Text-Based Scaling of Legislatures: A Comparison of Methods with Applications to the US Senate and UK House of Commons

Nick Beauchamp*
NYU Department of Politics

[DRAFT: Feb 24, 2011]

Abstract

Many models call for actors to occupy positions on a political spectrum, but although procedures for scaling legislatures with well-documented, poorly-disciplined voting are well established, scaling individuals in all other situations remains a challenge. A new approach is to use automated techniques to scale legislators based not on their votes, but on written and spoken text. This paper compares five text-based scaling methods, some familiar from the political science literature, some not: three supervised approaches that use reference texts to score speakers, and two unsupervised approaches that scale without user input. Their theoretical foundations are examined and despite their apparent dissimilarities, the three supervised methods are shown both analytically and via simulation to produce relatively similar results. The methods are then tested against a well-established scaling, DW-Nominate scores for the 2006 US Senate. Using the aggregate speech of all members of either party as two reference texts, the supervised approaches produce scores that closely correlate with DW-Nominate scores and with each other, while the unsupervised approaches do less well. When applied to a domain that cannot be scaled by votes alone, the UK House of Commons, the supervised approaches using aggregate party texts as references produce scalings that strongly separate members of the various parties. Furthermore, these scalings do endure across years, unless there has been a change in party control. In that case, removing party ministers and their abundant technical speech, or using more extreme or out-of-power parties, appears to alleviate the problem. Future users of automated text-based scaling may take from this a series of practical guidelines: that the supervised approaches seem to work better than the unsupervised; that of the former, the Bayesian approach is more theoretically secure than the popular "Wordscores," and the vector-projection method possibly more effective than either, but all three work similarly; and that when scaling legislatures, it works best to exclude leadership terminology or members, or to use two relatively extreme out-of-power parties. Having successfully replicated existing scales, we can also move on with greater confidence to the myriad other political dimensions accessible with properly chosen reference texts.

*Author email: nick.beauchamp@nyu.edu

1 The need for algorithmic text-based scaling

1.1 Introduction

Many models of political action begin with the assumption that the players have ideal points positioned in some shared policy space. Although vote-based methods are the standard for scaling legislators, situations in which vote information is both plentiful and informative are relatively rare. Even in the US context, where poor party discipline allows for voting that is informative about individual members (Poole & Rosenthal 1985, 1991, 2000*a*; Poole 2005), party realignment has recently led to the near-disappearance of the second dimension from DW-Nominate's vote-based scaling (Poole & Rosenthal 2000*b*, McCarty, Poole & Rosenthal 2005). Yet no one believes that these regional or racial issues have in fact vanished; they have just become more invisible in the voting data. In European countries such as the UK, tight party discipline is the rule rather than the exception, and vote-based scaling of individual legislators is difficult or impossible without extensive expert judgment (Norton 1975, Norton 1980, Cowley 2002, Spirling & McLean 2007). And in the legislatures of developing countries or US states, or in their committees, or in myriad other political settings, voting records are often sparse, nonexistant, or difficult to obtain. But what many such institutions often do record is speech. Though talk may be cheap – a cliché which remains to be tested – at least it is plentiful, and a method that can extract political positions from speech records can be of great use to all those scale-dependent theories of legislative behavior.

Given that there may be as many possible approaches to using speech to score political actors as there are speech acts, how should a researcher wishing to turn a set of texts into a scaling proceed? In this paper I examine a number of approaches both theoretically and practically, comparing mathematical formulations, simulated results, and real-world outcomes. These approaches can be grouped into two general classes, with particular focus here on the first: "supervised," where the researcher provides a reference set of pre-scaled documents and uses those to scale the rest;[1] and "unsupervised," where the collection of documents is scaled without any additional input or variation.[2] On the theoretical level, we will see that there are various reasons to prefer some methods over others, but despite their varied approaches, the mathematics and simulations suggest that the supervised scalings will be broadly similar under many circumstances. I then proceed to test these scalings in a familiar setting, the US Senate. Using the aggregated speech of all Democrats and all Republicans as the two reference texts for the supervised systems, the 2006 Senate is scaled and the results compared against the benchmark DW-Nominate scores. Many of the text-based techniques match the DW-Nominate scores closely, and in particular the supervised approaches work both well and similarly to each other. Finally, I apply the text-based approach to a domain much more in need of scaling, the UK House of Commons. The results clearly separate Labour members from Conservatives, as we would expect for any plausible scaling. Furthermore, party-aggregated reference texts appear to work well across years, and even across power transitions once the technical terminology of the leadership is excluded; and using a pair of relatively extreme, out-of-power parties (such as Liberal and Conservative in 1998) to scale legislators seems to produce

[1] Such as the well-known "Wordscores" (Laver, Benoit & Garry 2003).

[2] Such as the IRT-based approaches of Monroe and Maeda (2005) and the similar "Wordfish" of Slapin and Proksch (2007).

a scaling that best matches our expectations and works across years and power transitions.

1.2 Background and theory

Text-based scaling is of course nothing new, and already ranges from the most manual, expert-based evaluation, to highly automated, fully algorithmic approaches. Historically, techniques have generally progressed from the former to the latter. At one end we find expert surveys, where judgments about party or manifesto positions are made by experts in the field (Janda, Harmel, Edens & Goff 1995, Laver & Schofield 1998). A slightly more systematic and arguably more objective approach would be something like the long-running Comparative Manifesto Project (Budge, Robertson & Hearl 1987), where trained non-experts look at sentences or clauses within a manifesto, and judge them according to their valence and saliency regarding dozens of possible topics. Near the most algorithmic end of the spectrum, one finds Laver and Garry's dictionary-based approach (2000), where dictionaries of key terms are semi-automatically created from key texts (identified by experts), and are then used to rate the similarity to those key texts of new texts under examination. Such dictionaries may be pruned with more or less expert human attention, but in all cases, the advantage of the computerized rating lies in its speed and reliability relative to the laborious techniques characteristic of the CMP. At the far end of the text-analysis spectrum lies the almost fully automated 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. But even this last is still "supervised learning," inasmuch as it relies on an expert to provide the reference texts and scores. Most recently, an unsupervised approach has been developed by Monroe and Maeda (2005) and Slapin and Proksch (2007), which models documents and words by arranging them in a shared space such that, as in item-response (IRT) models, the frequency of words in a document best matches the distance between each word and the document. Here, apart from a few minor decisions, the algorithm is entirely automatic – which is both a strength and a weakness, inasmuch as whatever scaling you get is the only one possible, unless you return to expert supervision and specify a conceptual/linguistic domain *a priori*.

Outside of political science, highly-automated text analysis has been an active area of research for many years now. It is beyond the scope of this paper to summarize that work, but even standard textbooks on machine learning now contain numerous examples based on document classification (Bishop 2006). Because of the context in which these techniques developed, most are geared around classification rather than scaling, but the techniques are usually naturally extended to continuous values. For instance, the so-called "Naive Bayesian" approach – the basis of most spam-filters, for instance – establishes a likelihood that any given document belongs to a certain class, along with a threshold that determines whether the document is classified as class A or B. But as discussed in more detail below, the threshold can be easily omitted, allowing us to directly use the likelihood as a scale. Similarly, the vector-projection approach projects each document vector onto some line, along with a cutpoint for classification; but again, the cut-point is unnecessary. Both approaches are supervised, insofar as they rely on reference texts for their classification or scaling. At the fully unsupervised end we find techniques like principal component analysis (PCA) or IRT scaling, where classification applications are often augmented with cut-points or thresholds, but those additions

are not necessary.

1.3 Plan of the paper

Of these various approaches, in the second section of this paper I compare five of the most automated: 1) vector-based scaling, 2) "naive" Bayesian scaling, 3) Wordscores, 4) principal component analysis, and 5) an IRT-based model like that of Monroe and Maeda or Slapin and Proksch. I first lay out the theoretical bases of each, paying particular attention to the first three, since the latter two are already well documented in the political science literature. As we will see, the Bayesian and Wordscores approaches have much in common, and though the former is theoretically preferable to the latter, simulations suggest that their results will often be similar. The vector approach, despite appearing fundamentally quite different, also produces results that are broadly similar to those derived from Bayes/Wordscores. The unsupervised IRT and PCA approaches match each other quite closely but match the real-world scalings and the supervised scalings less well.

In the third section of this paper, I apply these scalings to a well-scaled context, the 2006 US Senate, and compare their results to the current gold standard of legislative scaling, DW-Nominate.[3] Of course, using such an existing scoring is a debatable approach, since a text-based scaling may well be measuring something different from, but just as important as, the extant vote-based scalings. However, much of the issue here will be resolving signal from noise; any purported scaling will assign numbers to each legislator, but only a tiny subset of such scalings will have any meaning whatsoever.[4] A researcher hoping to employ these techniques to a vote-poor (or highly disciplined) domain must be confident that these numbers are not random, and insofar as we continue to assume that the votes are the most essential act of the legislator, we want something that can at least capture that traditional dimension – even if later we might branch out into other dimensions. This is, in fact, one of theoretical advantages of the semi-supervised approaches over the unsupervised approaches like PCA or the Monroe-Maeda IRT: in the latter cases, one can only explore alternative dimensions by carefully culling a list of words to constrain the results to some specific domain. In the first three approaches, on the other hand, one simply specifies a small number of reference texts – as few as two – and simply uses those to scale the rest. And this is how the legislators are scaled: since DW-Nominate is almost entirely determined by party identity and loyalty (as discussed below), the natural reference texts are those based on the two parties: two documents, one comprising the entirety of the Democratic speech, one comprising the entirety of Republican speech. These two simple documents provide scalings that match the DW-Nominate one quite well; and should a researcher wish to explore some other dimension, she need only compose a different set of scaling documents.

The final challenge, of course, is applying this to a truly useful domain, where vote data does not

[3] Although DW-Nominate incorporates changes in position over time, whereas none of the approaches examined here do, if one believes that incorporating those changes results in more truthful estimates of positions, then that, rather than the a-temporal Nominate score, is the best metric to use as a measure of a scaling's "accuracy."

[4] Indeed, although Wordscores, for instance, has been applied to numerous party manifestos, there are generally so few parties that the resultant scaling can only be heuristically compared against existing expectations. Applying such a scaling to a large number of speakers is really the only way to assure that the scaling produces meaningful results at all, rather than simply a small set of numbers that may be optimistically interpreted by the researcher.

allow DW-Nominate-syle scaling. In the fourth part of this paper I apply the three semi-supervised techniques to the UK House of Commons (HOC), whose members cannot be scaled by traditional methods due to tight party discipline. The first and fundamental result is that all three techniques do appear to work, insofar as they produce (quite similar) scalings that clearly separate the members of the two major parties. But there are a number of other important questions this raises. Are such scalings specific to the moment in which they are conducted, or do they reflect longer-term political spectra? Does using parties as reference texts owe its success not to political difference, but to the difference between leadership and opposition, say? And in the multi-party context, how dependent is the scaling on exactly which parties are chosen as reference texts? As we will see, the scalings seem robust across time, but less so across significant changes in the political landscape, such as a change of power in the legislature. However, reference texts seem a provide a more universal scaling when ministers (and their technical jargon) are excluded, and work best when a pair of relatively extreme parties who are both out of power (such as Liberal and Conservative in 1998) are used. Finally, the resultant scaling is briefly compared to a few existing scalings of HOC members, which though imperfect, does provided evidence that the text-based scaling matches the vote-based attempts almost as well as it does in the US/DW-Nominate case.

The theoretical analysis and empirical results developed here will hopefully provide a useful set of guidelines for future researchers wishing to scale well-disciplined or vote-poor legislatures, or more generally, to scale any set of political actors where there exists speech and party (or other categorical) data. Partially supervised algorithms such as the projection, Bayesian, or Wordscores methods all seem to provide robust scalings that array political actors along plausible political dimensions, and offer the flexibility to provide scalings along many other conceivable dimensions beyond the simple vote-based ones. An enormous world of political modeling is only now opening up as more and more political text becomes readily available, and as these methods are honed and validated.

2 Scaling Methods

The five methods investigated here are vector projection, "naive" Bayesian, Wordscores, principal component analysis, and an IRT-like scaling, based on that in Monroe and Maeda and Slapin and Proksch. Because such approaches are still uncommon in political science, I will first explore the methodology of each before turning to the application of these approaches.

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.[5] 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,[6] each senator then has an associated 1000-dimensional

[5] 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.

[6] 1000 was chosen simply because it was the most that were computationally feasible, and previous work suggests

vector,[7] 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.[8] In addition to this 1000x100 matrix,[9] 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 best as possible the traditional political spectrum captured by vote-based, party-dominated methods such as DW-Nominate.

2.1 Vector Projection

In this method, 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 distances $||R - D|| = A$, $||S - R|| = B$ and $||S - D|| = C$ yields:[10]

$$Score_{proj} = (A - B + C)/2A \tag{1}$$

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

2.2 Naive Bayes

This method is somewhat similar to the vector projection, inasmuch as it takes two reference points and evaluates each text vector relative to them. However, as the name suggests, the approach is Bayesian rather than spatial: the purpose here is usually to estimate the likelihood that a document belongs to class R or class D, given that we are presented with a document S of unknown class.

---

that, all else being equal, these methods work better with more words. See the next note for a caveat to that, however. 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.

[7] 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.

[8] 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, and work any Bayesian weighting at the parameter estimation stage, if at all.

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

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

We wish to discover $p(R|S)$, ie, the probability that a speaker is Republican (say) given their speech document S.[11] From Bayes, we know that:

$$p(R|S) = \frac{p(S|R)p(R)}{p(S)} \tag{2}$$

Now, if the probability of word i given that the speaker is a Republican is $p(w_i|R)$, then the "naive" part is simply to 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 $p(w_i|R)$.[12] Thus we would say:

$$p(S|R) = \Pi_i p(w_i|R) \tag{3}$$

This is undoubtedly false (since words are correlated with each other; this is the "naive" part), but it seems to work fairly well in practice, as we will see. Combining these two, we get:

$$p(R|S) = \frac{p(R)}{p(S)}\Pi_i p(w_i|R) \tag{4}$$

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

$$p(D|S) = \frac{p(D)}{p(S)}\Pi_i p(w_i|D) \tag{5}$$

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:

$$\frac{p(R|S)}{p(D|S)} = \frac{p(R)}{p(D)}\Pi_i \frac{p(w_i|R)}{p(w_i|D)} \tag{6}$$

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:

$$log\frac{p(R|S)}{p(D|S)} = log\frac{p(R)}{p(D)} + \Sigma_i log\frac{p(w_i|R)}{p(w_i|D)} \tag{7}$$

Here I diverge from the standard approach, which as was said before, is generally geared towards classification instead of scaling. Since we are not interested classification, just scoring, I drop the

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

[12] 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.

middle quantity (it's a constant for all Senators, after all), and merely calculate the latter quantity for each Senator. That is:

$$Score_{Bayes} = \Sigma_i log \frac{p(w_i|R)}{p(w_i|D)} \tag{8}$$

Finally, for simplicity, I take $p(w_i|R)$ to be simply the percentage of word w_i in document R. This too is undoubtably an over-simplification, since $p(w_i|R)$ should at the very least include some 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. Indeed, in the IRT section below we will see something more along these lines. But as we will also see, this simplistic approach works quite well on its own, is computationally efficient, and allows easy comparison with the "Wordscores" method described next.

2.3 Wordscores

"Wordscores" was developed by Laver, Benoit and Garry (2003) 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*. The method appears similar to the naive Bayesian approach, and although the functional form is fundamentally different (as will be discussed shortly), the resultant scores will under many circumstances be quite similar.

Instead of beginning with $p(w_i|R)$, the independent probability of encountering word i given text R, they begin with $P_{iR} \equiv p(R|w_i)$,[13] 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:

$$p(R|w_i) = \frac{p(w_i|R)p(R)}{p(w_i)} = \frac{p(w_i|R)p(R)}{p(w_i|R)p(R) + p(w_i|D)p(D)} \tag{9}$$

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

$$p(w_i|R) = \frac{W_{iR}}{W_R} \text{ and } P(R) = \frac{W_R}{W_R + W_D} \tag{10}$$

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:

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

$$p(R|w_i) = \frac{W_{iR}}{W_{iR} + W_{iD}} \tag{11}$$

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

$$P_{iR} = \frac{\frac{W_{iR}}{W_R}}{\frac{W_{iR}}{W_R} + \frac{W_{iD}}{W_D}} \tag{12}$$

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;[14] 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):

$$S_i = A_R \cdot P_{iR} + A_D \cdot P_{iD} \tag{13}$$

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

$$S_V = \sum_i \frac{W_{iV}}{W_V} \cdot S_i \tag{14}$$

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

The basic idea here is fairly Bayesian. Each word is assigned a score that is essentially a weighted mean of the two *a priori* scalar values for the reference texts; the weight is simply determined by the relative frequency of that word in document R versus D. The overall score for a virgin document V is then simply the weighted sum of each of these word scores, where each individual word score is weighted by the relative frequency of that word in document V. The closeness of the overall score to one *a priori* polar value or the other is akin to the probability that a document V belongs to class R or D.

We can, however, already see that there is a significant divergence between this and the true Bayesian score: as the authors point out, if reference text R contains a word and the other does

[14]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 trivial.

not, that makes $P_{iR} = 1$ – ie, the probability that you have document R given word i is 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.[15]

We can characterize a bit more precisely the difference in the two scores. 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:[16]

$$S_V = \sum_i \frac{W_{iV}}{W_V} \cdot S_i \text{ and } B_V = \sum_i W_{iV} \cdot B_i \tag{15}$$

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:[17]

$$S_i = \frac{F_{iD} - F_{iR}}{F_{iD} + F_{iR}} \text{ and } B_i = \log\left(\frac{F_{iD}}{F_{iR}}\right) \tag{17}$$

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 Naive Bayesian method, we generally weight B_i by $\frac{W_{iV}}{W_V}$ rather than W_{iV}, which of course produces a result much

[15]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 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.

[16]Note that here, i indexes each different word i, whereas in the previous discussion of the Naive Bayesian approach, i denotes every single word, with a new index even for repeated words.

[17]By comparison, although the Projection method score correlates fairly highly with the other two, the formulation using matching notation is quite different:

$$P_V = \frac{\sqrt{\sum_i (F_{iD} - F_{iR})^2)} - \sqrt{\sum_i (F_{iV} - F_{iD})^2)} + \sqrt{\sum_i (F_{iV} - F_{iR})^2)}}{2\sqrt{\sum_i (F_{iD} - F_{iR})^2)}} \tag{16}$$

more similar to that of Wordscores.[18] 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 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.[19] 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, as we will see shortly.

Regardless of the similarity in result, however, there are good reasons to prefer the theoretically more well-grounded Bayesian approach. Current users of Wordscores may continue the practice without too much worry (particularly in light of the similar empirical results developed below), but at some point it would probably be better to build our political scalings on more secure theoretical foundations.

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. The results can be seen in Table 1. Two quantities of words (1000 and 2000) and two families of distribution functions for those words were examined. Approaches such as Monroe and Maeda implicitly assume an exponential decline in word frequency in a text, whereas much current research suggests that the frequency of words in texts has a much fatter tail, instead following a "power law" of the form x^{α}. To be thorough, I explore a variety of distributions to determine how closely we might expect these various scalings to correlate.

The result is that although the Bayesian and Wordscores methods are generally more tightly correlated than the projection method is with either, that correlation weakens as either the number of words increases, or the mass of the tail increases. The latter can be achieved in two ways: either by increasing the relative mass of the tail for a given distribution, or by switching to a fatter-tailed distribution. But with either a fatter or longer tail, or simply more words, the the divergence between the two base functions, as can be seen at the edges in Figure 1, plays an increased role (ie, low-frequency words have more of an effect) and the two functions consequently diverge more, as does the projection method from either. As we will see later, for our 1000-word vectors, as

[18] 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.

[19] 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.

in the simulation, the correlation between Bayes and Wordscores is quite close, but as we see in Table 1, that need not be so if different situations have different types of word distributions, or if we used longer vectors (1000 was the computational limit for the present study). 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 more words or heavier tails, the scorings begin to diverge significantly. Since there is no way to know *a priori* whether the documents we are dealing with are in the upper half of Table 1 or the lower, the safer course would be to go with the Bayesian or projection methods, although Wordscores will in many cases be adequate.

2.5 Principal Component Analysis

Although the primary focus here is on supervised scaling techniques, it is worth including a couple of unsupervised scaling methods for comparison. The first, principal component analysis, is well established in the field and requires no further technical explication. Each Senator's speech is taken as a vector as usual, and principal component analysis is evaluated for that matrix of vectors to find the subspace of dimensions that best account for the variance of the vectors.

2.6 Wordfish

"Wordfish" is another unsupervised scaling method, developed by Slapin and Prosksch (2007; hereafter, SP) based on the work of Monroe and Maeda (2005; hereafter, MM). Unlike PCA, which is entirely a-theoretic when applied to text scaling, the methods of SP and MM are based on a (somewhat) more explicit model of the text generation process. MM's original model is based on "item response theory" (IRT), which was originally developed by psychologists, taking a set of respondents and their answers to various questions, and seeking 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).[20]

Thus the model is:

[20]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.

$$\begin{aligned} y_{ij} &\sim Poisson(\lambda_{ij}) \\ ln(\lambda_{ij}) &= c + c_i^x + c_j^\alpha + \gamma_j(x_i - \alpha_j) \end{aligned} \tag{18}$$

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 underindentified, 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 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.

3 Scaling the US Senate

Having laid out these five approaches, and compared the first three with simulated data, it remains to test whether such text-based scalings can reflect existing political spectra. Although the application here is to legislatures with many members, it should be noted in passing that many applications of supervised scaling such as Wordscores have generally been applied to entire parties (via their manifestos), which are generally so few that there is no systematic way to test the results either against expert opinion or existing scalings. So although the primary interest here is establishing that these methods can work with large numbers of speakers and party data alone, it is also an important test of the validity of supervised political scaling more generally. Thus before moving on to difficult cases like the House of Commons, it is all the more important to begin with a familiar baseline like the 2006 US Senate.[21]

To begin, although it is not properly part of the test here, it is illuminating and a good intuition check to examine the top Democratic and Republican words. These are not the most common words, since many words are common to both parties, but rather the top most-Democratic and most-Republican words. I take the vector $R - D$ (where R and D are the vectors derived from the concatenated Republican and Democratic documents), and in Table 2 simply list the 40 highest-value (most Republican) and lowest-value (most Democratic) words. As can be seen, many of the most Republican words are quite procedural, while the Democratic words tend to be more substantive, and more clearly related to particularly Democratic concerns. This is because the Republicans are in the leadership in 2006, and the technical duties of the Senate tend to predominate

[21] In the future it may be worthwhile to examine other years (as was done with the House of Commons in section 4) as well as the House of Representatives, to better generalize these initial results.

in their speech. As we will see, this is also an issue that occurs in the House of Commons scalings, although in both cases, even with these procedural words dominating one side, the scalings do separate the parties in the expected way (see Figure 2). Where this issue matters practically is in scaling a year when party A controls the chamber via the data from a year when party B controls things. And on the more theoretical level, it also raises the question of whether we are truly measuring political difference rather than structural oppositions particular to the organization of a chamber. But as we will see in a moment, this is only a minor, though interesting, problem.

Table 3 presents the main results of this first test, the correlations between the various text scalings and DW-Nominate. Of main interest is column 1, the correlations of the various scalings with DW1, the first dimension DW-Nominate score. The Wordscores scaling has been omitted simply because, as the simulations suggested, it correlates with the Bayes score at over 95% in all cases; essentially, the two are the same measure for all practical purposes here. As can be seen in Table 3, the vector projection best reproduces the DW-Nominate score, correlating at 0.66.[22] And in fact, if the 10 or 20 procedural words that dominate the Republican vector are removed from the 1000-word corpora, this correlation with DW-Nominate rises to 0.73. That is, far from depending on the procedural aspect, the text-based scaling best matches DW-Nominate when these procedural words are removed, leaving only the more substantive words. However, as a practical matter, if one is scaling a body whose terminology may be unfamiliar (or is in an unknown language), the presence of such terminology seems not to harm the scaling too greatly here.[23] In any case, given how little attention is paid to talk in the Senate – almost invariably the epithet "cheap" is appended – it is remarkable that the degree to which one speaks like one's party as a whole is a strong predictor of how one votes. This may sound unsurprising to some, but there was no compelling reason to expect that such a crude textual analysis would recover so much of the vote-based scaling.

Before turning to the UK House of Commons, it is valuable to examine the interrelationships between these various scaling results. The Bayesian scaling does nearly as well as the projection approach, and correlates highly with the projection method. Intriguingly, the first principal component is effectively uncorrelated with DW1. But the second is quite correlated with it. This is surprising, because unlike the first two scalings, the PCA scaling has no notion of party built in, and is entirely unsupervised – it takes only the set of document vectors, and finds their highest-varying orthogonal dimensions. There is no *a priori* reason why any of these dimensions should have anything to do with politics; there are, after all, thousands of imaginable dimensions that speakers might be arrayed along.

If one looks at the Scree plot for the first four components of the PCA (Figure 3), one would generally assume that this is clearly a one-dimensional system, and would generally discard the

[22] One may wonder whether the entire practice of using only two reference texts is the best way to capture party-based ideology. And indeed, concatenating texts like that will tend to bias the reference texts towards the most talkative, perhaps leading to a domination of the party effects by those few individuals. Perhaps it would be better to instead take the reference texts as the simple average of every party member's vector, or even more simply, take the position of a scaled speaker as simply the weighted mean of his similarity to every other speaker, ie, $\Sigma_i d_i P_i$, where d_i is the distance between the unknown speaker and a known speaker, and P_i is the party of the latter, an indicator which is 1 if the party is Democrat, and -1 otherwise, say. But it turns out that a score calculated thusly correlates closely with the regular projection score and about as well as that scoring does with DW-Nominate. Perhaps more useful would be using, say, the 5 most liberal and conservative speakers to scale the rest, as we will see below.

[23] Though we will see that in the UK case, it may make a more substantial difference.

others; yet clearly that would be a mistake, since the second dimension is quite informative. But consider Figure 4, where DW-Nominate is plotted against the first principal component. As can be seen, there is no overall correlation. However, looking by party, we immediately see that PCA 1 does in fact correlate with DW-Nominate. Indeed, the within-party correlation is 0.35, rather high. So just as PCA 2 may be naturally capturing party differences, PCA 1 is capturing within-party structure, at least as measured by DW-Nominate. In any case, these correlations with DW1 are relatively weak; for instance, a regression of DW1 on the projection score along with PCA1 and PCA2 yields no significance for the latter two. On the other hand, regressing DW1 on the projection score for only one party or the other yields no significance for the projection score, whereas PCA1 does indeed correlate with within-party DW-Nominate scores.

Finally, the IRT-like scaling of MM and SP produces results that also match DW1 relatively well – about as well as PCA2, in fact. Indeed, if we look a bit more closely, we see that the IRT scaling is almost identical to the PCA2 component, correlating at 0.95. It is perhaps unsurprising that these two unsupervised approaches have ended up in similar places, although why the IRT scaling should have settled on the second principal component, and whether this is a general property, must remain the subject of future study.[24]

Stepping back for a moment, although much stock is put in the DW-Nominate scores, it is not clear exactly what they are measuring within-party. After all, most of the cutting planes (for roll-call votes) that the process relies upon are toward the middle of the two parties, and the distinctions between members at the centers of parties or towards the edges of the spectrum are much less reliable (Poole & Rosenthal 2000*b*). Indeed, if we regress DW-Nominate against a party dummy alone, that single variable explains 86% of the variance (for 2006). If we add to that regression the Congressional Quarterly 2006 measure of party loyalty (that is, how often each member votes with his or her party) and an interaction with the party dummy, those three variables together explain 95% of the variance in DW1. So there really isn't very much to the DW-Nominate score besides the most basic party effects.[25] It appears that most of the match between the text scaling and DW-Nominate scores is due to the first variable alone, the party ID dummy, and indeed that dummy does a better job matching DW-Nominate than the text scalings do, and regressing both against DW-Nominate does not improve model fit over the dummy alone. On the other hand, within-party correlations between the text- and vote-based scalings are clearly not zero (on the order of 0.2), so we know that the text scaling is reflecting more than mere party.

One way to boost intra-party matching might be to use, rather than the parties as a whole as reference texts, only a few of the most extreme members of each party, pooled into two presumably more extreme reference texts. This might potentially overcome the folding problem, where, say, Democrats to the left and right of the center of the party are both projected towards the 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

[24]Regarding the topic of dimensionality, is also 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.

[25]And looking at Figure 2, this might not be very surprising, given the near-perfect separation we see between the parties for the DW-Nominate score – a separation that may make sense in the voting context, but seems unlikely to reflect the true (if hidden) ideological diversity between and within the parties.

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 eliminates the intra-party increase; 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.

More importantly, rather than see the various mismatches between text-based scalings and DW-Nominate as drawbacks of text-based scaling, they can just as well be seen as drawbacks of DW-Nominate, which purports to measure ideology, but mostly (or perhaps entirely) just measures party ID and party loyalty. To the degree that the text-based scalings – even using the party aggregates as reference texts – diverge from DW-Nominate, they may be providing a better measure of ideology that it does. It may be that even in the purportedly low-discipline US congress, party effects overwhelm ideological nuances when it comes to individual votes, and thus the need for scaling in high-discipline or low-vote legislatures may be just as acute in the purportedly well-mapped US case.

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

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, we can at least replicate the analysis in Figure 2. Figure 5 therefore shows the members of the 1996 HOC as scaled by the Liberal and Conservative aggregate texts 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. (Indeed, we can see in Table 4 that the latter two correlate almost perfectly here.) More importantly, Figure 5 shows that these scalings well separate members of the two major parties, although perhaps more dubiously they place Liberals largely between the two. Since as we saw, DW-Nominate is largely a measure of party membership and loyalty, the strong party separation seen here shows that this is at least as plausible a measure of ideology as the such vote-based ones are, and potentially better, since the close connections between DW-Nominate scores and party suggest that even that measure is strongly shaped by strategic concerns and thereby removed from pure political preference.

But is this really a measure of ideology? We know that voting patterns do not change substantially from year to year, and in that sense appear to reflect the sorts of slow-shifting positions we associate with political ideal points, but does the same hold for text-based positions? The separation that we see in Figure 5 could just as well reflect the debates of the moment instead of long-held views, or simply a difference in language between leadership and opposition – particularly given the effects of the language of leadership that we saw in the US case and the preponderance of talk by leaders like Major. The most basic test of the temporal stability question is simply to compare the scores of members scaled in one year with members scaled in another year (obviously only including those members present in both years). Although comparing many year-pairs would be essential for securely establishing these results, the results shown in Table 5 are intriguing. In the first two lines, we see that members who continued from 1996 (the last year of Conservative rule) to 1998 (the first full year of Labour rule) have scores that correlate only weakly, at 0.135.[26] However, if we compare members who are present in 1998 and 1999, the correlation is much higher, at 0.625. This is not just a matter of two years versus one; instead, the change in leadership appears to mix up the scorings, whereas scores across years of the same leadership seem stable.

Is this due to a change in which party wields the language of leadership, or instead due to some sea-change in the ideological dimensions of debate after a new party has taken control? The next two lines of Table 5 show what happens when we remove the speech of the talkative and technical ministers from the reference texts. We see that, even within the same year, the correlation with the ministers-included method is fairly low, and indeed very low for 1996. Thus the degree to which a scaling is dominated by ministers is not a constant – Conservative rule seems to have been more influenced by technical speech than Labour rule. But more importantly, lines 5 through 7 shows that removing the ministers does boost 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.

Another important question is whether the ideological scale as determined by the Labour/Conservative pair of reference texts matches what we would find with other parties as references. Lines 8 and

[26] All scalings hereafter are done with the 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.

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. This suggests that the scalings 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 6 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, but Figure 6 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).[27]

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 (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 useful 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

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

the rebellion rates for all MPs.[28] 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 not 0, much as it was for the intra-party correlation in the US case.[29] Finally, if we look at more methodologically sophisticated vote-based scalings, such as a basic cluster or PCA approach,[30] we find a correlation of 0.55 with the text scaling. However, a closer examination of those 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 remain many other avenues of exploration here, such as using reference texts from one year to scale another as a measure of ideological stability over time (Figure 6 suggests that using 1999's reference texts to scale 1998 produces very similar results to within-year scaling). But this, and further verification of the results above, will have to wait until many more HOC years can be tested against each other. What we have seen, however, is quite instructive: First, scalings using opposing parties as reference texts do produce plausible separation 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 major 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

[28] Rebellion rates acquired from thepublicwhip.co.uk, 2011*a*

[29] 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.

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

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, but in theory, simulations, and practice, they tend to produce remarkably similar outcomes. The theory and simulations show, however, that in some cases their results can diverge (eg, with many words, or with fatter-tailed word distributions), so for these reasons the more theoretically secure Bayesian approach is preferable, although in many cases it may make little difference. In any case, the projection method produced the best correlation with the benchmark DW-Nominate scores, so until further tests are made, it may be the most informative – or at least, the most like existing measures.

Although the principal component and IRT approaches are appealingly free of any expert supervision, there are reasons to be wary of both. The correlation between PCA or IRT 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. In addition, the identification decisions that are made for any IRT scaling may be doing at least some of the supposedly unsupervised work in determining a spectrum. Finally, if one does go with the IRT approach, it may be better to use a more empirically validated likelihood function for word frequencies, the fatter tailed power law rather than the existing poisson distribution – but this too bears further research. In any case, for their flexibility and computational speed, the supervised approaches such as the Bayesian or projection scoring appear most useful.

In terms of actually scoring existing legislatures, the Bayesian, projection, and Wordscores 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 from about 65% to about 75% 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. But resolving such a question would require a deep trip into the meaning of ideology and its relation (strategic or otherwise) to speech and actions such as voting. Regardless of such theoretical issues, however, it is clearly important in future studies to attempt some form of error analysis. As it stands, we have no way to be certain that one scaling really matches the DW-Nominate scores better than another; bootstrapping on the level of speech acts or speakers might provide more certainty about which techniques are better, although their similarity even without error bars suggest that, at least as measured by the DW-Nominate benchmarks, the scores will probably be indistinguishable.

Finally, it might seem that applying these methods to a setting without any vote-based scalings for comparison would prove hopelessly untestable, but 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 (although again it seems to matter little which of the specific techniques is employed). 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 between the party members is quite strong and respects the ordering experts would expect *a priori*. But whatever the variant, this approach seems to work well as a scaling technique – a perhaps surprising result, given how "cheap" we often assume talk to be.

There are still a set of significant questions remaining about this technique. The first was raised just above: how does one scale multi-party legislatures? There are two approaches to this: the first is, as many have done with Wordscores, to simply assign a value to each party and then take the expectation value for each speaker; doing this is straightforward for Wordscores and would only require a little more math for the Bayesian approach, but might require a modification of the projection method.[31] The second 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 Labour-Liberal 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.

Such a decision becomes all the more essential outside of the simple legislative context, where one may use reference texts entirely external to the speakers to scale them, and where there may be no obvious existing scale to mirror. Even within legislatures, it is a truism that the bulk of lawmaking occurs before any voting occurs, when political identity may be much more fluid than the rigid strategies of public voting demand. In committees of all sorts, or any public gathering or collection of speech acts, there will be many different possible dimensions of political (or other) conflict, and all such situations will require that one work carefully from one's model to one's reference texts to best determine the present and changing positions of the players. There is a vast pool of political data only now opening up before us, and much remains to be done to allow our theories to engage with this new, complex, and cacophonous world.

[31] Perhaps using the by-individual scaling mentioned in a previous note, with each individual-as-reference-text weighted according to his party. One can also simply determine the distance between the n parties, embed that in a $n-1$ dimensional space, and then project the speakers onto this space, although selecting a single dimension to report would remain an issue.

References

2011*a*.
URL: *http://www.publicwhip.org.uk/index.php*

2011*b*.
URL: *http://ex-parrot.com/*

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.

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.

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*a*. *Congress: A political-economic history of roll call voting*. Oxford University Press, USA.

Poole, K.T. & H. Rosenthal. 2000*b*. *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.

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

Number of Words	Distribution of Words	Wordscores & Bayes	Projection & Bayes	Wordscores & Projection
1000	e^{-5x}	0.993	0.989	0.978
1000	e^{-20x}	0.884	0.770	0.862
1000	e^{-100x}	0.624	0.731	0.580
1000	$(5x)^{-2.2}$	0.995	0.917	0.890
1000	$(20x)^{-2.2}$	0.952	0.523	0.316
1000	$(100x)^{-2.2}$	0.939	0.158	0.109
2000	e^{-5x}	0.994	0.958	0.955
2000	e^{-20x}	0.987	0.940	0.903
2000	e^{-100x}	0.602	0.318	0.308
2000	$(5x)^{-2.2}$	0.991	0.868	0.814
2000	$(20x)^{-2.2}$	0.999	0.807	0.811
2000	$(100x)^{-2.2}$	0.594	0.995	0.625

[a] Correlations between these three methods generally decline with larger-tailed distributions, or more words. The correlation between Bayes and Wordscores is generally higher than that between either and the Projection method, but even that tight correlation weakens when the tail is long.

Table 2: Top 40 Democratic and Republican words for the 2006 US Senate

Democratic		Republican	
iraq	oil	consent	under
administration	than	ask	10
year	last	unanimous	meet
health	us	bill	court
families	troops	committee	judge
program	provide	senate	defense
care	nation	30	following
debt	trade	2006	district
women	need	border	consideration
veterans	congress	senator	minutes
help	cuts	vote	debate
americans	0	law	business
country	million	hearing	motion
children	medicare	authorized	united
new	american	further	amendment
education	cut	states	other
funding	billion	proceed	marriage
workers	bush	order	illegal
programs	companies	session	agreed
disaster		time	

[a] The 40 largest and smallest values from from the vector $R - D$, ie, the most Republican and Democratic words.

Table 3: Correlations of scalings of 2006 US Senate[a]

	DW1	PROJ	BAYES	PCA1	PCA2
PROJ	0.658				
BAYES	0.615	0.839			
PCA1	0.082	0.077	0.065		
PCA2	0.387	0.742	0.567	-0.162	
IRT	0.375	0.644	0.474	0.306	0.953

[a] DW1 = DW-Nominate score. PROJ = distance on line projected between R and D. BAYES = log likelihood ratio of text being in Republican / Democratic category. PCA1, PCA2 = principal components 1 and 2. IRT = scaling based on MM and implemented by "Wordfish." See text for details. Of particular interest are the correlations between DW1 and the other values.

Table 4: Correlations between different methods for scaling 1996 House of Commons[a]

	Projection	Bayes - Percent
Bayes - Percent	0.772	
Wordscores	0.769	0.999

[a] The three methods of scaling the 1996 House of Commons produce relatively similar results, with Bayes and Wordscores as usual producing almost identical scorings.

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 projection-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.

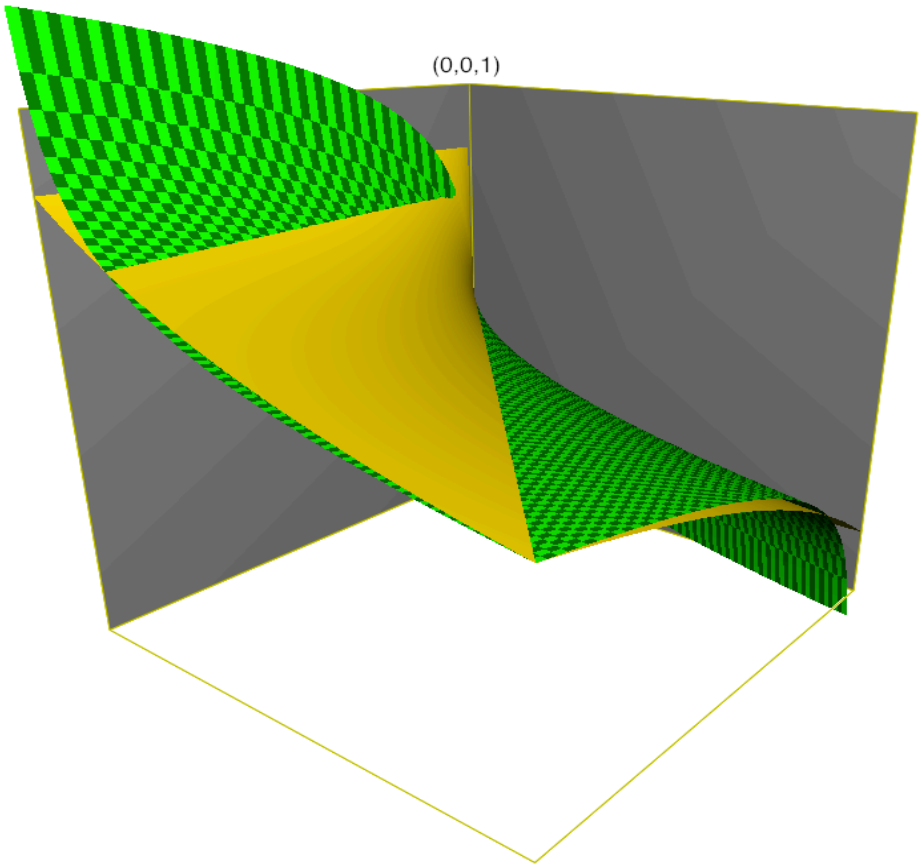

Figure 1: A comparison of the word weighting assigned by Wordscores (smooth) and the Bayesian method (checkered). The x and y axes correspond to the frequencies of some word i in the two reference documents (ie, F_{iR} and F_{iD} from equation 17), and the z axis corresponds to S_i or B_i. That is, z is the weight assigned to a given word, which is then multiplied by the frequency of that word in the virgin text to get the contribution of that word to the total Wordscore or Bayesian score (scales are arbitrary). As can be seen, despite the apparent dissimilarity 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.

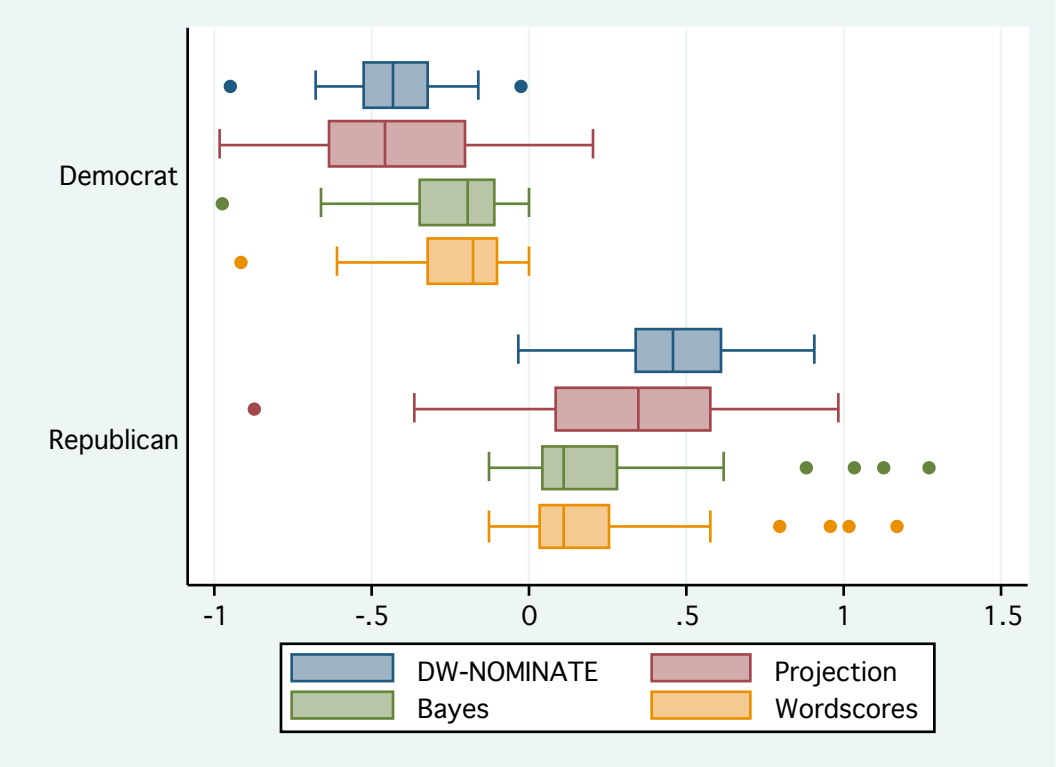

Figure 2: Box plots of the four scaling methods, grouped by Democratic and Republican parties. Axis scale is from DW-Nominate; the other scores have been scaled to match.

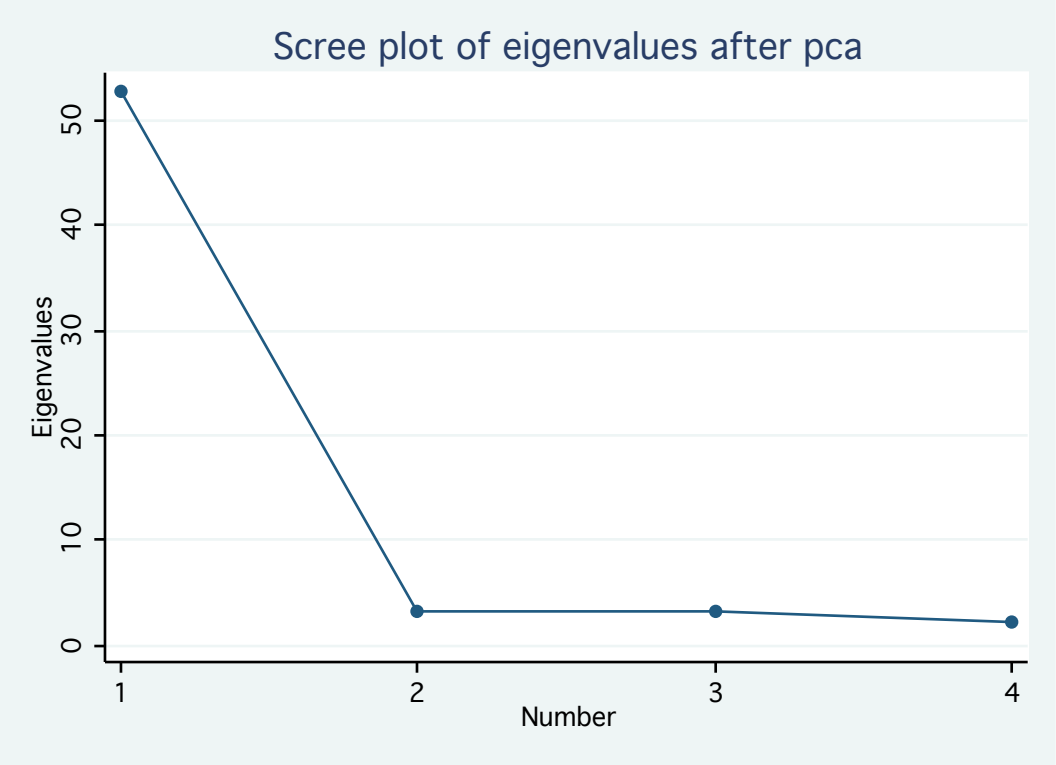

Figure 3: Scree plot of the first four principal components in the Senator scaling.

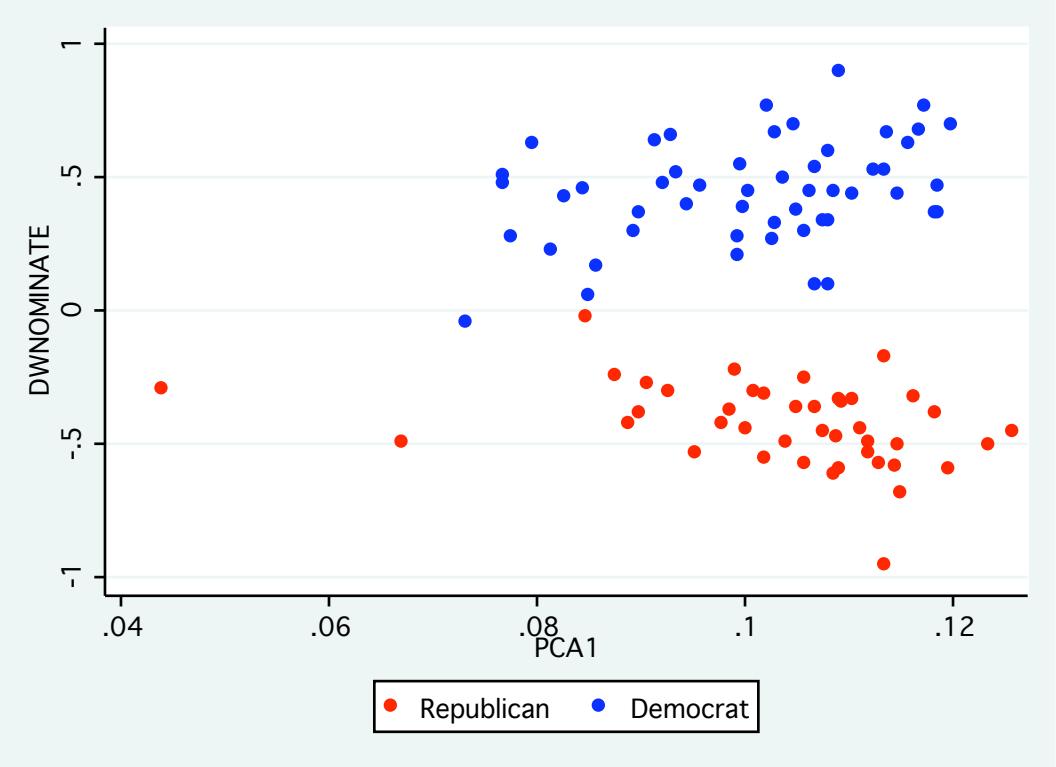

Figure 4: Plot of the first principal component against the first DW-Nominate dimension. Note that although there is no overall correlation between the DW1 score and the PCA1 score, within each party the scores are quire correlated.

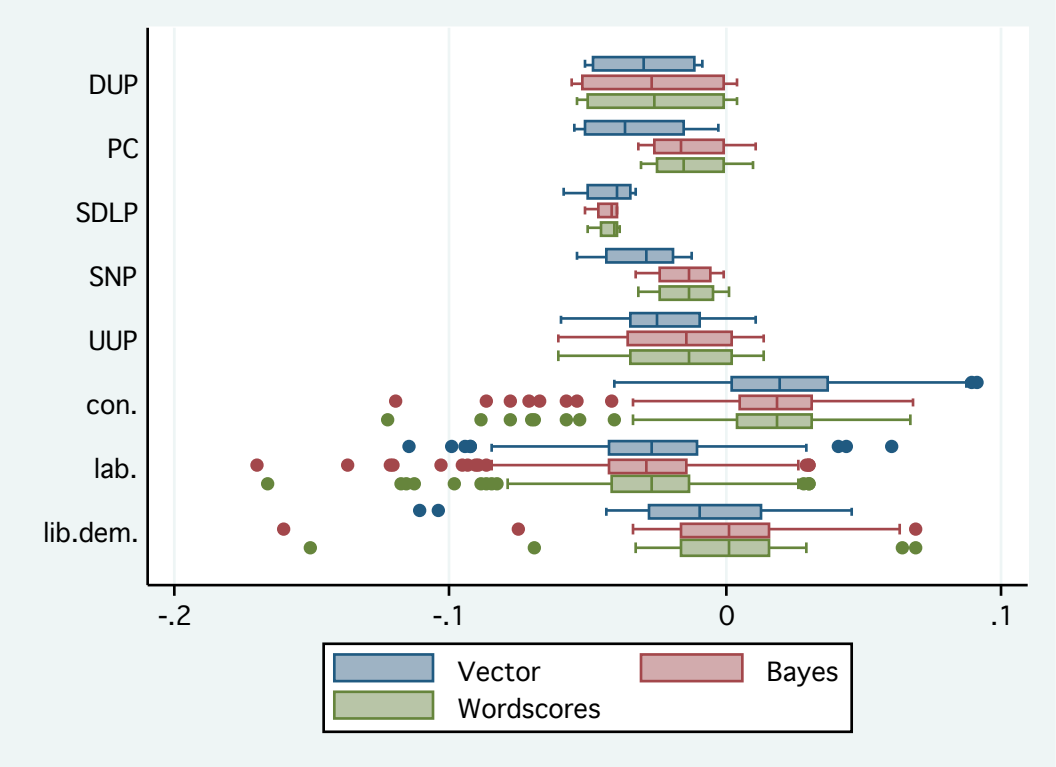

Figure 5: As in Figure 2, the main parties in the House of Commons as scaled by the three main techniques. Axis scale is from Bayes; the others have been scaled to match.

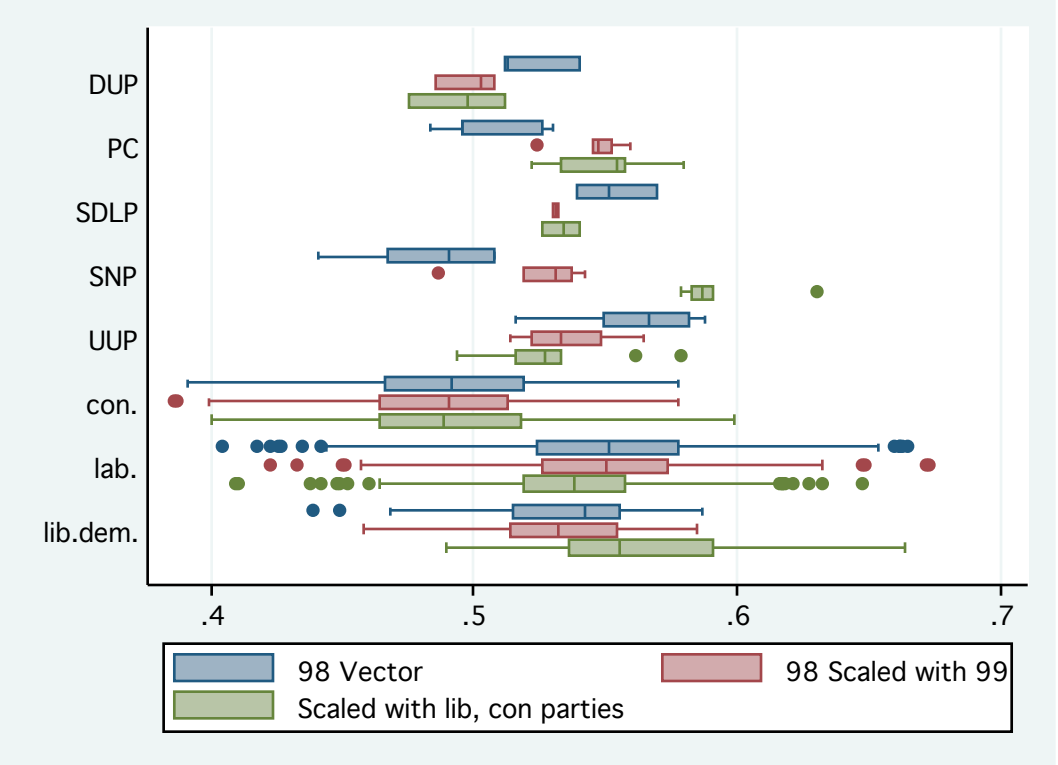

Figure 6: The 1998 House of Commons, scaled by various reference text pairs. "1998 Vector" is simply the standard Labour/Conservative scaling. "98 Scaled with 99" is scaling the 1998 HOC with reference texts based on the aggregated party speech from 1999. "Scaled with lib, con parties" is 1998 as scaled by those parties as reference texts.