German Politics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713635220

How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany

Sven-Oliver Proksch; Jonathan B. Slapin

Online Publication Date: 01 September 2009

To cite this Article Proksch, Sven-Oliver and Slapin, Jonathan B.(2009)'How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany',German Politics,18:3,323 — 344

To link to this Article: DOI: 10.1080/09644000903055799

URL: http://dx.doi.org/10.1080/09644000903055799

How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany

SVEN-OLIVER PROKSCH and JONATHAN B. SLAPIN

The statistical analysis of political texts has received a prominent place in the study of party politics, coalition formation and legislative decision making in Germany. Yet we still lack a thorough understanding of the conditions under which such analysis produces valid estimates of policy positions. This article examines the properties of the word scaling method 'Wordfish' and uses the technique to estimate party positions in Germany. Through Monte Carlo simulations, we investigate the effects of the choice of texts on party position estimates, including the number of documents included in the analysis and their length. Moreover, we present guidelines on how to process linguistic information for political scientists interested in using the technique, focusing specifically on German texts. Finally, we present an analysis of the German party system from 1969–2005 using the Wordfish algorithm. We demonstrate the robustness of the algorithm to extract left-right positions for various subsets of words, but show that agenda effects dominate when estimating a long-time series if the entire manifesto corpus is analysed.

INTRODUCTION

Quantitative studies of the German political system typically require researchers to estimate party positions in a political space. Studies of German politics rely upon estimates of party ideology to understand government spending,[1] law-making and the relationship between the *Bundestag* and *Bundesrat*,[2] and coalition formation on the national[3] and state level.[4] While some studies rely on estimates generated by expert surveys[5] and the hand-coded estimates of the *Comparative Manifestos Project*,[6] new advances in computer-based content analysis have greatly reduced the cost of treating both written and spoken text as data when studying ideology. This paper explores the word-frequency-based position estimation technique, Wordfish, paying particular attention to its applicability to German politics and to German language.[7] Our goal is to outline the assumptions of this technique, discuss both its limitations and advantages, and provide guidelines to researchers interested in using the technique. The first part of the paper presents the Wordfish technique and describes how it has been used to study party positions in Germany. The second part of the paper uses Monte Carlo simulations to examine the conditions under which Wordfish performs well, and discusses the steps required to process manifestos for use in the analysis of German politics. Lastly, using words from election manifestos as data we estimate policy positions for German parties from 1969 to 2005. We examine what the results tell us about German party politics and the application of Wordfish.

German Politics, Vol.18, No.3, September 2009, pp.323–344
ISSN 0964-4008 print/1743-8993 online
DOI: 10.1080/09644000903055799

THE WORDFISH POSITION ESTIMATION TECHNIQUE

The Wordfish technique treats ideology as a latent variable.[8] This means that ideology is not something that the researcher can directly observe, rather it must be indirectly estimated based upon observable actions taken by parties and their members. The observable action we are most concerned with here is the writing of election manifestos. In line with other manifesto-based content analytic methods, Wordfish assumes that the language used by political parties expresses political ideology. Ideology manifests itself in the word choice of politicians when writing party documents. More specifically, Wordfish assumes that parties' relative word usage within party documents conveys information about their positions in a policy space. To give an example, the technique assumes that if one party uses the word 'freedom' more frequently than the word 'equality' in a document on economic policy while another party uses 'equality' more often than 'freedom' in a similar document, these two words – 'equality' and 'freedom' – provide information about party ideology with regard to the economic policy dimension, and discriminate between the parties.

Other quantitative position estimation techniques, such as Wordscores,[9] make a similar assumption. However, Wordscores compares the relative frequencies of words in the documents under examination to words contained in reference texts. Researchers must first assign some texts to be 'right-wing' and others to be 'left-wing'. To do so, researchers choose reference texts that anchor the ends of the political spectrum. Wordfish, on the other hand, does not require researchers to anchor the ends of the political spectrum through reference texts. It also does not require the creation of dictionaries, even though it is possible to analyse subsets of words or sentences if the particular research question justifies such a choice. The interpretation of the estimated dimension in Wordfish is left to the researcher. In the above example, Wordfish does not tell the researcher whether 'equality' is a 'left-wing word' while 'freedom' is a 'right-wing word'. The algorithm will simply use the relative frequencies of these words as data to locate the manifestos on a scale, and it is up to the researcher to make an assessment about what constitutes 'left' and 'right' based upon her knowledge of politics. In fact, we may suspect that left-wing parties (e.g. the PDS/Die Linke) stress the importance of equality while right-wing parties (CDU and FDP) tend to stress the need to guarantee freedom.

Critics of word frequency-based approaches are quick to point out that such algorithms are ignorant of sentence structure and context. For instance, the expressions 'We are against lowering taxes, and for tax increases' and 'We are for lowering taxes, and against tax increases' use the exact same words with the same frequencies, even though the meaning is reversed. A word frequency approach used on only these statements, however, will provide identical estimates. While this may indeed be cause for concern for short statements, we believe that this is not problematic for the analysis of long texts such as election manifestos. Moreover, it has been pointed out 'that words are used in practice in the advocacy of particular policy positions.'[10] Furthermore, the German language seems particularly well-suited for word-based analysis. In contrast to English compound words, which are separated by spaces or hyphens, German allows the concatenating of nouns to form one long word and, in theory, there is no limit to the compounding of nouns. Thus, nouns as single words should contain significantly more

information than do nouns in English. For example, the manifestos we analyse contain 64 phrases that start with the word *Steuer* (tax). The German phrases *Steuersenkung* and *Steuererhöhung*, which appear repeatedly in German party manifestos, would translate in English as 'tax cut' and 'tax increase'. German compound nouns may therefore be more informative because they preserve some context that is lost in English. Nevertheless, other work has demonstrated that the word scaling algorithm works quite well for English and German.[11] Finally, for those who remain sceptical, a statistical approach to textual data offers numerous alternatives. Instead of using words (unigrams) one could use word pairs (bigrams), or in fact any *n*-gram, instead. Bigram frequencies could be scaled in the exact same fashion as those from unigrams.[12]

Estimation Process

Scaling techniques are a commonly used method to estimate latent ideal points. Poole and Rosenthal's NOMINATE technique, for example, estimates legislators' ideologies using roll call votes.[13] To do so, they construct an underlying expected utility model of voting, based upon a spatial model. More recently, item response models have been used to scale vote choices in legislatures and courts.[14] Wordfish works in a similar fashion to these item response models, the difference being that it scales word frequencies instead of dichotomous vote choices. Building upon the work of numerous linguists, Wordfish uses a *naïve Bayes* assumption.[15] In a naïve Bayes approach (also known as a bag-of-words approach), a text is represented as a vector of word counts or occurrences. Multiple document vectors are then put together in a term-document matrix, where each column represents a document and each row represents a unique word, or term. The cells of the matrix contain the number of times the unique word (term) is mentioned in each document. The order of words is lost and elements in the matrix simply represent the term frequency. Therefore, this approach assumes individual words are distributed at random throughout a text. It has been pointed out that 'while this assumption is clearly false in most real-world tasks, naïve Bayes often performs classification very well',[16] and it has become 'probably the most common way of representing texts for further computation'.[17] Scholars then have tried to determine statistical distributions which most accurately approximate word usage. Commonly used distributions include the Poisson,[18] the negative binomial,[19] and other Poisson mixtures[20] as well as zero-inflated (binomial) distributions.[21] All of these distributions are heavily skewed, as is the case of word usage.

Wordfish assumes that word frequencies are generated by a Poisson process, the simplest of these distributions. The systematic component of this process contains four parameters: document (party) positions, document (party) fixed effects, word weights (discriminating parameters), and word fixed effects.[22] Word fixed effects are included to capture the fact that some words need to be used much more often in a language. Such words may serve a grammatical purpose but they have no substantive or ideological meaning, such as conjunctions or definite and indefinite articles. The document fixed effect parameters control for the possibility that some documents in the analysis may be significantly longer than others. When using manifestos to estimate party positions, this can happen when some parties in some years write much longer manifestos. In Germany, the length of election manifestos has increased over time, but varies by party within elections.[23]

The parameters of the greatest interest are those capturing the position of the party documents, and the word discrimination parameters. The interpretation of the document positions parameters is clear. These are the positions of the parties relative to the other parties in the political space. The word discrimination parameters allow the researcher to analyse which words differentiate party positions. In the example above, 'equality' would have a high absolute value for its discrimination value and its usage would most likely be associated with left-wing parties. The word 'freedom' would also have a high absolute value but with the opposite sign because its usage would be associated with right-wing parties. This allows the researcher to estimate party positions and uncover the variations in political language that are responsible for placing parties on this dimension.[24]

Identification

As is the case with all item response models, the model as such is unidentified.[25] Typically, the ideal point literature offers two ways to identify one-dimensional models.[26] The first possibility is to identify the model by transforming all estimated positions to have a mean of 0 and a standard deviation of 1. This relative identification can be made absolute by prescribing a direction for the position (e.g. constraining the CDU in a particular year to be to the right of the PDS in that year). The second possibility is to simply choose two documents and assign fixed values to them. Then, all other positions will be estimated relative to these two anchors. Note that this is not the same thing as assigning 'reference values' à la Wordscores, because we remain agnostic and leave it up to the estimator to identify the extreme values on the scale as well as what texts represent those extremes. The two identification strategies should produce identical results albeit on different scales.[27]

CONDITIONS FOR USING WORDFISH: MONTE CARLO RESULTS

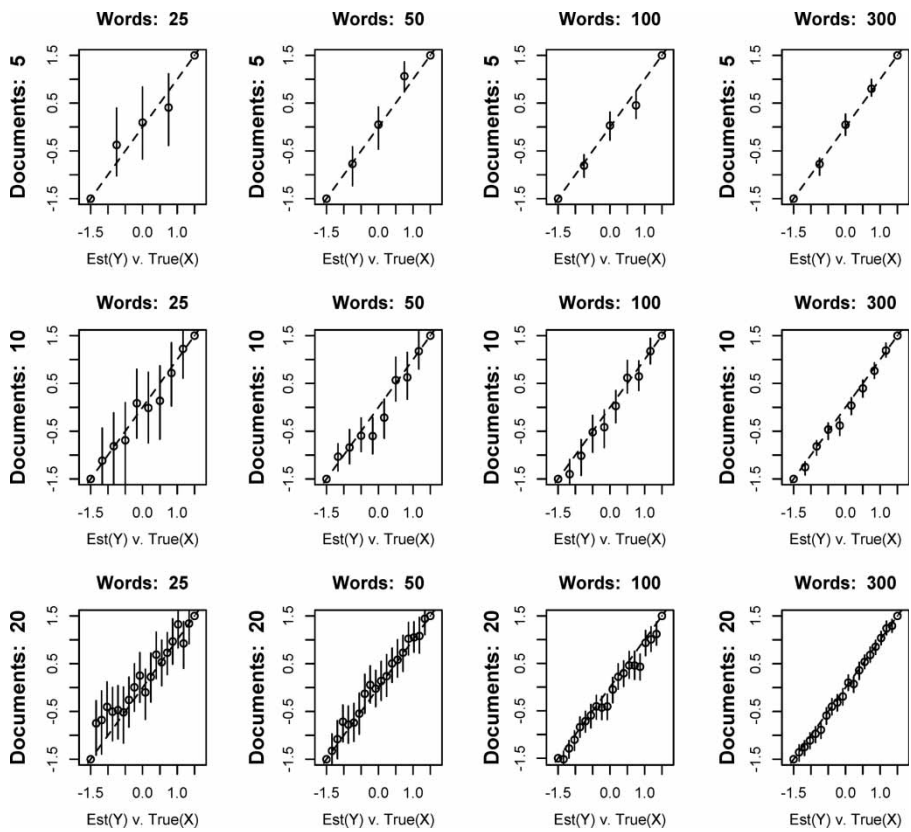
The model specification used by Wordfish works best as more data is available, meaning as more documents are used in the analysis and as those documents contain more unique words. If the documents do not contain a sufficient number of unique words, there will not be adequate information to estimate document parameters. It remains an open question, however, what constitutes a sufficient number of documents and words to run Wordfish. We attempt to answer this question through the use of Monte Carlo simulations. This involves fixing document positions and all other parameters and generating a term-document matrix assuming that the data generating process we specify is, in fact, the true data generating process that produces textual data.[28] To generate simulated term-document matrices we take draws from Poisson distributions with the 'true' parameter values that we set. Several term-document matrices are constructed, each time changing the number of documents and number of words they contain.

We examine simulated data for 5, 10 and 20 documents and 25, 50, 100 and 300 unique words.[29] After the different data sets have been generated, we run Wordfish on each dataset to examine how well our algorithm recaptures the true document positions. We also examine how the number of documents and words in the data affect the size of the confidence intervals around our estimates. To do so, we run a parametric

bootstrap that uses the estimated parameters to generate new datasets by drawing counts from the Poisson distribution. This procedure is repeated 500 times for each of our 12 simulations and the 95%-confidence intervals are calculated from the distribution of the 500 estimated positions. We would expect that as more unique words are added to the analysis the size of the confidence intervals should shrink. The results of the Monte Carlo simulations, presented in Figure 1, suggest that this is indeed the case. These graphs plot the 'true' values of the positions against our estimates with confidence intervals. As estimates lie closer to the dashed 45% line, they come closer to recapturing the true values as set by us.

A glance at Figure 1 reveals that more unique words shrink the confidence intervals. Simulations reveal that confidence intervals become smaller as both the number of documents and words increase. In simulations with 100 or more unique words, it is possible to discern statistically significant differences in the estimated positions as confidence intervals do not overlap. With only 25 or 50 unique words it is more difficult to find statistically significant differences among parties. Estimation also improves with the number of documents in the analysis. When we use only five

FIGURE 1
SIMULATIONS OF CONFIDENCE INTERVALS FOR VARIOUS NUMBERS OF TEXTS AND LENGTHS

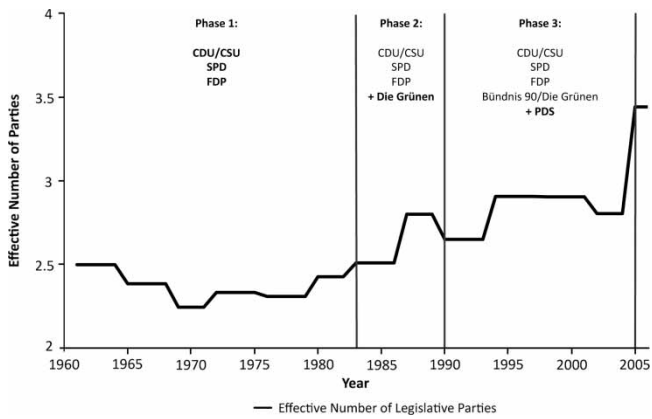


documents, there are at least two instances in which the true value for the position does not fall within (or lies on the border of) the calculated 95% confidence interval. As the number of documents increases, the estimated positions lie closer to the 45% degree line. Once we have 300 unique words and 20 documents, the results line up very tightly along the 45% line. This suggests Wordfish should perform quite well for most analyses of manifestos as they will use thousands of unique words and more than a dozen documents. However, it also means that if researchers have only a handful of short documents, Wordfish may not work as well.

POSITION ESTIMATION WITH WORDFISH: GERMANY, 1969–2005

We now discuss and demonstrate the application of the Wordfish technique to the estimation of left–right positions for German parties from 1969–2005. Choosing such a long time period poses additional challenges to the estimation. The German party system has undergone several transitions during this time, each time adding new parties to system. Figure 2 shows the effective number of parties in Germany since 1960 using the formula proposed by Laakso and Taagepera.[30] On the basis of these values, the system experienced two major junctures.[31] After the first phase of party system concentration between 1949 and 1961, the subsequent phase was dominated by three parties, the two mass parties CDU/CSU and SPD and the smaller FDP, up until the early 1980s. The fractionalization towards a multi-party system began in 1983, when the Green Party was first elected to the Bundestag and has been represented ever since.[32] The year 1990 marks a significant juncture for the German party system away from the West German system to the multi-party system of unified Germany. The system now included the PDS as the successor of the East German SED, raising the effective number of parties substantially. In 2005, an electoral alliance was formed

FIGURE 2
EFFECTIVE NUMBER OF PARTIES IN GERMANY, 1961–2006.



Note: The effective number of parties is calculated on the parliamentary seat shares (s) using the formula $1/\Sigma s_i^2$ by Markku Laakso and Rein Taagepera, 'Effective Number of Parties'. Authors' calculations based on data from http://www.bundeswahlleiter.de

between West German former SPD members (WASG), who had left the party out of protest against SPD chancellor Gerhard Schröder's reform policies, and the East German PDS (which had meanwhile renamed itself to Linkspartei.PDS). The electoral alliance formally merged into a new party called Die Linke in 2007. The question remains to what extent the increased fractionalization of the party system was accompanied by increased ideological polarization, and how election manifestos can be used to measure polarization. In previous research, Wordfish has been used to estimate party positions between 1990–2005.[33] The analysis demonstrated that the technique captured a socio-economic left–right dimension of politics. Extending the analysis back in time allows us to examine how the party positions may have shifted following reunification and the expansion of the German party system.

Election Manifestos: Definition and Selection

Party manifestos provide a great deal of information about parties' policy proposals and can potentially reveal information about party ideology. As one of the most impressive content analysis projects in political science up until today, the Comparative Manifestos Project has hand-coded party documents into more than 50 categories. The use of these data in comparative politics research in general, and in German politics research in particular, is ubiquitous. But not all source documents coded by the Manifesto Research Group/Comparative Manifestos Project (MRG/CMP) are actual manifestos. The MRG/CMP initially selected documents that contained a 'recognizable statement of policy, which has the backing of the leadership as the authoritative definition of party policy for that election'.[34] This included, as the primary source, election manifestos, even though parties in more than half of the original 19 MRG/CMP countries did not produce a manifesto as such.[35] But what sources can we use as election manifestos? Election manifestos have been defined as 'encyclopaedic documents dealing with a wide range of policy issues, [which are] published in a clearly-defined political context',[36] namely election campaigns. They are also 'strategic documents written by politically sophisticated party elites with many different objectives in mind'.[37]

In the German context, three types of officially-endorsed, programmatic party documents can be distinguished: basic programmes, action programmes, and election manifestos.[38] *Basic programmes* are 'adopted by party congresses after long discussions [...] with widespread participation. [They] are intended to give the party a basic orientation and philosophy for as much as a decade, and sometimes involve a radical reshaping of the party image [...]. While obviously central, such documents appear only at relatively long intervals and do not (and are not intended to) reflect election exigencies.' *Action programmes* are party documents published throughout a legislative term 'for specific purposes and areas.' Finally, *election programmes* or *election manifestos* 'assess the importance of current political problems, specify the party's position on them, and inform the electorate about the course of action the party will pursue when elected.'[39] Even though German parties nowadays use more channels to quickly communicate positions, in particular the internet, election programmes still constitute the most authoritative statement of parties' policy positions prior to elections.

Our goal is to demonstrate the effects of document selection and processing decisions on the estimation of one-dimensional positions from German election manifestos. Our

target period is the time between 1960, when the West German party system solidified into three main parties, and 2005, the date of the last available manifesto data. First, the time period selection requires that the source text data are comparable and similar in nature. Therefore, we only include documents that meet the definition of an election manifesto as an *encyclopaedic written statement of a party position during an election campaign*. This definition deviates from the CMP procedure of selecting manifestos for the analysis, or in their absence, selecting the 'nearest equivalent',[40] such as campaign speeches, election proclamations, newspaper reports, or press releases.

The table in the appendix presents the list of German party documents since 1961.[41] In addition, the table lists the type of document and the length in words. The indication of the type of document allows us to decide whether a document should be included in a quantitative word scaling exercise or not. First, our definition excludes simple election proclamations *(Wahlaufruf)*. Such proclamations are published during an election by the party and primarily call upon voters to go to the polls and cast a vote, rather than presenting them with a holistic discussion of policy issues and the party's priorities and proposed course of actions. As a result, such proclamations are rather short. The documents of the CDU-CSU in 1961 and of the FDP in 1972 fall in this category. These documents are in fact the shortest in the document collection. Second, our definition excludes party congress speeches. Such speeches were used by the CMP as a nearest equivalent, but pose substantive and technical problems for the analysis. They do not necessarily follow the structure of a manifesto, they use a different linguistic style, and they include selected speakers from the party who may not represent the party's position on the full set of issues. In two instances, the CMP used such party speeches (SPD 1961 and CDU-CSU 1965). Finally, the CMP also used an action programme for the FDP in 1965. As the FDP could not reach a consensus on a common election programme during that year,[42] the MRG initially excluded the FDP from the analysis, but later decided to include it by using an action programme which was published in 1967, two years after the election. As this does not qualify as an election manifesto and was published after the election, we also remove it from the analysis.

We decide to exclude the elections in 1961 and 1965 entirely from the analysis, as we would only be able to estimate one party position in each of them. Moreover, we exclude the FDP position in 1972 owing to the different document type. This leaves us with 44 party manifestos between 1969 and 2005. For users of quantitative content analysis of manifestos in general, we strongly advise considering the type of party document before including it in the analysis.

From Manifestos to the Term-Document Matrix

Prior to running the Wordfish code to extract ideal points, the documents of interest must be carefully pre-processed. In fact, in any statistical analysis of text, document processing is essential and possibly the most arduous task in the estimation process. In the following, we describe the steps and decisions involved in creating term-document matrices from German manifestos.

Text input and policy dimensionality. There are two primary considerations when selecting political texts. The first is the nature of the source documents, as described

above, and the second is the quality of the text data. Wordfish estimates a single policy dimension, and the information contained in this dimension depends upon the texts that the researcher chooses to analyse. Therefore, the selection of texts should depend on the particular policy dimension the researcher wishes to examine. For example, if one is interested in comparing foreign policy positions of parties in Germany, then only text exclusively containing German foreign policy statements should be included in the analysis. On the other hand, if the research question is to determine a general ideological position using all aspects of policy (e.g. left–right), then the analysis could potentially be conducted using all parts of a political text, assuming that the documents or speeches are encyclopaedic statements of policy positions. This is most likely the case for party manifestos and our selection criteria above followed this definition. The estimated single dimension will thus be a function of the selection of the text corpus. This is different from the Wordscores approach, which estimates different dimensions not by altering the text inputs but by changing the reference values assigned to reference texts.[43] This means, for example, that economic, social, and foreign dimensions are estimated in Wordscores on the full manifesto texts, even though only some of the sections relate to the policies of interest.

Thus, while the position estimation itself is left to the scaling algorithm in Wordfish, the text input needs to be carefully chosen by the user. Independent of whether scholars want to use estimates as a dependent variable and explain party position movement or use them as predictors for policy output and outcomes, policy dimensions are usually defined *a priori*.[44] Researchers can use any type of approach to demarcate the dimensions. Fortunately, most manifestos are rather structured documents with well-defined section titles that can easily be assigned to policy areas. In previous research, we have described one possible strategy to identify policy areas by parsing the manifestos into policy areas (economic, societal, foreign).[45] For instance, a paragraph describing the proposed tax policies of a party would be assigned to an economic policy area. If there are three main policy areas under investigation, this procedure results in three documents for each manifesto. Term-document matrices are then constructed separately for each area and the position estimation is performed separately on each of those matrices. This procedure constitutes only one of many ways for identifying policy areas in manifestos, which should always be mentioned explicitly for replication purposes.

While issues of dimensionality constitute important coding decisions, our focus here is on how to proceed once the texts have been chosen. We therefore focus simply on the whole of the manifesto texts to study the effects of different text manipulation methods on the position estimates.

Party manifesto processing. Once appropriate documents (or subsections of documents) have been selected, the researcher must ensure that they are in machine-readable format. If the document is a scanned version of the manifesto, converting it to a text file will most likely require running optical character recognition software over the documents, at which point additional error might be added to the data.[46] Despite the ubiquitous use of the CMP data,[47] digitized files of the analysed manifestos are not readily available. Currently, the only available electronic versions of the manifestos are archived at the *Zentralarchiv für Empirische Sozialforschung* by the

Comparative Electronic Manifestos Project.[48] While the electronic availability of the manifesto files would suggest an easy application of quantitative text analysis, the quality of the electronic versions of German manifestos varies significantly. As a result of the scanning process, some documents contain spelling errors and/or missing text. We therefore base our analysis on electronic manifesto texts that have been checked for several types of errors.[49] Having processed the documents, checking for mistakes, we construct the term-document matrix.[50]

Stemming words. One option when creating a term-document matrix is to count words exactly as they appear in the original manifestos, but another option is to count stemmed words. A stemmer algorithm removes morphological and inflexional endings from words and returns the stemmed words. The advantage is that essentially similar words will be captured as one. Moreover, the term-document matrix will have fewer unique words if words are stemmed, thus making the estimation more efficient. A potential disadvantage is that certain compound nouns might be reduced to a stem thus losing information. This may be particularly problematic in German, where words are often compounds. We demonstrate that stemming does not change the results, but makes the estimation more efficient. To make the estimation more efficient, we also remove so-called stop words from the matrix (e.g. *aber, der, die, das*, etc.). Stop words are very common words with minimal information value.[51]

The Challenge of Dynamic Estimation

Using text to estimate party positions over time creates an additional challenge. On the one hand, we would like to use as much information in the manifestos as possible. On the other hand, we would like to estimate position change over time. This is a trade-off. For example, if the political debate changes and new vocabulary enters the political lexicon in election t, then this will differentiate the manifestos at point t from those at point $t-1$. In fact, in this instance, we are likely to pick up a policy agenda shift in manifestos, whereas we are interested in party position change. Moreover, there is a danger that we might mistakenly confuse agenda-shift with ideological change, when they are two analytically distinct phenomena. This problem has not received much attention in the party position estimation literature. There are two potential routes to addressing this issue. One could model the election specific effects through a hierarchical model with election fixed-effects or through changing word weights, which would significantly complicate the statistical model (one would need to decide which words are allowed to shift weights) and increase the estimation time. The other route, and the one we describe here, is to carefully select the words that enter the analysis.

When using manifestos to estimate party positions, the Wordfish model treats each manifesto as a separate party position and all positions are estimated simultaneously. In other words, the position of party i's manifesto in election $t-1$ does not constrain the position of party i's manifesto in election t. If a party maintains a similar position from one election to the next, it means the party has used words in similar relative frequencies over time. On the other hand, if the model indicates that a party moves away from its former position and closer to the position of a rival, it implies that the party's new word choice more closely resembles that of the rival's than of its former self. This specification has the advantage that we do not impose prior knowledge on the

estimation. An alternate specification might assume that a party's position at time t is both a function of its word choice at time t and its position in previous elections. Such a specification might ensure smooth party movement over time, and the movement would both be a function of the word usage and the assumptions about the model's functional form.

A common issue with dynamic ideal point estimation is the anchoring of the dimension over time. Technically, identification of the model is achieved by either constraining the position estimates to have mean zero and standard deviation one or by setting two document positions as fixed and estimating the rest relative to these anchors.[52] Thus, if there is movement of parties, it can only be due to different word usage. This requires that the word data over time must be comparable at a minimum level. We have already pointed out that careful source document selection matters and that 'manifestos' that do not qualify as such (e.g. short election proclamations or party congress speeches) should be excluded. Assume that the political lexicon in the manifestos at election time t contains an issue that is no longer relevant at time $t + 1$, e.g. official relations with the GDR (East Germany). If all parties make a statement with regard to the GDR at point t but not at $t + 1$, then the words will not only distinguish parties at point t, but also distinguish the elections. As a result, if all words are counted, even the rare ones, the parties are more likely to be clustered by election.

This brings our attention to the word inclusion criteria. We opt to include only words that fulfil a minimum threshold criterion based on non-informative and informative priors, and examine the effects of varying this threshold. While there is a technical controversy over how to handle very uncommon words in term-document-matrices,[53] it is important to note that the choice of words in text scaling is related to the notion of dimensionality. Our first alternative term-document matrix includes words that are mentioned in a minimum number of documents (in at least 20%), thus essentially keeping words that are deemed important enough to be mentioned either over time by one party or by several parties. The other alternative matrix contains only those words that appear both pre- and post-1990. We chose this as our criteria because reunification added words to the German political lexicon that were not in it previously. Likewise, some words that were previously important likely fell out of use. If we do not control for this fact, we would see a large jump in all parties around 1990 as they all shift their word usage to account for new political realities. In essence, this change is probably a second dimension of politics that is unrelated to left–right (assuming all parties shift). By eliminating words unique to either the pre- or post-reunification period we hope to control for this dimensional shift.

RESULTS: PARTY POSITION ESTIMATES, 1969–2005

Table 1 presents a summary of the different German manifesto term-document matrices. The first 'naive' approach simply counts all words (dataset A). Each word, even if it occurs only once in one manifesto, is included in the analysis. The large majority of words, in fact, only occur once, which means, in theory, they can receive infinite discrimination weights.[54] The single words are therefore election specific terms, and the corresponding estimated party positions are likely to pick up

TABLE 1
TERM-DOCUMENT MATRICES: WORD EXCLUSION CRITERIA

Word Count Data Set	Word Exclusion Criterion	Stemming	Stopword Removal	Unique Words
A	All words			41,684
B	Words mentioned in at least 20% (n = 9) of all documents	X	X	3,455
C	Words mentioned pre- & post-1990		X	11,273
D	Words mentioned pre- & post-1990	X	X	8,178

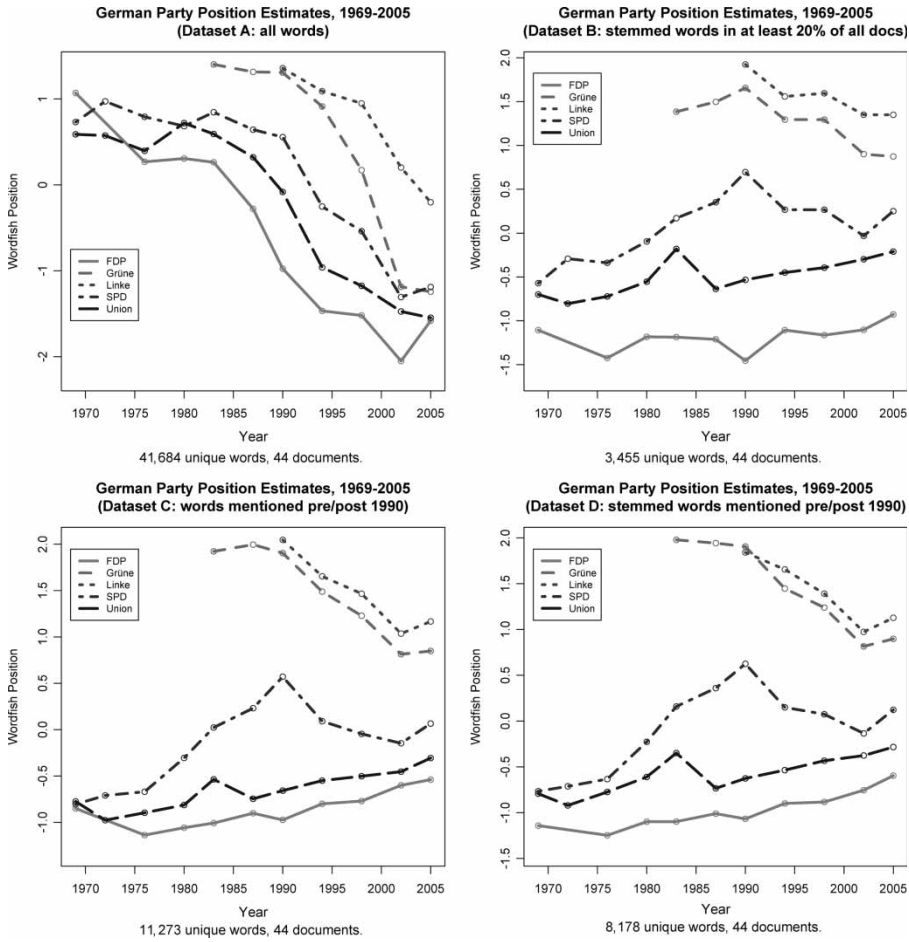
policy agenda change over time in addition to preference change of the parties. If this is the case, then the movements of party positions over time will be strongly correlated. In dataset B, we apply basic text processing and stem words and drop the most uncommon words. We keep only words that are mentioned in at least 20% of all manifesto texts (at least nine documents). This cut-off will guarantee that a word needs to be mentioned in at least two elections. This criterion does not require any prior knowledge of the party system. The final two datasets incorporate knowledge about reunification in 1990, the year in which the German party system has transformed from the three/four-party West German system to the current five party system. As an anchoring criterion, we only consider words that are mentioned at any point in any manifesto pre- *and* post-1990 and distinguish between words (dataset C) and word stems (dataset D).

Manifesto Analysis

Figure 3 presents the results of the estimation.[55] The top left plot shows the estimated positions using all words in the word count data set. The face validity of the rank ordering of the parties seems very plausible. The PDS/Linke is estimated to the left of the Greens. In the centre are the SPD and the CDU/CSU (Union), and to the right is the FDP. The rank order of the estimated positions remains remarkably stable. The exceptions are the elections in 1969, when the FDP is estimated to the left of the SPD and CDU/CSU (Union), in 1983, with the CDU and SPD being very similar, and in 2005 when the SPD and Greens also have almost identical positions. The most dominant trend in the plot is the fact that all party position estimates move together over time along the estimated dimension. This suggests strong policy agenda effects. All parties use words in a given election that were not used in the previous and subsequent elections. The political lexicon in the 2000s naturally differs from that in the 1970s, with new political terms having been incorporated and old ones removed. Moreover, using all unique words also places great weight on words mentioned only in one manifesto. If there are disproportionately more of such rarely mentioned words, which is the case here, then the party position estimation will be influenced heavily by them, resulting in agenda shifts.

A simple correction is done by stemming words and excluding the rare ones (dataset B). The estimated positions are significantly smoother and there is less change over time, but the rank ordering of the estimates remains very similar. The PDS and Greens, when they enter the *Bundestag* in 1983 and 1990 respectively, are

FIGURE 3
POSITION ESTIMATES IN GERMANY, 1969–2005



on one side of the spectrum, whereas the SPD and the Union are located in the centre. The FDP is located in all elections to the right of the CDU/CSU.

The final two plots present words mentioned pre/post 1990 (dataset C) and word stems mentioned pre/post 1990 (dataset D). The results provide almost identical results to the estimation based on dataset B. Again, while there is party movement over time, parties do not leapfrog each other. The only difference between the two versions is the election in 1969, in which the FDP is estimated to be very similar to the SPD and CDU/CSU, but different using word stems. Pinpointing the FDP in 1969 using manifestos seems to be quite dependent on the political lexicon used. This might not just affect our word frequency scaling algorithm, but other text-based position estimation techniques, such as Wordscores,[56] in particular if the FDP manifesto is used as a reference text.

What is the bottom line of these estimations? As suspected, agenda effects over time dominate the results when all words are used. Excluding rare words induces stability and the results are corroborated by their good face validity. Our robust results suggest that the ideological polarization in Germany significantly increased with the entry of the Greens and the PDS. In the first ten years following re-unification, the system polarization has gradually decreased. Recent developments suggest, however, that the relative spread of party positions is beginning to increase again, with the emergence of the new party Die Linke.

Word Analysis

In addition to estimating the positions of the documents, Wordfish locates the positions of the words in the same space. One possibility then is to analyse which words have the largest word weight, i.e. which words are located at the extremes of the political space. Since the PDS and the Greens are estimated on one side, we would expect words on this side of the dimension to represent words that these two parties emphasise in manifestos which the other parties do not. Similarly, words on the other side of the spectrum should be highlighted by the party estimated at the other extreme, the FDP, and the words should reflect the FDP manifesto. Figure 4 shows word stem estimates based on dataset B. The left plot presents the 50 words with the most positive word weight values (which corresponds to 'left' positions), and the right plots displays the 50 words with the most negative word weight values (which corresponds to the 'right' positions). The hundred words are plotted in proportion to their word weight value. In other words, larger words are more extreme than smaller words.

Even though these plots show only a very small subset of words, they help us to understand the dimension being estimated. Of the 'left words', we can identify party labels (PDS, Grün) and words related to policies of non-discrimination, way of life, direct democracy, economic redistribution, workers' rights and participation, and nuclear energy. Similarly, the 'right words' contain party labels (here the FDP), but then keywords from market economy, law and order policies, education, and a few

FIGURE 4
WORDS WITH THE MOST EXTREME WEIGHTS (BASED ON DATASET B).

Note: The left plot shows the words with the largest positive weights, the right plot shows the words with the largest negative weights. Size of words is proportional to the absolute value of word weights (the larger the word, the more extreme the word weight value). The words are shown in random order. Both plots have been created with http://www.wordle.net/advanced

words related to defence and foreign policy. Together, these areas do cover the political lexicon that is typically associated with left–right. The algorithm picks up the different use of economic vocabulary and places the parties correspondingly. For example, whereas the left parties highlight redistributive elements and economic disadvantaged groups ('redistribution', 'wealth', 'unemployed', 'work time reduction'), the right emphasises a free market economy ('market-driven', 'market distortion', 'performance-dependent'). Thus, the word stems receive estimated values that we commonly associate with left and right positions in German politics. The results further strengthen our argument that the word exclusion preserved words with strong political connotations.

One of the interesting words estimated on the right is the word 'liberal'. Liberal policies in Germany are typically associated with economic policies favouring lower taxes and social policies favouring individual rights (i.e. discarding traditional values and open to, for instance, increased minority rights, same sex marriage, and abortion). This means that the word itself has a two-dimensional meaning. Parties can mention liberal policies in their manifestos, but in fact talk about different kinds of liberal policies. Some examples of usage of the word in German manifestos include:

- FDP, 1994: 'The FDP stands for the protection and promotion of private property as the foundation of a liberal society.' *('Die FDP bekennt sich zum Schutz und zur Förderung privaten Eigentums als Grundlage einer liberalen Gesellschaftsordnung.')*
- Greens, 1998: 'A liberal society does not tell people how to live their lives.' *('Eine liberale Gesellschaft schreibt den Menschen nicht vor, wie sie ihr Leben zu gestalten haben.')*
- PDS, 1998: 'The penal system must adhere to a humane and liberal standard.' *('Der Strafvollzug muß einem humanen und liberalen Leitbild verpflichtet sein.')*
- SPD, 2002: 'Without our policies since 1998, Germany would be less modern, less social, and less liberal.' *('Ohne unsere Politik seit 1998 wäre Deutschland heute weniger modern, weniger sozial, auch weniger liberal.')*
- CDU/CSU, 2002: 'A liberal state must be able to defend itself, otherwise the free democratic order will not endure.' *('Ein liberaler Staat muss auch ein wehrhafter Staat sein, sonst hat die freiheitliche Demokratie keinen Bestand.')*
- FDP, 2005: 'Low, simple, and fair – these are the criteria for the liberal tax plan.' *('Niedrig, einfach, und gerecht – das sind die Kriterien für das liberale Steuerkonzept.')*

This shows that the word has a multidimensional meaning. Additionally, in Germany, the FDP refers to itself as *Die Liberalen*, putting a party label dimension into the word usage:

- FDP, 2005: 'The Liberals advocate cosmopolitanism and tolerance.' *('Die Liberalen stehen für Weltoffenheit und Toleranz.')*

In the actual estimation, the word stem 'liberal' receives the largest negative weight due to the fact that the FDP uses this term to describe its own proposals.

Thus, while we might expect that the word is used in a different context by other parties, the estimation places it on the right due to the 'word-ownership' of the FDP.

CONCLUSION

This paper has accomplished three main tasks. It has provided further guidance to researchers interested in using Wordfish to estimate policy positions from text data, paying particular attention to how this method can be applied to German politics. Second, it has further examined the conditions under which Wordfish is likely to accurately capture the policy space by using Monte Carlo simulations. Lastly, it has applied the Wordfish technique to German politics to examine party system changes from 1969–2005. This extends previous analyses of German party positions using Wordfish.

The paper demonstrates that computer-based position estimation works quite well in many instances and successfully estimates party positions for Germany. Our Monte Carlo analysis suggests Wordfish works well in most practical scenarios and the analysis of German politics shows that it can work even over a longer time period. However, we have also demonstrated that computer-based position estimation cannot replace the researcher's judgement. In particular, we have argued that researchers first need to assure the quality of the source documents. More care needs to be invested if positions are estimated from short documents and from documents other than party manifestos. While computer-based position estimation has been applied to speeches, more care needs to go into the discussion of the source document, the political language being used, and whether the language is comparable across documents. This is not a trivial task. For instance, we could very well code 100 cookbook recipes, count the word frequencies, apply the scaling algorithm, and place the recipes on a single dimension. While this is technically possible, we clearly would not want to conclude that we have uncovered an ideological dimension. Therefore, we caution against a premature application to political documents that do not follow the same data generating process as party statements. The nice feature of the latter is that parties can put different emphases on words from the political lexicon. If texts with different types of authorship are to be compared (e.g. party manifestos with laws or judicial decisions), we first need to make sure that the nature of the documents, in particular the word generating process, is in fact comparable across document type. Even when comparing documents of the same type other than manifestos, the researcher should ensure that authors of the documents could use words that express a political position, and that these documents do not simply follow a formulaic legal structure.

In addition, estimation improves when researchers are able to bring substantive knowledge to the task at hand. This is especially true when estimating long time-series and when language usage is likely to have changed. We have shown that it is difficult to disentangle agenda shifts and party movement if political rhetoric changes substantially over time. A simple correction, i.e. excluding rare words and eliminating words used by parties only before and only after 1990, improved the estimation results.

Computer-based content analysis provides a systematic way to estimate party positions from political texts. It continues to hold great promises for the study of German politics and its party system. However, like every other quantitative study, it requires the researcher to pay careful attention to the type of data being used. The

paper has added to the recent literature by pointing quantitative analysts of text to the potential pitfalls they could encounter, and how such pitfalls can be avoided.

ACKNOWLEDEGMENTS

The order of authors' names reflects the principle of rotation. Both authors have contributed equally to all work. A previous version of this article was presented at the Workshop on Estimating Policy Preferences hosted by the Mannheim Centre for European Social Research in June 2008.

NOTES

1. Kathleen Bawn, 'Money and Majorities in the Federal Republic of Germany: Evidence for a Veto Players Model of Government Spending', *American Journal of Political Science* 43/3 (1999), pp.707–36.
2. Thomas König and Thomas Bräuninger, 'The Checks and Balances of Party Federalism: German Federal Government in a Divided Legislature', *European Journal of Political Research* 36/2 (1999), pp.207–34; Thomas König, 'Bicameralism and Party Politics in Germany: an Empirical Social Choice Analysis' *Political Studies* 49 (2001), pp.411–37. See also *German Politics* 17/3 (2008) for a general discussion.
3. Sven-Oliver Proksch and Jonathan B. Slapin, 'Institutions and Coalition Formation: The German Election of 2005', *West European Politics* 29/3 (2006), pp.540–59.
4. Marc Debus, 'Party Competition and Government Formation in Multilevel Settings: Evidence from Germany', *Government and Opposition* 43/4 (2008), pp.505–38.
5. Kenneth Benoit and Michael Laver, *Party Policy in Modern Democracies* (London: Routledge, 2006).
6. Ian Budge, Hans-Dieter Klingemann, Andrea Volkens and Judith Bara, *Mapping Policy Preferences II: Estimates for Parties, Electors and Governments in Central and Eastern Europe, European Union and OECD 1990-2003* (Oxford: Oxford University Press, 2006).
7. Jonathan B. Slapin and Sven-Oliver Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts', *American Journal of Political Science* 52/3 (2008), pp.705–22.
8. Ibid. Wordfish has been implemented in the *R* statistical language. Current code is available at www.wordfish.org.
9. Michael Laver, Kenneth Benoit and John Garry, 'Extracting Policy Positions from Political Texts Using Words as Data', *American Political Science Review* 97/3 (2003), pp.311–32.
10. Ibid., pp. 329–330.
11. Sven-Oliver Proksch and Jonathan B. Slapin, 'Position Taking in European Parliament Speeches,' *British Journal of Political Science* (Forthcoming).
12. See Laver *et al.*, 'Extracting Policy Positions from Political Texts', p.329. Text mining software allows the fast counting of any *n*-gram in any language, such as the free text mining package TM for *R* (Ingo Feinerer, Kurt Hornik and David Meyer, 'Text Mining Infrastructure in R', *Journal of Statistical Software* 25/5 (2008), pp.1–54.
13. Keith Poole and Howard Rosenthal, 'A Spatial Model for Legislative Roll Call Analysis', *American Journal of Political Science* 29/2 (1985), pp.357–84.
14. Joshua Clinton, Simon Jackman and Douglas Rivers, 'The Statistical Analysis of Roll Call Data', *American Political Science Review* 98 (2004), pp.355–70; Andrew D. Martin and Kevin M. Quinn, 'Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999', *Political Analysis* 10 (2002), pp.134–53.
15. Susana Eyheramendy, David Lewis and David Madigan, 'On the naive Bayes model for text categorization', *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, (2003); David D. Lewis, 'Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval', *Proceedings of the 10th European Conference on Machine Learning* (1998), pp.4–15.
16. Andrew McCallum and Kamal Nigam, 'A Comparison of Event Models for Naive Bayes Text Classification', *AAAI-98 Workshop on Learning for Text Categorization*, (1998).
17. Feinerer *et al.*, 'Text Mining Infrastructure in R', p.10.

18. Frederick Mosteller and David L. Wallace, *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*, (Springer Verlag: New York, 1964).
19. Ibid.
20. Kenneth W. Church and William A. Gale, 'Poisson Mixtures', *Natural Language Engineering* 1/2 (1995), pp.163–90.
21. Martin Jansche, 'Parametric Models of Linguistic Count Data', *41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan (2003), pp.288–95.
22. Formally, $y_{ijt} \sim \text{Poisson}(\lambda_{ijt})$, where y_{ijt} is the count of word j in party i's manifesto at time t. The lambda parameter has the systematic component $\lambda_{ijt} = \exp(\alpha_{it} + \psi_j + \beta_j * \omega_{it})$, with α as a set of document (party-election year) fixed effects, ψ as a set of word fixed effects, β as estimates of word specific weights capturing the importance of word j in discriminating between manifestos, and ω as the estimate of party i's position in election year t (therefore it is indexing one specific manifesto). See also Slapin and Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts', for a more detailed discussion.
23. The appendix lists the length of manifestos in words.
24. Since the only data in the model are word counts (the 'dependent variable'), we cannot estimate the parameters on the right-hand side of the equation simultaneously. But given some starting values, we can estimate document parameters conditional on word parameters. This will yield new estimates for the document parameters, which are then used as data to re-estimate word parameters. Such an estimation procedure employed by Wordfish is an iterative process called an Expectation-Maximization algorithm: first party parameters are held fixed at a certain value while word parameters are estimated, then word parameters are held fixed at their new values while the party positions are estimated. This process is repeated until the parameter estimates reach an acceptable level of convergence. For a more detailed description of the estimation process, see ibid.
25. A transformation to the parameters can yield identical log-likelihoods.
26. Douglas Rivers, 'Identification of Multidimensional Spatial Voting Models', *Political Methodology Working Paper*, 2003, available at http://polmeth.wustl.edu; Joshua Clinton, Simon Jackman and Douglas Rivers, 'The Statistical Analysis of Roll Call Data', *American Political Science Review* 98/2 (2004), pp.355–70.
27. Both identification strategies are implemented in the latest release of Wordfish.
28. To do this, document positions are fixed and range between −1.5 and 1.5. The word fixed effects and word discrimination parameters are drawn from normal distributions, and document fixed effects are set as a sequence of values. *R* code to run the simulation is available upon request from the authors.
29. Note that we are varying the number of unique words and not the number of total words in each document. The number of total words is determined by the parameter values we set to generate the data. We fix two values of *omega* to identify the model and fix the extreme *omega*s at −1.5 and +1.5. Therefore, there are no confidence intervals for the two extreme positions as they are excluded from the estimation.
30. Markku Laakso and Rein Taagepera, 'Effective Number of Parties: A Measure With Application to Western Europe', *Comparative Political Studies* 12/1 (1979), pp.3–27.
31. Thomas Saalfeld, 'The German Party System: Continuity and Change', *German Politics* 11/3 (2002), pp.99–130; Charles Lees, 'The German Party System(s) in 2005: A Return to Volkspartei Dominance', *German Politics* 15/4 (2006), pp.361–75.
32. In 1990, the West German Greens failed to surpass the electoral threshold. In contrast, the East German party Bündnis 90/Die Grünen gained parliamentary seats.
33. Slapin and Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts'.
34. Ian Budge, 'The Internal Analysis of Election Programmes', in Ian Budge, David Robertson and Derek Hearl (eds.), *Ideology, Strategy, and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies* (Cambridge University Press, 1987), p.18.
35. Ibid.
36. Michael Laver and Kenneth Benoit, 'Locating TDs in Policy Spaces: Wordscoring Dail Speeches' *Irish Political Studies* 17 (2002), p.65.
37. Michael Laver and John Garry, 'Estimating Policy Positions from Political Texts', *American Journal of Political Science* 44/3 (2000), p.620.
38. Hans-Dieter Klingemann, 'Electoral Programmes in West Germany 1949–1980: Explorations in the Nature of Political Controversy', Budge *et al.*, *Ideology, Strategy, and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*, pp.294–323.
39. Ibid., p.300.
40. Ibid.
41. The list is based on appendix V included with the CMP dataset, see Ian Budge, Hans-Dieter Klingemann, Andrea Volkens and Judith Bara, *Mapping Policy Preferences II: Estimates for Parties, Electors*

and Governments in Central and Eastern Europe, European Union and OECD 1990–2003 (Oxford: Oxford University Press, 2006).

42. Klingemann, 'Electoral Programmes in West Germany', p.301.
43. Laver *et al.* 'Extracting Policy Positions from Political Texts' but see Sven-Oliver Proksch and Jonathan B. Slapin, 'Institutions and Coalition Formation: The German Election of 2005' for an alternative strategy that uses only policy-area specific subsets of the reference manifestos.
44. We agree with the Wordscores approach on this need.
45. Slapin and Proksch, 'A Scaling Model for Estimating', pp.712–713.
46. Making sure the file is saved as unicode (UTF-8) ensures cross-platform compatibility and that non-English characters, such as German vowels with umlauts, are preserved in a readable format. One should be careful to read the documents after scanning them to ensure that characters were encoded correctly and that all parts of the document were properly scanned.
47. Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara and Eric Tanenbaum (eds.), *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998* (Oxford: Oxford University Press, 2001); Budge *et al. Mapping Policy Preferences II.*
48. The Comparative Electronic Manifestos Project is directed by Paul Pennings and Hans Keman, Vrije Universiteit Amsterdam, in collaboration with the Zentralarchiv für Empirische Sozialforschung, Universität zu Köln.
49. We thank Thomas König and Bernd Luig for making these files available to us. Each text file includes the full text of the manifesto listed in the appendix. Some texts include data that researchers may prefer to remove prior to the estimation. Examples include the listing of speakers or party names, self-reference of party names, headers and footers, enumeration, bullets, section headings, etc. This can either be done manually or with the help of pattern-matching using customized PERL or PYTHON scripts. Additionally, software specifically designed for text processing is available to perform many of these pre-processing tasks. This software will remove punctuation, numbers, and stop-words (i.e. words defined by the researcher that should be systematically removed from the text). In addition, these programs can change all capital letters to lower case letters, so words are not counted differently merely due to capitalization. In addition, researchers should also ensure that the spelling of words is consistent across documents. This may be particularly problematic in German given the recent reform of German spelling. In the version that we use, titles, candidate-oriented preambles, headers, footers, and indices were removed, and spelling and grammar corrected.
50. The construction of the word count matrix is done using the *R* text mining package TM. This text mining package allows the removal of customized stopwords, the use of dictionaries, and the consideration of bigrams instead of unigrams. Alternative word count procedures include Jfreq or Yoshikoder.
51. Feinerer *et al.*, 'Text Mining Infrastructure in R', p. 25. The TM package includes a list of 264 German stop words, but only a small proportion is contained in the manifestos.
52. In the Congressional and US Supreme Court ideal point estimation literature, legislators' positions over time either are a polynomial function of previous positions (D-NOMINATE, see Keith T. Poole and Howard Rosenthal, *Congress: A Political-Economic History of Roll Call Voting* (New York: Oxford University Press, 1997) or defined by a random walk process, see Andrew D. Martin and Kevin M. Quinn, 'Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999', *Political Analysis* 10 (2002), pp.134–53; Andrew D. Martin and Kevin M. Quinn, 'Assessing Preference Change on the U.S. Supreme Court', *Journal of Law, Economics, and Organization* 23 (2007), 303–325). Wordfish does not place any such constraint on the data.
53. Burt L. Monroe, Michael P. Colaresi and Kevin M. Quinn, 'Fightin' Words: Lexical Selection and Evaluation for Identifying the Content of Political Conflict', *Political Analysis* (Forthcoming).
54. A prior on the distribution of word weights in Wordfish constrains the range of estimated values.
55. The model is identified by fixing the mean position at 0 and the standard deviation at 1 and by constraining the FDP in 1990 to have a smaller value than the PDS in 1990.
56. Laver *et al.*, 'Extracting Policy Positions from Political Texts'.

APPENDIX 1
GERMAN PARTY DOCUMENTS, 1961–2005

Election year	Party	Type	Document	Length (words)
1961	CDU-CSU	Election proclamation	Kölner Manifest 1961	426
	SPD	Party congress speeches	Das Regierungsprogramm der SPD. Ausserordentlicher Kongress der SPD, 28.4.1961. Reden von Carlo Schmid, Erich Ollenhauer, Herbert Wehner, Willy Brandt und Waldemar von Knoeringen.	6,059
	FDP	Election manifesto	Aufruf zur Bundestagswahl 1961	2,364
1965	CDU-CSU	Party congress speeches	Die Düsseldorfer Erklärung der CDU, Rednerdienst, Mai 1965.	9,744
	SPD	Election manifesto	Tatsachen und Argumente. Erklärungen der SPD Regierungsmannschaft 1965	22,033
	FDP	Action programme	Ziele des Fortschritts: Aktionsprogramm der Freien Demokratischen Partei (107 Thesen), April 1967	4,224
1969	CDU-CSU	Election manifesto	Sicher in die 70er Jahre: Wahlprogramm der CDU 1969–1973	2,262
	SPD	Election manifesto	Regierungsprogramm der SPD 1969	2,969
	FDP	Election manifesto	Praktische Politik für Deutschland – Das Konzept der FDP, Juni 1969	4,224
1972	CDU-CSU	Election manifesto	Wir bauen den Fortschritt auf Stabilität, CDU Regierungsprogramm 1972	3,620
	SPD	Election manifesto	Mit Willy Brandt für Frieden, Sicherheit und eine bessere Qualität des Lebens. Wahlprogramm der SPD, Oktober 1972	11,765
	FDP	Election proclamation	Vorfahrt für Vernunft: Wahlaufruf 1972	821
1976	CDU-CSU	Election manifesto	Aus Liebe zu Deutschland. Für die Freiheit, die wir lieben. Für die Sicherheit, die wir brauchen. Für die Zukunft, die wir wollen. Das Wahlprogramm der CDU und CSU 1976	6,434
	SPD	Election manifesto	Weiter Arbeiten am Modell Deutschland: Regierungsprogramm 1976–1980	15,178
	FDP	Election manifesto	Die F.D.P.: Die liberale Alternative. Wahlprogramm, Mai 1976	7,685
1980	CDU-CSU	Election manifesto	Für Frieden und Freiheit: Das Wahlprogramm der CDU/CSU 1980	10,657
	SPD	Election manifesto	Sicherheit für Deutschland: Wahlprogramm 1980	8,834
	FDP	Election manifesto	Unser Land soll auch morgen liberal sein: Wahlprogramm 80	23,040
1983	CDU-CSU	Election manifesto	Arbeit, Frieden, Zukunft. Miteinander schaffen wir's. Das Wahlprogramm der CDU/CSU 1983	4,865
	SPD	Election manifesto	Das Regierungsprogramm der SPD 1983–1987	9,906
	FDP	Election manifesto	Wahlaussage zur Bundestagswahl 1983	7,117

(*Continued*)

APPENDIX 1
CONTINUED

Election year	Party	Type	Document	Length (words)
	Grüne	Election manifesto	Diesmal: Die Grünen – warum? Ein Aufruf zur Bundestagswahl 1983	4,150
1987	CDU-CSU	Election manifesto	Das Wahlprogramm von CDU and CSU für die Bundestagwahl 1987	16,501
	SPD	Election manifesto	Zukunft für alle – arbeiten für soziale Gerechtigkeit und Frieden: Regierungsprogramm 1987–1990 der Sozialdemokratischen Partei Deutschlands	9,239
	FDP	Election manifesto	Die Liberalen: Wahlplattform '87, September 1986	5,642
	Grüne	Election manifesto	Farbe bekennen: Bundestagswahlprogramm 1987	16,911
1990	CDU-CSU	Election manifesto	Ja zu Deutschland. Ja zur Zukunft. Wahlprogramm der CDU Deutschlands zur gesamtdeutschen Bundestagswahl am 2. Dezember 1990	5,269
	SPD	Election manifesto	Regierungsprogramm 1990–1994. Der Neue Weg: ökologisch, sozial, wirtschaftlich stark	7,240
	FDP	Election manifesto	Wahlaufruf der F.D.P. zur Bundestagswahl am 2.12.1990	25,503
	Bündnis90/ Die Grünen	Election manifesto	Bündnis 90/Die Grünen: Wahlplattform 1990 [Note: manifesto of East German Bündnis 90]	4,089
	Linke Liste/ PDS	Election manifesto	Wahlprogramm der Linken Liste/ PDS	8,978
1994	CDU-CSU	Election manifesto	Regierunsprogramm von CDU und CSU 1994	10,930
	SPD	Election manifesto	Das Regierungsprogramm der SPD 1994: Reformen für Deutschland	13,761
	FDP	Election manifesto	Liberal denken. Leistung wählen: Wahlprogramm 1994	38,548
	Bündnis90/ Die Grünen	Election manifesto	Nur mit uns: Programm zur Bundestagswahl 1994	28,875
	PDS	Election manifesto	Wahlprogramm der PDS 1994	6,477
1998	CDU-CSU	Election manifesto	Wahlplattform von CDU und CSU	7,814
	SPD	Election manifesto	Arbeit, Innovation und Gerechtigkeit: SPD-Programm für die Bundestagswahl 1998	12,966
	FDP	Election manifesto	Es ist Ihre Wahl. Das Wahlprogramm der F.D.P. zur Bundestagswahl 1998	22,354
	Bündnis90/ Die Grünen	Election manifesto	Neue Mehrheiten nur mit uns, 1998–2002. Vier Jahre für einen politischen Neuanfang [Note: short version of election manifesto, the CMP used the long version]	3,987
	PDS	Election manifesto	Programm der PDS zur Bundestagswahl 1998. Für den politischen Richtungswechsel! Sozial und solidarisch – für eine gerechte Republik!	14,257

(*Continued*)

APPENDIX 1
CONTINUED

Election year	Party	Type	Document	Length (words)
2002	CDU-CSU	Election manifesto	Leistung und Sicherheit. Zeit für Taten. Regierungsprogramm 2002–2006 von CDU und CSU	18,851
	SPD	Election manifesto	Erneuerung und Zusammenhalt – Wir in Deutschland. Regierungsprogramm 2002–2006.	19,123
	FDP	Election manifesto	Bürgerprogramm. Programm der FDP zur Bundestagswahl 2002	30,022
	Bündnis90/ Die Grünen	Election manifesto	Grün wirkt! Unser Wahlprogramm 2002–2006	21,384
	PDS	Election manifesto	Es geht auch anders: Nur Gerechtigkeit sichert Zukunft! Programm der PDS zur Bundestagswahl 2002	12,977
2005	CDU-CSU	Election manifesto	Deutschlands Chancen nutzen. Wachstum. Arbeit. Sicherheit. Regierungsprogramm 2005–2009	9,972
	SPD	Election manifesto	Vertrauen in Deutschland – Das Wahlmanifest der SPD	11,351
	FDP	Election manifesto	Arbeit hat Vorfahrt. Deutschlandprogramm 2005	20,069
	Grüne	Election manifesto	Eines Für Alle: Das Grüne Wahlprogramm 2005	26,504
	Linkspartei/ PDS	Election manifesto	Für eine neue soziale Idee	7,939