Using Text to Scale Legislatures with Uninformative Voting

Nick Beauchamp*
NYU Department of Politics

[DRAFT: June 27, 2011]

Abstract

Many models of legislative behavior require knowing the positions of individual legislators, but while such positions can be derived from rollcall votes when party discipline is weak, few legislatures exhibit such informative voting. This paper shows how legislators' written and spoken text can be used to scale individuals even in the absence of informative votes, by constructing two reference texts out of the aggregated speech of all members from each of two major parties. Although the popular Wordscores method can be used with this approach, this paper develops a Bayesian scaling that is more theoretically sound and which produces empirically similar results to Wordscores; this paper also develops a vector-based scaling that works better than either. The unsupervised Wordfish scaling is also tested, but is found to do worse than the supervised approaches, and no better than a quick principal component analysis of the text. These scalings are first successfully validated in the US Senate against the benchmark vote-based DW-Nominate scores, and are then tested in the UK House of Commons. In the latter case, 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. Of practical importance, removing the technical language of legislation improves results greatly, as does using more extreme and out-of-power parties. Given that vote-based scalings capture little more than party and party loyalty, the text-based scaling developed here both matches what we already have 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.

*Author email: nick.beauchamp@nyu.edu. For their valuable contributions at various stages of this paper's development, the author would like to thank: Mik Laver, Jonathan Nagler, Neal Beck, Andrew Martin, Ken Benoit, Will Lowe, Larry Mead, Sandy Gordon, Brandon Stewart, and Alex Herzog.

1 Introduction

Most models of legislative behavior assume that members have ideal points positioned in some shared policy space. In order to test such models, the positions of dozens or hundreds of members must first be estimated, a procedure that is relatively straightforward given well-documented and sufficiently varying roll call voting data (Poole & Rosenthal 1985, 1991, 2000; Poole 2005). But although informative voting data are plentiful in the US,[1] this is more the exception than the rule: more often, party discipline means that all members of a party tend to vote alike, and thus voting behavior reveals little about the ideology of the legislators. But while informative voting may be rare outside the US, what we do have in increasingly plentiful quantities is the recorded speech and writings of legislators, a vast and growing archive of varied and revealing behavioral data. In addition, we generally have party membership information, and often a general sense of the overall political positions of the parties. As we will see, these two data sources allow us to scale legislators. But there have been other approaches to text-based scaling previously: the key thing here is to discover exactly which methods work best – a nontrivial problem when member positions are not already known.

To do that, I examine first theoretically and then empirically a number of existing techniques, and also introduce two that are new to political science. On the theory side, I show (1) that the well-established "Wordscores" method is essentially a flawed Bayesian approach, though it produces results that are often similar to what a properly Bayesian technique does, and (2) that a vector-based approach developed here should also produce somewhat similar results despite having entirely different underpinnings.

[1] Though perhaps becoming less so in recent years as Congress becomes more polarized (McCarty, Poole & Rosenthal 2005).

But the proof is in the pudding, and the bulk of this paper is empirical. The best test of these various techniques is to compare them in a setting where we already have vote-based scalings, such as the US Senate. The current dominant scaling is Poole and Rosenthal's DW-Nominate (Poole & Rosenthal 1985, 1991, 2000; Poole 2005), which I take as the gold standard against which to compare the text-based scalings. Roughly speaking, the fundamental approach behind these text-based scalings is to create two aggregate texts – all Democrat speech and all Republican speech, in the US case – and use them to scale each individual legislator by their similarity to one or the other of these reference documents. I find that under a variety of assumptions, the techniques described above all work well – Wordscores, the proper Bayesian approach, and the vector-based approach – although the latter appears to work best of all. I also compare the other major text-based scaling, "Wordfish," along with a principal-component based approach, but I find that both of these work less well than the first three (though similar to each other). Going one step farther, I show that all of these approaches work nearly as well as DW-Nominate on its raison-d'etre, predicting roll-call votes, and thus may be considered equivalent if not better measures of ideology. Lastly, I show that an important part of successfully scaling legislators is to eliminate, as well as possible, the technical language of legislation (hereby, unanimous, committee, etc), which appears to muddle ideology with party leadership.

In the second half of the empirical section, I turn from the benchmark US context to an important case where vote-based scaling is impossible: the UK House of Commons. Despite first appearances, there are a number of ways that the text-based scaling of these legislators might be tested against what we already know. I find (1) that these approaches are well able to distinguish the members of different parties based purely on what they say; (2) using the Liberal and Conservative

parties as "references" works best to scale everyone, including members of other parties; (3) again, eliminating technical language or ministerial speech is essential; (4) these scalings work well even across years, though less so across changes of party control; and (5) these approaches work as well as anything extant in predicting what little "revolt" there is in UK roll call voting.

Having validated this approach, it can now be used to scale legislators wherever a few basic facts about party membership and ideology are known. Moreover, the party-derived reference documents could plausibly be used to determine the ideological position of other speakers and documents in a wide variety of settings: local legislatures, candidates, appointed officials, or indeed anyone with a speech record obtainable via a search engine. While there have been other efforts to expand scaling beyond the setting of informative voting, speech records and and the tools developed here may have the widest applicability for scaling political actors and testing models with ideological positions.

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 (Janda, Harmel, Edens & Goff 1995, Laver & Schofield 1998) to non-expert judgments 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. Finally, the most automated scaling approach is currently the Wordfish algorithm of Slapin and Proksch (2007), based on work by Monroe and Maeda (2005), which scales text without any human intervention at all. While Wordscores is a "supervised" method (where humans set the positions of reference documents), Wordfish is "unsupervised," requiring no human supervision at all. But as we will see, at some point the freedom from human supervision comes with the cost of poorer performance.

More broadly, although scalings are the primary desideratum in political science, machine learning more generally has often focussed on classification rather than scaling. This paper introduces a couple of new scaling methods to political science, both of which are based on existing classification algorithms which have been modified to produce continuous quantities: First, the naive Bayesian classifier (the basis of most spam filters, for instance) has been modified into a Bayesian scaler. And second, the ubiquitous Support Vector Machine (SVM) has been modified into a vector-based scaler. The first, as we will see, turns out to be similar to Wordscores, which was designed in a broadly, if imprecisely, Bayesian way. And the second turns out to work better than either Wordscores or the Bayesian method. Finally, as will be shown, all three superivised methods work better than the unsupervised Wordfish or an empirically similar but more efficient Principal Component Analysis (PCA).

Thus in total five different methods will be investigated here: three supervised methods – Wordscores, a Bayesian scaling, and the vector projection method – and two unsupervised methods – Wordfish, and a principal component approach. 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.[2] Each text file is then transformed into a vector, where each word corresponds to a dimension, and the frequency of that word in the document is the position in that dimension. Since only the top 1000 words are retained,[3] each senator then has an associated 1000-dimensional vector,[4] where for each word value w_{ij} (where i is the word number, j is the senator number) the values are normalized such that $\Sigma_i w_{ij} = 1$. In short, each w_i value is the percent of the Senator's entire speech that consists of that word.[5] In addition to this 1000x100 matrix,[6] two additional 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. Again, other reference texts are possible, but the current goal is simply to replicate as well as possible the traditional political spectrum captured by vote-based, party-dominated methods such as DW-Nominate.

[2]Of course, almost all speech acts take place in exchanges on the floor, and the Congressional Record does not divide information by single speech act. So each text element, consisting of perhaps dozens of exchanges, must be chopped up by Senator and added to the correct concatenated files.

[3]1000 was chosen simply because it was the most that were computationally feasible, and previous work suggests that, all else being equal, these methods work better with more words. Also, it is common practice to exclude a set of about 100 "stop-words" at the outset – uninformative words like "the," "of," "and," and so forth – and since those are often the most common words, the vector actually consists of the 100th to 1100th most common words, approximately.

[4]Though many of these values may be 0 for a given individual. This is another reason not to go far beyond 1000 words, since with a larger set, a much larger percentage of any individual's vector will be 0s.

[5]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 often better to work directly with frequencies to begin with, and work any Bayesian weighting at the parameter estimation stage, if at all.

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

2.1 Bayesian scaling

Although the Bayesian approach here was developed in part to remedy some of the theoretical problems with Wordscores, the exposition works better to 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 ratio of the likelihoods 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).

Thus we wish to discover $p(R|S)$ and $p(D|S)$, ie, the probabilities that a speaker is Republican 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 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.[7]

[7]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, given 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, particular 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

[Figure 1 about here]

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 references documents 1 and 2 (x and y axes). Of note, despite the different formulations, the functions are surprisingly similar, except when frequencies are near zero. As can be seen, when the 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 proper priors) incorrectly weight words that occur in one document and not the other simply by chance. (See Appendix for more details.)

2.3 Vector Projection

This third supervised scaling is derived from Support Vector Machine classifiers, but discard much of the mechanism of classification, working directly with the vector space. Each Senator is considered as a point in 1000-dimensional space (corresponding to his/her vector end-point), and each point is projected onto a line between two fixed points. Given the close connection between party ID and traditional scalings, the reference texts here are the total Democrat vector, and the total Republican vector. 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, an especially simple approach given the

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.

distances $||R - D|| = a$, $||S - R|| = b$ and $||S - D|| = c$ yields:[8]

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

This value is calculated for each Senator, and without further transformation is taken as their vector score.

2.4 A comparison of simulated results from the first three methods

Although the base functions of Wordscores and naive Bayes appear similar in Figure 1, it is not always the case that their results will be so alike. To better understand the interrelation between the three supervised techniques, a series of simulated "texts" were created and scaled. 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, much current research suggests that the frequency of words in texts has a fatter tail, following instead a "power law" of the form x^{α} (Newman 2005).

[Table 1 about here]

[8] We could also use basic vector projection to get the same result, or we could dispense with lengths and only employ the angle between vectors, using the cosine similarity metric.

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 (1000 words was the computational limit for the empirical work here). 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 will in many cases be adequate.

2.5 Unsupervised scaling: Wordfish

"Wordfish" is an unsupervised scaling method, developed by Slapin and Prosksch (2007; hereafter, SP) based on the work of Monroe and Maeda (2005; hereafter, MM). MM's original model is based on item response theory (IRT), which originally considered a set of respondents and their answers to various questions, and sought to place the questions and respondents in a shared space. For instance, if rightward corresponds to harder questions and more able respondents, then the further to the right a respondent's position is relative to a question's position, the more we would expect the respondent to get that question correct (modeled via, say, a logistic function of that relative separation).

In the scaling context, words are analogous to questions and documents to respondents, where the likelihood of a word appearing in a document is analogous to the likelihood of a respondent

answering a question correctly. But since we are dealing with (almost) continuous quantities (word frequencies), we don't need to employ logistic functions with cut-points. Rather, we just estimate the likelihood of a word based on its distance from the document, that is, $p(w_i|S)$ is simply some function of the distance of word w_i from Senator S_j. Following MM and SP, $p(w_i|S)$ would be taken as a poisson function of the distance between the Senator and the word (akin the exponential function used in the preceding simulation).[9]

Thus the model is:

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

Where i indexes documents, j indexes words, c is a constant, c_i^x are a document-specific constants, c_j^α are word-specific constants, and $(x_i - \alpha_j)$ is the distance between a document position x_i and a word position α_j. An additional parameter γ_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 the reverse when a speaker/document is to the left of the word) and this would correspond to a positive γ for that word; conversely, for a left-wing word, we would expect the reverse, with a negative γ. Perhaps unsurprisingly given three separate word parameters, the model is under-indentified, so MM jettison the word position parameters α_j (setting them all to 0) and interpret the γ_j parameter as something like word position, where larger positive values correspond, say, to more right-wing

[9]Following Zipf and Newman (2005), a fatter-tailed distribution such as a "power law" might be more appropriate here, where the likelihood goes as d^k rather than k^d (where d is distance and k is some constant). Preliminary testing suggest that this makes little practical difference, however.

words, etc. 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. MM and Wordfish do this via expectation maximization.

2.6 Unsupervised scaling: Principal Component Analysis

A much simpler and more quickly estimated unsupervised scaling is to take the matrix of documents (speakers as columns and word frequencies as rows), and simply find its eigenvectors. 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 eigenvectors by more elaborate means.

3 Benchmark: Scaling the US Senate

Having shown that the supervised scaling methods theoretically ought to agree in a general way, it remains to test whether such text-based scalings can reflect existing political spectra. 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. The advantage of working with an entire legislature is that there are enough data not just to adjudicate between different methods, but to validate whether indeed any of these methods have statistical correlations with trusted scalings like DW-Nominate. Once that has been established, we can move on to the

more uncertain grounds of the UK House of Commons.

[Table 2 about here]

The data consist of the entire Congressional Record for the US Senate in 2006, separated by speaker. As discussed earlier, for the three supervised methods, the first step is to generate two reference vectors, one based on the entirety of Republican speech, the other based on the entirety of Democratic speech. The unsupervised methods work solely with the individual speech vectors. The correlations between these text-based scalings and the benchmark DW-Nominate (DW1) are presented in Table 2. All of the methods produce results that correlate significantly with the vote-based DW-Nominate score.[10] The Wordscores scaling is omitted because, as suspected, it correlates with the Bayesian method at over 0.95 here. Clearly the supervised approaches do better than the unsupervised methods, and of the supervised approaches, the vector method seems to do best, although all three correlate reasonably tightly 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, suggesting the latter has, somehow, simply picked up the second eigenvector through its procedure.[11] Whether this is the usual outcome of Wordfish, though, must remain for later study.

[Table 3 about here]

[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.

[11] 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.

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[12]. Obviously it will not always be possible to cull such language from the corpus, but as we will see again the the UK case, doing so makes the text-based scaling work considerably better.

While establishing correlations with DW-Nominate does show that these methods work in some basic sense, one important question is whether we are getting more out of these scalings than the party information that we put in. The unsupervised scalings, after all, do not know anything about party membership, whereas the supervised scalings work with that information: is the scaling of the latter any better than party by itself? One way to resolve this is to ask whether these scalings correlate with scores within parties. In that case, both the vector projection and Wordfish methods do sometimes show significant correlations, depending on the party and whether we use a cardinal or ordinal correlation test, but the Bayes method does not – perhaps reflecting its roots as a classifier rather than scaler.[13] The one method that does systematically correlate

[12]While the Bayes and Wordscore increases only slightly, interestingly.

[13]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.

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]

A deeper question is how well DW-Nominate itself captures within-party ideological differences. Users of this vote-based scaling often take it for granted that it distinguishes left-wing from right-wing Democrats, say, but this distinction is rather weak. 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, what percent of the time a Senator votes with his party), that variable explains 95% of the variance of DW1 (even with more stringent ordinal correlation).[14] Clearly there is not much room in DW-Nominate for more than party and party loyalty, at least with the limited data examined here.[15]

But we can take it one step further and compare these scalings directly with DW-Nominate's ultimate justification, the roll call voting record itself. The simplest test of this is to predict votes out of sample using a straightforward logit model. For each of 1000 runs, the observations

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. Similarly, although one could perhaps boost the match with DW-Nominate by using the full set of DW-Nominate scores in a SVM or Bayesian machine learning approach, this would need heavy out-of-sample validation to show you weren't just using DW-Nominate scores to predict DW-Nominate scores, and of course it is also useless for scaling legislatures with uninformative voting.

[14] If we similarly construct a variable that is the product of party and PCA1, we get an 0.83 (ordinal) correlation; if we do the same and the vector score, we get a 0.78 (ordinal) correlation. (Ordinal is the correct measure here, since cardinal correlations will vary depending on the choice of dummy values.)

[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.

(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 DV is the rollcall vote and the IV 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).

[Table 4 about here]

The mean out-of-sample prediction accuracy for each of our various scalings is presented in Table 4. As we can see, DW1 does do the best, but only barely; a simple PCA scaling of the rollcall matrix produces almost exactly the same accuracy (0.88), and the simple party-times-loyalty variable does nearly as well at 0.85. The supervised text-based scalings also do quite well (and better than party alone), at about 0.82, while the unsupervised approaches unsurprisingly do a bit worse. There are two important upshots here: First, DW1 does not seem to do much more than a basic measure of party loyalty or a simple eigenvector of the rollcall matrix. Its reputation for measuring ideology notwithstanding, very little of it seems to comprise any intra-party ideological difference apart from party loyalty (at least with these data). Second, the text based scalings do almost as well as DW1 on its home turf, predicting rollcall votes.[16] Since they also measure many other aspects of political behavior (as the words in Table 3 attest), they may in fact be a better measure of ideology than DW-Nominate, and the failure to correlate perfectly with that measure may be less weakness than strength, evading some of the the effects of party strategy even in the US case, and going beyond mere vote-prediction without even much cost in vote-prediction accuracy.

[16]Regarding the topic of dimensionality, is worth noting that none of these scalings correlate very well with the second, DW2 dimension at all. All correlations are less that 0.2, and since DW2 in fact correlates with DW1 at around the same level, none of these scalings seem to be capturing anything of the second voting-based dimension. This may be an especially low-dimension time period though, since the out-of-sample vote prediction is not increased by adding the second DW-Nominate dimension to the logistic model; nor does adding the second roll-call PCA dimension aid the first; nor the second text-based PCA dimension boost the accuracy of the first alone.

4 Uninformative voting: scaling the UK House of Commons

Although comparisons of text-based results with established ones like DW-Nominate shows that these scalings are nonrandom and similar to existing measures of ideology, clearly the vote-based methods will remain the standard for some time to come – when appropriate voting data are available. But in legislatures with strong party discipline, while the vote data may exist, party-line voting makes it almost impossible to place individual members on a spectrum (Norton 1975, Norton 1980, Cowley 2002, Spirling & McLean 2007). Indeed, the United States aside, vote-based scaling methods are largely ineffectual for most developed democracies, and the data are largely unavailable for many others. For that reason, text-based scaling may be essential for testing many of the myriad theories concerning the behavior of individual politicians.

However, apart from simply running the speech data through the same procedure as before, there are a number of issues raised by text-based scaling that need resolving. First, even with the reassurance of the US case, how do we know that a new scaling is measuring something like ideal points in a policy space? Second, even if we can be assured that it is measuring ideology, how can we be certain that that supposed ideological dimension isn't merely a refection of the disagreements of the hour, or the terminology of leadership and opposition, rather than reflecting long-term, deeply held views? And third, in a multi-party context, there is the related question of how much a single policy dimension is shared by everyone. In a legislature with three major parties, such as the UK, would the scaling generated by a Labour/Conservative dimension match a Liberal/Conservative one, and do these different scalings hold across time and power changes? I will address these questions in turn using the UK House of Commons (HOC) as the testbed.[17]

[17]The author would like to thank Ken Benoit for providing the Hansard House of Commons debate archive data.

The HOC is particularly suited because, although it has been long-scrutinized and the data are plentiful, the voting is largely uninformative, and thus the need for a complete scaling is strong.

[Figures 3 and 4 about here]

The first and fundamental question is whether scaling the HOC with the two major parties result in legislator scores that are meaningful measures of ideology. Although there is no benchmark comparison as before, the first thing we can do is see how well the text-based scaling distinguishes members of the left parties from the right. Figure 3 shows the results of this for the simpler case of the US Senate, where we see that the scores of Democrats and Republicans are strongly separated using the vote-based DW-Nominate, but are also well separated using the vector projection scores and (to a somewhat lesser degree) using the Wordscores and Bayesian scores. Figure 4 presents similar results for the members of the 1996 HOC as scaled by the two largest parties, Labour and Conservative, using the three supervised methods, with members grouped by party. Once again, it is clear that all three methods produce quite similar results, particularly Wordscores and Bayes.[18] More importantly, Figure 4 shows that these scalings distinguish members of the two major parties.

[Table 5 about here]

If this is really a measure of ideology, another test would be whether these scalings simply reflect the concerns of a single year, or whether the supposedly ideological reference texts can measure ideology across multiple years, including time periods not used to construct the reference texts themselves. The first two lines of Table 5 show that members who continued from 1996 (the

[18]Specifically, the Bayes and Wordscores correlate at the 0.99 level here, and the two correlate with the vector approach at 0.77.

last year of Conservative rule) to 1998 (the first full year of Labour rule) have scores that correlate only weakly, at 0.135.[19] However, if we compare members who are present in 1998 and 1999, the correlation is much higher, at 0.625. Thus it appears that these positions are comparable across time, but only when the same party is in power; when the parties have changed power, the entire terms of debate might shift so much that cross-temporal textual comparison is difficult.

Harking back to the US case, it is important to determine whether it is just the fact that one party controls most of the ministerial positions and thus uses most of the technical legislative language that allows us to distinguish the parties. The next two lines of Table 5 shows that if we remove the speech of the ministers from the reference texts, the correlation with the ministers-included method is fairly low. But lines 5 through 7 shows that removing the ministers boosts the correlation across the change in party control, suggesting that much of the non-correlation across this time period is due to the change in which side uses the technical language. As in the US case, the practical upshot is that ministers[20] and, where possible, ministerial language should be removed for a more consisent ideological scaling.

[Figure 5 about here]

As mentioned earlier, the UK case differs importantly for the US in the number of parties. How does the choice of Labour/Conservative as the reference texts differ from using other parties? Lines 8 and 9 of Table 5 show that using the Liberal and Conservative parties as reference correlates fairly closely with scalings obtained by using the Labour and Conservative references.

[19] All scalings hereafter are done with the vector projection method, simply because it produced the best correlation with DW-Nominate previously. However, none of these results are substantively changed with the Bayesian or Wordscores methods.

[20] This is especially the case when, as in the UK, members are often recorded separately both in their role as minister and in their role as individual MP, with the former identity containing all of the misleading technical language.

This suggests that the scalings that are not deeply dependent on the choice of reference pairs, and not merely picking up characteristics of only those two parties, or only of leadership-opposition. Lastly, Figure 5 suggests that using the Liberal and Conservative parties as reference texts may offer another improvement over the Labour/Conservative pairing. Not only does using the Liberal and Conservative pairs have the advantage of automatically lacking the ministerial speech (in 1998), but Figure 5 shows that, unlike using the two main parties, the Labour/Conservative pair orders the members of all three of the largest parties correctly (or at least, in keeping with our prior expectations).[21] One final validation of the superiority of using Liberal and Conservative as the scales is that we have a rather nice out-of-sample test here, in the positions of all the other parties. In Figure 5 the contest is essentially between the "98 Vector" scaling (Labour/Conservative) and the Liberal/Conservative scaling. The latter correctly orders the members of the three northern Irish parties, putting members of the Democratic Unionist Party (DUP) and (to a lesser degree) the Ulster Unionist Party (UUP) to the right of the Social Democratic and Labour Party (SDLP). It also correctly situates the Scottish National Party (SNP) towards the left[22] and plausibly puts the Plaid Cymru (PC) party towards the center. This is a purely out-of-sample test that provides a nice, if minor, validation of the method and the reference texts.

Finally, although we lack anything like the touchstone DW-Nominate scaling against which to compare these results, there are other more indirect approaches that might partially validate this approach. The members of the House of Commons have been exhaustively studied even without a fully revealing vote record, and their positions have been estimated in various ways both intuitive and systematic, perhaps most prominently by Norton (1975, 1980) and later Cowley

[21] This "correct" ordering also occurs when scaling 1996 by its Liberal and Conservative parties.

[22] This party differs the most between the Liberal and Labour reference texts, perhaps reflecting its continuing internal divisions arising in part of the party's origin in the amalgamation of the left-of-centre National Party of Scotland and the right-of-centre Scottish Party.

(2002, e.g.). In both cases, the relatively infrequent occurrences where party members vote against their leadership are carefully examined to determine ideological position or, alternately, distinctive voting blocks. Although the approaches may vary, the fundamental data are these rebellion rates, which by themselves constitute a measure of party loyalty and thus, perhaps, ideology. Of course, the problem is the same as before – one may revolt from the left or the right of one's party – but it is often assumed that rebellions by members of the ruling Labour party (say) are generally from the left. Thus Quinn & Spirling (2010), in motivating their dirichlet clustering algorithm, select five MPs – four Labour ranked by rebellion rate and a Conservative on the right – in order to demonstrate the failure of traditional vote-based rankings, which of course conflate the Conservative opposition "rebellion" with the left-wing rebellions. They show that various vote-based scalings – Optimal Classification, Nominate, or PCA, for instance – fail to get this order correct, generally grouping the more liberal rebels with the Conservative. Since such scalings fail, they argue, the cluster-based approach, though not exactly what was desired, is the best available tool for objective analysis.

Although the resultant clusters are successfully teased from the data and are a very useful tool in their own right, this motivating example somewhat undercuts things, since although naive vote-based scalings that pool Labour and Conservative members do fail by conflating left and right, we know this not by some esoteric expert judgment, but simply because (they assume) we can rank most Labour members from left to right depending on their rebellion level – a perfectly serviceable vote-based scaling, if perhaps limited to the ruling party. So given this reasonable (if nonideal) measure of ideology, how does the text-based scaling compare? To begin, using the vector scaling of the 1998 data as our basis, of the five MPs Spirling and Quinn employ as their test case,

four are ranked "correctly" by the text-based scaling, with the one incorrect placement being the Labour "loyalist" John Prescott, who unlike all the rest, spoke during this time period primarily in his capacity as "The Secretary of State for the Environment, Transport and the Regions." Most importantly, the other Labour loyalist is placed between the Conservative and the two rebellious Labour leftists.

But of course, five data points tell us nothing certain. The next step is to compare the scaling with the rebellion rates for all MPs.[23] In this case, the results are similar to the US case. The correlation between the text scaling and a party dummy (examining only Labour and Conservative members) is 0.55. The correlation between the text scaling and rebellion rate for Labour members is about 0.17 – weak, but significantly different from 0, much as it was for the intra-party correlation in the US case.[24] Finally, if we look at more methodologically sophisticated vote-based scalings, such as a basic cluster or PCA approach,[25] we find a correlation of 0.55 with the text scaling. However, a closer examination of those vote-based scalings finds that, just as with DW-Nominate, over 95% of their variance can be explained with a combination of a party dummy and the rebellion rate, so a match between those scalings and the text-based one provides not much additional validation. All of this simply goes to show that the need for non-vote-based scaling is strong (even if the scaling is based on an expert interpretation of rebellion rates), and that the text-based approach may provide at least as good a window into ideology as rebellion-based scalings or clusters.

There are a number of important conclusions to draw from this analysis of the UK House of Commons. First, scalings using opposing parties as reference texts do produce plausible separation

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

[24] The correlation is essentially zero for Conservative or other parties, but it increases to 0.22 for other parties (primarily Liberals) if you use the liberal-conservative aggregates as reference texts.

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

between the positions of members of opposing parties, based purely on the content of their speech. Second, these scalings are robust across time, but less so across power transitions. Third, the scaling are more robust across power transitions when the technical language of ministers has been excluded from reference texts. Fourth, using more extreme pair pairs as reference texts, such as Liberal and Conservative in this case, appears to have the significant advantage of ordering members of all the parties correctly, and may potentially be ordering members within parties more correctly as well. And fifth, the text-based scaling not only orders parties correctly, but provides a small but distinct correlation with intra-party rebellion-based scalings. Whether divergences between this scaling and our previous expectations are deficiencies or simply alternatives bears much further thought.

5 Discussion

Political settings that lack informative voting data vastly outnumber those with informative voting. And as we have seen, even the dominant vote-based scaling technique for the US Senate tells us little more than a member's party membership and loyalty. So there remains an enormous need for political scalings are are not dependent on strategy-constrained voting information. The problem, of course, is validating a proposed scaling, and demonstrating that it is not just an arbitrary set of number assignments. This paper has approached that problem from a number of directions: theoretical; a comparison with a "known" domain; and an exploration of a new domain much in need of scaling. From this there follow a number of clear guidelines for researchers seeking to scale large numbers of speakers, whether in legislatures, other political arenas, or even beyond the traditional left-right political spectrum altogether.

On the theoretical side, it was shown that although they apparently derive from quite different models, the vector projection, Bayesian, and Wordscores approaches in practice will deliver similar results. In particular, the Bayesian scaling is in many ways the "corrected" version of Wordscores' approximately Bayesian approach, though in theory, simulations, and practice, they tend to produce similar outcomes. But since theory and simulations show that in some cases their results can diverge, the more theoretically secure Bayesian approach is preferable. In any case, the vector method produced the best correlation with the benchmark DW-Nominate scores, so until further tests are made, it appears to be the most informative scaling.

Although the unsupervised Wordfish and principal component approaches are appealingly free of any expert supervision, there are reasons to be wary of both. The correlation between PCA or Wordfish and the benchmark DW-Nominate scores was lower than for the supervised approaches. More importantly, what you get is what you get with these methods; there can be no tweaking of the reference texts, either to strengthen the signal-to-noise ratio for a given dimension of interest, or to select political spectra other than the most dominant one. Whereas with the supervised approaches, different reference texts allow one to score speakers along any dimension – such as economic, social, or even politeness – to do the same with the unsupervised methods requires in fact much more supervision, making dozens or hundreds of selections to delimit a specific vocabulary to scale. If one does not have party or other information to work with, the unsupervised approach may be the necessary one; in that case, the quick PCA seems to offer much the same results as the more theoretically elaborate IRT-based Wordfish method – although there are questions about dimensionality that remain unresolved.

In terms of actually scoring existing legislatures, the Bayesian, vector projection, and Word-

scores approaches all seem to work fairly well in matching an existing scoring like DW-Nominate, when the aggregate party texts are used as references to scale the speakers. The match improves by 10-20% when technical terminology is removed, indicating that this might be a good general practice when the context and language are sufficiently well understood. Indeed, since 95% of the variance in DW-Nominate scores are explained by a combination of party ID and party loyalty, the text-based scoring may in fact be doing a more thorough job in capturing true ideological positions than the party-heavy DW-Nominate scores. And since they do almost as good a job on DW-Nominate raison d'etre, predicting rollcall voting, a case can be made that text scalings may be superior.

Finally, while it might have appeared that applying these methods to a setting without any vote-based scalings for comparison would prove hopelessly untestable, that is not the case. Using the aggregate Labour and Conservative texts as references, scaling the House of Commons produces a clear distinction between the members of the two parties. Nor is this merely a case of using existing party data to "predict" that same data: the reference texts from one year can successfully scale another, although the success of that scaling (as measured against the within-year scaling) declines if the two years span a party transition. But even in that case, when the terms of debate have shifted so significantly, scalings across years work fairly well as long as ministerial speech is excluded, and it appears to work even better if one uses a pair of parties both out of power and relatively extreme, such as the Liberal and Conservative parties. In that case, the separation betweenst the party members is quite strong, and the ordering off all parties – even "out-of-sample" ones such as the regional ones – matches existing expectations.

This does not entirely resolve the question of scaling multi-party legislatures. Going beyond

the two-reference approach, one could assign values to each reference party and take the expectation value for each speaker; doing this is straightforward for Wordscores and would only require a little more work for the Bayesian and vector approaches. The alternative is simply to accept that there may be no single "real" political dimension, and that the Labour-Conservative dimension may in fact be different from the Liberal-Conserivative dimension. In that case, one must use one's theory to choose what type of scale is relevant, and which texts (aggregate or otherwise) are best suited to picking out that dimension.

But although the choice of reference parties is a tricky decision in multi-party situations, the payoff may be quite large. Not only can speech and party information be used to scale legislators, it can be straightforwardly expanded to scale any manner of political actors as long as they have produced documented speech, even in languages unknown to the researcher. While a variety of new methods have expanded scaling beyond legislatures with informative voting, the methods developed here go far beyond existing ones, encompassing virtually any subject within the wide ambit of a search engine.[26] As we have seen, the legislative voting record is dominated by party strategy even in the low-discipline US setting; similarly, speech too is a highly strategic behavior, both within legislatures and outside them. Thus 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 spatial theories, and to develop new theories explicitly designed to model the cacophony of talk surrounding every political action.

[26]For instance, in recent exploratory work, reference texts were constructed out of the top 10 results for google searches for "democrat" and "republican." These reference texts were then used to scale each senator, using the top 10 results of a google search for their name. The resultant scaling correlates with DW1 at 0.67, and correctly classifies the party membership of 91% of the Senators – strong results for a preliminary test.

A Appendix

A.1 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:[27]

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

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 $p(w_i|R)$.[28] 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:

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

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

[28] 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.

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:

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

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:

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

□

A.2 Wordscores

Instead of beginning with $p(w_i|R)$, the independent probability of encountering word i given text R, LBG begin with $P_{iR} \equiv p(R|w_i)$,[29] 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 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:

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

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

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;[30] 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 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

[30]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.

overall score B_V. We then have similar formulations:[31]

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

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:[32]

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

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.[33] 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

[31] Note that here i indexes only unique words.

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

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

[33] 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.

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 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.[34] 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.

[34]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

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.

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

Electoral-Vote.com. 2011.

URL: *http://www.electoral-vote.com/*

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

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. & N. Schofield. 1998. *Multiparty Government: The Politics of Coalition in Europe*. University of Michigan Press.

Lightfoot, Chris. 2011.

URL: *http://ex-parrot.com/*

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

McCarty, N., K.T. Poole & H. Rosenthal. 2005. "Polarized America: The dance of ideology and unequal riches.".

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.

Poole, K.T. & H. Rosenthal. 2000. *Congress: A political-economic history of roll call voting*. Oxford University Press, USA.

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: Out-of-sample rollcall vote prediction accuracy, 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 PCA1	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 DV and the scaling as IV, 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 & 1996 w/o ministers	0.164
1998 & 1998 w/o ministers	0.496
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 (including ministers) 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 $p < 0.05$.

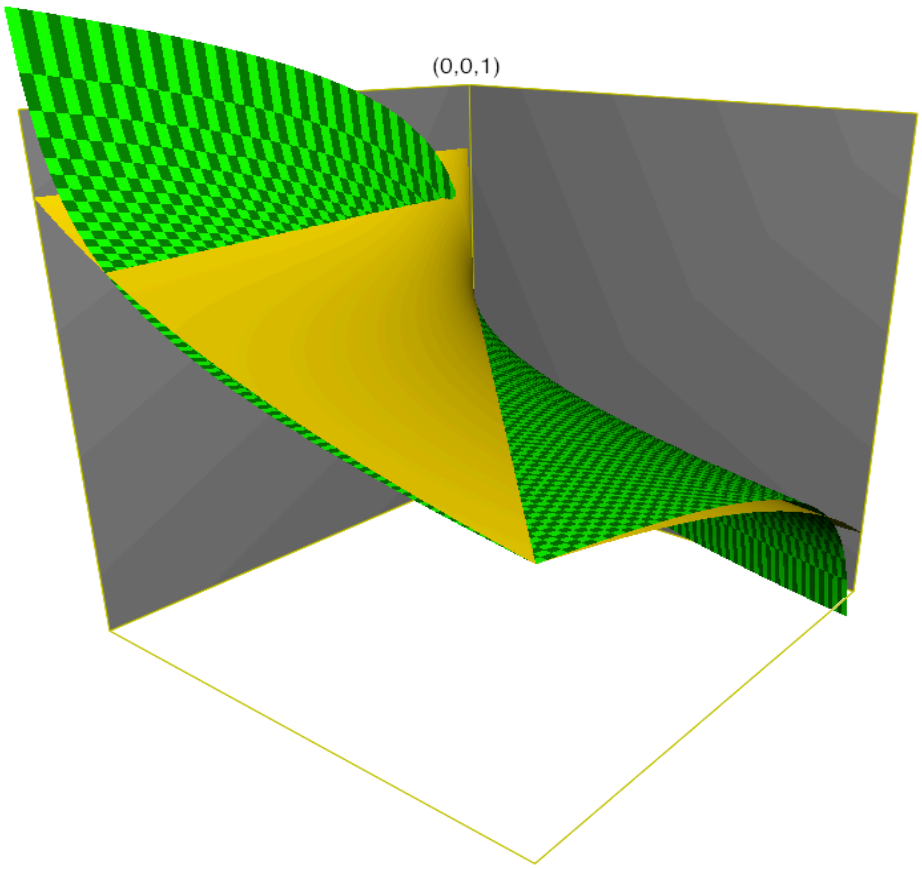

Figure 1: A comparison of the word weighting assigned by Wordscores (smooth) and the Bayesian method (checkered). The x and y axes correspond to the frequencies of some word i in the two reference documents (ie, F_{iR} and F_{iD} from equation 17), and the z axis corresponds to S_i or B_i. (See A.2.) 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.

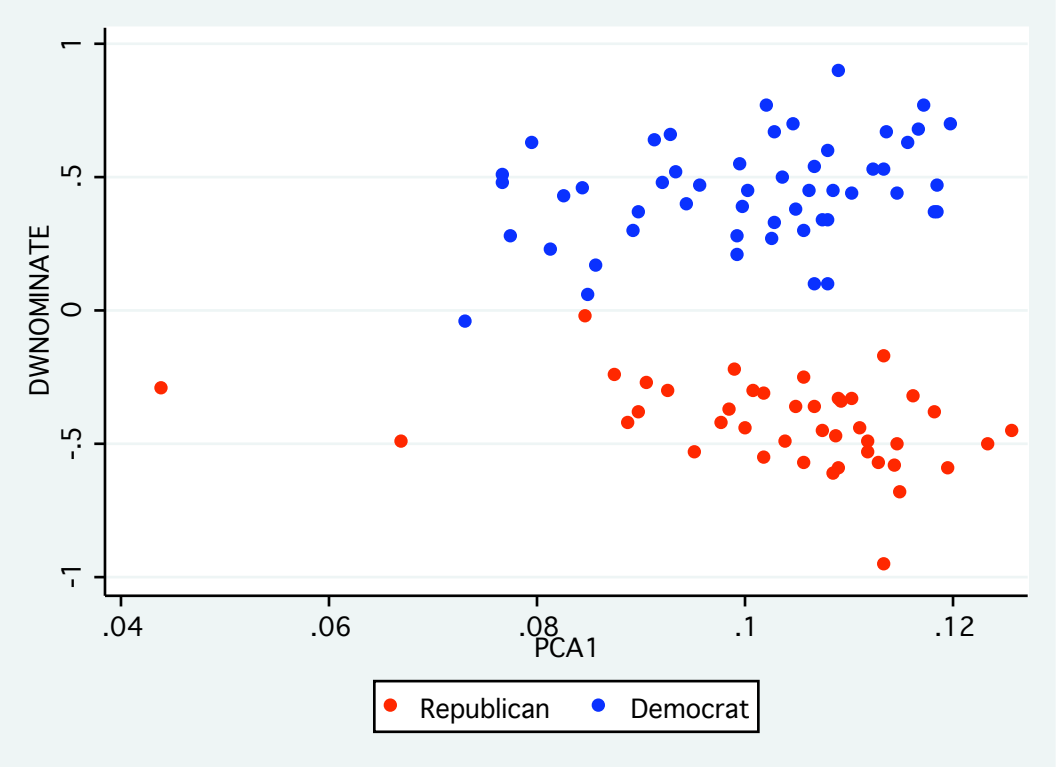

Figure 2: Plot of the first principal component against the first DW-Nominate dimension. Note that although there is no overall correlation between the DW1 score and the PCA1 score, within each party the scores are quire correlated.

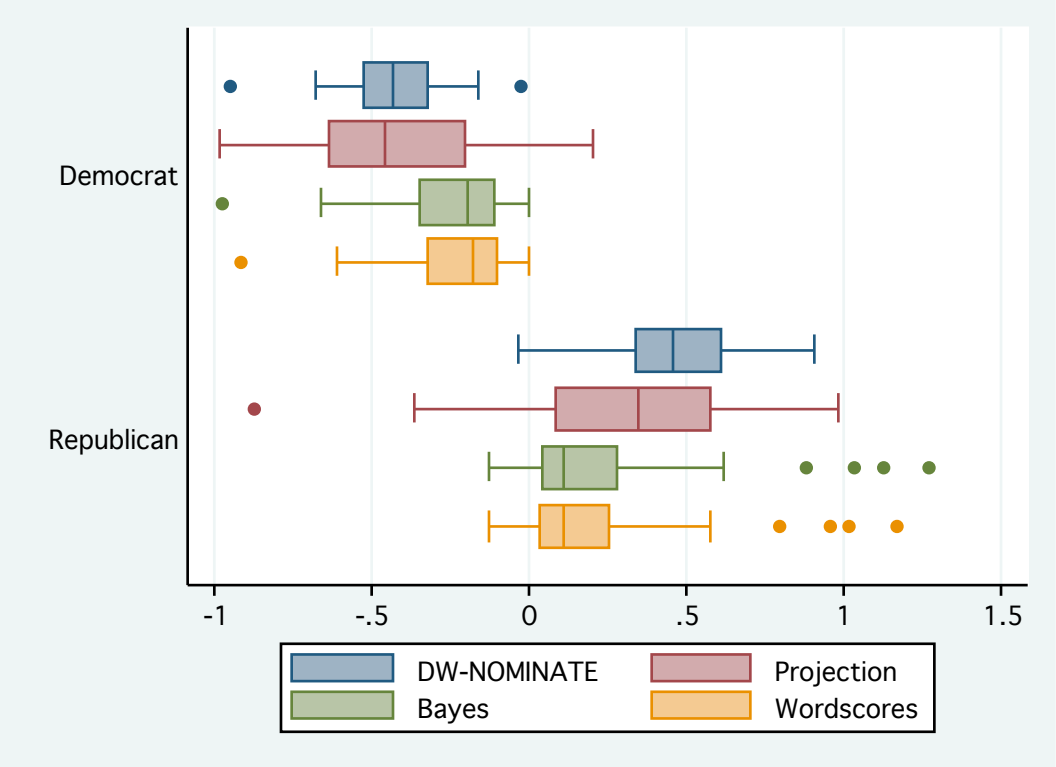

Figure 3: Box plots of the four scaling methods, grouped by Democratic and Republican parties. Axis scale is from DW-Nominate; the other scores have been rescaled to match.

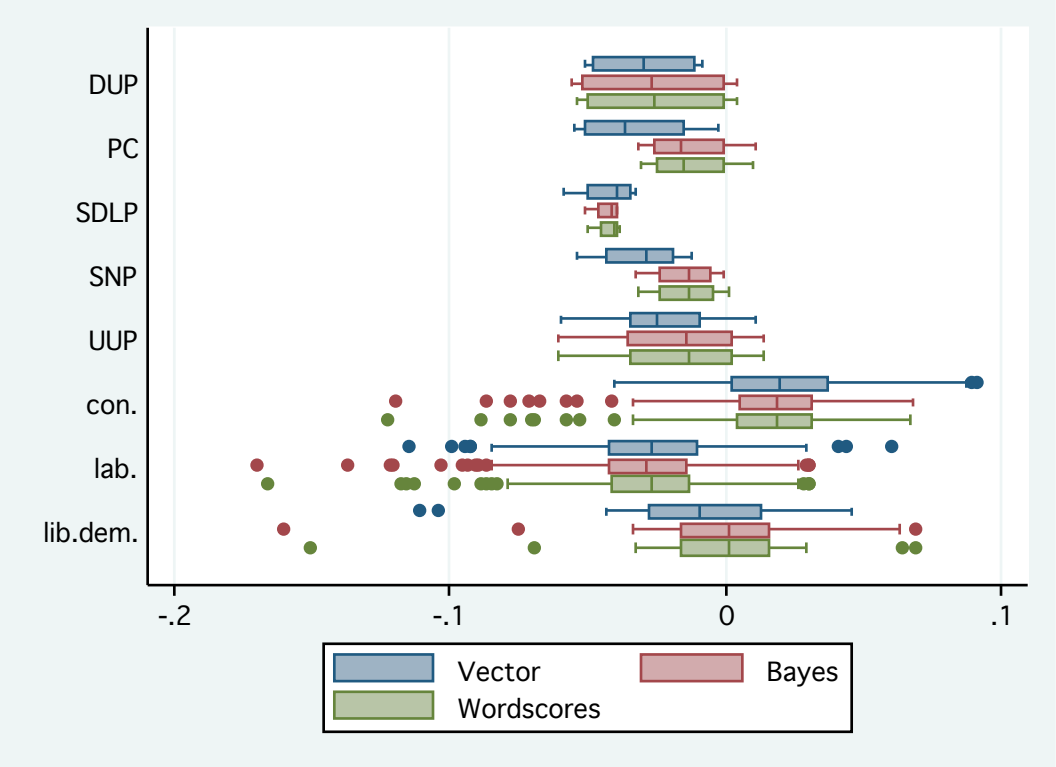

Figure 4: As in Figure 3, the main parties in the House of Commons as scaled by the three main techniques. Axis scale is from Bayes; the others have been rescaled to match.

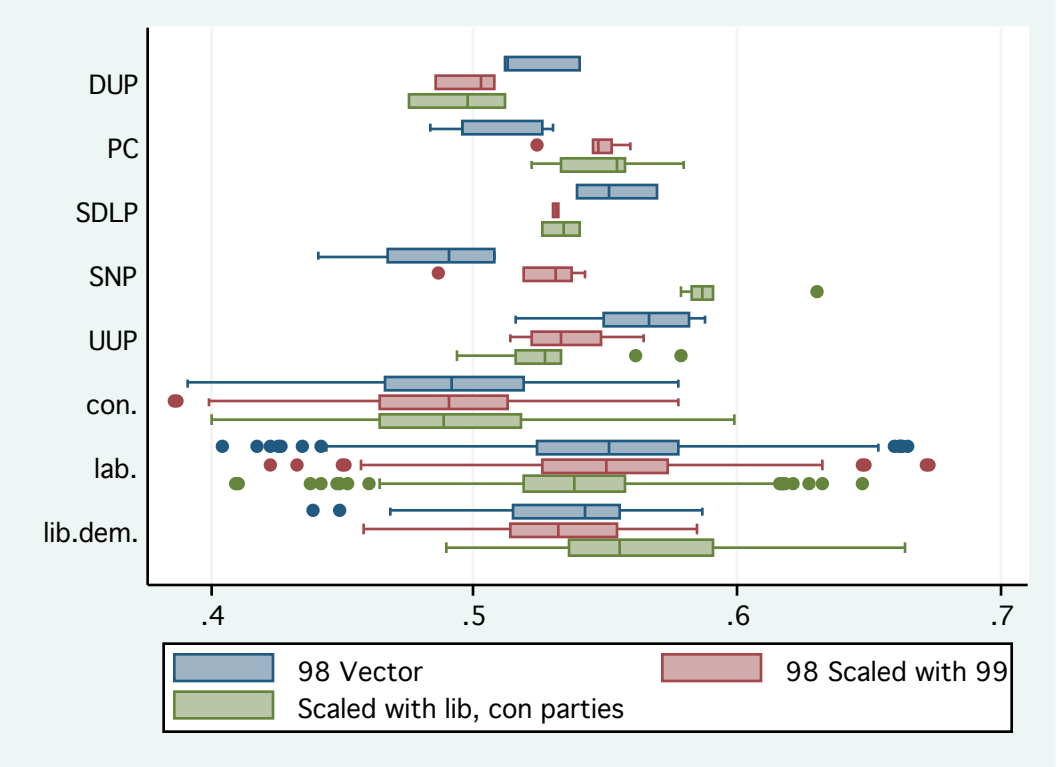

Figure 5: The 1998 House of Commons, scaled by various reference text pairs. "1998 Vector" is simply the standard Labour/Conservative scaling. "98 Scaled with 99" is scaling the 1998 HOC with reference texts based on the aggregated party speech from 1999. "Scaled with lib, con parties" is 1998 as scaled by those parties as reference texts.