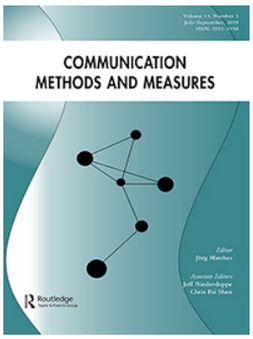


Routledge
Taylor & Francis Group

Communication Methods and Measures

ISSN: 1931-2458 (Print) 1931-2466 (Online) Journal homepage: https://www.tandfonline.com/loi/hcms20

Validating Wordscores: The Promises and Pitfalls of Computational Text Scaling

Bastiaan Bruinsma & Kostas Gemenis

To cite this article: Bastiaan Bruinsma & Kostas Gemenis (2019) Validating Wordscores: The Promises and Pitfalls of Computational Text Scaling, Communication Methods and Measures, 13:3, 212-227, DOI: 10.1080/19312458.2019.1594741

To link to this article: https://doi.org/10.1080/19312458.2019.1594741

Published online: 17 Apr 2019.

COMMUNICATION METHODS AND MEASURES
2019, VOL. 13, NO. 3, 212–227
https://doi.org/10.1080/19312458.2019.1594741

Routledge
Taylor & Francis Group

Validating Wordscores: The Promises and Pitfalls of Computational Text Scaling

Bastiaan Bruinsma [a] and Kostas Gemenis [b]

[a]Department of Political and Social Sciences, Scuola Normale Superiore, Florence, Italy; [b]Max Planck Institute for the Study of Societies, Cologne, Germany

ABSTRACT

Wordscores is a popular computational text analysis method with numerous applications in communication research. *Wordscores* claims to scale documents on specified dimensions without requiring researchers to read or even understand the language of the input text. We investigate whether *Wordscores* delivers this claim by scaling the Euromanifestos of 117 political parties across 23 countries on 4 salient dimensions of political conflict. We assess validity by comparing the *Wordscores* estimates to expert surveys and other judgmental measures, and by examining the *Wordscores*'s estimates ability to predict party membership in the European Parliament groups. We find that the *Wordscores* estimates correlate poorly with expert and judgmental measures of party positions, while the latter outperform *Wordscores* in the predictive validity test. We conclude that *Wordscores* does not live up to its original claim of a "quick and easy" language blind method, and urge researchers to demonstrate the validity of the method in their domain of interest before any empirical analysis.

Computational text analysis is a rapidly growing research field with many applications in political communication research. From using Twitter data to identify the political preferences of citizens (Ceron, Curini, Iacus, & Porro, 2014; Temporão, Vande Kerckhove, van der Linden, Dufresne, & Hendrickx, 2018), issue agendas (Zheng & Shahin, 2018), and the degree of polarization across topics (Hemphill, Culotta, & Heston, 2016), to studying the tone or the mediated associations in news coverage (Arendt & Karadas, 2017; Young & Soroka, 2012), and comparing the communication strategies of political actors through their press releases (Sagarzazu & Klüver, 2017) or social media posts (Stier, Bleier, Lietz, & Strohmaier, 2018; Van Dalen, Fazekas, Klemmensen, & Hansen, 2015), researchers have employed a wide variety of computational text analysis methods. Sentiment analysis, especially on social media data, has brought a resurgence of methods based on dictionaries (Nulty, Theocharis, Popa, Parnet, & Benoit, 2016; Van Dalen et al., 2015; Young & Soroka, 2012), although more recent work features increasingly more sophisticated classifiers originally developed in other fields of study (Grimmer & Stewart, 2013, p. 275–291).

Text scaling methods have been particularly popular among researchers when the variables of interest can be measured on continuous scales. Such methods used to measure latent features in texts include *Wordfish* (Slapin & Proksch, 2008), *Wordshoal* (Lauderdale & Herzog, 2016), the class affinity model (Perry & Benoit, 2017), applications of multidimensional scaling and correspondence analysis to text (Brier & Hopp, 2011), as well as other well-known classifiers modified to work with continuous outcomes (Beauchamp, 2012). The most prominent, however, is *Wordscores* originally proposed by Laver, Benoit, and Garry (2003), hereafter referred to as LBG. The *Wordscores*

CONTACT Kostas Gemenis gemenis@mpifg.de Max-Planck-Institut fur Gesellschaftsforschung, Cologne, Germany
Replication files and Supplementary material can be accessed at the Harvard Dataverse, doi:10.7910/DVN/GOSVG6.

algorithm (described in detail in the supplementary material Appendix A), can be seen as an application of correspondence analysis to words as data (Lowe, 2008, p. 366–368) with a roughly Bayesian derivation (Beauchamp, 2012, p. 8–10), and as a supervised scaling model (Grimmer & Stewart, 2013, p. 292) where the vocabulary of a set of "reference" texts for which the position on the dimension of interest is known is used as a training set for estimating the unknown positions of another set of "virgin" texts. Like other text scaling methods, *Wordscores* makes the "bag-of-words" assumption by treating individual words as "data" irrespective of their syntactic context, and assumes that the relative frequencies of specific words provide manifestations of the latent underlying scale (Klemmensen, Hobolt, & Hansen, 2007, p. 748).

Following recent calls to examine the validity of computational text analysis methods in both communication science (van Atteveldt & Peng, 2018, p. 86–87) and political science (Grimmer & Stewart, 2013, p. 271), we take on the task of validating the *Wordscores* method in the most rigorous way to date. We begin with a comprehensive literature review that attests to the continuing popularity of *Wordscores* as a computational text analysis method, outline the previous attempts to validate the method, and describe the design of our study. We use *Wordscores* to estimate the ideology of 117 political parties across 23 countries on several dimensions of interest using documents prominent in the communication flows during election campaigns. We conclude with findings that have important implications for those who use *Wordscores* for empirical analyses in political communication research and beyond.

Literature review

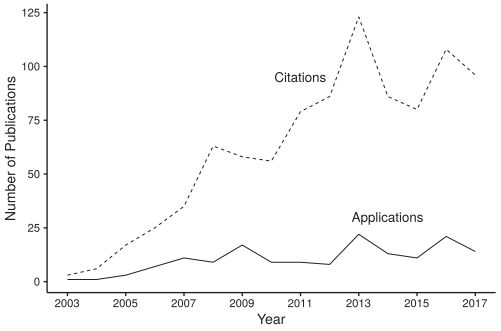
Citation analysis

Wordscores has proven to be highly popular due to its ease of use and implementation in statistical software. As of December 2018, Google Scholar gives 1,328 citations to LBG who introduced *Wordscores*, with plenty of empirical applications in political communication. Applications include Twitter data analysis to scale journalists and media outlets from liberal to conservative (Barberá & Sood, 2015), measuring message clarity of website statements of congressional candidates (Chapp, Roback, Johnson-Tesch, Rossing, & Werner, 2018), and creating a democracy index from newspaper articles (Marzagão, 2017). Moreover, *Wordscores* has been used to analyze latent measures of ideology from speeches at the level of governments (Hakhverdian, 2009), delegates (Benoit, Laver, Arnold, Pennings, & Hosli, 2005), governors (Weinberg, 2010), legislators (Bernauer & Bräuninger, 2009; Laver & Benoit, 2002), and countries (Baturo, Dasandi, & Mikhaylov, 2017).

In addition to media content and speeches, *Wordscores* has been used on a diverse range of texts, such as public statements by US Senators justifying their votes (Bertelli & Grose, 2006), advocacy briefs in the US Supreme Court (Evans, McIntosh, Lin, & Cates, 2007), interest group contributions (Klüver, 2009), and open-ended questions in surveys (Baek, Cappella, & Bindman, 2011). Moreover, some of the most prominent applications of *Wordscores* involve analysis of documents produced by political parties to communicate policy positions and promises during election campaigns. Researchers have therefore applied *Wordscores* to election manifestos to scale parties and empirically test wide range of questions, such as the formation of government coalitions (Linhart, Debus, & Bräuninger, 2010; Proksch & Slapin, 2006), the fulfillment of electoral promises (Debus, 2008), the success of bills in legislatures (Brunner & Debus, 2008), or the choice of putting the EU's constitutional treaty to a referendum (Hug & Schulz, 2007b).

Despite the breadth of applications, one could argue whether *Wordscores* can be still considered a prominent method given the recent advances in computational text analysis (Grimmer & Stewart, 2013; Welbers, Van Atteveldt, & Benoit, 2017). To investigate this possibility, we performed citation analysis on all citations to LBG as captured by Google Scholar.[1] Our citation analysis revealed a total of 156 uses of *Wordscores* in empirical analyses, from 2003–2017, 79 of which have been published in peer-review journals, with the remaining appearing in monographs, chapters in edited volumes,

(a) Citations compared to empirical applications.



(b) Empirical applications published in peer-reviewed journals compared to other outlets.

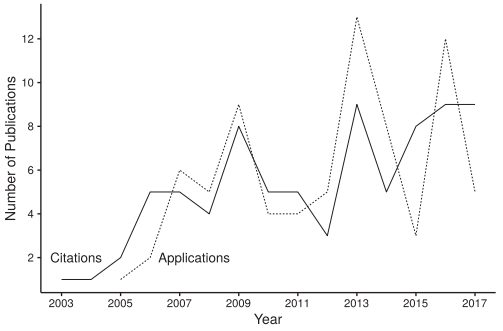


Figure 1. Analysis of citations to Laver et al. (2003).

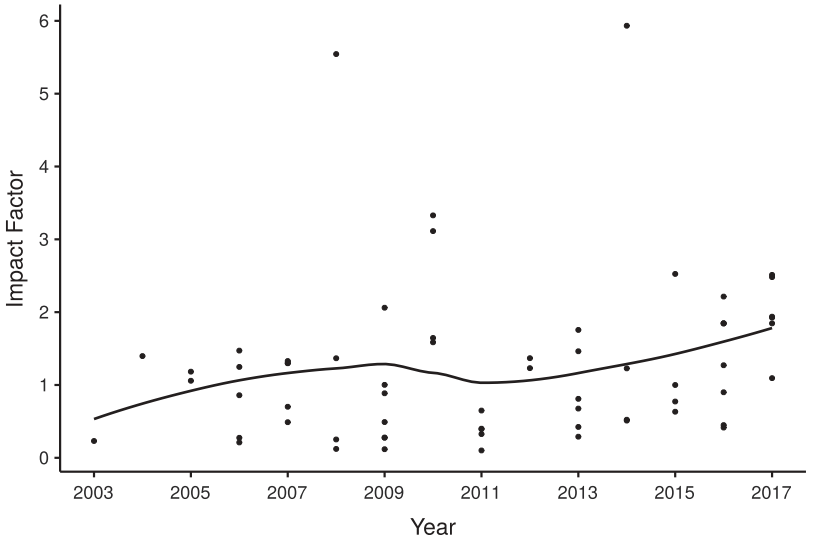


Figure 2. Impact factor trend of journals using *Wordscores* in empirical analyses (LOESS, bandwidth = .7).

working papers, and conference papers. Interestingly, as Figure 1 shows, the publication of empirical analyses using *Wordscores* constitute a relatively stable fraction of the total citations to LBG, whereas the trend of the publications in peer-review journals closely mirrors the trend of publications in other outlets. Finally, as shown in Figure 2, the citation analysis shows no evidence that empirical analyses using *Wordscores* are now published in lesser quality journals (at least judging from their impact factor) compared to previous years. We therefore conclude that despite the advent of more sophisticated methods of computational text analysis, *Wordscores* deserves a rigorous evaluation in its own right as it remains highly popular in the literature.

Validation studies

To position documents, *Wordscores* makes a series of assumptions regarding the distribution of reference documents across the latent trait of interest, the distribution of words across reference documents, and of the use of words as data more generally (Lowe, 2008). However, as Lowe (2008) notes, most of these assumptions do not usually hold in practice so the question is not whether

computer-assisted methods satisfy assumptions with regards to how language works and texts are generated, but to evaluate methods on the basis of "their ability to perform some useful social scientific task" (Grimmer & Stewart, 2013, p. 270). In this respect, researchers have called for studies that demonstrate the validity of computational text analysis methods (Grimmer & Stewart, 2013; van Atteveldt & Peng, 2018). Nevertheless, despite these calls, our review of the empirical analyses using *Wordscores* revealed very few validation attempts. Moreover, most of the validation studies have been rather limited in terms of scope.

Naturally, the first attempt to validate the *Wordscores* output was presented in LBG. In particular, LBG used the 1992 election manifestos of British and Irish parties as reference texts and assigned to them reference scores from expert surveys conducted in 1992 in order to estimate party positions in the 1997 election manifestos at both economic and social policy dimensions. LBG assessed the criterion validity of the estimates by comparing the *Wordscores* output against the estimates of an expert survey conducted in 1997. LBG also used a similar approach to estimate party positions for the German election of 1994 but, due to the lack of comparable expert survey data, assessed the German estimates only in terms of face validity. We argue that if *Wordscores* aims to be a useful tool, its validity needs to be evaluated beyond such simple "proof of concept" demonstrations, especially when these demonstrations are not particularly robust. Our replication of the LBG validation analysis (available in the supplementary material, Appendix B), reveals inconsistencies between the definitions in the article and the way *Wordscores* is implemented in statistical software, as well as evident that comparing *Wordscores* estimates to expert surveys can be manipulated by adding or removing manifestos from the set of virgin texts.

Budge and Pennings (2007) performed an independent validation exercise by employing *Wordscores* on the election manifestos of British parties across time and comparing the estimates on the left–right dimension to those of the Manifesto Project. Their results proved to be unfavorable for *Wordscores* since they found that the method produced flat scores across time compared to the Manifesto Project estimates. Benoit and Laver (2007a) responded by dismissing these findings arguing that *Wordscores* had not been properly implemented (Budge & Pennings merged several manifestos before using them as reference texts), and that the Manifesto Project estimates should not be used as a benchmark because they are controversial with respect to their own validity.

Klemmensen et al. (2007) performed a similar evaluation by using *Wordscores* to estimate the positions of Danish parties on the left–right dimension. Although their article has been widely cited as a successful validation of *Wordscores*, a closer investigation of the results shows that this is not actually the case. The correlations reported by Klemmensen et al. show that *Wordscores* performed worse than the Manifestos Project estimates when compared to a common benchmark (expert surveys). If the proponents of *Wordscores* argue that the Manifesto Project estimates are problematic in terms of their validity because they exhibit rather low correlations with expert surveys (Benoit & Laver, 2007a, 2007b), then it should follow that *Wordscores* estimates are worse than the oft-criticized Manifesto Project estimates.

Most recently, Hjorth, Klemmensen, Hobolt, Hansen, and Kurrild-Klitgaard (2015) repeated this exercise for political parties in both Denmark and Germany, by comparing the *Wordscores* output to placements as given by experts and voters, using rank order correlations. Their results indicated that the *Wordscores* estimates correlated better with independent measures of party positions compared to the estimates derived from *Wordfish*. However, the rank order correlations examined by the authors provided a far too lenient test on a method which claims to deliver interval level measurements (point estimates with 95% confidence intervals).

The most comprehensive validation conducted so far has been Bräuninger, Debus, and Müller (2013) who used *Wordscores* to estimate the left–right positions of parties in 13 West European countries between 1980 and 2010. Their study specifically aimed to assess the validity of the technique and the results were mixed, concluding that *Wordscores* estimates correlated well with the Manifesto Project in some countries, but not in others. We note that the results of this

comparative study were far more cautious compared to the earlier investigations based on single countries (including the original proof of concept in LBG).[2]

In addition, we identified several studies that examined the validity of *Wordscores* in contexts beyond the policy positions of parties. Klüver (2009) applied *Wordscores* on contributions by interest groups and found that the estimates correlated very moderately with hand-coding. Baek et al. (2011) found that *Wordscores* performed acceptably in distinguishing negative from positive responses in open-ended survey questions when 7% of the text corpus is used as reference texts, while Evans et al. (2007) found that *Wordscores* performed well in classifying legal briefs submitted to the US Supreme Court as liberal or conservative. The latter two instances, however, involved binary classification which is a very lenient test for a scaling method such as *Wordscores*.

Our study addresses all the limitations in the available validation studies by examining *Wordscores* on multiple dimensions, using different sets of reference scores and transformation methods, and by assessing the estimates across different types of validity using appropriate statistical measures.

Study design

We examine the validity of *Wordscores* in political communication research by focusing on the "Euromanifestos" of political parties as pertinent illustrations. Euromanifestos are documents that are produced and disseminated by political parties prior to elections to the European Parliament with the aim of communicating their policy positions to the public. Although parliamentary debates, campaign speeches, and media reports (Bernauer & Bräuninger, 2009; Helbling & Tresch, 2011; Laver & Benoit, 2002) can also be analyzed with regards to the positions of political actors, election manifestos are usually given preference as authoritative statements on party policy because they are usually ratified in party conventions and represent the organization as a whole (Klingemann, Volkens, Bara, Budge, & McDonald, 2006, p. 164–165). Consequently, Euromanifestos have been repeatedly content analyzed both in terms of their visual and textual content (Popa & Dumitrescu, 2017; Spoon, 2012; Wüst, 2009), and have been combined with mass survey and media content data to advance research in political communication (Adams, Ezrow, & Somer-Topcu, 2014; Schuck, Xezonakis, Elenbaas, Banducci, & De Vreese, 2011). Our choice of European Parliament elections over national elections aims to improve the comparability of estimates across countries. National elections contain more idiosyncratic parameters in the campaigning and use of political texts compared to elections to the European Parliament which take place at the same time and within a shared political context. This also implies that, by avoiding comparisons across time (Bräuninger et al., 2013; Hakhverdian, 2009; Hjorth et al., 2015) which could be affected by changes in the political discourse, we provide a favorable context to test the validity of *Wordscores*.

Unlike the previous studies we do not limit our validation to the general left–right dimension but also estimate party positions on three more specific dimensions that have been used extensively to analyze party competition in the European Parliament context: the European integration pro–against dimension, the economic left–right dimension, and the socio-cultural liberal-conservative dimension (Hix, 1999; Hix, Noury, & Gérard, 2006; Hooghe & Marks, 1999; Hooghe, Marks, & Wilson, 2002; McElroy & Benoit, 2007).

Document selection

Our study focuses on 117 political parties across 23 EU member-states.[3] We use the 2009 Euromanifestos as virgin texts and the 2004 Euromanifestos as reference texts which we harvested from the Euromanifestos Project Schmitt2018.[4] Our analysis includes only parties for which a Euromanifesto was available for both 2004 and 2009. A list of the names of parties and documents used in our study can be found in the supplementary material (Appendix C).

Table 1. Party position data sources used in this study.

Source type	Used for reference scores (2004)	Used for the validation (2009)
Expert survey	BL 2003	-
Expert survey	CHES 2002	CHES 2010
Judgmental coding	EMP 2004	EMP 2009
Judgmental coding	-	EUP 2009

As in the case of the Manifesto Project (Gemenis, 2012; Hansen, 2008), the Euromanifestos Project collection includes, along with "genuine" manifestos, many documents of questionable comparability such as small pamphlets that do not present a broad policy profile and documents that contain irrelevant and/or misleading sections (e.g. references to the positions of *other* parties). Although such documents are problematic for computational text analysis (Proksch & Slapin, 2009), we nevertheless decided to use the Euromanifestos Project collection in order to test the *Wordscores* method in a realistic context as researchers are more likely to rely on established off-the-shelf collections for cross-country comparative analyses.

Pre-processing and estimation

We performed the pre-processing and *Wordscores* estimation using the quanteda package (version 1.3.4) in R (Benoit et al., 2018; Welbers et al., 2017). Following Grimmer and Stewart (2013, p. 272–273), we pre-processed the documents to make them suitable for computational text analysis by removing numbers, symbols, and words drawn from language-specific lists of stopwords. In our analyses, pre-processing by removing the 20 most frequent words instead of the stopwords (Ruedin, 2013a) produced near identical results, but we acknowledge that different pre-processing choices are likely to affect the substantive conclusions in multivariate models (Denny & Spirling, 2018; Greene, Ceron, Schumacher, & Fazekas, 2016).We do not use stemming as this decreases the effectiveness of the method (Ruedin, 2013b) and because it is not beneficial for all languages. This is especially the case for languages in which compound words are common, such as in German or Finnish, where stemming may lead to a reduction of information.

Unlike previous studies, we used a variety of sources for reference scores and to compare the *Wordscores* estimates in the validation exercise. Following Benoit and Laver (Benoit & Laver, 2007a, 2007b), we did not use the party position estimates of the Manifesto Project as their reliability and validity has often been the object of criticism in the literature (Gemenis, 2013b). Instead, we used expert survey estimates as did LBG and most of the empirical applications involving *Wordscores*. Of course, expert surveys have their own problems, so we cross-validated our use of expert surveys with estimates from an approach that can be best described as judgmental estimation (Gemenis, 2015, p. 2293–2296). Instead of aggregating the judgments of multiple experts, this approach involves a single coder who is asked to provide an informed judgment after reading a party manifesto or a similar document. We further examined the sensitivity of our findings by employing two different data sources within each approach as shown in Table 2. For expert surveys, we used the 2003 Benoit and Laver (2006) and the 2002 and 2010 Chapel Hill Expert Surveys (Bakker et al., 2015; Hooghe et al., 2010),[5] and for judgmental coding, the *overall* judgment made by the coders of the Euromanifestos Project on the basis of the whole document (Schmitt, Braun, Popa, Mikhaylov, & Dwinger, 2018), as well as the estimates from the 2009 EU Profiler dataset (Trechsel, 2013) using the scales proposed by Gemenis (2013a). The exact wording of questions and scales used in our study is available in the supplementary material (Appendix D).

Finally, and also unlike previous studies, we examined the sensitivity of our findings by employing two different transformations for each set of *Wordscores* estimates: the transformation originally proposed by LBG, and the alternative transformation proposed by Martin and Vanberg (2008) hereafter referred to as MV. Following LBG, we used all available documents for 2004 as reference

texts when using the LBG transformation, and the highest and lowest scoring documents for the MV transformation (the default option in quanteda).

Validation strategy

As Grimmer and Stewart (2013, p. 271), validation in supervised methods such as *Wordscores* should involve demonstrating that the computer-assisted method can reproduce the results in a set of documents for which the true scores of the quantity of interest are known. Since true scores of the ideological positions of parties do not exist, we follow two different validation approaches. Firstly, we compare the *Wordscores* estimates to human judgments of party positions (Lowe & Benoit, 2013) and, secondly, we use the *Wordscores* estimates to predict other known quantities of interest. These two strategies correspond to criterion and construct validity in the assessment framework as proposed by Carmines and Zeller (1979), or convergent and predictive validity as proposed by Grimmer and Stewart (2013).

Results

Criterion/convergent validity

We begin by examining the degree to which the *Wordscores* estimates correlate with other known measures of party positions as outlined in the study design section. Carmines and Zeller (1979) refer to this type of validity as criterion validity while other authors use the term concurrent validity. Our analysis uses the concordance correlation coefficient (CCC) defined as the product between Pearson's product-moment correlation coefficient that measures dispersion (i.e. the degree of random measurement error) and a bias correction factor that measures the deviation from the 45° line of perfect concordance (Lin, 1989). We use CCC as both Pearson's and Spearman's correlation coefficients are likely to overestimate the degree of validity in case of the presence of systematic measurement error.[6]

As CCC requires both measures to be on the same scale, we rescaled the *Wordscores* estimates using the empirical scale minimum and maximum per country. Moreover, to ensure that the validation results are not driven by this rescaling or the bias correction factor introduced by the CCC, we also report in the figures the simple Pearson's r correlation coefficients. In order to establish an objective criterion, we compare the correlation coefficients of *Wordscores* estimates with other measures to the correlation coefficients between each pair of measures shown in the last column of Table 1. In this way we avoid the entirely subjective criteria used in previous validation studies that reported correlations as being "strong," "moderate," etc. without using an objective reference category.

The results clearly show that the *Wordscores* party position estimates have much lower correlation to other estimates that the correlation of other estimates to one another. Figure 3 presents the results for the left–right dimension. While the CCC between *Wordscores* and other estimates range (depending on which reference scores were used) between 0.2 and 0.55, the CCC's among the other measures (represented by the three vertical lines) range between 0.75 and 0.87. Moreover, the Pearson correlations for *Wordscores* are similar for the LBG transformation, slightly higher for the MV transformation, but never as high as the correlations among the other measures.

The results are quite similar for the estimates on the European integration dimension (Figure 4) but the difference between CCC's and Pearson's correlations is more pronounced for the *Wordscores* estimates. This is mainly due to a centrist bias in the *Wordscores* estimates due to the scoring of many non-informative words with respect to the dimension of interest. We explore this problem of at some length among British Euromanifestos in the supplementary material (Appendix E), and we find that the LBG transformation is a, rather unsuccessful attempt to offset this centrist bias as it does not always pull the estimates toward the correct direction.

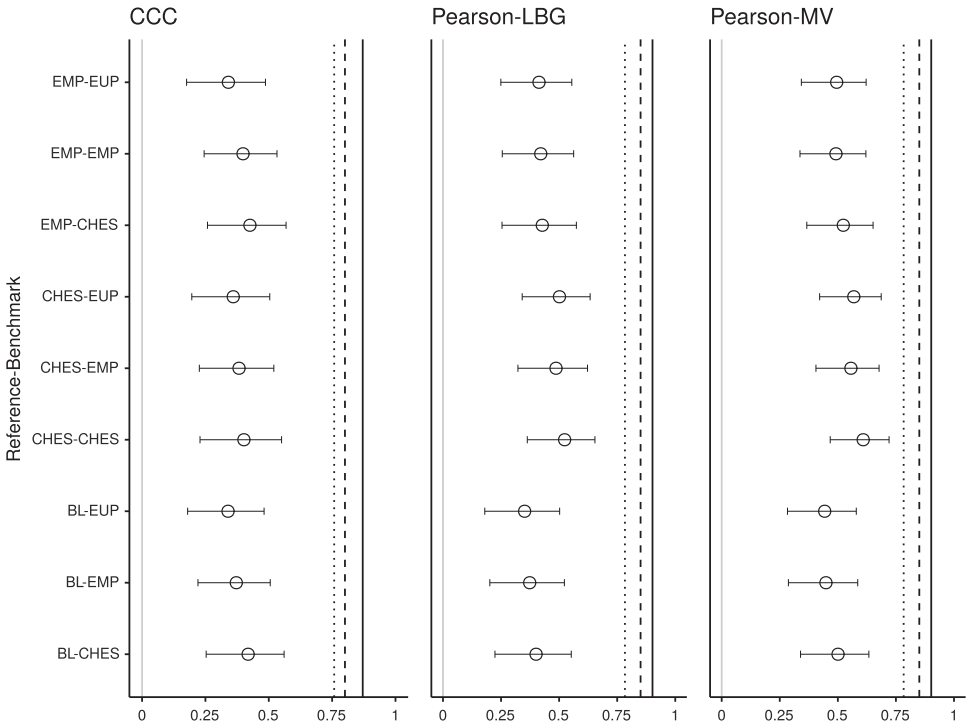


Figure 3. Correlations between *Wordscores* estimates and other measures on the left–right dimension. Note: Vertical lines represent the correlations between CHES/EMP (solid), CHES/EUP (short dash), EMP/EUP (long dash) measures.

The results for the socio-economic dimension (Figure 5) are largely similar to the left–right dimension, but for the socio-cultural dimension (Figure 6) the correlations among other estimates are more dispersed. This can be attributed to the poor quality of the Euromanifesto Project estimates in this particular dimension. In fact, when the Euromanifesto Project estimates were used as reference scores, the correlations of *Wordscores* estimates to other measures are indistinguishable from zero.

Overall, the correlation coefficients show that while all other measures share about 50% of their variance, the *Wordscores* estimates share only some 5–25% of their variance with other measures. Moreover, the MV transformed scores perform slightly better than the LBG transformed scores, and in some cases the *Wordscores* estimates are highly dependent on the quality of the reference scores.

To be sure, we also considered the possibility that the difference in performance between the more general left–right dimension on the one hand, and the more specific European integration, socio-economic and socio-cultural dimensions on the other, might be due to the multidimensionality of the text in terms of presenting policy positions. Parsing documents into sections representing different policy areas is required in text scaling methods such as *Wordfish* which assume that the text is unidimensional (Slapin & Proksch, 2008), so the same logic can be applied to *Wordscores*. Because the European integration dimension is reflected more generally in the Euromanifestos, we considered parsing the Euromanifestos and re-estimating positions on the socio-economic and socio-cultural dimensions on the respective sections of the parsed text (see also Proksch & Slapin, 2006). We report on some evidence in parsing British Euromanifestos in the supplementary material (Appendix F), where we show that the *Wordscores* estimates are statistically indistinguishable for the socio-economic dimension when the parsed text is used, while there seems to be some slight improvement in the socio-cultural dimension with some of the parties moving closer to expert survey estimates.

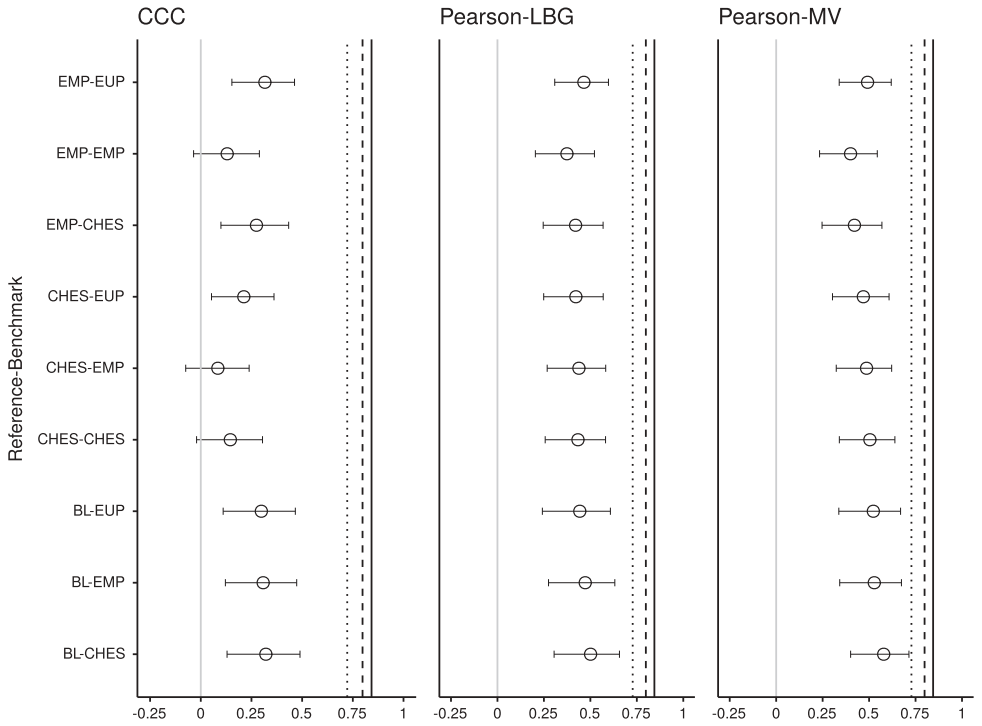


Figure 4. Correlations between *Wordscores* estimates and other measures on the European integration dimension. Note: Vertical lines represent the correlations between CHES/EMP (solid), CHES/EUP (short dash), and EMP/EUP (long dash) measures.

Construct/predictive validity

In addition to criterion/concurrent validity, we also examine what Carmines & Zeller (1979) call construct validity, and others refer to as predictive validity. Construct/predictive validity refers to the extent to which a measure behaves as expected within a given theoretical context. Following, Klingemann et al. (2006, p. 36–39) and McElroy and Benoit (2007) who use a similar approach to validate Manifesto Project and expert survey data respectively, we hypothesize that we can use the ideological positions of parties to predict membership in the political groups of the European Parliament (EP). To do so, we estimate a multinomial regression model where the dependent variable takes eight values, one for each of the seven party groups in the EP (as of 2009) with non-attached parties forming the eighth group. The predictors are party positions on the socio-economic, socio-cultural, and European integration dimensions.

We estimate a total of nine models: six models using *Wordscores* estimates with all possible configurations between the three sets of reference scores and two transformations, as well as three

Table 2. Predicting European Parliament party group membership using different measures of party positions.

Model	Reference scores	Transformation	McFadden's pseudo-R^2	Krippendorff's α
Wordscores	BL	LBG	.094	.151
Wordscores	BL	MV	.140	.201
Wordscores	CHES	LBG	.105	.172
Wordscores	CHES	MV	.106	.172
Wordscores	EMP	LBG	.066	−.012
Wordscores	EMP	MV	.093	.093
CHES 2010	-	-	.741	.771
EMP 2009	-	-	.388	.439
EUP 2009	-	-	.489	.525

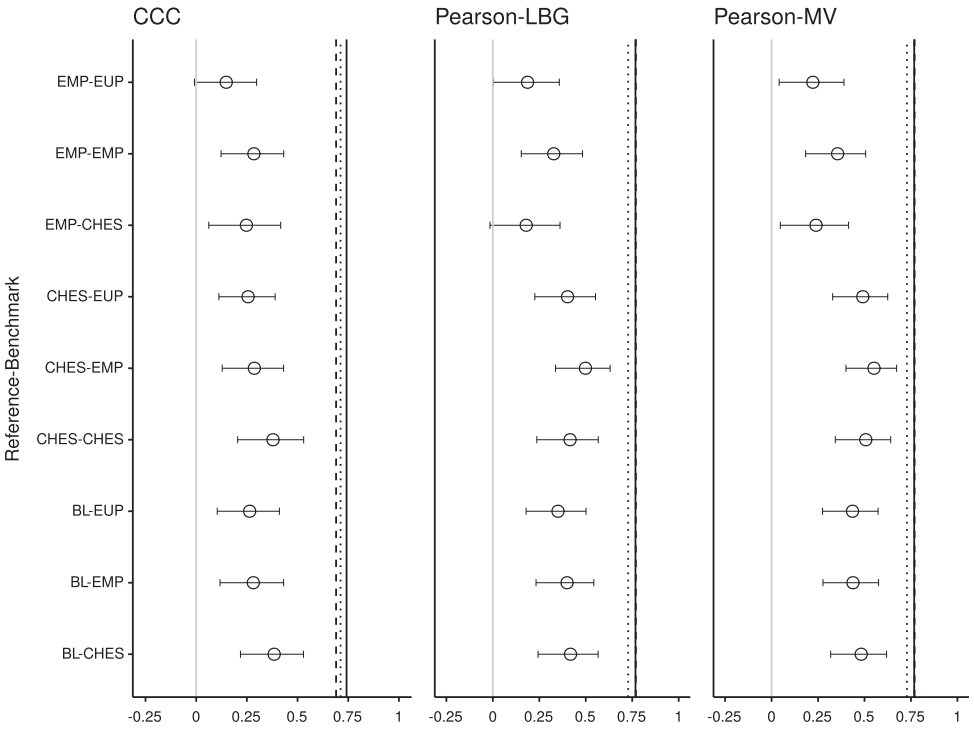


Figure 5. Correlations between *Wordscores* estimates and other measures on the socio-economic dimension. Note: Vertical lines represent the correlations between CHES/EMP (solid), CHES/EUP (short dash), and EMP/EUP (long dash) measures. In two of the plots the solid and long dash lines coincide.

models using the benchmark datasets shown in Table 1 (the 2010 Chapel Hill Expert Survey, and the judgmental coding of the Euromanifestos Project and EU Profiler). To assess the explanatory power of each measure we use McFadden's pseudo-R^2 which compares the explanatory power added by the independent variables compared to a model that includes only the intercept, as well as the agreement between the actual values and each model's predicted values using the nominal version of Krippendorff's alpha coefficient (Hayes & Krippendorff, 2007).

As can be seen in the estimates presented in Table 2, in none of the cases do the *Wordscores* estimates perform better compared to other measures in predicting membership in the EP party groups. The best performing model using *Wordscores* estimates (using the BL expert survey reference scores and the MV transformation) yields a McFadden's pseudo-R^2 at .14 and a Krippendorff's α at .2 which indicate poor predictions by social science standards. To further illustrate this, we present some of the model predictions in heatmaps plotting the predicted against the actual values. Figure 7 shows clearly that *Wordscores* estimates make very noisy predictions, especially when compared to the predictions made by the 2010 CHES expert survey data. For instance, *Wordscores* misclassified all of the parties in the ECR, EFD, and NI groups, while other methods got most, if not all, of the parties in these groups correct. Even if we consider that the expert survey predictions might be inflated because experts might be taking EP group membership into account when giving their responses to the survey, *Wordscores* also performed poorly in comparison to other types of judgment. As the case of the EMP data shows, the quick overall judgments made by individual students after reading the Euromanifestos provide better predictions in terms of EP group membership compared to *Wordscores*.

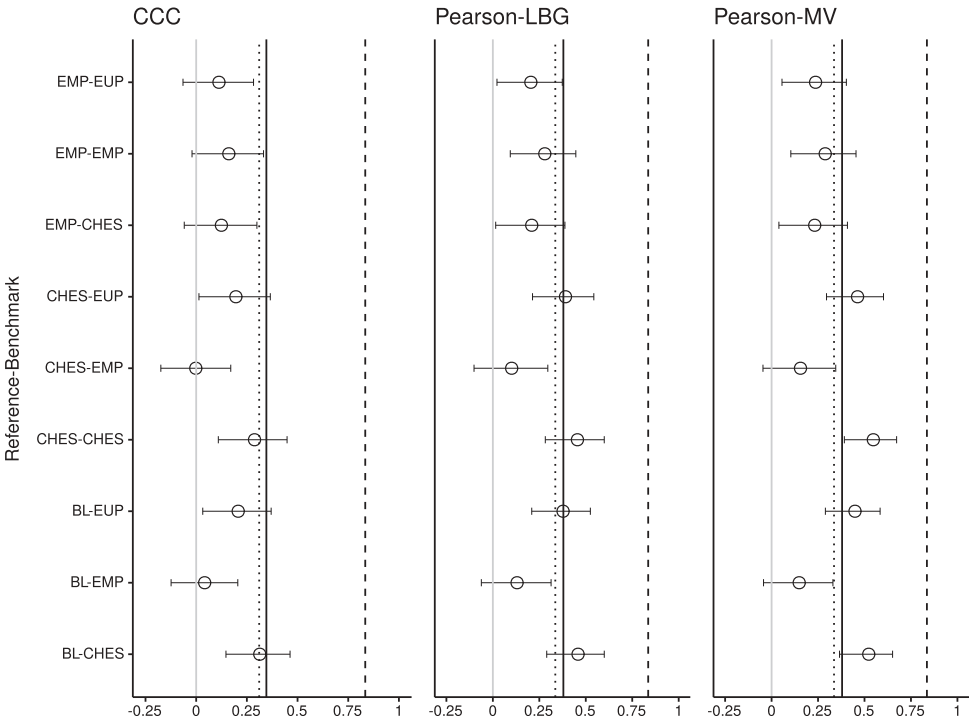


Figure 6. Correlations between *Wordscores* estimates and other measures on the socio-cultural dimension. Note: Vertical lines represent the correlations between CHES/EMP (solid), CHES/EUP (short dash), and EMP/EUP (long dash) measures. The grey line indicates a zero correlation.

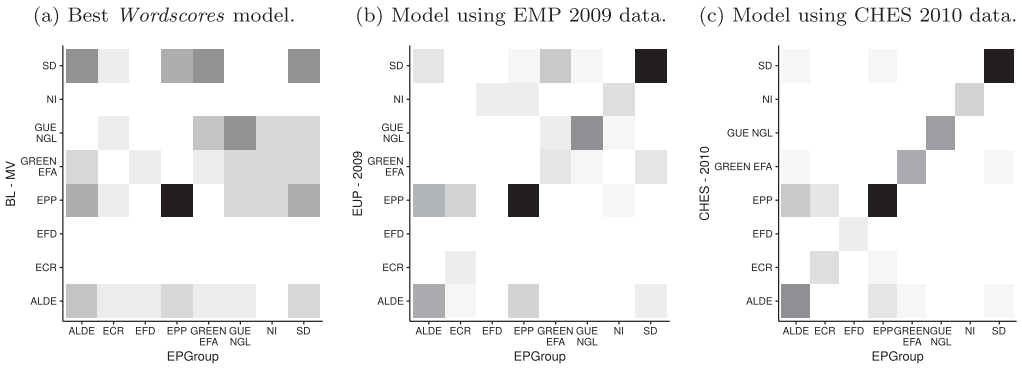


Figure 7. Predicted vs. actual values in European Parliament party group membership.

Conclusions

In their proof-of-concept, Laver et al. (2003, p. 329) claim that *Wordscores* can deliver, in a matter of seconds, "effective" estimates of the policy positions of political actors. Our replication of the central analysis in this proof-of-concept showed that the validity of the results was highly dependent on the selection of virgin texts as well as particularities in the software implementation of the proposed algorithm. Following recent calls to evaluate the validity of computational text analysis methods (Grimmer & Stewart, 2013; van Atteveldt & Peng, 2018) we subjected *Wordscores* to rigorous

validation in a political communication context, and in a way consistent with past applications of the method.

In contrast to LBG claims, our findings indicate that the *Wordscores* estimates had a much lower correlation to measures of party positions than the correlation of other measures to one another. In addition,, our findings indicated that the *Wordscores* estimates had a significantly lower predictive power compared to other measures of party positions when used to predict party membership in EP groups. The results did not differ greatly with respect to the dimension of interest but they were sometimes highly dependent on the quality of the reference scores, while the MV transformation produced slightly better results. Furthermore, our assessment of *Wordscores* remained unchanged when different statistical measures were used to assess criterion/convergent and construct/predictive validity.

Moreover, we obtained these results after a rigorous pre-processing of the source documents where all stop words were removed in order to address a well-known problem in *Wordscores* whereby the presence of non-informative words pushes the estimates toward the middle of the scale (Lowe 2008).[7] There is some evidence that further parsing the documents in order to use only their most informative parts with regards to the dimension of interest could further improve the *Wordscores* estimates, but at the expense of considerable investment in time and resources. Researchers would need to read and parse the documents manually, resulting in several forking paths in analytic decisions (Denny & Spirling, 2018), which in turn negate the original claim of a quick, easy, and language blind method (Laver et al., 2003, p. 226, 312).

We are aware that our findings could be driven by the use of Euromanifestos rather than national election manifestos in our study design. However, the most comprehensive validation study using national election manifestos also found mixed results (Bräuninger et al., 2013). It seems that the problem is not so much the context in which the documents are produced, but rather the quality of the documents as sources of the quantity which is being scaled (Proksch & Slapin, 2009). Given that computational text analysis methods are meant to augment humans and not to replace them (Grimmer & Stewart, 2013), researchers should consider examining the content of the documents that they analyze. Researchers should also consider the type of document before including it in the analysis, and drop or replace documents as needed (Proksch & Slapin, 2009, p. 329–330)

Furthermore, we urge researchers who are interested in using *Wordscores* to demonstrate the validity of the method in their domain of interest before proceeding with the empirical analysis. As has been demonstrated for text analysis (Benoit, Mikhaylov, & Laver, 2009, p. 505–510) and argued for in computational communication science more generally, failure to do so, "can introduce systematic biases in subsequent multivariate analysis and threaten the validity of statistical inference" (van Atteveldt & Peng, 2018, p. 86). Consequently, we advise researchers to use different methods, including *Wordscores*, to assess the robustness of their multivariate statistical analyses. Similarly, we urge researchers to refrain from using *Wordscores* in order to validate other computational text analysis methods. Like other computational text analysis methods (Grimmer & Stewart, 2013), the validity of *Wordscores* should not be taken for granted, and researchers should be cautious when using computational text analysis methods, especially ones that promise to deliver lots with minimal human input.

Notes

1. A spreadsheet with the details of the citation analysis can be found in the replication materials.
2. Ruedin (2013a) and Hug and Schulz (2007a) compared *Wordscores* estimates to several other methods aiming to measure the positions of parties. Their comparisons, however, did not focus on *Wordscores* as such but rather showed how results might differ across different methods.
3. Our study excludes Luxembourg and Malta where no appropriate reference scores were available for 2004. For estimation purposes, we treat Northern Ireland/Great Britain and Flanders/Wallonia as four different party systems.
4. The collection can be accessed at http://www.ees-homepage.net.

5. The Chapel Hill Expert Survey data are available at http://www.chesdata.eu.
6. Pearson's and Spearman's coefficients are based on the presumption of linearity ($Y = bX$) which is different than agreement between two measurements ($Y = X$) making it possible therefore for two measures to correlate perfectly using such coefficients without them being identical (Krippendorff, 1970, p. 144).
7. It is also worth mentioning that this problem has been accounted for by the *Wordfish* scaling method which applies weights "capturing the importance of [words] in discriminating between party positions" (Slapin & Proksch, 2008, p. 709).

Acknowledgements

Earlier versions of the article were presented at the Text in Politics Workshop at Politicologenetmaal, Leiden University (June 1-2, 2017), the Text Analysis Lecture Series, University of Amsterdam (May 12, 2016), the POLSENT Workshop on Politics and the Media, Dublin City University (July 15, 2015), and the 5th Annual General Conference of the European Political Science Association, Vienna (June 25-27, 2015). We would like to thank Sven-Oliver Proksch, Gijs Schumacher, Oliver Treib, as well as the editor and anonymous reviewers of Communication Methods and Measures for their helpful comments. Any errors or omissions remain our responsibility.

Disclosure

No potential conflict of interest was reported by the authors.

ORCID

Bastiaan Bruinsma http://orcid.org/0000-0002-2556-4940
Kostas Gemenis http://orcid.org/0000-0002-3973-5675

References

Adams, J., Ezrow, L., & Somer-Topcu, Z. (2014). Do voters respond to party manifestos or to a wider information environment? An analysis of mass-elite linkages on European integration. *American Journal of Political Science*, *58* (4), 967–978. doi:10.1111/ajps.12115

Arendt, F., & Karadas, N. (2017). Content analysis of mediated associations: An automated text-analytic approach. *Communication Methods and Measures*, *11*(2), 105–120. doi:10.1080/19312458.2016.1276894

Baek, Y. M., Cappella, J. N., & Bindman, A. (2011). Automating content analysis of open-ended responses: Wordscores and affective intonation. *Communication Methods and Measures*, *5*(4), 275–296. doi:10.1080/19312458.2011.624489

Bakker, R., de Vries, C., Edwards, E., Hooghe, L., Jolly, S., Marks, G., Polk, J., Rovny, J., Steenbergen, M., Vachudova, M.A. (2015). Measuring party positions in Europe: The Chapel Hill Expert Survey trend file, 1999–2010. *Party Politics*, *21*(1), 143–152. doi:10.1177/1354068812462931

Barberá, P., & Sood, G. (2015, June). *Follow your ideology: Measuring media ideology on social networks*. Paper presented at the Annual Conference of the European Political Science Association, Vienna, Austria.

Baturo, A., Dasandi, N., & Mikhaylov, S. J. (2017). Understanding state preferences with text as data: Introducing the UN general debate corpus. *Research & Politics*, *4*(2). doi:10.1177/2053168017712821

Beauchamp, N. (2012). *Using text to scale legislatures with uninformative voting*. Retrieved from http://nickbeauchamp.com/work/Beauchamp scaling current.pdf

Benoit, K., & Laver, M. (2006). *Party Policy in Modern Democracies*. London, UK: Routledge.

Benoit, K., & Laver, M. (2007a). Benchmarks for text analysis: A response to Budge and Pennings. *Electoral Studies*, *26* (1), 130–135. doi:10.1016/j.electstud.2006.04.001

Benoit, K., & Laver, M. (2007b). 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., Arnold, C., Pennings, P., & Hosli, M. O. (2005). Measuring national delegate positions at the convention on the future of Europe using computerized word scoring. *European Union Politics*, *6*(3), 291–313. doi:10.1177/1465116505054834

Benoit, K., Mikhaylov, S., & Laver, M. (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

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Mueller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774–778. doi:10.21105/joss.00774

Bernauer, J., & Bräuninger, T. (2009). Intra-party preference heterogeneity and faction membership in the 15th German bundestag: A computational text analysis of parliamentary speeches. *German Politics*, *18*(3), 385–402. doi:10.1080/09644000903055823

Bertelli, A. M., & Grose, C. R. (2006). The spatial model and the senate trial of president clinton. *American Politics Research*, *34*(4), 535–559. doi:10.1177/1532673X05282848

Bräuninger, T., Debus, M., & Müller, J. (2013). *Estimating policy positions of political actors across countries and time*. Mannheim, DE: Mannheimer Zentrum für Europäische Sozialforschung. (MZES Working Paper No. 153).

Brier, A., & Hopp, B. (2011). Computer assisted text analysis in the social sciences. *Quality & Quantity*, *45*(1), 103–128. doi:10.1007/s11135-010-9350-8

Brunner, M., & Debus, M. (2008). Between programmatic interests and party politics: The German Bundesrat in the legislative process. *German Politics*, *17*(3), 232–251. doi:10.1080/09644000802300288

Budge, I., & Pennings, P. (2007). Do they work? Validating computerised word frequency estimates against policy series. *Electoral Studies*, *26*(1), 121–129. doi:10.1016/j.electstud.2006.04.002

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Thousand Oaks, CA: Sage.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society*, *16*(2), 340–358. doi:10.1177/1461444813480466

Chapp, C., Roback, P., Johnson-Tesch, K., Rossing, A., & Werner, J. (2018). Going vague: Ambiguity and avoidance in online political messaging. *Social Science Computer Review*. doi:10.1177/0894439318791168

Debus, M. (2008). Unfulfilled promises? German social democrats and their policy positions at the federal and state level between 1994 and 2006. *Journal of Elections, Public Opinion and Parties*, *18*(2), 201–224. doi:10.1080/17457280801987926

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189. doi:10.1017/pan.2017.44

Evans, M., McIntosh, W., Lin, J., & Cates, C. (2007). Recounting the Courts? Applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, *4*(4), 1007–1039. doi:10.1111/j.1740-1461.2007.00113.x

Gemenis, K. (2012). Proxy documents as a source of measurement error in the comparative manifestos project. *Electoral Studies*, *31*(3), 594–604. doi:10.1016/j.electstud.2012.01.002

Gemenis, K. (2013a). Estimating parties' positions through voting advice applications: Some methodological considerations. *Acta Politica*, *48*(3), 268–295. doi:10.1057/ap.2012.36

Gemenis, K. (2013b). What to do (and not to do) with the comparative manifestos project data. *Political Studies*, *61* (S1), 3–23. doi:10.1111/1467-9248.12015

Gemenis, K. (2015). An iterative expert survey approach for estimating parties' policy positions. *Quality & Quantity*, *49*(6), 2291–2306. doi:10.1007/s11135-014-0109-5

Greene, Z., Ceron, A., Schumacher, G., & Fazekas, Z. (2016). *The nuts and bolts of automated text analysis: Comparing different document pre-processing techniques in four countries*. Retrieved from https://osf.io/4z5z3/

Grimmer, J., & Stewart, B. M. (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

Hakhverdian, A. (2009). Capturing government policy on the left-right scale: Evidence from the United Kingdom, 1956–2006. *Political Studies*, *57*(4), 720–745. doi:10.1111/j.1467-9248.2008.00763.x

Hansen, M. E. (2008). Back to the archives? A critique of the Danish part of the manifesto dataset. *Scandinavian Political Studies*, *31*(2), 201–216. doi:10.1111/j.1467-9477.2008.00202.x

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. doi:10.1080/19312450709336664

Helbling, M., & Tresch, A. (2011). Measuring party positions and number salience from media coverage: Discussing and cross-validating new indicators. *Electoral Studies*, *30*(1), 174–183. doi:10.1016/j.electstud.2010.12.001

Hemphill, L., Culotta, A., & Heston, M. (2016). # Polar Scores: Measuring partisanship using social media content. *Journal of Information Technology & Politics*, *13*(4), 365–377. doi:10.1080/19331681.2016.1214093

Hix, S. (1999). Dimensions and alignments in European Union politics: Cognitive constraints and partisan responses. *European Journal of Political Research*, *35*(1), 69–106. doi:10.1111/1475-6765.00442

Hix, S., Noury, A., & Gérard, R. (2006). Dimensions of politics in the European Parliament. *American Journal of Political Science*, *50*(2), 494–511. doi:10.1111/j.1540-5907.2006.00198.x

Hjorth, F., Klemmensen, R., Hobolt, S., Hansen, M. E., & Kurrild-Klitgaard, P. (2015). Computers, coders, and voters: Comparing automated methods for estimating party positions. *Research & Politics*, *2*(2), 1–9. doi:10.1177/2053168015580476

Hooghe, L., Bakker, R., Brigevich, A., de Vries, C., Edwards, E., Marks, G., Rovny, J., Steenbergen, M., Vachudova, M. (2010). Reliability and validity of the 2002 and 2006 Chapel Hill expert surveys on party positioning. *European Journal of Political Research*, *49*(5), 687–703. doi:10.1111/j.1475-6765.2009.01912.x

Hooghe, L., & Marks, G. (1999). The making of a polity: The struggle over European integration. In H. Kitschelt, P. Lange, G. Marks, & J. Stephens (Eds.), *Continuity and change in contemporary capitalism* (pp. 70–100). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139175050.005

Hooghe, L., Marks, G., & Wilson, C. (2002). Does left/right structure party positions on European integration? *Comparative Political Studies*, *35*(8), 973–976. doi:10.1177/001041402236310

Hug, S., & Schulz, T. (2007a). Left-right positions of political parties in Switzerland. *Party Politics*, *13*(3), 305–330. doi:10.1177/1354068807075938

Hug, S., & Schulz, T. (2007b). Referendums in the EU's constitution building process. *Review of International Organizations*, *2*(2), 177–218. doi:10.1007/s11558-007-9020-3

Klemmensen, R., Hobolt, S. B., & Hansen, M. E. (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

Klingemann, H.-D., Volkens, A., Bara, J., Budge, I., & McDonald, M. D. (2006). *Mapping Policy Preferences II: Estimates for parties, electors and governments in central and Eastern Europe, European Union and OECD, 1990–2003*. New York: Oxford University Press.

Klüver, H. (2009). Measuring interest group influence using quantitative text analysis. *European Union Politics*, *10*(4), 535–549. doi:10.1177/1465116509346782

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70. doi:10.1177/001316447003000105

Lauderdale, B. E., & Herzog, A. (2016). Measuring political positions from legislative speech. *Political Analysis*, *24*(3), 374–394. doi:10.1093/pan/mpw017

Laver, M., & Benoit, K. (2002). Locating TDs in policy spaces: The computational text analysis of Dáil speeches. *Irish Political Studies*, *17*(1), 59–73. doi:10.1080/714003143

Laver, M., Benoit, K., & Garry, J. (2003). Estimating the policy positions of political actors using words as data. *American Political Science Review*, *97*(2), 311–331. doi:10.1017/S0003055403000698

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*(1), 255–268. doi:10.2307/2532051

Linhart, E., Debus, M., & Bräuninger, T. (2010). The 2009 elections in Schleswig-Holstein: Polarised electoral campaign, exceptional election results, and an unspectacular process of government formation. *German Politics*, *19*(2), 237–253. doi:10.1080/09644008.2010.481713

Lowe, W. (2008). Understanding Wordscores. *Political Analysis*, *16*(4), 356–371. doi:10.1093/pan/mpn004

Lowe, W., & Benoit, K. (2013). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, *21*(3), 298–313. doi:10.1093/pan/mpt002

Martin, L. W., & Vanberg, G. (2008). A robust transformation procedure for interpreting political text. *Political Analysis*, *16*(1), 93–100. doi:10.1093/pan/mpm010

Marzagão, T. (2017). Automated democracy scores. *Brazilian Review of Econometrics*, *37*(1), 31–43. doi:10.12660/bre.v37n12017.58160

McElroy, G., & Benoit, K. (2007). Party groups and policy positions in the European Parliament. *Party Politics*, *13*(1), 5–28. doi:10.1177/1354068806068593

Nulty, P., Theocharis, Y., Popa, S. A., Parnet, O., & Benoit, K. (2016). Social media and political communication in the 2014 elections to the European Parliament. *Electoral Studies*, *44*, 429–444. doi:10.1016/j.electstud.2016.04.014

Perry, P. O., & Benoit, K. (2017). *Scaling text with the class affinity model*. Retrieved from https://arxiv.org/abs/1710.08963

Popa, S. A., & Dumitrescu, D. (2017). National but European? Visual manifestations of Europe in national parties' Euromanifestos since 1979. *Party Politics*, *23*(5), 526–537. doi:10.1177/1354068815610963

Proksch, S.-O., & Slapin, J. B. (2006). Institutions and coalition formation: The German election of 2005. *West European Politics*, *29*(3), 540–559. doi:10.1080/01402380600619884

Proksch, S.-O., & Slapin, J. B. (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

Ruedin, D. (2013a). Obtaining party positions on immigration in Switzerland: Comparing different methods. *Swiss Political Science Review*, *19*(1), 84–105. doi:10.1111/spsr.12018

Ruedin, D. (2013b). The role of language in the automatic coding of political texts. *Swiss Political Science Review*, *19*(4), 539–545. doi:10.1111/spsr.12050

Sagarzazu, I., & Klüver, H. (2017). Coalition governments and party competition: Political communication strategies of coalition parties. *Political Science Research and Methods*, *5*(2), 333–349. doi:10.1017/psrm.2015.56

Schmitt, H., Braun, D., Popa, S. A., Mikhaylov, S., & Dwinger, F. (2018). *European parliament election study 1979–2014, Euromanifesto study [computer file]*. Cologne, DE: GESIS Data Archive. doi:10.4232/1.12830

Schuck, A. R., Xezonakis, G., Elenbaas, M., Banducci, S. A., & De Vreese, C. H. (2011). Party contestation and Europe on the news agenda: The 2009 European Parliamentary elections. *Electoral Studies*, *30*(1), 41–52. doi:10.1016/j.electstud.2010.09.021

Slapin, J. B., & 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

Spoon, -J.-J. (2012). How salient is Europe? An analysis of European election manifestos, 1979–2004. *European Union Politics*, *13*(4), 558–579. doi:10.1177/1465116512448123

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. *Political Communication*, *35*(1), 50–74. doi:10.1080/10584609.2017.1334728

Temporão, M., Vande Kerckhove, C., van der Linden, C., Dufresne, Y., & Hendrickx, J. M. (2018). Ideological scaling of social media users: A dynamic lexicon approach. *Political Analysis*, *26*(4), 457–473. doi:10.1017/pan.2018.30

Trechsel, A. H. (2013). *EU-Profiler: Positioning of the parties in the European elections [computer file]*. Cologne, DE: GESIS Data Archive. doi:10.4232/1.11689

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2–3), 81–92. doi:10.1080/19312458.2018.1458084

Van Dalen, A., Fazekas, Z., Klemmensen, R., & Hansen, K. M. (2015). Policy considerations on Facebook: Agendas, coherence, and communication patterns in the 2011 Danish parliamentary elections. *Journal of Information Technology & Politics*, *12*(3), 303–324. doi:10.1080/19331681.2015.1061398

Weinberg, M. (2010). Measuring governors' political orientations using words as data. *State Politics and Policy Quarterly*, *10*(1), 96–109. doi:10.1177/153244001001000105

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures*, *11*(4), 245–265. doi:10.1080/19312458.2017.1387238

Wüst, A. M. (2009). Parties in European Parliament elections: Numbers, framing, the EU, and the question of supply and demand. *German Politics*, *18*(3), 426–440. doi:10.1080/09644000903055849

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*(2), 205–231. doi:10.1080/10584609.2012.671234

Zheng, P., & Shahin, S. (2018). Live tweeting live debates: How twitter reflects and refracts the US political climate in a campaign season. *Information, Communication & Society*, 1–21. doi:10.1080/1369118X.2018.1503697