legislatures and the many other domains where textual data are plentiful.

*Email: nick.beauchamp@nyu.edu; web: nickbeauchamp.com.

# 1 Introduction

Most models of legislative behavior assume that members have ideal points positioned in some shared policy space. For empirical work, these ideal points must be estimated for dozens or hundreds of individuals, and the standard approach is to use the roll-call voting data to do so, using a variety of algorithmic methods (Poole & Rosenthal 1985, Poole & Rosenthal 1991, Poole 2005, Jackman 2001). However, although this approach works well in domains like the US Congress, US state legislatures, and even multi-member courts – where divisions among members on different issues often vary – in the vast majority of non-US legislative domains party discipline means that vote cleavages are almost always the same, with all members of a given party voting together. Because of this, established vote-based scaling methods fail, and although alternative methods using vote-based data have been proposed (Quinn & Spirling 2010), a general and widely-applicable method for scaling legislators in high-discipline legislatures remains lacking.

But although vote-based data is often insufficient, in recent years vast quantities of another sort of data generated by individual legislators have become available: text databases recording the speeches of individual legislators. While there is already some evidence that vote-based scaling results can be predicted using text algorithms trained with those same vote-based scalings (Diermeier, Godbout, Yu & Kaufmann 2012), the practical need is for a scaling method that can be employed precisely where vote-based scalings are impossible. This paper develops a new technique for using text to do just this. It shows that by using only party-membership information plus the speech data, one can scale every member of a legislature in ways that may be just as useful as the undisciplined-voting-based approaches.

This scaling method works by aggregating the speech of all members of each of two major parties – one from the left, one from the right – and then scaling the individual speakers using these aggregated reference texts. This scaling does more than merely recover existing party information using textual data, in part because it also captures degrees of similarity to the party centers. As we will see, since vote-based measures even in the US often boil down to little more than party ID plus party loyalty, the similar structure of the text-based scaling means that these scores closely resemble vote-based results. Of course, the usefulness of this approach is in domains where vote-based measures are lacking, so it will be important to benchmark these results as best we can in a disciplined legislature – a task which raises fundamental questions about the nature of ideological spaces in the absence of effects on voting.

The procedure is as follows: First, the textual scaling approach is developed. Two new scaling methods are devised, one inspired by the popular Wordscores method (Laver, Benoit & Garry 2003), but put on a more theoretically sound Bayesian footing; and the second inspired by spatial scalings such as Wordfish (Slapin & Proksch 2007) or support vector machines (SVM).[1] Next, I show via simulations that these three methods (the Bayesian, the spatial, and Wordscores) actually produce similar results in many cases, although the Bayesian and Wordscores are the most similar, and the spatial method diverges from the others for more realistic word distributions – important because later we will see that that spatial method seems to often work best, empirically.

Before turning to the more useful application in disciplined legislatures, it is essential to validate this approach in a domain where vote-based scaling is available. I apply these methods to the US Senate, and show that the resultant scalings match the established DW-Nominate scores

[1] In both cases, the methods are constructed to deal primarily with two reference points – the aggregated speech of the members of a left-wing and right-wing party – although the spatial method can be easily extended to multidimensional scaling.

quite closely; this match is best with the spatial method, and is further improved by removing the technical language of legislation wherever possible. By comparison, unsupervised textual scalings such as Wordfish or principal component analysis (PCA) do not work nearly as well, since the latent dimensions along which speech varies the most in legislatures, although arguably a form of ideology, do not seem to closely reflect the kind of ideology that affects voting. But DW-Nominate is not the last word in benchmarking ideology: more fundamentally, DW-Nominate exists to predict and explain roll-call voting, and we can also test the text-based scalings directly against the votes. Using strict out-of-sample testing, we see that the text-based approach does somewhat better than party ID alone in predicting votes, and that DW-Nominate itself only does a little better than party ID plus a measure of party loyalty. The text-based scaling seems to work by capturing, in addition to party information, something like the loyalty measure that itself accounts for much of the advantage of vote-based scaling over party ID alone.

But although this gives us some grounds to believe this procedure might work in non-US settings, how might we validate it elsewhere? This is a hard question that raises deep issues about the meaning of ideological scaling. To test this, I turn to the UK House of Commons (HOC), which has strong party-loyalty of just the sort that makes vote-based scaling difficult, but has also been closely studied with numerous expert-based scalings, as well as a number of attempts to scale using what little variance there is in the roll-call record. The most basic validation is to simply show that using, for instance, Conservative and Labour aggregate speech as references, one can clearly distinguish speakers from the two parties. Furthermore, as before, this distinction between parties is improved by dropping the technical language of legislation (quite easy when, as in the UK, legislators are identified by whether they are speaking in their role as representative

or Minister). These results also seem to carry across years, but not as well across the change in party control in 1997, reflecting the fundamental shift in the terms of debate then. But even more effective is using not the Labour and Conservative parties, but the more extreme Liberal-Democrat (Lib-Dem) party instead of Labour, which orders all the parties in close conformity with expert-based scalings such as the Comparative Manifesto Project (Budge, Klingemann, Volkens, Bara & Tanenbaum 2001, Klingemann, Volkens, Bara, Budge & McDonald 2006). Finally, looking at what few individual-level vote-based scaling attempts there are, the text-based results seems to match them well, although as before, these vote-based measures are largely reducible to a combination of party ID and party loyalty. Thus both the speech-based scalings and vote-based scalings, even in undisciplined legislatures, are comprised largely of party ID and loyalty, with a residual of "ideology" which is presumably quite complex and multi-dimensional.

Once this approach has been developed and validated, it can be applied not just to legislatures with uninformative voting, but even to other domains where only speech data exist. Looking forward, I briefly describe a new online tool which allows one to use legislative or other reference documents to scale anyone via the textual data derived from a search-engine query of their name. Although this more experimental application requires further validation, it illustrates the power and generality of these methods as useful tools both for scaling legislatures with uninformative voting, as well as the many other domains where political actors reveal their ideologies through speech.

## 2 Theory

While automated text-based scaling is relatively new, in a sense, text-based scaling has long been the norm – albeit by experts rather than machines. There has been a gradual progression from more subjective, expert-based scaling to more automated approaches over the last few decades, ranging from expert surveys (Laver & Hunt 1992, Janda, Harmel, Edens & Goff 1995, Benoit & Laver 2006) to evaluations of manifestos sentence-by-sentence (Budge, Robertson & Hearl 1987) to computer dictionary-based approaches (Laver & Garry 2000). Perhaps the first deeply automated political scaling was the Wordscores method of Laver, Benoit and Garry (2003), which takes a series of reference texts, asks experts to score the political positions of those texts on a 1-dimensional scale, and then automatically scores "virgin" texts according to their similarity to the reference texts. Currently, one of the most automated scaling approaches is the Wordfish algorithm of Slapin and Proksch (2007), based on work by Monroe and Maeda (2005), which can scale text without any human intervention at all. While Wordscores is a "supervised" method, where humans set the positions of reference documents, Wordfish is "unsupervised," and can be run with little or no human supervision. But as we will see, for specific tasks such as ideological scaling, unsupervised scalings may not work as well as those shaped to the task at hand.

In this case, that task is scaling legislators when party information is available, but voting information is uninformative or non-existent. The basic idea (described in more detail shortly) is to create two reference texts by aggregating the total speech of everyone of party L and party R, and use those two references to scale individuals, where each individual is represented by the total recorded speech of that individual in the legislature that year. This supervised scaling can be done without further adjustment using the popular Wordscores package, but that is not ideal. First, because

although the construction of Wordscores was clearly inspired by Bayesian methods, it diverges from a properly Bayesian approach in important ways. Thus I construct here a Bayesian approach that resembles Wordscores, but puts the scaling on a more secure theoretical basis. Another reason not to simply employ Wordscores or the Bayesian method is that, as we will see, their results may not work quite as well, empirically, as a different supervised scaling method I construct, based on the spatial models underlying Wordfish or SVMs. This vector-based spatial method is in fact algorithmically similar to the Wordscores and Bayesian methods, but presumes that speakers' positions in high-dimensional word-space reflects their ideological positions in ideology-space, and thus mapping them onto the line connecting the two reference documents in word-space will reflect their ideological positions in unidimensional ideological space. This spatial approach is theoretically fundamentally different from the Bayesian (or Wordscores) approach, but as we will see, its results are similar, and at times empirically superior. In the remainder of this section, the Bayesian and spatial methods will be described in greater detail, contrasted with the Wordscores method, and compared using simulated data, illustrating where they overlap and diverge. When we turn to the first empirical section, the US Senate, these three supervised methods will also be contrasted with unsupervised approaches like Wordfish, illustrating their advantages in legislative scaling.

The initial steps are the same for all methods. For clarity, this will be discussed in terms of the US Senate scaling, but the same logic applies whatever the context. First, the entire 2006 Senate congressional record is processed so that every speech delivered by a given Senator is concatenated into a single text file. The text files are then summarized by matrix $W_{i,j}$, where each value is the proportion of Senator $j$'s speech consisting of word $i$, although here only the top 1000 words are

retained.[2] In addition to this 1000x100 matrix,[3] two additional word-proportion vectors are also calculated: a vector produced from the entire speech output of all Democrats, and a similar vector for all Republicans. These last two vectors are used as reference texts in the first three, supervised methods. All the scaling methods to be discussed next make use of the text matrix $W_{i,j}$, but only the supervised methods make use of the two aggregate-speech vectors.

## 2.1 Bayesian scaling

The Bayesian approach I develop here was designed in part to remedy some of the theoretical problems with Wordscores. I will begin with the Bayesian logic, and then describe Wordscores and where it diverges from that logic. The key idea is to assume there are two classes of documents – Republican and Democratic – and then to simply assign to each legislator a score based on the relative likelihood of belonging to the R and D classes, where the likelihood of belonging to a document class is simply a function of the similarity between the speaker's text (S) and the reference document's text (R or D). Obviously the degree to which one's speech more resembles the paradigmatic Republican or Democrat is only a proxy for ideological position, but as we will see, this proxy in fact does a better job that party ID alone, since it can distinguish centrist from doctrinaire Democrats, for instance.

Thus we wish to discover $p(R|S)$ and $p(D|S)$, ie, the probabilities that a speaker is Repub-

[2] 1000 was chosen for mainly computational feasibility, and because increasing beyond this increases the effects of very rare words, particularly for the Bayesian approach. In addition, a standard list of "stop words" was omitted – words like "the," "of," "and," etc. Another common approach is to reweight the word frequency matrix using "tf-idf" (term frequency–inverse document frequency) weights. This essentially gives more weight to words that are frequent in a document but infrequent in the larger corpus. However, although the motivations for using it are Bayesian, it is theoretically cleaner to work directly with frequencies to begin with, and work any Bayesian weighting at the parameter estimation stage, if at all.

[3] Actually 1000x96, since some Senators did not speak enough in 2006 to be usable.

lican and Democrat given their speech document S. From this we construct the likelihood ratio, $p(R|S)/p(D|S)$, and that (or more practically, its log) is the Bayes score. What we know for each document is $p(w_i|X)$ (where X may be S, D or R), that is, the probability of any given word $w_i$ given that we have encountered document X. From this we directly build our likelihood ratio, in the manner described in detail in the Appendix.

This produces the simple result:

$$\text{Bayesscore} = B_V = \Sigma_{i \in S} \log \frac{p(w_i|R)}{p(w_i|D)} \tag{1}$$

$p(w_i|R)$ is simply taken to be the percentage of word $w_i$ in document R. This is undoubtably a simplification, since $p(w_i|R)$ should perhaps include priors about the distribution of $w_i$ (conforming to some Poisson process, say), which in turn could depend on various parameters concerning word ideal points, document ideal points, word "informativeness," and much else. But as we will see, this simplistic approach works quite well on its own, is computationally efficient, and allows easy comparison with the Wordscores method described next.

### 2.2 Wordscores and Bayesianism

"Wordscores" was developed by Laver, Benoit and Garry (2003; hereafter LBG) specifically in the political context, although it can be extended to any scaling of a "virgin" text with respect to reference texts whose positions are given *a priori*. As discussed in further detail in the Appendix, the derivation is roughly Bayesian, though it diverges in a number of important ways. More interestingly, although the final formulation of the Wordscore is quite different from the Bayesian

score, in practice (analytically, via simulation, and empirically) the results will often be quite similar.

Essentially, the Wordscore of a *word* is the weighted mean of that word's scores in each of the individual reference documents, where each reference document's score is given *a priori* by some expert, and each word in that document inherits that score. When there are two reference documents, the weight for each can be seen as an approximation of the probability that the virgin document belongs to class X given word $i$. The Wordscore for a virgin document is simply the sum of each of the scores of the words in it, weighted by the frequency of each of those words in the virgin document.

Following LBG, if we define $W_{iX}$ as the count of word $i$ in document X, and $W_X$ as the total number of words in document X, then the Wordscore of a virgin document given two reference documents is:

$$\text{Wordscore} = S_V = \sum_i \frac{W_{iX}}{W_X} \cdot S_i \tag{2}$$

where $S_i$, the score of word $i$, is

$$S_i = A_R \cdot P_{iR} + A_D \cdot P_{iD} \tag{3}$$

$A_X$ is the *a priori* score given to reference document X. $P_{iX}$ is (approximately) the probability of word $i$ given document X, and LBG state that, for instance,

$$P_{iR} = \frac{\frac{W_{iR}}{W_R}}{\frac{W_{iR}}{W_R} + \frac{W_{iD}}{W_D}} \tag{4}$$

This is not quite Bayesian, as is explained in more detail in the Appendix. More importantly,

while the word score is the weighted sum of the reference scores, the final document score is not directly tied to the probability that it belongs to the reference classes, except in an approximate way. Again, see the Appendix for details, but we can easily note one obvious area of divergence. As LBG point out, if reference text R contains a word and the other reference document does not, that makes $P_{iR} = 1$. From the Bayesian point of view, if that word $i$ then occurs even once in the test document, we know for certain that that document belongs to class R (the score as devised above goes to + or - infinity). For Wordscores, however, we only add $W_{iV}/W_V \cdot A_R$ to the running total.[4]

Figure 1 illustrates the difference between the word score assigned by the Bayesian approach and the Wordscores approach, given the frequency of that word in reference documents 1 and 2 (x and y axes). Despite the different formulations, the functions are similar except when word frequencies are near zero. As can be seen, when a word frequency is low in one document but high in the other, the Bayesian method correctly infers that the unknown word is very likely to belong to the second class, whereas the Wordscores method tops out. However, there is an advantage to this limitation, in that the Bayesian approach can (without smoothing priors) over-weigh words that occur in one document and not the other simply by chance. (See Appendix for more details.)

[Figure 1 about here]

[4]Lowe (2008) makes a similar point about the flaws inherent in Wordscores, showing that words unique to a single document are erroneously given the score of the document. Depending on the choice of prior, this may even be going too far in the opposite direction, giving a word an overly mild contribution to the scoring of the reference text. In any case, the use of aggregation to define the two reference documents minimizes this problem, since both reference texts tend to share almost all their words in common, just at different frequencies. Lowe also shows that Wordscores – as with the Bayesian approach used here – fails to distinguish between informatively centrist words and uninformative words that on expectation reside in the center. But while this may collapse scores centerward, it does not bias scores, so if one (as here) is unworried by a set of tightly clustered scores (by some measure), this is no large problem, particularly when there are only a left and right pair of reference documents. Finally, he also points out an important resemblance between Wordscores and an approximation of an ideal-point model, although he shows that this approximation may be poor when word positions or informativeness are unevenly distributed. Determining whether this theoretical problem is of practical import is directly addressed by the following empirical sections here.

### 2.3 Vector Projection

The alternative supervised scaling method developed here is derived from the spatial models underlying Wordfish and SVMs, but differs in important ways from either. Each Senator is considered as a point in a 1000-dimensional space, where each dimension corresponds to the proportion of a given word in that Senator's speech. The two aggregated-speech vectors are also points in this space (Democrats and Republicans), and the vector scaling is derived by projecting a Senator's position onto the line between the two party vectors. The underlying model makes very modest assumptions: that the presumed unidimensional ideological space can be embedded in the high-dimensional word space such that the proportion of each word $i$ in speaker $j$'s speech is a linear function of speaker $j$'s position in this unidimensional space plus some error: $w_{i,j}|x(S_j) = f_i(x(S_j)) + \epsilon$. If we assume these errors are i.i.d., then to place the line $x(S_j), \forall j$ in word space, we would simply find the OLS line through all points $w_{i,j}, \forall j$. This is similar to what unsupervised approaches such as principal component analysis and Wordfish (effectively) do, as discussed below. But if instead we hypothesize that a better estimate of a unidimensional ideological space is one that more closely aligns with partisan difference, we can take the line that runs through the D and R aggregates and project speakers onto that instead. Although this line through the party aggregates may not capture the greatest variance of the cloud of speaker points in word space, as we will see, it is empirically closer to the dimension of partisan disagreement that is reflected in vote-based scaling (where such scalings are possible).[5]

[5] Again, this procedure borrows some of the framework from classification algorithms, this time from nearest centroid classifiers, which classify unknown points according to whether they are closer to the centroid of class A or class B (for two classes). But here, the relative distances, rather than the binary classification, is the outcome.

Alternatively, although an SVM with two categories could also supply a scaling (replacing the DW-Nominate scores in Diermeier et al. (2012) with a binary class), because SVM's focus on the margin between the two classes, they are less well suited to scaling all the non-marginal (ie, non-centrist) cases.

Obviously there is no inherent zero point or scale to this, so we can simply normalize the score such that p(R) = 0 and p(D) = 1. If we are projecting onto the vector $R - D$, and wish to know the distance of some third point S from R as projected onto that line, a computationally efficient approach given the distances $||R - D|| = a$, $||S - R|| = b$ and $||S - D|| = c$ is simply:[6]

$$\text{Vectorscore} = \frac{a - b + c}{2a} \tag{5}$$

This value is calculated for each speaker, and without further transformation is taken as their vector score. Note that unlike the Wordscore or Bayesian score, speakers can be projected onto the ideological space beyond the segment between D and R, so that in principal extremists can be distinguished from centrists.

### 2.4 Comparative results with simulated textual data

Although the base functions of Wordscores and the Bayesian score appear similar in Figure 1, in more realistic circumstances it is not the case that their results will always be so alike. To better understand the interrelation between the three supervised techniques, I create and scale a series of simulated "texts." For each scaling, two reference vectors were randomly created, along with a third to be scaled according to the three different methods; this process was repeated 1000 times, and the scores between those three datasets were examined for correlation. Two quantities of "words" (1000 and 2000) and two families of distribution functions for those words were examined. While approaches such as Wordfish implicitly assume an exponential decline in word frequency in a text,

[6] We could also use inner products to get the same result, or we use the cosine similarity metric, although the implied geometry of that is slightly different. All these metrics produce similar results.

much current research suggests that the frequency of words in texts has a fatter tail, following instead a "power law" of the form $x^{\alpha}$ (Newman 2005).

As Table 1 shows, although the Bayesian and Wordscores methods are generally more tightly correlated than the vector projection method is with either, that correlation weakens as either the number of words or the mass of the tail decreases. In those cases, the relative frequency of words occurring in only one document increases, contributing to the divergence between the two functions, as seen at the edges in Figure 1. The upshot is that, although we will see that in the case of the US Senate the empirical scores are quite similar, in documents with fewer words or thinner tails, the scorings begin to diverge significantly. Since it is difficult to know *a priori* what the exact word-distribution for a corpus will be, the safer course would be to go with the Bayesian or vector methods, although Wordscores may in many cases be adequate.

[Table 1 about here]

### 2.5 Unsupervised scaling

Currently the most popular unsupervised scaling method is "Wordfish," developed by Slapin and Prosksch (2007; hereafter, SP) based on the work of Monroe and Maeda (2005; hereafter, MM).[7] In this model, $p(w_i|S)$ is a poisson function of the distance of word $w_i$ from Senator $S_j$:

$$\begin{aligned} y_{ij} &\sim Poisson(\lambda_{ij}) \\ ln(\lambda_{ij}) &= c + c_i^x + c_j^{\alpha} + \gamma_j(x_i - \alpha_j) \end{aligned} \tag{6}$$

[7]This scaling was originally inspired by item response theory, which in psychology takes a set of respondents and their answers to various questions, and placing the questions and respondents in a shared space.

Where $i$ indexes documents, $j$ indexes words, $c$ is a constant, $c_i^x$ are a document-specific constants, $c_j^\alpha$ are word-specific constants, and $(x_i - \alpha_j)$ is the distance between a document position $x_i$ and a word position $\alpha_j$. An additional parameter $\gamma_j$ measures the "discrimination" effect of word $j$: for instance, in a left-right political dimension, when words are right-wing we would expect increasing distance to the right of a word to result in greater use of that word, and this would correspond to a positive $\gamma$ for that word. Perhaps unsurprisingly given three separate word parameters, the model is under-indentified, so MM jettison the word position parameters $\alpha_j$ (setting them all to 0) and interpret the $\gamma_j$ parameter as something like word position, where larger positive values correspond to more right-wing words (and vice versa for negative/left-wing values). Once the likelihood has been established for any given set of word and document positions, given the word frequency data, it's only a matter of maximizing that likelihood.

An alternative unsupervised approach, both simpler and more quickly estimated than Wordfish, is just to take the matrix $w_{i,j}$ and find its principal component. These are fairly large matrices, but the Nipals algorithm employed here can very quickly determine the principal components of large matrices. The result, as we will see, is something that produces quite similar results to Wordfish, suggesting that the latter is perhaps simply finding the same vector by more elaborate means.[8]

[8]For a fast online implementation of this PCA method, which is standardized to scale both words and documents in the same space, see the "AutoScale" tool at nickbeauchamp.com.

## 3 Benchmark: Scaling the US Senate

Having devised this general approach to legislative scaling using textual data and party aggregates, and having developed two specific techniques using Bayesian analysis and vector projection, it is important to test these methods against well-established benchmarks before turning to the harder case of a disciplined legislature. Scalings like Wordscores and Wordfish have generally been applied to parties (eg, party manifestos) rather than individuals, but because there are usually so few of these, it has been difficult to validate the results statistically, rather than just checking for vague confirmation with expert opinion. Working instead with an entire legislature such as the US Senate allows much more reliable results. As we will see, scaling with party aggregates works well both in matching existing vote-based scalings like DW-Nominate, and in explaining the political behavior upon which that scaling is based, the roll-call votes themselves. Conversely, not being designed for the task, existing unsupervised approaches do considerably less well.

For the test-bed I take the entire Congressional Record for the US Senate from 2006, which records speeches, debates, and material that has been added directly to the record by Senators. Once again, for the three supervised methods, the first step is to generate two reference vectors, each based on the entirety of Republican and Democratic speech, respectively; by contrast, the unsupervised methods work solely with the aggregated speech vectors for each Senator. As a first test, Table 2 shows the correlations between these text-based scalings and the DW-Nominate scores (DW1).[9] Remarkably, using only textual data, all of the methods produce results that correlate significantly with the vote-based DW-Nominate score.[10] Clearly the supervised approaches do

[9]This shows only the first DW-Nominate dimension, DW1. Correlations with DW2 are negligible.

[10]These are cardinal correlations. If one believes that the rank order rather than the positions per se are more important to get right, we might employ ordinal correlations. In that case, correlation coefficients rise above 0.7 for all supervised methods. Wordscores results are omitted because, as before, they correlate with the Bayesian method

better than the unsupervised methods, and of the supervised approaches, the vector method seems to do best, although all three correlate fairly closely with each other. Also notable is that it is the second principal component, not the first, that correlates with DW-Nominate, and that second component in fact correlates highly with Wordfish,[11] suggesting the latter has, somehow, simply picked up the second component through its procedure.[12]

[Table 2 about here]

One intuition check for Wordscores and Wordfish is that both methods provide scores for words as well as documents. We can do the same for the vector projection method if we construct the vector $R - D$ and select the highest and lowest scoring words. The results are presented in Table 3. Clearly the Democratic words are as expected, but notably, the Republican ones are dominated by the technical language of legislation, because they were the majority party at the time and responsible for procedural matters. If we remove those words from our corpus and rerun the vector scaling, the correlation between DW1 and the text vector scaling goes from 0.64 to 0.73.[13] Obviously it will not always be easy to cull such language from the corpus, but doing so significantly improves the text-based scaling; in the UK case, as we will see, this culling is actually considerably easier, because a member is documented separately according to whether she is speaking in her

at over 0.99.

[11]Results like this are in part why Wordfish is in practice rarely used without supervision: generally a certain amount of document or word selection is done prior to a final scaling to ensure the relevance and subject-specificity of the results. This makes Wordfish in practice more of a semi-supervised than fully unsupervised method.

[12]Exploratory work suggests there is a general tendency for the second principal component to be more substantive than the first. The first eigenvector often picks up the "junk" in a a corpus: scaling novels from gutenberg.org, the first dimension is dominated by gutenberg's own prefatory words; with screenplays or TV ads, the first dimension will contain stage directions or character names; with web pages, it will contain any leftover html code you haven't cleaned up; and so on. Similarly, MM found that the first dimension in their IRT scaling of congress picked up differences in speaking style rather than political speech. Proving this general tendency is beyond the scope of this article, but it is good to be aware that, with unsupervised scaling, the second dimension may the one of most substantive interest.

[13]The words to be removed were selected based on a review of all top words, and the selection was made once, without modification, to prevent post-hoc fitting of results to DW-Nominate. The vast majority of randomly deleted word lists only worsen fit, of course.

capacity as representative or minister.

[Table 3 about here]

While establishing correlations with DW-Nominate does show that these methods are measuring ideology to some degree, an important question is whether we are getting more out of these scalings than the party information that we put in. If we look at correlations between these scalings and DW-Nominate within a single party, the vector projection and Wordfish methods do show slight but significant correlations at about the 0.2 level, but the Bayesian method does not, perhaps reflecting its roots as a classifier rather than a scaler.[14] Interestingly, the one method that does more strongly correlate within-party is the first principal component, the one that did not appear to correlate with DW-Nominate at all. If we look at the plot of DW1 against PCA1 in Figure 2, we see why: PCA1 does not distinguish the parties, but does distinguish within the parties.

[Figure 2 about here]

Although these within-party results are slightly weaker than for the full set, they raise a deeper question about how significantly ideological we believe within-party DW-Nominate scores

[14]One way to boost intra-party matching could be to use, rather than the parties as a whole as reference texts, only a few of the most extreme members of each party, aggregated into two presumably more extreme reference texts. This might potentially overcome the "folding" problem, where members to the left and right of their party's center are both projected towards the overall center. But this approach is no panacea: First, it is unsuited to the fundamental purpose here, since it cannot be used to scale legislatures that do not already have scalings available. Second, it appears not to work as well as one might hope: using the 5 left- and right-most Senators (based on DW-Nominate) as the two reference texts, one does see a rise in intra-party correlations (to about 0.3), along with an unsurprising drop in the pooled correlation (to about 0.5, due if nothing else to the reference data being fewer). However, omitting the 10 Senators used as reference texts from the scaling output eliminates this rise; essentially, the scaling appears to become noisier with skimpier reference texts, with no concomitant gain in intra-party matching. This isn't to say that a careful choice of reference texts might not succeed, just that it offers no easy improvements, as well as being unsuited to the scaling of hitherto unscaled legislatures. This is of course the same problem in using DW-Nominate scores to scale text using an SVM (Diermeier et al. 2012).

really are. For instance, if we construct a variable that is the product of a {-1, 1} dummy for party times the Congressional Quarterly measure for party loyalty (ie, the proportion times a Senator voted with his/her party in 2006), that variable explains 95% of the variance of DW1 (using the more stringent and correct ordinal correlation). At least based on the results here, it appears that DW-Nominate itself accounts for little more than party ID and party loyalty.[15]

Perhaps more substantively, we can take things one final step, and test these scalings again the fundamental behavioral justification for DW-Nominate and many other ideological scalings, the roll call votes themselves. The most direct test of this is to predict votes out of sample, employing the various scalings in a straightforward logit model. For each of 1000 runs, the observations (Senators) are randomly divided into an in-sample and out-sample set (80%/20%). For each roll call vote, a logit is estimated on the in-sample, where the dependent variable is the rollcall vote and the independent variable is a given scaling, and then that logit is used to predict the out-of-sample votes using the associated scaling numbers (essentially by choosing a cut-point value). The mean out-of-sample prediction accuracy for each of the various scalings is presented in Table 4.

[Table 4 about here]

As can be seen in Table 4, although DW1 best predicts the roll-call vote, getting 88% of the individual votes correct, this is only by a very small margin, and in fact a simple PCA scaling of the roll-call matrix produces almost exactly the same accuracy (0.88). By contrast, using the party-

[15] An important alternative to DW-Nominate are expert-based scalings. These too are often based on votes, but generally a select subset of them that reflect the concerns of, for instance, a specific interest group. A number of these vote-based interest group ratings can be found at http://www.electoral-vote.com/ (2011), which provides both the ratings of seven liberal interest groups, and a mean rating that aggregates them all. This aggregate rating correlates at 0.94 with DW-Nominate. For the other scalings, the correlations are: Vector: 0.71; Bayes/Wordscores: 0.61; PCA2/Wordfish: 0.37. So the results largely mirror those with DW-Nominate itself, though this is not surprising since these ratings are based on votes rather than more holistic expert judgments.

times-loyalty variable variable does almost as well, at 0.85; all the machinery behind DW-Nominate produces only a 3 point gain over party-times-loyalty (and indeed in this test, adding the second DW dimension offers no additional out-of-sample improvement). Party ID by itself allows us to predict 79% of the votes, so the additional loyalty measure adds about 6 points onto this, with another three added by DW1. By contrast, the best text-based scaling adds about 4 points over pure Party ID,[16] showing that while it does not work as well as adding loyalty or DW1's scaling, when neither of those measures are available (ie, when the voting is disciplined), the textual scaling does offer significantly more vote-correlated information than the party ID that went into it. And since the textual scalings are also a measure of many other aspects of political behavior (as the words in Table 3 attest), they may in fact be a more robust measure of ideology than DW-Nominate, going beyond mere vote-prediction without even much cost in vote-predicting accuracy.

## 4 Uninformative voting: scaling the UK House of Commons

Although the text-based scalings capture both voting and speaking behavior, and match established measures of ideology like DW-Nominate quite well, clearly vote-based scalings will remain the standard for some time to come – when such voting data are available. But in legislatures with strong party discipline, while roll-call data may exist, party-line voting makes vote-based scaling difficult or impossible (Norton 1975, Norton 1980, Cowley 2002, Spirling & McLean 2007). While there are other measures, such as expert surveys (Benoit & Laver 2006, Benoit & Laver 2007), party manifesto analysis (Budge et al. 2001, Klingemann et al. 2006), or using what little voting variance there is to extract ideological information (Quinn & Spirling 2010), these methods either

[16] Although this difference is quite statistically significant, a better verification will be to compare these results across many more years.

operate on the party rather than individual level, or depend on very little data per individual. Thus text-based scaling could be essential for any empirical work requiring individual measures of legislator ideology.

The utility of the UK House of Commons is that, although the voting data are less informative, it is nevertheless a well-studied domain with numerous benchmarks to check our results against. To a certain degree, however, this is not just an empirical undertaking, but one that raises fundamental questions about the meaning and existence of underlying ideological dimensions. Should we expect there to be just one? Does it stay the same over time or across changes in leadership and agendas? Does a dimension that applies to, say, the Labour and Conservative parties, also apply to the other parties in this multi-party chamber? Do scalings that appear to work on the level of the entire legislature also work for individuals within the same party? And to what degree, if any, do these scalings reflect the sorts of behaviors that shape what little rebellious voting there is? As we will see, we can glean a significant amount of insight into the plausibility of these scalings without actually having informative votes or vote-based scalings to validate against.

Since we are using party aggregates to scale individuals, the most basic question is simply whether we can distinguish legislators from the two aggregated parties based purely on their speech data. If we box-plot the text scaling scores for Senators from the US case, grouped by whether they belong to the Democratic or Republican parties, we can see in Figure 3 that members of the two parties are quite distinct from each other in all the various scalings, both text and vote-based. Similarly, if we do this for the 1996 House of Commons (HOC) using the two largest parties, Labour and Conservative, as the reference texts, we again see a good separation between the members of these two parties, showing that we are able to distinguish at least the partisan component of

ideology based only on speech (Figure 4).

[Figures 3 and 4 about here]

A deeper question is whether this difference in partisan speech really reflects ideology, in the sense that it reflects deep differences in political outlook rather than temporary issues or idiomatic speech differences. Although it does not definitively answer the question, one test of this is to compare scalings for the same members from one year to another. If we are just picking up on trivial topics of the time, a scaling from one year is unlikely to carry over to another; if on the other hand we are just using idiomatic differences to distinguish members, then we should do about as well across any combinations of years. A useful test of this is to examine the years near the transition from Conservative to Labour rule in 1997. The two lines of of Table 5 shows the correlation in member scores between 1996 and 1998 (across the transition), versus the correlation in member scores from 1998 to 1999 (within the same Labour regime).[17] Interestingly, the results are neither consistently high, nor consistently low, but quite naturally match a reasonable story: that within a given regime, the terms of debate and thus the scalings remain relatively constant, but across regimes, the agenda and nature of debate (which is largely controlled by the ruling party) changes so significantly that it is difficult to track from one era to the other. However, it is not quite as difficult as line 1 might have us believe: if once again we remove the technical language by removing ministers from the aggregates (lines 3-5),[18] the correlation across regimes goes significantly up, though never as high as within a regime. More than just a result about plausibility

[17] All further scaling has been done using the vector projection method, due to its slight superiority in US benchmark tests, but results are substantively unchanged with the Bayesian or Wordscores methods.

[18] In the UK, like many legislatures but unlike the US, members are recorded separately in the record depending on whether they are speaking in their capacity as minister or representative; this makes removing the confounding technical language quite simple, and avoids questions of cherry-picking (or not even understanding) the vocabulary.

of this textual scaling, this also reflects the sense in which speech can capture ideological differences that are masked by the strategic imperatives imposed by party competition even in legislatures with weak discipline. To some degree left and right remain the same when the left or right switch power, but in important ways, the entire domain of conflict shifts, and there is an advantage to a scaling which reflects this.

[Table 5 about here]

Perhaps more troublesome for the notion of a simple unidimensional ideological spectrum is that this is a mutli-party system with numerous competing interests. But even if we attempt to reduce things to a single left-right dimension, the ordering of the parties in Figure 3 leaves something to be desired, for instance placing the Liberal-Democrats and the SNP in between the Labour and Conservative parties. The mistake here is in using the two largest parties, rather than two large but more relatively extreme parties. If we instead use the Liberal Democrats and Conservatives as the references, not only do these scalings correlate reasonably well with the Labour/Conservative one (lines 6-7 of Table 5), they also correlate better across regimes, perhaps because at least one end (post 1997) naturally lacks ministerial speech. Figure 5 shows that, unlike using the two main parties, the Lib-Dem/Conservative pair orders the members of all three of the largest parties "correctly".

One important validation of the benefit of using Liberal and Conservative as the references is that we have a convenient out-of-sample test using the positions of all the other parties. In Figure 5, Lib-Dem/Con orders the other parties very plausibly, putting the Democratic Unionist Party (DUP) and the Ulster Unionist Party (UUP) to the right of the Social Democratic and Labour

Party (SDLP), Plaid Cymru (PC), and Scottish National Party (SNP) parties. More precisely, if we compare these results with established measures such as Benoit and Laver (BL) for 2001 (Benoit & Laver 2006), the Comparative Manifesto Project (CMP) for 1997 (Budge et al. 2001), and the Chapel Hill Expert Survey Series (CHES) for 1999 (Hooghe, Bakker, Brigevich, De Vries, Edwards, Marks, Rovny, Steenbergen & Vachudova 2010), we find these measures have Spearman rank correlations with the text scaling of 0.88, 0.83, and 0.90, respectively (all are significant at $p < 0.05$).[19] Given that these established measures correlate amongst themselves at 0.82, 0.60, and 0.80, the fact that the text scaling correlates with all of them better than any of them do with each other is a solid validation of party-level results.

[Figure 5 about here]

Finally, if we are to examine the scaling of individual legislators, we must turn to existing individual-level scalings. Scholars of the HOC have done their best to tease out ideological information using expert knowledge and what few "rebellions" do occur, most prominently Norton (1975, 1980) and later Cowley (2002, e.g.). Of course, as with DW-Nominate, distinguishing left-right from simple party loyalty is challenging for any vote-based approach. Quinn & Spirling (2010), for instance, motivate their dirichlet clustering algorithm by showing how existing scalings (such as Optimal Classification, Nominate, or PCA) fail to order five Labour members correctly, conflating rebellion from the left with that from the right. By comparison, when the Lib-Dem/Conservative text scaling is applied to these five members, four are ordered "correctly," with the one incorrect placement being the Labour "loyalist" John Prescott, who unlike all the rest, spoke during this

[19] Data acquired April 2012 from http://www.tcd.ie/Political_Science/ppmd/ for BL, from http://www.nsd.uib.no/macrodataguide/set.html?id=62&sub=1 for CMP, and from http://www.unc.edu/g̃wmarks/data_pp.php for CHES.

time period primarily in his capacity as "The Secretary of State for the Environment, Transport and the Regions." More importantly, the other Labour loyalist is placed between the Conservative and the two rebellious Labour leftists. If we look more broadly at rebellion rates (ie, party loyalty), correlations between the text scaling and individual rebellion rates ranges between 0.17 and 0.22 within-party – small, but statistically significant, and comparable to the US case.[20] But if we consider two more methodologically sophisticated vote-based HOC scalings, which employ clustering or PCA approaches,[21] we find a correlation of 0.55 with the text scaling. However, since (as with DW-Nominate) these vote-based scalings are 95% explained by party ID plus loyalty, this primarily shows that it is very challenging to extract ideological information from the vote record that goes beyond loyalty rates, whatever the degree of discipline in the legislature. The text-based scaling both picks up on this loyalty measure, and also captures all the substantive disagreement that strategic voting alone masks.

## 5 Conclusion

Political settings that lack informative voting data vastly outnumber those with informative voting. We have seen here that a new approach using only textual data plus party information allows us to scale speakers in a manner that resembles vote-based results and allows us to predict roll-call votes where applicable; these results go beyond party ID to capture both party loyalty and all the other aspects of political difference reflected in speech. In addition to this general scaling approach, two new estimation methods are developed – Bayesian and vector projection – which are

[20] Rebellion rates acquired from thepublicwhip.co.uk, 2011

[21] Ibid. and Lightfoot (2011), respectively. These two scaling techniques produce results that themselves correlate at the 0.98 level.

shown to be both better theoretically and empirically than existing methods, whether supervised or unsupervised. And even in the challenging setting of disciplined legislatures, the results are validated against existing expectations based on party and individual positions.

In multi-party systems, these results appear to work best when large, more extreme parties are chosen as the reference points. Of course, many scholars assume that these multi-party legislatures are also ideologically multidimensional – and indeed the apparent unidimensionality of the US system may be more a reflection of the two-party system than the actual ideological space that legislators think and (in domains other than voting) also behave in. The spatial text-scaling method in particular can be easily extended to a multidimensional setting.[22]

Another direct extension of this approach is to domains outside of legislatures altogether. For instance, a new online scaling tool "Gscale" using the vector scaling method allows one to automatically use as reference texts the results of any Google search, and to automatically scale either a set of documents or the results of other Google searches. A initial test took Google searches of "democrat" and "republican" as reference texts, and used those to scale each current senator again using only the text from Google searches of their names; remarkably, the resultant scaling correlates with DW1 at 0.67, and correctly classifies the party membership of 91% of Senators. This tool can also be easily used to scale other politicians, journalists, or anyone else with a presence on the internet – although validating these results will be a significant undertaking. Much work remains to be done to establish how portable a text-based scaling is across domains and time periods, particularly as we venture farther from verifiable settings like legislatures. But the already vast pool of political text data is only growing, offering up enormous opportunities to test existing

[22]For instance, knowing the distance between three reference parties allows us to place them in a shared two-dimensional space, after which each individual can be projected onto that plane knowing the distances to those three parties. See nickbeauchamp.com for tools implementing this and Gscale.

spatial theories, and to develop new theories explicitly designed to model the cacophony of talk surrounding and affecting every political action.

## Appendix A: Bayesian scaling

We wish to discover $p(R|S)/p(D|S)$ given $p(w_i|R)$, $p(w_i|D)$ and $p(w_i|S)$, where R and D are the two reference documents, S is the document (speaker) of unknown ideology, and $p(w_i|X)$ is the probability of encountering word $i$ given document X. From Bayes, we know that:[23]

$$p(R|S) = \frac{p(S|R)p(R)}{p(S)} \tag{7}$$

If the probability of encountering word $i$ given that the speaker is a Republican is $p(w_i|R)$, then we might naively assume that $p(S|R)$ – ie, the probability of an entire document S given that the speaker is a Republican – is simply the probability of each event $p(w_i|R)$, considered as independent of all other events $p(w_i|R)$.[24] Thus we would say:

$$p(S|R) = \Pi_{i \in S} p(w_i|R) \tag{8}$$

This is undoubtedly false (since words are correlated with each other), but it seems to work fairly well in practice. Combining these two, we get:

$$p(R|S) = \frac{p(R)}{p(S)} \Pi_{i \in S} p(w_i|R) \tag{9}$$

If we assume that a speaker is either a Republican or not (=D), then we also have:

$$p(D|S) = \frac{p(D)}{p(S)} \Pi_{i \in S} p(w_i|D) \tag{10}$$

Taking the ratio of these last two, we can cancel out $p(S)$ and get a likelihood ratio, which is what we are really interested in:

$$\frac{p(R|S)}{p(D|S)} = \frac{p(R)}{p(D)} \Pi_{i \in S} \frac{p(w_i|R)}{p(w_i|D)} \tag{11}$$

[23] This exposition is drawn from Bishop (2006) and http://en.wikipedia.org/wiki/Naive_Bayes_classifier.

[24] Note that the notation employed in this section and the following, while not quite standard, was chosen in order to facilitate comparison with the Wordscores approach and the somewhat idiosyncratic notation it employs. One important difference in notation between the two sections, however, is that here "$i$" denotes each occurrence of a word, with a new number even for repetitions of the same word, whereas in the Wordscores section, $i$ indexes each word uniquely.

It is trivial to go from this ratio back to $p(R|S)$, but the ratio itself is an equivalent score. In practice, given that the right-most quantity will be quite small, we calculate the log ratio:

$$log\frac{p(R|S)}{p(D|S)} = log\frac{p(R)}{p(D)} + \Sigma_{i\in S} log\frac{p(w_i|R)}{p(w_i|D)} \tag{12}$$

As was discussed earlier, the Bayesian approach has generally been designed for classification instead of scaling. Since we are not interested classification, just scoring, we can drop the middle quantity (it's a constant for all Senators, after all), and merely calculate the latter quantity for each Senator. That is:

$$\text{Bayesscore} = B_V = \Sigma_{i\in S} \text{log}\frac{p(w_i|R)}{p(w_i|D)} \tag{13}$$

□

## Appendix B: Wordscores

Instead of beginning with $p(w_i|R)$, the independent probability of encountering word $i$ given text $R$, LBG begin with $P_{iR} \equiv p(R|w_i)$,[25] the probability that a text is of type $R$ given an encounter with word $i$. Again from Bayes (sticking with two reference texts for simplicity), we have:

$$p(R|w_i) = \frac{p(w_i|R)p(R)}{p(w_i)} = \frac{p(w_i|R)p(R)}{p(w_i|R)p(R) + p(w_i|D)p(D)} \tag{14}$$

Call $W_R$ the total number of words in document R, and likewise for $W_D$; call $W_{iR}$ the number of occurrences of word $i$ in document R, and likewise for $W_{iD}$. Then, as before, we have

$$p(w_i|R) = \frac{W_{iR}}{W_R} \text{ and } P(R) = \frac{W_R}{W_R + W_D} \tag{15}$$

ie, the probability of word $i$ given document R is just the percentage of document R made up of word $i$, and the probability of document R given that we're reading either R or D is simply the

[25] Laver, Benoit, and Garry use $P_{wr}$, with $w$ as the $w$th word instead of $i$; for consistency, the $i$ notation is retained here.

percentage of total words that make up R. Thus from the Bayesian approach, one gets:

$$p(R|w_i) = \frac{W_{iR}}{W_{iR} + W_{iD}} \tag{16}$$

Laver, Benoit, and Garry instead present a slightly different formulation:

$$P_{iR} = \frac{\frac{W_{iR}}{W_R}}{\frac{W_{iR}}{W_R} + \frac{W_{iD}}{W_D}} \tag{17}$$

That is, the probability that you have document R given word $i$ is the percentage of document R made of word $i$ divided by the sum of the respective percentages of R and D made up of word $i$. If $P_{iR} \equiv p(R|w_i)$, this is false, but when $W_R \approx W_D$ (as in the example in their paper), these two formulations will be nearly the same.

At this point, their method becomes somewhat less Bayesian. Each virgin document is assigned an *a priori* scalar value $A_R$ and $A_D$;[26] for instance, if, as is the case here, we consider R and D to be two poles on a linear spectrum, we might assign $A_R = -1$ and $A_D = 1$, although any two numbers would produce essentially equivalent scalings. Every possible word is then assigned a score $S_i$, where (sticking to two reference texts):

$$S_i = A_R \cdot P_{iR} + A_D \cdot P_{iD} \tag{18}$$

And finally, to construct an overall score for a virgin document, $S_V$, we have

$$S_V = \sum_i \frac{W_{iV}}{W_V} \cdot S_i \tag{19}$$

Where of course the fraction $W_{iV}/W_V$ is simply the percentage of our virgin document V made up of word $i$.

We can characterize a bit more precisely the difference between Wordscores and the Bayesian

[26]The authors actually allow for scores on multiple dimensions, corresponding to different values $A_{Rd}$, but for simplicity and for parity with previous explications, only a single dimension is employed here; the extension is straightforward.

approach. If Wordscores assigns a scalar $S_i$ to each word $i$, and an overall score $S_V$, we can analogously say that the Bayesian approach similarly assigns a score $B_i$ to each word, and an overall score $B_V$. We then have similar formulations:[27]

$$S_V = \sum_i \frac{W_{iV}}{W_V} \cdot S_i \text{ and } B_V = \sum_i W_{iV} \cdot B_i \tag{20}$$

If we assign values of $A_D = +1$ and $A_R = -1$ to the two reference texts in the Wordscores method, and we denote $F_{iR} \equiv \frac{W_{iR}}{W_R}$ and similarly for $F_{iD}$, we have:[28]

$$S_i = \frac{F_{iD} - F_{iR}}{F_{iD} + F_{iR}} \text{ and } B_i = \log\left(\frac{F_{iD}}{F_{iR}}\right) \tag{22}$$

Thus $S_i$ and $B_i$ correspond to the weight assigned by each method to each word $i$, which is then multiplied by the frequency of that word in the virgin text and summed over all words $i$ to get the final score, as in equation (15). The formulas in (17) appear quite different, but in fact the results are fairly similar (see Figure 1). The overall values $S_V$ and $B_V$ are often even more similar for two reasons: First, as mentioned before, and as can be seen in the figure, the formulas for $S_i$ and $B_i$ differ most when either $F_{iR}$ or $F_{iD}$ is low, but in those cases $W_{iV}$ tends also to be low, lessening the impact of the different values. Second, when actually applying the Bayesian method, we generally weight $B_i$ by $\frac{W_{iV}}{W_V}$ rather than $W_{iV}$, which of course produces a result much more similar to that of Wordscores.[29] The reason for this is that the latter multiplier correctly utilizes information about the length of various texts to estimate the score: if texts of type R are generally longer than those of type D, the Bayes method makes use of that information. However, in the current context, we are interested fundamentally in the content of the texts, not their length; although the length might well be correlated with the ideology of the speaker, in the legislative context, the amount of

[27] Note that here $i$ indexes only unique words.

[28] Although the vector projection method score correlates fairly highly with the other two, the formulation using matching notation is quite different:

$$P_V = \frac{\sqrt{\sum_i (F_{iD} - F_{iR})^2)} - \sqrt{\sum_i (F_{iV} - F_{iD})^2)} + \sqrt{\sum_i (F_{iV} - F_{iR})^2)}}{2\sqrt{\sum_i (F_{iD} - F_{iR})^2)}} \tag{21}$$

[29] That said, there are still cases where a word only appears in one of the two reference texts. To prevent the Bayesian approach from automatically assigning a document with that word to that reference document's position (or from encountering worse problems when a test document has words unique to both reference texts), some smoothing prior must be applied. It turns out that the results are almost identical no matter what gentle prior is used, whether it is uniform or based on the frequency of words in the larger world.

text a speaker manages to get into the record will depend heavily on which party is in power, the seniority of the speaker, his/her party position, and so forth.[30] Although all this might correlate with the ideological content of their speech, of course, overall more noise is eliminated by effectively normalizing all documents to the same length.

[30] As only one example, John Major, when Prime Minister, had nearly 5 times as many words entered in the House of Commons record than any other member, which clearly reflects much more than mere ideology.

## References


Benoit, K. & M. Laver. 2006. *Party policy in modern democracies.* Vol. 19 Taylor & Francis.

Benoit, K. & M. Laver. 2007. "Estimating party policy positions: Comparing expert surveys and hand-coded content analysis." *Electoral Studies* 26(1):90–107.

Bishop, C.M. 2006. *Pattern recognition and machine learning.* Springer.

Budge, I., D. Robertson & D. Hearl. 1987. *Ideology, Strategy and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies.* Cambridge University Press.

Budge, I., H.D. Klingemann, A. Volkens, J. Bara & E. Tanenbaum. 2001. "Mapping Policy Preferences, Estimates for Parties, Governments and Electors 1945-1998.".

Cowley, P. 2002. *Revolts and rebellions: Parliamentary voting under Blair.* Politico's.

Diermeier, D., J.F. Godbout, B. Yu & S. Kaufmann. 2012. "Language and ideology in Congress." *British Journal of Political Science* 42(1).

*Electoral-Vote.com.* 2011.
**URL:** *http://www.electoral-vote.com/*

Hooghe, L., R. Bakker, A. Brigevich, C. De Vries, E. Edwards, G. Marks, J. Rovny, M. Steenbergen & M. Vachudova. 2010. "Reliability and validity of measuring party positions: The Chapel Hill expert surveys of 2002 and 2006." *European Journal of Political Research* 42(4).

Jackman, S. 2001. "Multidimensional analysis of roll call data via Bayesian simulation: identification, estimation, inference, and model checking." *Political Analysis* 9(3):227.

Janda, K., R. Harmel, C. Edens & P. Goff. 1995. "Changes in Party Identity: Evidence from Party Manifestos." *Party Politics* 1(2):171.

Klingemann, H.D., A. Volkens, J. Bara, I. Budge & M. McDonald. 2006. "Mapping Policy Preferences II. Comparing 24 OECD and 24 CEE Countries, 1990-2003.".

Laver, M. & J. Garry. 2000. "Estimating Policy Positions from Political Texts." *American Journal of Political Science* 44(3):619–634.

Laver, M., K. Benoit & J. Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(02):311–331.

Laver, M.J. & W.B. Hunt. 1992. *Policy and party competition.* Routledge and Kegan Paul.

Lightfoot, Chris. 2011.
**URL:** *http://ex-parrot.com/*

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

Monroe, B.L. & K. Maeda. 2005. "Talk's Cheap: Text-Based Estimation of Rhetorical Ideal-Points." *Working paper* .

Newman, M.E.J. 2005. "Power laws, Pareto distributions and Zipf's law." *Contemporary Physics* 46(5):323–351.

Norton, P. 1975. *Dissension in the House of Commons 1945-74*. London: Macmillan.

Norton, P. 1980. *Dissension in the House of Commons, 1974-1979*. Clarendon press.

Poole, K.T. 2005. *Spatial models of parliamentary voting*. Cambridge Univ Pr.

Poole, K.T. & H. Rosenthal. 1985. "A spatial model for legislative roll call analysis." *American Journal of Political Science* 29(2):357–384.

Poole, K.T. & H. Rosenthal. 1991. "Patterns of congressional voting." *American Journal of Political Science* 35(1):228–278.

Quinn, K. & A. Spirling. 2010. "Identifying Intra-Party Voting Blocs in the UK House of Commons." *Journal of the American Statistical Association. Forthcoming* .

Slapin, J.B. & S.O. Proksch. 2007. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *Working Paper* .

Spirling, A. & I. McLean. 2007. "UK OC OK? Interpreting optimal classification scores for the UK house of commons." *Political Analysis* 15(1):85.

*The Public Whip*. 2011.
**URL:** *http://www.publicwhip.org.uk/index.php*

Table 1: Correlations between Bayes Wordscores and Vector Projection methods for simulated data.[a]

| Number of Words | Distribution of Words | Wordscores & Bayes | Vector & Bayes | Wordscores & Vector |
|---|---|---|---|---|
| 1000 | $e^{-5x}$ | 0.993 | 0.989 | 0.978 |
| 1000 | $e^{-20x}$ | 0.884 | 0.770 | 0.862 |
| 1000 | $(5x)^{-2.2}$ | 0.995 | 0.917 | 0.890 |
| 1000 | $(20x)^{-2.2}$ | 0.952 | 0.523 | 0.316 |
| 2000 | $e^{-5x}$ | 0.994 | 0.958 | 0.955 |
| 2000 | $e^{-20x}$ | 0.987 | 0.940 | 0.903 |
| 2000 | $(5x)^{-2.2}$ | 0.991 | 0.868 | 0.814 |
| 2000 | $(20x)^{-2.2}$ | 0.999 | 0.807 | 0.811 |

[a] Correlations between these three methods generally decline with thinner-tailed distributions or fewer words. The correlation between Bayes and Wordscores is quite high, but even that tight correlation weakens when many words occur in only one document. All correlations significant at $p < 0.05$.

Table 2: Correlations of scalings of 2006 US Senate[a]

| | DW1 | VECT | BAYES | PCA1 | PCA2 |
|---|---|---|---|---|---|
| VECTOR | 0.644* | | | | |
| BAYES | 0.615* | 0.839* | | | |
| PCA1 | 0.082 | 0.077 | 0.065 | | |
| PCA2 | 0.387* | 0.742* | 0.567* | -0.162* | |
| WORDFISH | 0.375* | 0.644* | 0.474* | 0.306* | 0.953* |

[a] * indicates correlations significant at $p < .05$. Note that Wordscores values are omitted because they correlate with Bayes at 0.99 in this case. Of particular interest are the correlations between DW1 and the other values.

Table 3: Top 20 Democratic and Republican words for the 2006 US Senate[a]

| Democratic | Republican | Rep, no tech.[b] |
|---|---|---|
| iraq | consent | border |
| administration | ask | law |
| year | unanimous | states |
| health | bill | court |
| families | committee | judge |
| program | senate | defense |
| care | border | district |
| debt | senator | business |
| women | vote | united |
| veterans | law | marriage |
| help | hearing | illegal |
| americans | authorized | think |
| country | states | very |
| children | proceed | system |
| new | order | lot |
| education | session | iran |
| funding | time | amnesty |
| workers | meet | good |
| programs | court | judiciary |
| disaster | judge | circuit |

[a] The 20 largest and smallest values from from the vector $R - D$, ie, the most Republican and Democratic words.

[b] Top Republican words when the technical language, which dominates R words when R is in majority, is eliminated from the corpus.

Table 4: Proportion of yea/nay votes guessed correctly out-of-sample using the specified scaling, 2006 Senate.[a]

| Scaling | Mean (s.d.) |
|---|---|
| Party | 0.794 (0.022) |
| Party * Loyalty | 0.851 (0.023) |
| DW1 | 0.881 (0.015) |
| DW2 | 0.707 (0.038) |
| Rollcall PCA1 | 0.878 (0.015) |
| Rollcall PCA2 | 0.674 (0.031) |
| Vector | 0.814 (0.030) |
| Bayes | 0.829 (0.024) |
| Wordfish | 0.722 (0.038) |
| Word PCA1 | 0.672 (0.030) |
| Word PCA2 | 0.724 (0.038) |
| DW1 & 2 | 0.878 (0.015) |
| RC PCA1 & 2 | 0.878 (0.015) |
| Word PCA1 & 2 | 0.734 (0.040) |

[a] For each run, observations are randomly divided 80% in-sample, 20% out-sample. For each roll-call vote, a logit is estimated in-sample with the vote as dependent variable and the scaling as independent variable, and then the vote is predicted out-of-sample. This is done for all rollcall votes, and repeated for 1000 samples.

Table 5: Correlations between different HOC scalings[a]

| Scalings | Correlation |
|---|---|
| 1996 & 1998 | 0.135 |
| 1998 & 1999 | 0.625 |
| 1996 & 1998 w/o ministers | 0.249 |
| 1998 & 1996 w/o ministers | 0.523 |
| 1998 w/o ministers & 1996 w/o ministers | 0.306 |
| 1998 & 1998 con/lib refs | 0.494 |
| 1996 & 1996 con/lib refs | 0.602 |

[a] Unless otherwise specified, "year A & year B" means the correlation between the vector-based scalings of the members of year A, using year A's Labour and Conservative aggregate texts as reference, with the members who are also present in year B, as scaled by year B's Labour and Conservative reference texts. "199x w/o ministers" means that all of that year's members have been scaled with that year's Labour and Conservative aggregate party texts not including the speech of the ministers. "199x con/lib refs" means that that year's reference texts are the Conservative and Liberal parties. All correlations significant at $p < 0.05$.

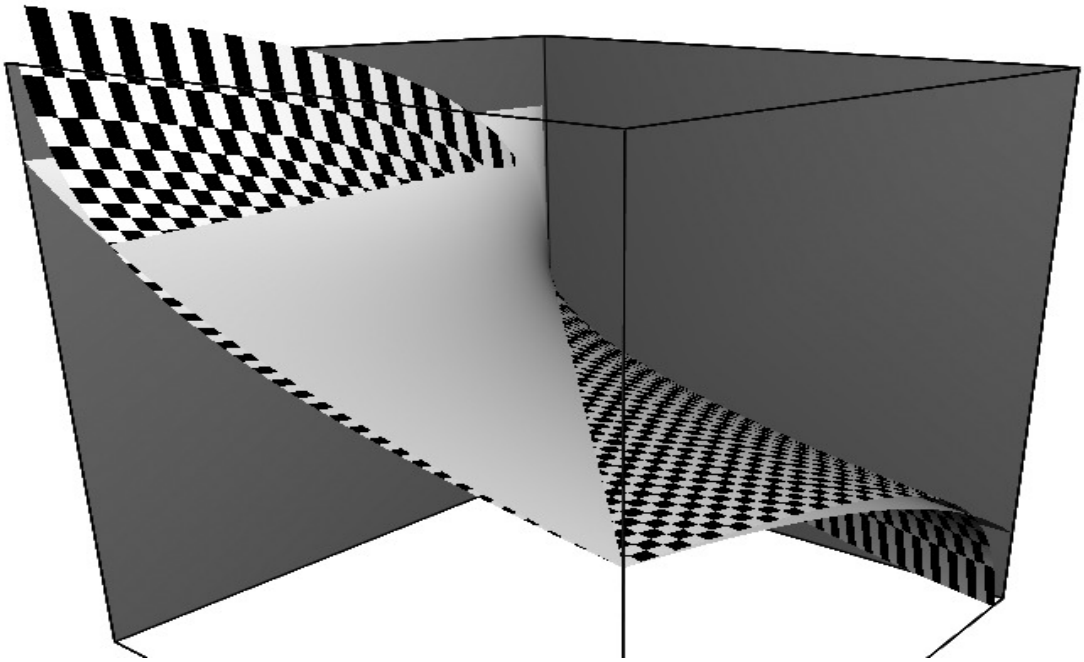

Figure 1: A comparison of the word weighting assigned by Wordscores (smooth) and the Bayesian method (checkered). The $x$ and $y$ axes correspond to the frequencies of some word $i$ in the two reference documents (ie, $F_{iR}$ and $F_{iD}$ from equation 17), and the $z$ axis corresponds to $S_i$ or $B_i$. (See Appendix B.) As can be seen, despite the apparent dissimilarities in equation (17), the two functions are quite similar, diverging mainly for low values of $F_{iR}$ or $F_{iD}$, where the Bayesian weight correctly goes to infinity when the word frequency is 0 in only one of the two reference texts.

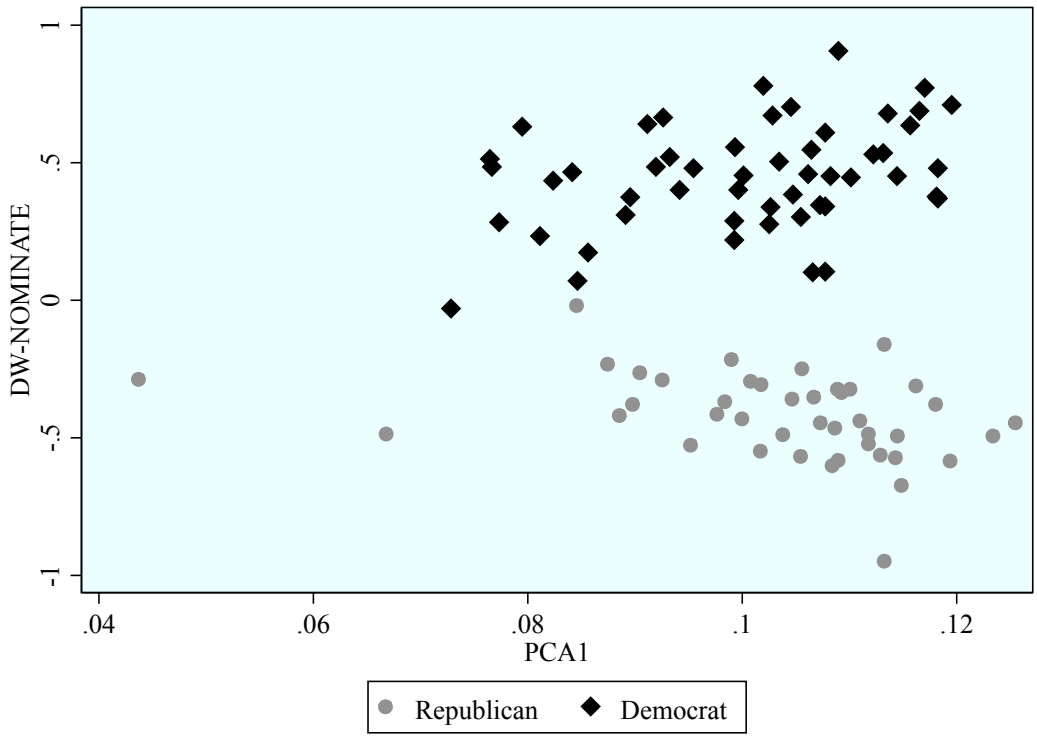


Figure 2: Plot of the first principal component against the first DW-Nominate dimension. Note that although there is no overall correlation between the DW1 score and the PCA1 score, within each party the scores are quire correlated.

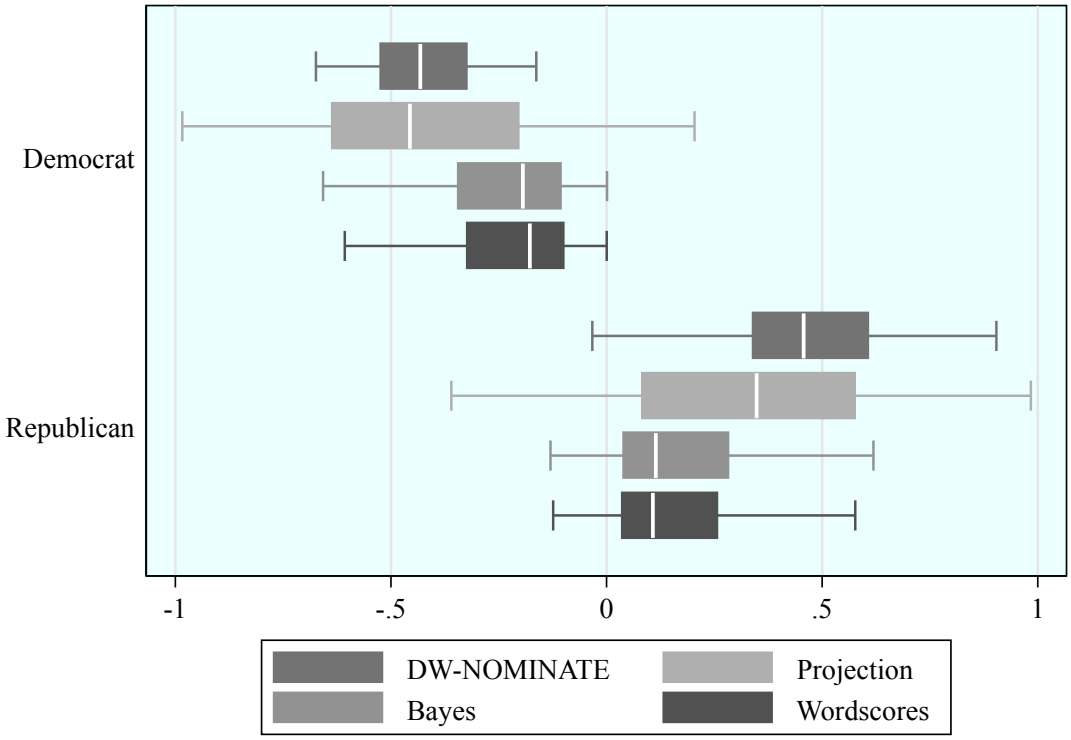


Figure 3: Four methods of scaling members of the US Senate, displayed as grouped by Democratic and Republican parties. Axis scale is from DW-Nominate; the other scores have been rescaled to match.

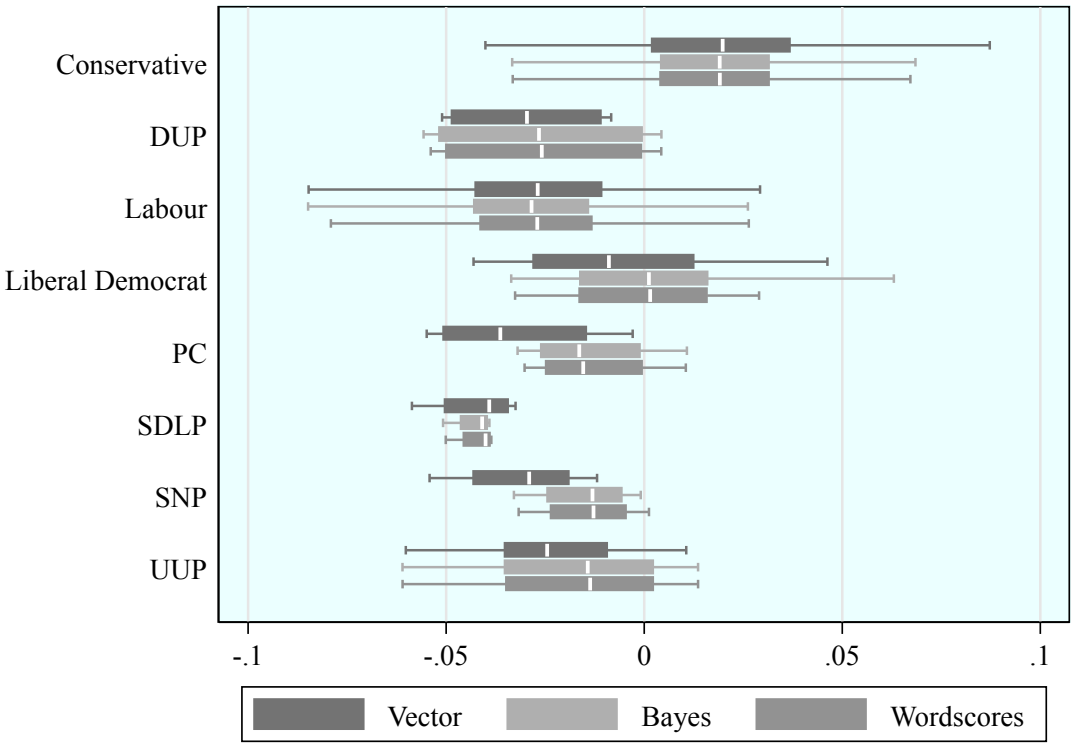


Figure 4: As in Figure 3, the members of the House of Commons, scaled by the three main techniques, grouped by party. Axis scale is from Bayes; the others have been rescaled to match.

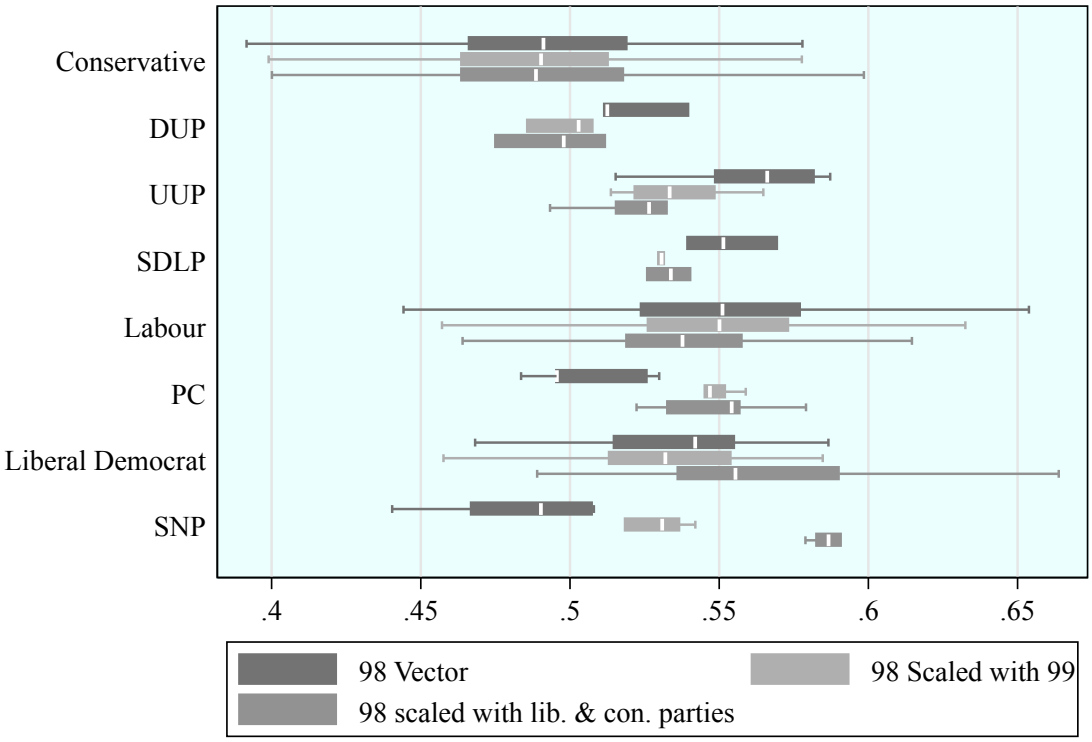


Figure 5: The 1998 House of Commons members scaled by different reference text pairs. "98 Vector": the basic Labour/Conservative scaling. "98 Scaled with 99": scaling the 1998 members with reference texts based on the aggregated party speech from 1999. "98 scaled with lib. & con. parties": 1998 members as scaled by Liberal Democrat and Conservative aggregates as reference texts. Y axis: parties sorted by third scaling.