

Computers, coders, and voters: Comparing automated methods for estimating party positions

Frederik Hjorth¹, Robert Klemmensen², Sara Hobolt³,
 Martin Ejnar Hansen⁴, Peter Kurrild-Klitgaard⁵

Abstract

Assigning political actors positions in ideological space is a task of key importance to political scientists. In this paper we compare estimates obtained using the automated *Wordscores* and *Wordfish* techniques, along with estimates from voters and the Comparative Manifesto Project (CMP), against expert placements. We estimate the positions of 254 manifestos across 33 elections in Germany and Denmark, two cases with very different textual data available. We find that *Wordscores* approximately replicates the CMP, voter, and expert assessments of party positions in both cases, whereas *Wordfish* replicates the positions in the German manifestos only. The results demonstrate that automated methods can produce valid estimates of party positions, but also that the appropriateness of each method hinges on the quality of the textual data. Additional analyses suggest that *Wordfish* requires both longer texts and a more ideologically charged vocabulary in order to produce estimates comparable to *Wordscores*. The paper contributes to the literature on automated content analysis by providing a comprehensive test of convergent validation, in terms of both number of cases analyzed and number of validation measures.

Keywords

Text as data, automated content analysis, party positions, *Wordscores*, *Wordfish*, CMP

Introduction

Assigning party positions to political actors in a policy space is a task of key importance to political scientists. Since spatial models of political behavior are the preferred workhorse of many branches of empirical political science, obtaining valid party position estimates is essential (Cox, 1999). Recently, the use of textual data has regained prominence with the advent of techniques for automated text analysis (Grimmer and Stewart, 2013). By using automated content analysis methods, the huge costs traditionally associated with coding text are dramatically diminished, which opens a host of new avenues for data collection. However, whenever a new content analysis method is proposed, validation is needed in order to ensure that it can be appropriately applied to texts and contexts of interest.

In this paper we evaluate two prominent automated content analysis techniques. The first of these is the *Wordscores* approach developed by Laver et al. (2003). The second

method is the more recently developed *Wordfish* approach (Slapin and Proksch, 2008). Our study contributes to the literature by providing the first systematic test of the validity of each method's estimates of party policy positions based on the same manifestos. We use a very rich set of textual data, allowing us to estimate a total of 254 party positions across 33 elections. In addition to testing each method on a large number of party manifestos, we cross-validate the estimates

¹University of Copenhagen, Denmark

²University of Southern Denmark, Denmark

³London School of Economics, UK

⁴Brunel University, UK

⁵University of Copenhagen, Denmark

Corresponding author:

Frederik Hjorth, Department of Political Science, University of Copenhagen, Øster Farimagsgade 5, opgang E, Copenhagen K, DK-1353, Denmark.

Email: fh@ifs.ku.dk



against three different benchmarks, which provides assurance that our findings are robust to the specific benchmark measure used.

Important past scholarship has demonstrated that legislative speech can be used to estimate the positions of members of the US Senate and the UK House of Commons. Following Beauchamp (2012), we find that Wordscores correlates more strongly with benchmark measures than Wordfish. We seek to test empirically which method, that of Wordscores or Wordfish, is best at predicting the party positions expressed in party manifestos in multiparty systems over long periods of time. In doing so, we are able to provide guiding principles for researchers in search of valid estimates that can be derived from manifestos or similar textual sources.

The paper proceeds as follows. Each of the methods is presented in the following section. Thereafter we discuss the data and turn to a comparison between the obtained scores from the automated procedures and positions obtained through expert surveys, voter placements of parties, and the CMP RILE measure, which is obtained through hand-coding of party manifestos. In this context, we conceive of ideology as a latent variable. The best way to validate the automated techniques is therefore to compare their estimates with a number of independent measures. As such, although we can never be 100 percent certain that we are measuring ideology correctly, we can have good reason to believe we are capturing this latent variable if we have several different measures using independent sources of data that correlate highly.

We conclude that, across contexts, the supervised Wordscores is the most promising automated content analysis method for placing political actors, achieving estimates with high cross-validity with our three independent data sources. However, when party manifestos are long and when their vocabularies are politically polarized, the unsupervised Wordfish produces estimates on a par with those of Wordscores. Consequently, the choice of which method to use should hinge on the quality of the textual data available vis-a-vis the availability of prior information to produce estimates.

Data

Our analysis relies on the manifestos of Danish and German political parties. We cover 24 elections from 1945 to 2007, which amounts to 212 manifestos in total. In the German case we analyze a period spanning from 1980 to 2009 over nine elections and 42 manifestos. The set of analyzed manifestos exactly matches those analyzed by the CMP.

Table 1 provides summary statistics for the textual data analyzed.

The cases of Denmark and Germany are suitable for our analysis for several reasons. Here, we highlight four. First of all, the Danish party system underwent dramatic change during the period we are investigating (Pedersen 1979). In contrast, the German party system has exhibited considerable continuity for the past four decades, even accounting

Table 1. Summary stats for German and Danish manifesto data.

	Germany	Denmark
Elections	9	24
Avg. manifestos per election	4.7	8.2
Avg. manifesto length (no. of words)	10,306	1,232.1
Std. dev. manifesto lengths	5,502.5	1,377.7

for unification (Saalfeld 2002). This allows us to test the ability of the different techniques to detect dramatic changes in a political space.

Second, as shown in Table 1, the average number of parties per election in Denmark is 8.2, far higher than the 4.7 average in Germany. As the number of parties increases, the a priori difficulty of correctly placing parties within any given election increases factorially.¹ This should, *ceteris paribus*, make the Danish data a harder test of each method's ability to retrieve the correct left-right ordering of parties within each election.

Third, as shown in Table 1, Danish manifestos are approximately eight times shorter than the German manifestos. Accordingly, it should be difficult for the automated techniques to place the parties correctly on a political dimension. Proksch, Slapin, and Thies (2011) show that Wordfish is capable of placing Japanese political parties correctly relative to CMP measures despite the manifestos being relatively short. By using a new source of relatively short manifestos, this study provides an additional test of each technique's ability to estimate positions under conditions of limited information.

Fourth, the two cases differ in terms of the ideological polarization of political lexica. Grimmer and Stewart (2013) suggest that one reason why they are unable to reproduce valid left-right estimates applying Wordfish to the US Congressional record is that the vocabulary used in the German manifestos is more ideologically charged. We revisit the question of the role of ideologically charged political lexica at the end of the article.

Methods

Automated methods compared: Wordscores and Wordfish

Several methods have been developed to place political parties. Mair (2001) and Benoit and Laver (2007) provide excellent discussions of the various methods, and these authors also discuss the pros and cons of using different data sources. In this article we confine ourselves to discussing methods based on content analysis of text, specifically party manifestos.

A key difference between the two techniques analyzed here is that Wordscores is a *supervised* method; that is, it requires prior information to produce estimates. Wordfish,

on the other hand, is *unsupervised*; that is, it produces estimates using only the information available in the textual data itself.

The Wordscores method was the first automated but supervised method of content analysis to gain influence within political science (Laver et al., 2003). The starting point is to assign position scores to a set of “reference texts” whose positions are known. Since the frequencies of words contained in these texts are also known, it is possible to calculate the value of each word that occurs in the reference texts. The values of these words are calculated as averages of reference document scores, weighted by the posterior probability of each document given that the word in question occurs within it. Policy position estimates for the “virgin texts”—the documents whose positions are to be estimated—are then computed as the mean of the scores of the words in the reference text, weighted by their relative frequencies within the virgin texts (Lowe, 2008). The Wordscores technique therefore requires *ex ante* available estimates of the positions of the reference texts on the policy dimension under investigation.

A more recent approach to placing political parties is the Wordfish approach. The Wordfish method has one considerable advantage over the Wordscores method: It does not depend on documents with *ex ante* assigned reference scores. Position estimates derived using Wordfish are based only on the information in the texts. This lack of an *ex ante* defined dimensionality is a double-edged sword: while Wordfish scales texts independently of prior information, it renders uncertain the exact nature of the dimension being estimated. One important drawback of unsupervised algorithms is thus that the nature of the dimensions produced requires intensive validation before they can be applied across different sets of texts and contexts (Grimmer and Stewart, 2013).

The only input required by Wordfish is a word frequency matrix that lists the frequency of each word across all documents. The Wordfish procedure then fits the frequency of each word j in party i 's manifesto to a standard Poisson count model, where the mean and variance λ_{ij} in each document is assumed to be a function of the policy position of the document, represented by the parameter ω_i .² Along with the policy position parameter, the model includes fixed-effects terms for words and document length.

Regardless of the specific approach used, there are several reasons why computer-based estimation procedures are preferable to those based on human coders. For one, the reliability of the obtained scores is dramatically enhanced (Grimmer and Stewart, 2013; Mikhaylov et al., 2012). Furthermore, the costs of conducting the content analysis are reduced dramatically, which in turn vastly increases the text universe and the number of actors for which positions can be estimated. This leaves the crucial question of whether the computer-based procedures can produce valid measurements.

In order to assess the validity of the two automated approaches to measuring party positions, we compare each

method with three alternative measures based on experts and voter surveys and the widely used CMP dataset, which is also based on manifestos.³

Validation measures: experts, voters, and CMP estimates

We analyze only the general left–right dimension. The reason for this choice is mainly empirical. It is the dominant dimension of party competition in European politics; moreover, it is widely accepted that party competition in Danish politics has been predominantly uni-dimensional and that the economic dimension has been very salient in Danish politics (Hansen, 2008; Klemmensen et al., 2007). Similarly, while lower-order dimensions can be identified, the economic dimension dominates German party politics (Debus, 2008a; Proksch and Slapin, 2009). In principle, the analysis could be extended to cover more dimensions.

As we conceive ideology to be a latent variable, the best way to validate the text-based techniques is to compare several independent measures. The first measure we use is expert survey estimates. As Benoit and Laver (2006) argue, average expert judgments can be interpreted as a local consensus on the relative positions of the parties in a party system. We follow Klemmensen et al. (2007) and use the party expert positions reported by Damgaard (2000). For the manifestos from the 2005 and 2007 elections, we use the Chapel Hill expert survey (Hooghe et al., 2010). The expert surveys we use do not provide interval-level point estimates of party positions, only ordinal rankings of party positions. Consequently we rely on the rank order correlation measure of Spearman's ρ as the measure of association.

The second measure is the aggregated voter left–right placement of parties. We use the question used in Eurobarometer surveys where respondents are asked to place themselves on a 1–10 left–right scale. For each party, the voter-assigned position is the value of this scale averaged across all respondents intending to vote for that party. For each election year, we use estimates from the nearest Eurobarometer survey prior to the election.

Finally, we validate the two automated techniques against the positions retrieved by the CMP data set's RILE measure (Volkens, 2013). The positions have been retrieved by hand-coding manifestos from 55 countries since 1945. For decades, the CMP data has had close to monopoly status in the comparative study of political processes, and consequently we believe that they serve as a sensible benchmark for the automated techniques.

Assigning Wordscores reference values

As discussed above, the Wordscores method requires that we assign scores to a set of texts that serve as reference texts, providing information on which values should be

Table 2. Summary stats for reference and virgin texts.

	Denmark	Germany
Reference text election year(s)	1947, 1975	1998
Avg. manifesto length (no. of words), reference texts	861.5	11984.4
Avg. manifesto length (no. of words), virgin texts	1264.4	10079.2

assigned to the individual words. For the Danish case we follow Klemmensen et al. (2007) and use the expert estimates for 1947 and 1975 elections as references.⁴ For the German case, we use the estimates from Debus (2008b) for Germany's 1998 election as references.⁵ In both cases, reference texts are excluded from the set of virgin texts estimated by Wordscores such that they do not artificially increase the correlations.

Table 2 shows the descriptive statistics for the reference texts and the virgin texts in their full and reduced forms. As can be seen, the German texts are considerably longer than their Danish counterparts, making it possible for us to assess the robustness of Wordscores to the use of reference texts of varying length.

Pre-processing

Both sets of manifestos are preprocessed according to standard procedures in quantitative text analysis: Numbers, punctuation, and white space are removed from each document, all words are converted to lowercase, and the most common words, so-called "stop words," are removed. Finally, words are stemmed using Danish and Germanic adaptations for the Porter Stemmer algorithm (Porter, 1980).⁶

Results

The Spearman's ρ correlation measure we apply to the methods is rather forgiving: A method only has to get the ordinal ranking of the parties right—in the sense of being equivalent to the benchmark measure—in order to achieve a perfect correlation. Hence, using rank order correlations ensures that the observed correlation measures do not rely on unreasonable assumptions, while remaining a conservative test of the validity of each method.

Figures 1 and 2 plot the Spearman's ρ correlation between Wordscores and Wordfish and each of the three benchmark measures for Denmark and Germany respectively. The vertical black line in each plot represents the average correlation across years. For Wordfish, which does not automatically produce a left–right direction of the estimates, we have chosen the direction that maximizes the average correlation. Table 3 summarizes average correlations and reports the computation time spent producing the estimates.

Figure 3 provides an alternative perspective on the data, yet reveals a similar pattern. The figure shows all estimates

from CMP, voters, Wordfish and Wordscores plotted against expert estimates across all elections. In order to ease comparisons across different scales, all measures are standardized within country-years.

In the case of Denmark (top row in Figure 3), Wordfish stands out from the rest, with a noticeably noisier association with expert estimates compared to the other three measures. In other words, across all observations in the Danish data, Wordfish estimates exhibit weak cross-validity with expert estimates compared to the other measures, Wordscores included.

However, results in the case of Germany (bottom row) are quite different. Wordfish and Wordscores are both strongly associated with expert estimates in the German case, and equally so. In other words, Wordfish performs at least as well as Wordscores when applied to the German data.

Difference in cross-validity: manifesto length vs. ideological strength

The difference in cross-validity obtained for Wordfish when comparing Denmark and Germany is probably attributable to the two key differences between the manifesto data sets described above: The greater length of the German manifestos and the ideological strength of the language. In order to assess the relative contribution of each, we re-ran the analysis on a set of artificially "laconic" German manifestos reduced by randomly sampling a number of words from each manifesto such that the average length is comparable to the Danish data. The results, reported in the online appendix, show that a substantial part, but not all, of the relative advantage of the German data vanishes when using the subsampled manifesto data. As a rough approximation, this indicates that around half of the relative advantage for the German Wordfish estimates is attributable to the difference in average manifesto length, the other half stemming from more ideologically charged language.

Discussion and conclusion

This study set out to test the validity of two prominent automated methods for estimating party policy positions on two sets of manifesto text data that differed in the number of party positions to be estimated, their average text length, and the strength of their ideological lexicon.

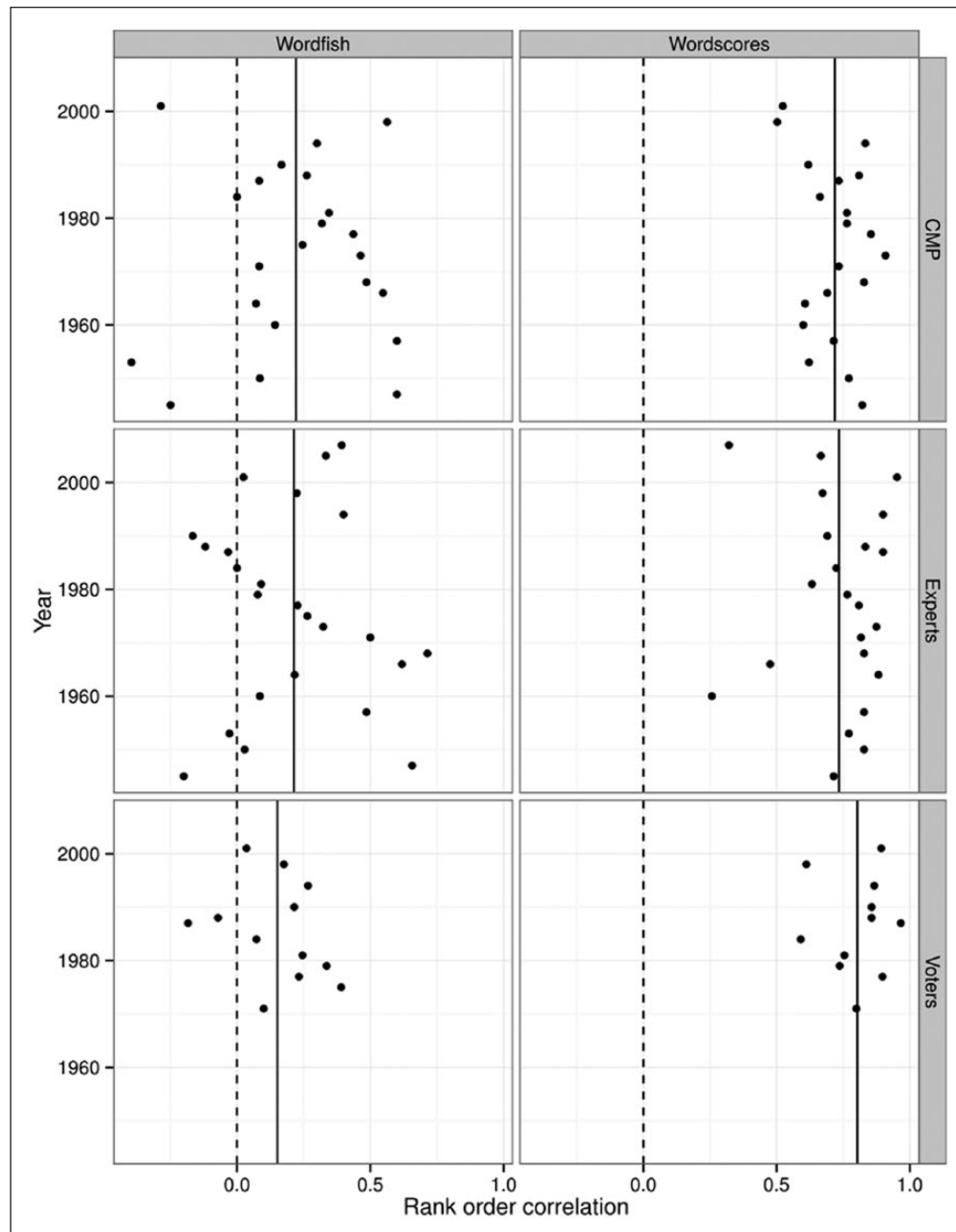


Figure 1. Wordfish and Wordscores estimates' rank order correlations with CMP, expert and voter estimates for each election year in the Danish sample. Vertical lines signify average rank order correlation across years.

While each method was tested against three benchmarks, the results were remarkably consistent across all three: When tested against the Danish data, Wordscores clearly outperforms Wordfish, exhibiting noticeably stronger rank order correlations with each of the three benchmark measures. In the German case, however, Wordscores and Wordfish perform equally well, correlating strongly with each of the three benchmarks.

How to explain this rather stark difference across the two cases? One likely explanation is that the lower

threshold for number of words necessary for Wordfish to estimate positions accurately lies somewhere in the considerable gap between the (comparatively short) Danish manifestos and the (comparatively long) German manifestos. Yet, as indicated by the analysis using the artificially “laconic” German manifestos, the difference is not due to length alone. The relatively better performance of Wordfish on German manifestos even when holding length constant illustrates the informational value (for estimation purposes) of a politically polarized vocabulary.

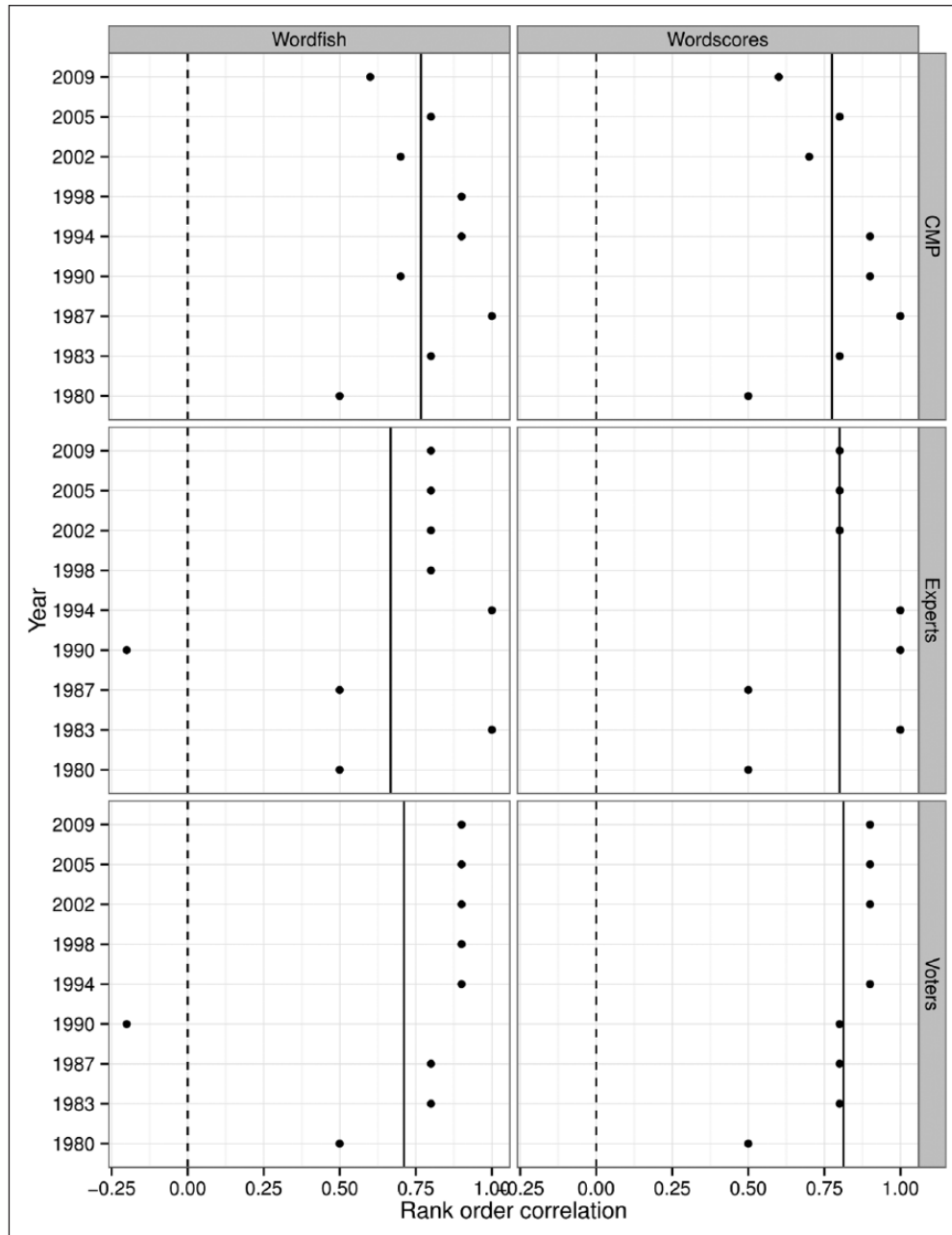


Figure 2. Wordfish and Wordscores estimates' rank order correlations with CMP, expert and voter estimates for each election year in the German sample. Vertical lines signify average rank order correlation across years.

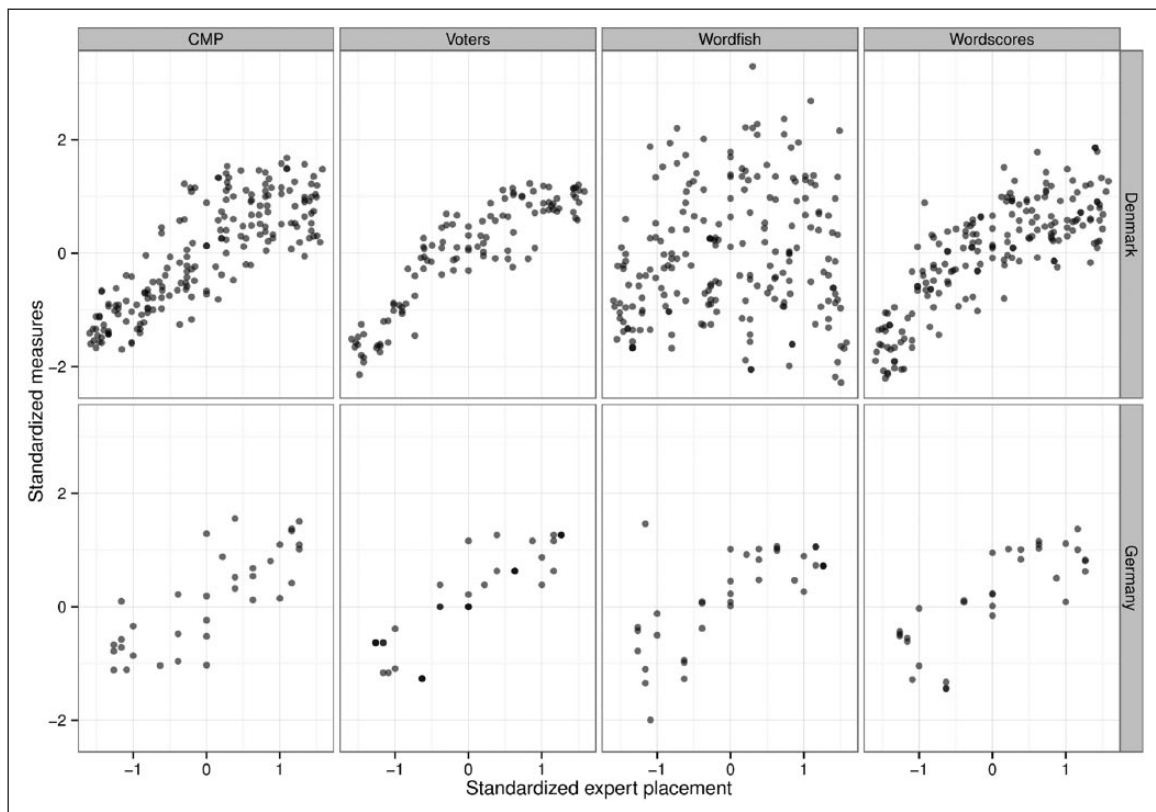
It is arguably an important advantage of Wordfish that it does not require prior information in order to estimate text positions. This feature has the double advantage of making Wordfish applicable in contexts where no such priors are available and rendering the estimation process fully data-driven rather than partly dependent on researcher input. However, as the results of this study show, this advantage comes at a cost. When the available textual data is less plentiful and less ideologically polarized than is the case for German manifestos, the validity of Wordfish estimates

suffers while Wordscores estimates remain reasonably valid.

In Table 4 we summarize our recommendations for researchers based on the results of this study. The recommendations should be read as a summary of the findings presented here. As such, their applicability in novel contexts is uncertain. With this caveat in mind, they may provide some guidance for researchers. We recommend that researchers use Wordscores if some *ex ante* position estimates are available; if not, use Wordfish, provided text

Table 3. Average rank order correlations with benchmark measures, Wordscores and Wordfish.

Country	Method	Estimation time (secs)	Benchmark	Avg. ρ
Denmark	Wordfish	75.6	CMP	0.2
			Experts	0.2
			Voters	0.1
	Wordscores	0.01	CMP	0.7
			Experts	0.8
			Voters	0.8
Germany	Wordfish	904.3	CMP	0.8
			Experts	0.7
			Voters	0.7
	Wordscores	0.04	CMP	0.8
			Experts	0.8
			Voters	0.8

**Figure 3.** Expert estimates plotted against standardized estimates from CMP, voter surveys, Wordfish, and Wordscores. Estimates are standardized within each country-year. Dots are jittered and semitransparent in order to make varying dot densities clearer.**Table 4.** Summary of recommendations.

Conditions	Recommendation
Some <i>ex ante</i> position estimates	Wordscores
No <i>ex ante</i> position estimates	Long and ideologically polarized texts
	Short and ideologically similar texts
	Gather more data

length and ideological polarization are at least on par with the German data tested here. If this latter condition is not fulfilled, researchers should seek more data.

Based on this study, neither of the two methods can be said to be superior regardless of context. Researchers would be wise to take careful account of the context and quality of the textual data available if and when settling on an automated method for estimating party positions.

Acknowledgements

The authors would like to thank Marc Debus, Zoltán Fazekas, and the two anonymous reviewers for helpful comments on earlier drafts of this manuscript.

Declaration of conflicting interest

The authors declare that there is no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Supplementary material

The online appendix is available at: <http://rap.sagepub.com/content/by/supplemental-data>

Notes

1. Specifically, in the typical German election, the probability of randomly ordering parties correctly is equal to $4!^{-1}$ or 1 in 24; in the typical Danish election, the equivalent probability is equal to $8!^{-1}$ or 1 in 40,320.
2. Slapin and Proksch (2008) denote this parameter ω_{it} . This suggests that Wordfish explicitly models temporal variation, which is not the case. Hence, we denote the position parameter ω_i .
3. We have also used various scaling techniques on more than 8000 roll calls. We get the same results as presented here but, due to the impossibility of estimating the position of governing parties with such data, we have not included the analysis in this paper. The results are available upon request.
4. The two years are chosen to account for the 1973 “earthquake” election. In 1947, parties are assigned the following scores: Communists 0, Social Democrats 3, Social Liberals 5, Justice Party 7, Conservatives 8, Liberals 10. In 1975, parties are assigned the following scores: Communists 0, Left Socialists 1, Socialists 2, Social Democrats 3, Social Liberals 5, Center Democrats 6, Christian People’s Party 7, Justice Party 7.5, Liberals 8, Conservatives 9, Progress Party 10. We have run estimations using different reference texts and obtained similar results to the ones presented here, provided that we include manifestos prior and subsequent to the 1973 election.
5. Parties are assigned the following scores: PDS 3.9, Green Party 8.8, SPD 9.4, CDU/CSU 14, FDP 17.5.
6. Results when removing rare words are available upon request. Both Wordscores and Wordfish are estimated using the package Austin for the statistical software R (Lowe, 2011).

References

- Bakker R, De Vries CE, Edwards EE, et al. (2012) Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010. *North*. Available at: http://www.unc.edu/~hooghe/data_pp.php.
- Beauchamp N (2012) Using text to scale legislatures with uninformative voting. Working paper. Available at: http://nick-beauchamp.com/work/Beauchamp_scaling_current.pdf
- Benoit K and Laver M (2006) *Party Policy in Modern Democracies*. London: Routledge.
- Benoit K and Laver M (2007) Estimating party policy positions: Comparing expert surveys and hand-coded content analysis. *Electoral Studies* 26(1): 90–107, doi:10.1016/j.electstud.2006.04.008.
- Benoit K, Laver M and Mikhaylov S (2009) Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science* 53(2): 495–513, doi:10.1111/j.1540-5907.2009.00383.x.
- Budge I and Farlie D (1983) *Explaining and Predicting Elections: Issue Effects and Party Strategies in Twenty-three Democracies*. London: Allen & Unwin.
- Cox G (1999) Electoral rules and the calculus of mobilization. *Legislative Studies Quarterly* 24(3): 387–419, doi:10.2307/440350.
- Damgaard E (2000) Denmark: The life and death of government coalitions. In: Müller WC and Strøm K (eds) *Coalition Governments in Western Europe (Comparative Politics)*. New York: Oxford University Press, pp.231–264.
- Debus M (2008a) Unfulfilled promises? German Social Democrats and their policy positions at the federal and state level between 1994 and 2006. *Journal of Elections, Public Opinion & Parties* 18(2): 201–224, doi:10.1080/17457280801987926.
- Debus M (2008b) Party competition and government formation in multilevel settings: Evidence from Germany 1. *Government and Opposition* 43(4): 505–538, doi:10.1111/j.1477-7053.2008.00267.x.
- Grimmer J and Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3): 267–297, doi:10.1093/pan/mps028.
- Hansen ME (2008) Back to the archives? A critique of the Danish part of the manifesto dataset. *Scandinavian Political Studies* 31(2): 201–116, doi:10.1111/j.1467-9477.2008.00202.x.
- Klemmensen R, Hobolt SB and Hansen ME (2007) Estimating policy positions using political texts: An evaluation of the Wordscores approach. *Electoral Studies* 26(4): 746–755, doi:10.1016/j.electstud.2007.07.006.
- Laver M and Garry J (2000) Estimating policy positions from political texts. *American Journal of Political Science* 44(3): 619–634, doi:10.2307/2669268.
- Laver M, Benoit K and Garry J (2003) Extracting policy positions from political texts using words as data. *American Political Science Review* 97(2): 311–331, doi:10.2307/3118211.
- Lowe W (2008) Understanding Wordscores. *Political Analysis* 16(4): 356–371, doi:10.1093/pan/mpn004.
- Lowe W (2011) Austin: Do things with words. Software: R package version 2.0.
- Mair P (2001) Searching for the positions of political actors: A review of approaches and a critical evaluation of expert

- surveys. In: Laver M (ed) *Estimating the Policy Position of Political Actors*. London: Routledge, pp.10–30.
- Mikhaylov S, Laver M and Benoit KR (2012) Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis* 20(1): 78–91, doi:10.1093/pan/mpr047.
- Mikhaylov S, Laver M and Benoit KR (1979) The dynamics of European party systems: Changing patterns of electoral volatility. *European Journal of Political Research* 7(1): 1–26, doi:10.1111/j.1475-6765.1979.tb01267.x.
- Porter MF (1980) An algorithm for suffix stripping. *Program* 3(14): 130–137.
- Proksch S-O and Slapin JB (2009) How to avoid pitfalls in statistical analysis of political texts: The case of Germany. *German Politics* 18(3): 323–344, doi:10.1080/09644000903055799.
- Proksch S-O, Slapin JB and Thies MF (2011) Party system dynamics in post-war Japan: A quantitative content analysis of electoral pledges. *Electoral Studies* 30(1): 114–124, doi:10.1016/j.electstud.2010.09.015.
- Saalfeld T (2002) The German party system: Continuity and change. *German Politics* 11(3): 99–130, doi:10.1080/714001303.
- Slapin JB and Proksch S-O (2008) A scaling model for estimating time-series party positions from texts. *American Journal of Political Science* 52(3): 705–722, doi:10.1111/j.1540-5907.2008.00338.x.
- Volgens A (2013) *The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2013b*. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB).