Validating Wordscores: The Promises and Pitfalls of Computational Text Scaling

-

Supplementary Material

Version 1.0

Appendix A: *Wordscores* estimation and assumptions

The estimation process begins with the researcher defining a set of reference texts that have positions on a dimension that we can assume with some confidence (for example, when they are obtained by an expert survey). Reference texts therefore need to be informative with regards to their content (words), and need to have a known position on the dimension of interest. *Wordscores*, implemented as a user-written package in Stata and R, begins by counting the frequency of words in each reference text and assigns a score to each of these words. To do so, *Wordscores* calculates the probability P that a word w appears in reference text r as follows:

$$P_{wr} = \frac{F_{wr}}{\sum_r F_{wr}} \tag{1}$$

where F_{wr} is the frequency of word w in reference text r. Using these probabilities, *Wordscores* calculates a score for each word w on each dimension of interest d as follows:

$$S_{wd} = \sum_r P_{wr} A_{rd} \tag{2}$$

where A_{rd} is the known position of reference text r on dimension d. To score each virgin text v on dimension d, *Wordscores* use the word scores S_{wd} obtained from reference texts as follows:

$$S_{vd} = \sum_w F_{wv} S_{wd} \tag{3}$$

According to Laver, Benoit & Garry (2003, 316), F_{wv} in equation 3 denotes 'the relative frequency of each virgin text word [w], as a proportion of the *total number of words in the virgin text* [v]' (emphasis added). However, all the statistical packages that have been written to implement *Wordscores*,[1] use a different definition of F_{wv}. Here the relative frequency of each virgin text word w is taken as a proportion of the total number of words *co-occurring between the reference and the virgin texts*. This inconsistency between the Laver et al. article and the software implementations is of no particular concern to how *Wordscores* work, but it does challenge the proof-of-concept

validation presented in the Laver et al. article as we will see in the following section.

Nevertheless, irrespective of how one defines F_{wv}, the S_{vd} scores only indicate the relative position of virgin texts to each other on dimension d. To be able to compare the scores of virgin texts to the scores of reference texts, we need one more step. *Wordscores* will transform the raw scores back to the original metric used in the scores used in the reference texts, as this allows us to compare the raw scores of the virgin texts with the assigned scores of the reference texts. In their original paper, Laver et al. suggest the following transformation:

$$S^*_{vd} = (S_{vd} - S_{\bar{vd}}) \left(\frac{SD_{rd}}{SD_{vd}} \right) + S_{\bar{vd}} \tag{4}$$

Here, S^*_{vd} is the transformed score, S_{vd} the raw score, $S_{\bar{vd}}$ the average raw score of the virgin texts, and SD_{rd} and SD_{vd} the standard deviations of the reference and virgin text scores respectively. This metric preserves the mean of the virgin text scores, but equals their variance to that of the reference text scores, thus allowing for comparison.

Lowe (2008) points out that the LBG transformation assumes that the raw virgin text scores have the correct mean, but the incorrect variance. However, due to the large amount of overlapping words, the virgin score mean is invariably close to the reference text mean—an effect called shrinkage. These overlapping words are often words as 'the' or 'and', and as they occur frequently in all documents, they get centrist scores. As such, the distances between the virgin texts are shrunken, and all texts bounce towards the middle of the scale. Laver et al. fix this by recouping the original variance, but falsely assume that the newly derived mean is correct. This is no problem when the variance and mean are expected to be the same for both reference and virgin texts. However, as Lowe (2008, 359–360) notes, increasing polarisation between parties, or joint movement to the sides of a set of parties, makes it hard, if impossible, to discern whether the mean of the virgin texts is centrist due to the reference scores or a shrinkage artifact.

Martin & Vanberg (2008, 95–97) agree with the above criticism and note several more shortcomings of the Laver et al. transformation method. First, as the transformation uses the standard deviation of the virgin text raw scores it depends on the set of virgin texts themselves. This makes the scores non-robust with regard to the virgin texts, and any difference in the set of reference texts automatically leads to a difference in the scores. This way, a researcher could obtain different positions in the virgin texts solely because of a different selection in the reference texts. Second, despite what Laver et

al. claim, their method fails to recover the accurate relative distance ratios and therefore put the transformed scores and the virgin scores on the same metric. This is due to shrinkage, as we pointed out above. To combat these problems, Martin & Vanberg (2008) provide a new transformation based on the idea of relative distance ratios S_i:

$$S_i = \frac{S_i - S_{R1}}{S_{R1} - S_{R2}} \tag{5}$$

where two 'anchoring texts' S_{R1} and S_{R2} are chosen, and the placement of all other texts are expressed in relation to this 'standard unit' (Martin & Vanberg 2008, 97). They then use these ratios to construct a new transformation:

$$S^*_{vd} = \left((S_{vd} - S_{R1}) \frac{A_{R2} - A_{R1}}{S_{R2} - S_{R1}} \right) + A_{R1} \tag{6}$$

Here, S^*_{vd} is the transformed score, S_{vd} the raw score, A_{R1} and A_{R1} are the assigned scores to reference texts $R1$ and $R2$ (where $R1$ is located to the left of $R2$), and S_{R1} and S_{R2} are the reference texts' raw scores. In their article, Martin & Vanberg use two reference texts, or 'anchor texts' located to the left and right of virgin texts. As seen in equation (6) above, both the assigned scores for the reference texts are recovered, and the virgin texts are thus placed on the original metric. However, as soon as more than two reference texts are used—as Laver, Benoit & Garry (2003) strongly advise—not all the original exogenous scores of the reference texts can be recovered exactly, as only two texts can be used to define the metric. MV thus suggest a change to the transformation:

$$S^*_{vd} = \left((S_{vd} - S_{Rmin}) \frac{A_{Rmax} - A_{Rmin}}{S_{R2} - S_{R1}} \right) + A_{Rmin} \tag{7}$$

Here A_{Rmin} and A_{Rmax} denote the lowest and highest placed reference text on the original metric. The positions of these texts will be recovered exactly, while the scores of the other texts will be distorted as the relative distance ratios of the raw scores do not correspond to the relative distance ratios of the reference scores. Comparison between reference and virgin texts thus becomes difficult and researchers face a trade-off between increased accuracy of the dictionary and internal consistency, and the ability to make valid comparisons (Martin & Vanberg 2008) (see Appendix G).

To conclude, while the transformation by Laver, Benoit & Garry (2003) depends on the virgin texts and is indifferent to the composition of the reference texts, the transformation by Martin & Vanberg (2008) depends on the reference texts and is indifferent to the composition of the set of virgin texts (Lowe 2008, 360). Moreover, Laver et al. assume that the variances of both the set of reference texts and virgin texts are the same, while the Martin & Vanberg transformation does not do so (Benoit & Laver 2008, 110). In this paper, we use both transformation methods as we have no use for the raw scores and neither of the scores has until now proven to be the most appropriate in all circumstances.

More generally, Lowe (2008) criticized *Wordscores* for its heavy dependence on reference texts. Lowe (2008, 366–368) views *Wordscores* as an approximation to correspondence analysis and goes on to treat the method as a statistical ideal point model for words. In doing so, he identified six conditions that *Wordscores* needs to fulfill in order to ensure consistent and unbiased estimation of the parameters of the ideal point model:

1. The word scores of the virgin texts need to be equally spaced and extend over the whole range of word scores for the reference texts

2. The word scores of the virgin texts need to be spaced relative to the informativeness term (all texts are thus informative)

3. The reference scores of the reference texts need to be equally spaced and extend past each word score of the virgin texts in both directions

4. The word scores of the reference texts need to be spaced relative to the informativeness term (all texts are thus informative)

5. All the words need to be equally informative

6. The probability of seeing a word needs to be the same for all words

According to Lowe (2008, 369), conditions 5 and 6 will never hold for word count data because any text exhibits a highly skewed word frequency distribution, regardless of the genre, and contain many uninformative words. Nevertheless, we can significantly reduce these problems by filtering out uninformative words such as stop words, function words that do not convey meaning but primarily serve grammatical functions, very uncommon words, and words which appear in less than 1% and more than 99% of documents in the corpus (Grimmer & Stewart 2013). Doing this makes the probability of seeing a word more equal, and removes non-informative words.

Conditions 1 and 2 will be less likely to hold when there is not enough overlap between word distributions between the reference documents. However, by using many documents as reference texts (as Laver et al. advised), the conditions might be well approximated. Condition 2, however, suffers from the fact that some documents are small, and thus contain very little to no information. This does not only increase the confidence intervals around the estimates, but also creates a large bias in the estimates, negatively influence the validity of the virgin documents scores.

Conditions 3 and 4 are similar to 1 and 2, but as words are more plentiful then texts, the changes of insufficient overlap are considerably lower, and the conditions are thus less important. Lowe even states 'we might hope that they [words] may relatively evenly spread out across a policy dimension' (Lowe 2008, 369), which makes the conditions even more plausible. Last, Lowe (2008, 369) considers that conditions 1 and 3 can never hold simultaneously, as this would require an infinite data set—and thus concludes that bias in *Wordscores* is inevitable.

Apart from the differences in transformation, the software implementation of the *Wordscores* algorithm in Stata and R introduced three further changes in the way *Wordscores* estimation is implemented compared to the formulas and illustration in Laver et al.

The first concerns the definition of F_{wv} in equation 3 in Laver et al. According to Laver, Benoit & Garry (2003, 316), F_{wv} denotes 'the relative frequency of each virgin text word [w], as a proportion of the total number of words in the virgin text [v]' (emphasis added). Most of the software implementations of *Wordscores* however (including 'v0.36' in Stata and all the implementations in R, namely 'austin' and 'quanteda'), define F_{wv} as the relative frequency of each virgin text word w is taken as a proportion of the total number of words co-occurring between the reference and the virgin texts. The only exception to this is the '23-June-2009' implementation in Stata which follows the definition of Laver et al. As we show in the Appendix B, the two definitions lead to considerable differences in the results.

The second difference refers to the presence of a smoothing parameter. Smoothing works by adding a constant (e.g. 1) in the data frequency matrix to account for words that are being used only once in one document and never in other documents, on the presumption that such words may reflect choices in language usage that do not carry as much information as to the underlying quantity of interest. From quanteda 1.3 the smoothing parameter default is zero to match the description in Laver, Benoit & Garry (2003).

The third difference refers to whether the reference texts are included in the set of virgin texts that are used in estimation. While Laver, Benoit & Garry (2003) do not treat reference texts together with virgin texts and the

implementation of *Wordscores* in Stata made their inclusion optional, the implementations of *Wordscores* in R included reference texts in the estimation as a default. With quanteda 1.1.1 it became possible to exclude reference texts from estimation by using the "include_reftexts" argument, but from 1.3 it is only possible to exclude reference texts by specifying the "newdata" option accordingly.

Appendix B: Reanalysis of Laver, Benoit & Garry (2003)

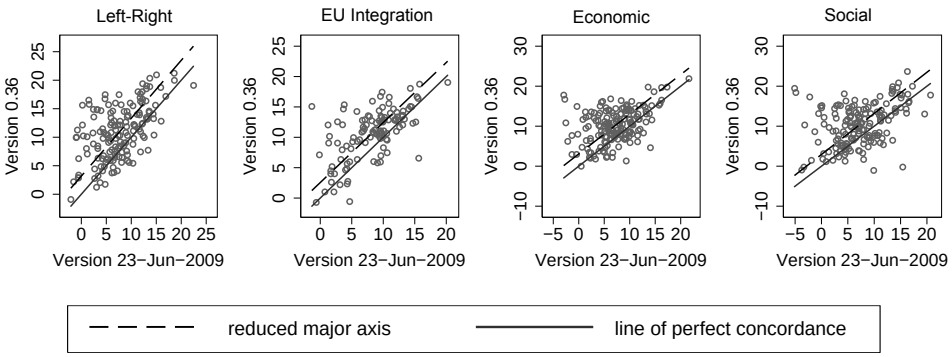
Much of the initial validation for *Wordscores* rested on scoring the 1997 Irish manifestos on a social and economic dimension using the 1992 manifestos as reference texts (Laver, Benoit & Garry 2003). We attempted to replicate the findings in the paper using the manifestos, code, and reference scores as available on the *Wordscores* website `http://www.tcd.ie/Political_Science/wordscores/index.html`. Unfortunately, we were not able to replicate the results published in Laver et al. using the materials from the website. Upon closer examination we realized that replication is not possible for two reasons.

First, the reference texts provided in the *Wordscores* website are not the same as the ones used in the Laver et al. article. As is clear from the number of words, the documents provided in the website have been cleaned differently compared to the documents used in the Laver et al. article. This cleaning refers to the removal of numbers, special characters, document formatting content (tables of contents, headers, footers), and occasionally stop words which is an important step in computer-assisted text analysis. Moreover, the website, includes in the set of reference texts the manifestos of two additional parties (Greens and Sinn Fein), unlike the Laver et al. article which uses as reference texts the manifestos of only five parties.

Second, and most importantly, the current (as of December 16, 2019) '23-June-2009' version of `wordscores` for Stata gives different results than the older version 'v0.36' that was used to produce the results in the Laver et al. article. The differences in the output given by these two versions can be attributed to changes in the code with regards to how F_{wv} (equation 3 in the main text) is calculated. According to Laver, Benoit & Garry (2003, 316), F_{wv} denotes 'the relative frequency of each virgin text word [w], as a proportion of the total number of words in the virgin text [v]' (emphasis added). This is what has been implemented in the '23-June-2009' version of the Stata `wordscores` package. Conversely, 'v0.36' and the two packages that can implement Wordscores in R ('austin' and 'quanteda'), define Fwv as the relative frequency of each virgin text word w is taken as a proportion of the total number of words co-occurring between the reference and the virgin texts. In an e-mail communication, Kenneth Benoit clarified that the 'correct' implementation of *Wordscores* is in the R packages and 'v0.36' version of `wordscores` for Stata. This implies that the definition of F_{wv} given in Laver et al. is incorrect. It also implies that all those who used the '23-June-2009' version in their (published) papers got the 'wrong' *Wordscores* results. In our communication, Kenneth Benoit also indicated that the change in how

F_{wv} is defined does not make much difference as the results correlate highly.

Figure 1: Comparing the results of the two implementations of F_{wv} in `wordscores` for Stata.



We tested this claim by implementing the two versions of `wordscores` (v0.23 and '23-June-2009') for Stata across all the parties in our analysis for four different dimensions (left-right, European integration, economic, social) using the Benoit & Laver (2006) expert survey for the reference text scores and the LBG transformation. Figure A1 shows the results which clearly contradict the claim that the results of the two implementations correlate would highly ('about .97'). The concordance between the two scores measured by the concordance correlation coefficient are .44 (left-right), .53 (European integration), .33 (economic), .32 (social). The respective Pearson correlation coefficients are .55, .62, .41, .38. The correlations are similar when different sources for the reference text scores were used. This is clear evidence that changing the definition of F_{wv} changes the *Wordscores* estimates radically.

Nevertheless, the most important point here is that the inconsistency between the Laver et al. article and the software implementations challenges the proof-of-concept validation presented in the Laver et al. article. In the figures presented in Table 1 below, we show how the *Wordscores* estimates for Irish party positions vary when one uses different sets of documents for reference texts (five parties as in the Laver et al. article versus seven parties as in the replication material found in the *Wordscores* website) and different implementations of `wordscores` for Stata ('v0.36' versus '23-June-2009') lead to substantially different results.

The results in the top left quartile of Table 1 attempt to replicate the findings of Laver et al. by using the manifestos of five Irish parties (FF, FG, Labour, DL, PD) and the 'v0.36' `wordscores` for Stata (which is identical to the `wordscores` and `quanteda` packages in R). They are almost identical

save some minor differences due to the way the documents were cleaned for the analysis in Laver et al. As pointed out in that article, the results look reasonable and consistent with how the parties have been placed in expert surveys (e.g. DL and Labour on the economic left, the other parties on the economic right).

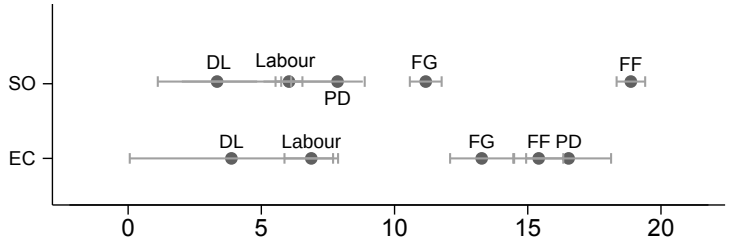
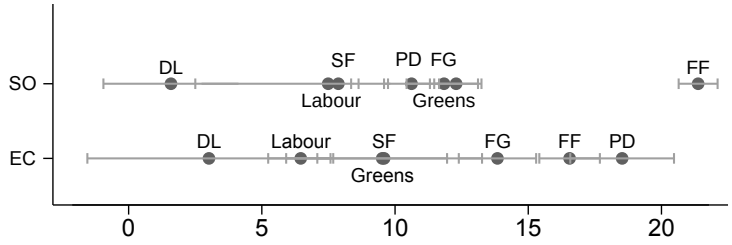
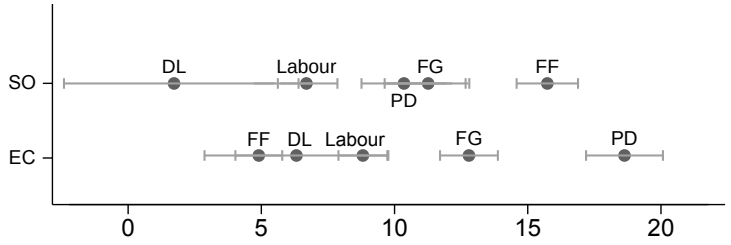
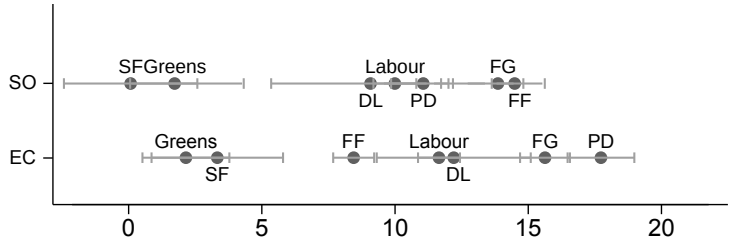
However, when we change the definition of v from 'the total number of words in the virgin text' as stated in the original article Laver, Benoit & Garry (2003, 316) to 'the proportion of the total number of words co-occurring in the virgin and reference texts' as was done in the '23-Jun-2009' version of `wordscores` for Stata, we get the much different results presented in the bottom left quartile. It is clear from the figure that changing the definition of v produces estimates that move parties in a way that does not make much sense (for instance, Fianna Fail as the most economically left party) and otherwise makes it impossible to distinguish between the parties given the confidence intervals of the estimates.

The change in the definition of v that was implemented on 23 June 2009 will produce party positions that appear reasonable and intuitive only if one adds the manifestos of Greens and Sinn Fein in the set of reference texts as shown in the bottom right quartile. However, if we add these two manifestos in the set of reference texts, but keep the definition of v as in the Laver et al. article, we will get the results in the top right quartile. Again, these results do not make much sense, since the confidence intervals overlap significantly and many of the point estimates are rather implausible (e.g. the Greens and Sinn Fein are in the middle of both scales.

We find it strange that the documents for Greens and Sinn Fein were not included in the APSR article, but were included in the replication of the article as implemented in the *Wordscores* website which contained a different Stata `wordscores` code. Why did the authors not include the SF and Greens documents in their original analysis as presented in the APSR article? We believe that this was not done because the addition of these two parties in 2003 under the alternative definition of v which is used in R and is favoured by Kenneth Benoit (as per our e-mail communication) would have given results that are inconsistent with expert surveys. Similarly, when the `wordscores` code was changed and the results appeared to be implausible, the two documents were added as reference texts in the replication materials in the *Wordscores* to improve the validity of the results. Since the positions of parties under the Laver et al. transformation (which is used in the APSR article) are sensitive to the inclusion/exclusion of virgin texts as shown by Martin & Vanberg (2008), we ask whether the exclusion of SF and the Greens from the analysis in Laver et al. but their inclusion in the 'replication' of the analysis in the *Wordscores* website does not constitute an attempt to 'cherry

pick' among different possible results in a way that supports the argument in favour of *Wordscores*.

Table 1: Replication of the original scores

Stata Version	Number of Parties: 5 parties	7 parties
0.36		
	Laver et al. (2003)	
23-Jun-2009		
		Laver et al. (2003) Replication Material

Appendix C: Documents used in the analysis

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
AT	2004	FPO	Freiheitliche Partei Österreichs	Türkei in der EU?	2015	897
AT	2009	FPO	Freiheitliche Partei Österreichs	Echte Volksvertreter statt EU-Verräter	1198	542
AT	2004	GRUNEN	Die Grünen – Die Grüne Alternative	Bestimmen Sie! Ihre Zukunt in Europa	6445	2072
AT	2009	GRUNEN	Die Grünen – Die Grüne Alternative	Vorwarts Grün!	6672	2090
AT	2004	OVP	Österreichische Volkspartei	Europa-Manifest zur Europawahl 2004	4021	1280
AT	2009	OVP	Österreichische Volkspartei	Wahlmanifest Zur Europawahl 2009	7596	2115
AT	2004	SPO	Sozialdemokratische Partei Österreichs	Österreich Muss Wieder Gehört Werden!	2036	712
AT	2009	SPO	Sozialdemokratische Partei Österreichs	Wahlmanifest SPÖ	4392	1439
BE(FR)	2004	CDH	Centre Démocrate Humaniste	Programme européen 2004 du CDH	19073	3530
BE(FR)	2009	CDH	Centre Démocrate Humaniste	Un autre monde, une autre Europe!	26043	4269
BE(FR)	2004	ECOLO	Ecolo	Projet pour l'Europe	8049	2054
BE(FR)	2009	ECOLO	Ecolo	Programme Ecole Élections 2009	15104	3079
BE(FR)	2004	MR	Mouvement Réformateur	25 Propositions pour l'Europe	7176	1734
BE(FR)	2009	MR	Mouvement Réformateur	Le Programme Complet du Mouvement Réformateur élections 2009	24308	3640
BE(FR)	2004	PS	Parti Socialiste	Programme du PS pour les élections européennes	27059	4111
BE(FR)	2009	PS	Parti Socialiste	Programme Union Européenne 2009	22967	3838
BE(NL)	2004	CDV	Christen-Democratisch en Vlaams	Europees verkiezingsprogramma CDV 13 Juni 2004	10437	2229
BE(NL)	2009	CDV	Christen-Democratisch en Vlaams	Europa op maat van de globalisering	5915	1479

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
BE(NL)	2004	GROEN	Groen!	Europa kan zoveel beter - Jij beslist!	12761	2970
BE(NL)	2009	GROEN	Groen!	Groene wegen voor een beter Europa	26645	4932
BE(NL)	2004	NVA	Nieuw-Vlaamse Alliantie	Verkiezingsprogramma N-VA Europese verkiezingen 13 juni 2004	3071	909
BE(NL)	2009	NVA	Nieuw-Vlaamse Alliantie	NVA Europees programma 2009	19376	3508
BE(NL)	2004	OPENVLD	Vlaamse Liberalen en Democraten	Programma VLD - Vlaamse en Europese verkiezingen 13 juni 2004	8017	2190
BE(NL)	2009	OPENVLD	Vlaamse Liberalen en Democraten	Top 15 van de Europese Liberalen voor de verkiezingen van het Europees parlement	1575	632
BE(NL)	2004	SPA	Socialistische Partij Anders	Europees programme 13 juni 2004	12384	2562
BE(NL)	2009	SPA	Socialistische Partij Anders	Mensen op 1 - Een eerlijke koers voor Europa	10027	1984
BE(NL)	2004	VB	Vlaams Belang	Vlaamse Staat, Europese Natie	29405	5159
BE(NL)	2009	VB	Vlaams Belang	Dit is ons land	20434	4200
CY	2004	AKEL	Ανορθωτικό Κόμμα Εργαζόμενου Λαού	Προγραμματικη Διακηρυξη	3020	1613
CY	2009	AKEL	Ανορθωτικό Κόμμα Εργαζόμενου Λαού	Στην Ευρώπη Διεκδικητές και όχι Χειροκροτητές	1849	711
CY	2004	DIKO	Δημοκρατικό Κόμμα	Ισχυρή Κύπρο στην Ευρώπη!	2375	1285
CY	2009	DIKO	Δημοκρατικό Κόμμα	Στείλε καθαρό *μήνυμα* στην Ευρώπη	1760	669
CY	2004	DISY	Δημοκρατικός Συναγερμός	Η καλύτερη ομάδα	1217	690
CY	2009	DISY	Δημοκρατικός Συναγερμός	Πρόταση Πολιτικής	2932	1092
CY	2004	EDEK	Κίνημα Σοσιαλδημοκρατών ΕΔΕΚ	Έχουμε θέση στην Ευρώπη	688	414
CY	2009	EDEK	Κίνημα Σοσιαλδημοκρατών ΕΔΕΚ	Ομιλία Γιαννάκη Ομήρου στην Κεντρική Συγκέντρωση	2037	734

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
CZ	2004	CSSD	Ceská strana sociálne demokratická	Za Evropu bezpecí, míru, prosperity a sociálních jistot	1629	712
CZ	2009	CSSD	Ceská strana sociálne demokratická	Jistota 2009	1237	653
CZ	2004	KDUCSL	Krestanská a demokratická unie – Ceskoslovenská strana lidová	Evropský volební program KDU - CSL	3660	1507
CZ	2009	KDUCSL	Krestanská a demokratická unie – Ceskoslovenská strana lidová	Volební Program Pro Volby Do EP 2009-2014	2960	1365
CZ	2004	KSCM	Komunistická strana Cech a Moravy	S vámi a pro vás doma i v EU	2130	1499
CZ	2009	KSCM	Komunistická strana Cech a Moravy	Otevrený volební program KSCM pro volby do - Evropského parlamentu 2009	979	549
CZ	2004	ODS	Obcanská demokraticktá strana	Stejnté usance pro vvsechny - Program pro volby do Evropského Parlamentu	2146	1018
CZ	2009	ODS	Obcanská demokraticktá strana	Volebntí Program ODS	9645	3435
DE	2004	B90GRUNEN	Bündnis 90/Die Grünen	Europa Besser Machen - Du Entscheidest!	2399	872
DE	2009	B90GRUNEN	Bündnis 90/Die Grünen	Für ein besseres Europa!	48171	8646
DE	2004	CDU	Christlich Demokratische Union Deutschlands	Europa-Manifest der CDU	3076	1067
DE	2009	CDU	Christlich Demokratische Union Deutschlands	Starkes Europa – Sichere Zukunft	6863	1977
DE	2004	CSU	Christlich-Soziale Union in Bayern e. V.	Für ein starkes Bayern in Europa	3863	1269
DE	2009	CSU	Christlich-Soziale Union in Bayern e. V.	CSU-Europawahlprogramm 2009	5582	1625

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
DE	2004	FDP	Freie Demokratische Partei	Wir können Europa besser! - Für ein freies und faires Europa	11411	2787
DE	2009	FDP	Freie Demokratische Partei	Ein Europa der Freiheit - für die Welt des 21. Jahrhunderts	14840	3842
DE	2004	LINKE	Partei des Demokratischen Sozialismus - DIE LINKE	Alternativen sind machbar: Für ein soziales, demokratisches und friedliches Europa!	22582	5396
DE	2009	LINKE	Partei des Demokratischen Sozialismus - DIE LINKE	Solidarität, Demokratie, Frieden - Gemeinsam für den Wechsel in Europa!	19355	4830
DE	2004	SPD	Sozialdemokratische Partei Deutschlands	Europamanifest der SPD	3422	1075
DE	2009	SPD	Sozialdemokratische Partei Deutschlands	Europamanifest	10078	2651
DK	2004	A	Socialdemokraterne	Socialdemokraternes Visioner for Fremtidens Europa	4700	1346
DK	2009	A	Socialdemokraterne	Fae llesskab	5077	1401
DK	2004	B	Det Radikale Venstre - Danmarks social-liberale parti	Program til Europa-Parlamentsvalg 2004	5040	872
DK	2009	B	Det Radikale Venstre - Danmarks social-liberale parti	Europa	4332	1304
DK	2004	C	Det Konservative Folkeparti	Sund konservativ fornuft i Europa	1704	576
DK	2009	C	Det Konservative Folkeparti	Konservatives EP-valgprogram	5374	1468
DK	2004	F	Socialistisk Folkeparti	Fremtidens Europa - SFs valgprogram til Europaparlamentsvalg 2004	7372	1873

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
DK	2009	F	Socialistisk Folkeparti	Et ansvarligt Europa	892	379
DK	2004	O	Dansk Folkeparti	Den Europae iske Union	1344	539
DK	2009	O	Dansk Folkeparti	Den Europae iske Union	2662	875
DK	2004	V	Venstre, Danmarks liberale parti	En stae rk stemme i det ny Europa – Venstres Valgprogram til EP valg 2004	4735	1420
DK	2009	V	Venstre, Danmarks liberale parti	Venstres handlingsprogram til Europa-Parlamentsvalget 2009	7893	1826
EE	2004	ER	Eesti Reformierakond	Reformierakonna Platvorm Euroopa Parlamendi Valimisteks	1179	587
EE	2009	ER	Eesti Reformierakond	Plaan Eesti Majanduskasvu Taastamiseks	2412	1066
EE	2004	KESK	Eesti Keskerakond	Eesti Keskerakonna Valimisprogramm Euroopa Parlamendi Valimisteks	1205	734
EE	2009	KESK	Eesti Keskerakond	Eesti Vajab Vahetust!	1600	899
EE	2004	RESP	Erakond Res Publica	Res Publica Teekaart Euroopas	7078	2935
EE	2009	RESP	Isamaa ja Res Publica Liit	Isamaa Ja Res Publica Liidu Programm Europarlamendi Valimistel	1304	677
EE	2004	SDE	Rahvaerakond Mõõdukad-Sotsiaaldemokraatlik Erakond	Sotsiaaldemokraatliku Erakonna Põhimõtted Ja Lubadused Tööks Euroopa Parlamendis	1390	759
EE	2009	SDE	Rahvaerakond Mõõdukad-Sotsiaaldemokraatlik Erakond	Inimesed Eelkõige: Uus Suund Euroopale	2083	1152

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
ES	2004	PNVEAJ	Partido Nacionalista Vasco-Euzko Alderdi Jeltzalea	Una Nueva Europa Ampliada Abierta A Las Personas Y Al Mundo	58494	5174
ES	2009	PNVEAJ	Partido Nacionalista Vasco-Euzko Alderdi Jeltzalea	Programa Electoral Europeas-09	13678	2780
ES	2004	PP	Partido Popular	Programa Electoral Elleciones Europeas	12857	2261
ES	2009	PP	Partido Popular	Programa Electoral Extenso Elecciones Al Parlamento Europeo	33886	4772
ES	2004	PSOEPSC	Partido Socialista Obrero Español	Manifiesto Europeas 2004	8096	1810
ES	2009	PSOEPSC	Partido Socialista Obrero Español	Manifiesto-Programa Electoral Psoe 'Europeas 2009'	11049	2294
FI	2004	KD	Suomen Kristillisdemokraatit	Kristillisdemokraattien	473	317
FI	2009	KD	Suomen Kristillisdemokraatit	Tehtävä EU:ssa	7944	3511
FI	2004	KESK	Suomen Keskusta	Keskustan Eurooppa-kannanotto	3662	1797
FI	2009	KESK	Suomen Keskusta	Urhoutta Eurooppaan	5580	3070
FI	2004	KOK	Kansallinen Kokoomus	"Jotta Suomella menisi paremmin" - Kokoomuksen eurovaalijulistus	2682	1374
FI	2009	KOK	Kansallinen Kokoomus	Kokoomuksen eurovaaliohjelma 2009	1483	831
FI	2004	RKPSFP	Suomen ruotsalainen kansanpuolue/Svenska folkpartiet i Finland	Eurooppa Koskee Sinua	1955	1069
FI	2009	RKPSFP	Suomen ruotsalainen kansanpuolue/Svenska folkpartiet i Finland	Moninaisuus tuo lisäarvoa. RKP – yhteinen tekijä	1183	677
FI	2004	SDP	Suomen Sosialidemokraattinen Puolue	Ihmisten Eurooppaan	2115	1138
FI	2009	SDP	Suomen Sosialidemokraattinen Puolue	Euroopan Parlamentin Vaalien - Vaaliohjelma 2009	4140	1710

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
FI	2004	VAS	Vasemmistoliitto	Meidän Eurooppa	839	497
FI	2009	VAS	Vasemmistoliitto	Parempi Eurooppa on mahdollinen	2157	1126
FI	2004	VIHR	Vihreä liitto	Vihreän liiton EU-ohjelma	334	227
FI	2009	VIHR	Vihreä liitto	Green new deal - uusi vihreä sopimus Euroopalle	3032	1574
FR	2004	FN	Front National	Les Abberations de l'Europe	2714	862
FR	2009	FN	Front National	«Leur» Europe N'est Pas La Notre ! Voila L'europe Que Nous Voulons	2343	865
FR	2004	PCF	Parti communiste français	L'Europe: oui. Mais pas celle-là!	3923	1386
FR	2009	PCF	Front de Gauche (PCF)	Déclaration de principes du Front de Gauche pour Changer d'Europe	2729	921
FR	2004	PS	Parti socialiste	Une Ambition Socialiste pour L'Europe	7426	2465
FR	2009	PS	Parti socialiste	L'Europe face à la crise: la relance des socialistes	1994	694
FR	2004	UDF	Union pour la Démocratie Française	Nous avons besoin d'Europe	14972	4295
FR	2009	UDF	Mouvement Démocrate (UDF)	Nous l'Europe	3270	1096
FR	2004	UMP	Union pour un mouvement populaire	Avec l'Europe, Voyons la France en Grand!	1862	1076
FR	2009	UMP	Union pour un mouvement populaire	30 Propositions pour une Europe Qui Protège et Qui Agit	3470	1211
GB	2004	CON	Conservative Party	Putting Britain First	13185	2153
GB	2009	CON	Conservative Party	Vote for Change	11295	1934
GB	2004	LABOUR	Labour Party	Britain is working	8175	1390
GB	2009	LABOUR	Labour Party	Winning the fight for Britain's future	10130	1529

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
GB	2004	LIBDEM	Liberal Democrats	Making Europe Work For You	14942	2733
GB	2009	LIBDEM	Liberal Democrats	Stronger Together, poorer apart	12392	2096
GB	2004	PC	Plaid Cymru – the Party of Wales	Fighting Hard For Wales	4387	1059
GB	2009	PC	Plaid Cymru – the Party of Wales	European Manifesto	8214	1399
GB	2004	SNP	Scottish National Party	Vote for Scotland	6525	1357
GB	2009	SNP	Scottish National Party	We've got what it takes	7441	1364
GR	2004	KKE	Κομμουνιστικό Κόμμα Ελλάδας	Διακηρυξη Της Κεντρικης Επιτροπης Του ΚΚΕ	4951	1660
GR	2009	KKE	Κομμουνιστικό Κόμμα Ελλάδας	Διακηρυξη Της Κεντρικης Επιτροπης Του Κκε Για Τις	7307	2273
GR	2004	ND	Νέα Δημοκρατία	Πολιτικα Κειμενα	15972	7010
GR	2009	ND	Νέα Δημοκρατία	Νεα Δημοκρατία Η Αυθεντική Ευρωπαϊκή Επιλογή	577	323
GR	2004	PASOK	Πανελλήνιο Σοσιαλιστικό Κίνημα	Ευρωεκλογες 2004 - Το Όραμα, Οι Θέσεις, Οι Δεσμεύσεις μας	3688	1106
GR	2009	PASOK	Πανελλήνιο Σοσιαλιστικό Κίνημα	Ψηφιζουμε Για Την Ευρώπη - Αποφασιζουμε Για Την Ελλάδα	4003	1982
GR	2004	SYRIZA	Συνασπισμός Ριζοσπαστικής Αριστεράς - Ενωτικό Κοινωνικό Μέτωπο	Συνασπισμός Της Αριστεράς Των Κινημάτων Και Της Οικολογίας	4403	1380
GR	2009	SYRIZA	Συνασπισμός Ριζοσπαστικής Αριστεράς - Ενωτικό Κοινωνικό Μέτωπο	Διακηρυξη Για Τισ Ευρωεκλογεσ	2053	1097
HU	2004	FIDESZ	Fidesz – Magyar Polgári Szövetség	Csak egyrütt sikerülhet!	39536	9932
HU	2009	FIDESZ	Fidesz – Magyar Polgári Szövetség	Elôszó	102353	19041
HU	2004	MDF	Magyar Demokrata Fórum	„A normális Magyarországért!"	2384	1132

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
HU	2009	MDF	Magyar Demokrata Fórum	Miért IGEN az MDF listájára június 7-én?	4965	1896
HU	2004	MSZP	Magyar Szocialista Párt	A Sikeres Európai MagyarországértM	755	402
HU	2009	MSZP	Magyar Szocialista Párt	Újult erovel	468	317
HU	2004	SZDSZ	Szabad Demokraták Szövetsége	Egy Új, Kibovített Európa, Mely Nyitott Állampolgárai	12263	3877
HU	2009	SZDSZ	Szabad Demokraták Szövetsége	200 001 Szabad, Demokrata Szavazó	687	436
IE	2004	FF	Fianna Fáil	Fianna Fáil 2004	8425	1611
IE	2009	FF	Fianna Fáil	Europe, we are better working together	15260	2432
IE	2004	FG	Fine Gael	Fine Gael European Parliament Elections 2004	10409	1984
IE	2009	FG	Fine Gael	Securing Ireland's Future in Europe	12341	2134
IE	2004	GREEN	Green Party	Manifesto 2004 - European and Local Elections	6530	1693
IE	2009	GREEN	Green Party	A Green New Deal for Europe	4067	1124
IE	2004	LABOUR	Labour Party	Making the Difference in Europe	5412	1177
IE	2009	LABOUR	Labour Party	Putting people, jobs and fairness at the heart of Europe	8365	1672
IE	2004	SF	Sinn Féin	An Ireland of Equals in a Europe of Equals	20944	2736
IE	2009	SF	Sinn Féin	Europe '09	8405	1446
IT	2004	FI	Forza Italia	Elezioni Per Il Parlamento Europeo	5896	1483
IT	2009	FI	Il Popolo della Libertà (FI)	Elezioni 2009: Manifesto del Partito Popolare Europeo	1367	591

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
IT	2004	LN	Lega Nord	Programma Per Le Elezioni Europee 2004	9823	2673
IT	2009	LN	Lega Nord	Proposte e Obiettivi	36290	6373
IT	2004	PD	Uniti nell'Ulivo (PD)	L'Europa contro le nostre paure	22737	4130
IT	2009	PD	Partito Democratico (PD)	L'Europa che Conviene	5199	1670
IT	2004	PRC	Partito della Rifondazione Comunista	La Sinistra, L'altra Europa	45971	7341
IT	2009	PRC	Altra Italia (PRC)	Programma Unitario Per Le Elezioni Europee	3667	1234
LT	2004	LICS	Liberalu ir centro sajunga	"Padarykime Europa Naudinga Lietuvai"	4867	1891
LT	2009	LICS	Liberalu ir centro sajunga	Liberalu Ir Centro Sajungos Rinkimu I Europos Parlamenta	6322	2365
LT	2004	LSDP	Lietuvos socialdemokratu partija	Su Europa - Už Lietuva Veikime Kartu!	3625	1567
LT	2009	LSDP	Lietuvos socialdemokratu partija	Lietuvos Socialdemokratu Partijos Rinkimu I Europos Parlamenta 2009 Metais Programa	6981	2541
LT	2004	TS	Tevynes Sajunga	Tevynes Sajungos Rinkimu I Europos Parlamenta Programa	8864	3040
LT	2009	TS	Tevynes sajunga - Lietuvos krikšščionys demokratai	Tevynes Sajungos-Lietuvos Krikščioniu Demokratu Rinkimu I Europos Parlamenta Programines Nuostatos	1311	664

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
LV	2004	JL	Jaunais Laiks	Jaunais laiks priekšvelešanu programma 2004.gada Eiropas Parlamenta velešanam	574	333
LV	2009	JL	Jaunais Laiks	Jaunais laiks priekšvelešanu programma 2009.gada Eiropas Parlamenta velešanam	586	320
LV	2004	LKS	Par cilveka tiesibam vienota Latvija (LKS)	Politisko organizaciju apvieniba "Par cilveka tiesibam vienota Latvija" priekšvelešanu programma 2004.gada Eiropas Parlamenta velešanam	572	332
LV	2009	LKS	Par cilveka tiesibam vienota Latvija (LKS)	PCTVL - Par cilveka tiesibam vienota Latvija priekšvelešanu programma 2009.gada Eiropas Parlamenta velešanam	583	322
LV	2004	LPPLC	Latvijas Celš	Savieniba "Latvijas celš" priekšvelešanu programma 2004.gada Eiropas Parlamenta velešanam	591	339
LV	2009	LPPLC	Latvijas Pirma partija/Latvijas Celš	Partija "LPP/LC" priekšvelešanu programma 2009.gada Eiropas Parlamenta velešanam	663	335
LV	2004	TBLNNK	Tevzemei un Brivibai/LNNK	Apvieniba "Tevzemei un Brivibai"/LNNK priekšvelešanu programma 2004.gada Eiropas Parlamenta velešanam	659	380

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
LV	2009	TBLNNK	Tevzemei un Brivibai/LNNK	Apvieniba "Tevzemei un Brivibai"/LNNK priekšvelešanu programma 2009.gada Eiropas Parlamenta velešanam	663	418
LV	2004	TP	Tautas Partija	Tautas partija priekšvelešanu programma 2004.gada Eiropas Parlamenta velešanam	543	328
LV	2009	TP	Tautas Partija	Tautas partija priekšvelešanu programma 2009.gada Eiropas Parlamenta velešanam	631	372
NL	2004	CDA	Christen-Democratisch Appèl	Verkiezingsmanifest CDA 2004	1778	607
NL	2009	CDA	Christen-Democratisch Appèl	Kracht en Ambitie	14856	2423
NL	2004	CUSGP	ChristenUnie - Staatskundig Gereformeerde Partij	Geloofwaardige keuzes - Manifest voor Christelijke politiek in Europa	15338	3283
NL	2009	CUSGP	ChristenUnie - Staatskundig Gereformeerde Partij	Samenwerking Ja, Superstaat Nee	18370	3136
NL	2004	D66	Democraten '66	Een succesvol Europa	6569	1570
NL	2009	D66	Democraten '66	Europa gaat om mensen!	20195	3424
NL	2004	GL	GroenLinks	Eigenwijs Europees	33733	5741
NL	2009	GL	GroenLinks	Nieuwe Energie voor Europa	27565	4894
NL	2004	PVDA	Partij van de Arbeid	Een Sterk en Sociaal Europa	10467	2186
NL	2009	PVDA	Partij van de Arbeid	Verkiezingsprogramma Europees Parlement 2009-2014	16522	3115
NL	2004	SP	Socialistische Partij	Wie zwijgt stemt toe!	17110	3303

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
NL	2009	SP	Socialistische Partij	Een Beter Europa Begint in Nederland	13533	2597
NL	2004	VVD	Volkspartij voor Vrijheid en Democratie	Een nieuw, Uitgereid Europa, open voor zijn burgers en open voor de wereld	14590	2416
NL	2009	VVD	Volkspartij voor Vrijheid en Democratie	Voor een werkend Europa	3463	1057
NI	2004	DUP	Democratic Unionist Party	The DUP's Vision for Europe	7970	1739
NI	2009	DUP	Democratic Unionist Party	Strong Leadership in Challenging Times	12810	2772
NI	2004	SDLP	Social Democratic and Labour Party	SDLP for Europe - Best record - Best agenda	9673	1875
NI	2009	SDLP	Social Democratic and Labour Party	A Vision For Europe - Ambition For You	11811	2509
NI	2004	SF	Sinn Féin	An Ireland of Equals in a Europe of Equals	20944	2736
NI	2009	SF	Sinn Féin	Sinn Féin European Election Manifesto 2009	17642	3118
NI	2004	UUP	Ulster Unionist Party	Ulster Unionists Manifesto	322	211
NI	2009	UUP	Ulster Unionist Party	Vote For Change	10607	1943
PL	2004	PIS	Prawo i Sprawiedliwosc	Deklaracja Krakowska	759	450
PL	2009	PIS	Prawo i Sprawiedliwosc	Nowoczesna Solidarna Bezpieczna Polska	91964	17150
PL	2004	PO	Platforma Obywatelska	Program Europejski Platformy Obywatelskiej	1464	762

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
PL	2009	PO	Platforma Obywatelska	Projekt dokumentu wyborczego EPL 2009r.	20407	6711
PL	2004	PSL	Polskie Stronnictwo Ludowe	Zadbamy O Polske !	1158	555
PL	2009	PSL	Polskie Stronnictwo Ludowe	Narodowe Priorytety Europejskiej Polityki PSL	5246	1635
PL	2004	SLDUP	Sojusz Lewicy Demokratycznej	Manifest Europejski SLD	792	477
PL	2009	SLDUP	Sojusz Lewicy Demokratycznej	Po pierwsze, czlowiek	8640	2693
PT	2004	BE	Bloco de Esquerda	Refundar a Europa Mudar Portugal	3387	1052
PT	2009	BE	Bloco de Esquerda	Compromisso Eleitoral Da Candidatura Do Bloco Às Europeias	5905	1674
PT	2004	CDU	Partido Comunista Português/Partido Ecologista "Os Verdes"	Declaração Programática2004	9583	1831
PT	2009	CDU	Partido Comunista Português/Partido Ecologista "Os Verdes"	Declaração Programática do PCP para as Eleições Europeias de 2009	10085	1880
PT	2004	PS	Partido Socialista	Pela Europa, pelos portugueses	9718	2197
PT	2009	PS	Partido Socialista	As Pessoas Primeiro - Um Novo Rumo Para A Europa	7305	1742
PT	2004	PSD	Partido Social Democrata	Força Portugal	4146	1268
PT	2009	PSD	Partido Social Democrata	Pelo Interesse Nacional	1235	495
SE	2004	C	Centerpartiet	Smalare men vassare!	4867	1385
SE	2009	C	Centerpartiet	Europas förenta krafter	2229	797
SE	2004	KD	Kristdemokraterna	Inför valet till Europaparlamentet 13 juni 2004	13390	3296
SE	2009	KD	Kristdemokraterna	Ett tryggt Europa – vår väg dit	1545	633

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
SE	2004	M	Moderata samlingspartiet	Europasamarbetet kan göra Sverige bättre	2531	856
SE	2009	M	Moderata samlingspartiet	Tid för ansvar	3462	1069
SE	2004	MP	Miljöpartiet de Gröna	Ja till samarbete, nej till EU-stat - för ett grönt och solidariskt Europa	6250	1795
SE	2009	MP	Miljöpartiet de Gröna	Valmanifest - Grönt Klimatval 2009	816	399
SE	2004	S	Sveriges Socialdemokratiska arbetarpart	Valmanifest 2004	1080	467
SE	2009	S	Sveriges Socialdemokratiska arbetarpart	Valmanifest - Jobben först	1375	487
SE	2004	V	Vänsterpartiet	Vänsterpartiets EU-Valplattform	2923	1088
SE	2009	V	Vänsterpartiet	Valplattform inför EU-parlamentsvalet	4291	1455
SI	2004	NSI	Nova Slovenija – krščanska ljudska stranka	Volitve V Evropski Parlament	1250	645
SI	2009	NSI	Nova Slovenija – krščanska ljudska stranka	Nova Slovenija Kršèanski Ljudska Stranka	1354	631
SI	2004	SD	Socialni demokrati	V Evropi za dobro Slovenije!	2675	1424
SI	2009	SD	Socialni demokrati	Manifest Stranke evropskih socialdemokratov	9261	2657
SI	2004	SDS	Slovenska demokratska stranka	Spletna Stran - Program	2698	1117
SI	2009	SDS	Slovenska demokratska stranka	Nova pot - 20 let slovenske pomladi	27987	6871
SI	2004	SLS	Slovenska ljudska stranka	»Vec Slovenije V Evropi, Vec Evrope V Sloveniji«	4307	1548

Country	Year	Party	Full Name	Title	Total Words*	Unique Words*
SI	2009	SLS	Slovenska ljudska stranka	SLO: SLS + SKD Slovenska Ljudska Stranka	400	215
SK	2004	KDH	Krestanskodemokratické hnutie	Volebný program KDH do volieb do Európskeho parlamentu	2131	1037
SK	2009	KDH	Krestanskodemokratické hnutie	Volebný program KDH do Európskeho parlamentu	2562	1174
SK	2004	LSHZDS	Ludová strana - Hnutie za demokratické Slovensko	Odpovede na otázky: Irena Belohorská, kandidátka na poslanca EP za HZDS	1366	698
SK	2009	LSHZDS	Ludová strana - Hnutie za demokratické Slovensko	Slovensko – Stabilné Srdce Európy	7704	2700
SK	2004	SDKUDS	Slovenská demokratická a krestanská únia - Demokratická strana	Manifest SDKÚ pre novú Európu	2600	1052
SK	2009	SDKUDS	Slovenská demokratická a krestanská únia - Demokratická strana	Za Prosperujúce Slovensko V Silnej Európe	8174	2383
SK	2004	SMERSD	Smer – sociálna demokracia	Silnejšie Slovensko v sociálnej Európe	3153	1149
SK	2009	SMERSD	Smer – sociálna demokracia	Sociálna Európa – Odpoved Na Krízu	678	329
SK	2004	SMKMKP	Strana madarskej komunity - Magyar Közösség Pártja	Helyünk Európában	3769	1581
SK	2009	SMKMKP	Strana madarskej komunity - Magyar Közösség Pártja	Navsa budúcnost v Európe	5664	2184

* Refers to the number of words after the documents were cleaned

Appendix D: Data sources and question wording

	LR - Left-Right	EU - EU Integration	EC - Economic	SO - Social
Benoit & Laver Expert Survey (Benoit & Laver 2006)	Left-Right - Please locate each party on a general left-right dimension, taking all aspects of party policy into account	†EU Authority (AT, BE, UK, DK, FI, DE, GR, IT, NL, NI, PT, ES, SE), EU Larger & Stronger (FR), †EU Strengthening (IE)	Economic (Spending vs. Taxes)	Social
	Left (1)	Favours (1)	Promotes raising taxes to increase public services (1)	Favours liberal policies on matters such as abortion, homosexuality, and euthanasia (1)
	Right (20)	Opposes (20)	Promotes cutting public services to cut taxes (20)	Opposes liberal policies on matters such as abortion, homosexuality, and euthanasia (20)
		Countries excluded are CZ, EE, HU, LV, LT, PL, SK, SI, CY		
Chapel Hill Expert Survey 2002 (Hooghe et al. 2010)	LRGEN = position of the party in 2002 in terms of its broad ideological stance, where	POSITION = overall orientation of the party leadership towards European integration in 2002, where	LRECON = position of the party in 2002 in terms of its ideological stance on economic issues (role of government in economy), where	GALTAN = position of the party in 2002 in terms of its ideological stance on democratic freedoms and rights (role of government in life choices), where

	LR - Left-Right	EU - EU Integration	EC - Economic	SO - Social
	0 indicates that a party is at the extreme left of the ideological spectrum	1 = Strongly opposed to European integration	0 indicates that a party is at the extreme left of the ideological spectrum	0 indicates that a party is at the extreme left of the ideological spectrum
	5 means that it is at the center	4 = Neutral, no stance on the issue of European integration	5 means that it is at the center	5 means that it is at the center
	10 indicates that it is at the extreme right	7 = Strongly in favour of European integration	10 indicates that it is at the extreme right	10 indicates that it is at the extreme right
Chapel Hill Expert Survey 2010 (Bakker et al. 2015)	LRGEN = position of the party in 2010 in terms of its overall ideological stance	POSITION = overall orientation of the party leadership towards European integration in 2010	LRECON = position of the party in 2010 in terms of its ideological stance on economic issues	GALTAN = position of the party in 2010 in terms of its ideological stance on democratic freedoms and rights
	0 = extreme left (-) 5 = center (-) 10 = extreme right	1 = strongly opposed (-) 4 = neutral (-) 7 = strongly in favour	0 = extreme left (-) 5 = center (-) 10 extreme right	0 = extreme left (-) 5 = center (-) 10 extreme right

	LR - Left-Right	EU - EU Integration	EC - Economic	SO - Social
Euromanifestos Project 2004 (Braun, Mikhaylov & Schmitt 2010)	LEFT - placement of Euromanifesto according to the coder on a left-right scale	†EU - placement of Euromanifesto according to coder on a pro-anti-EU-integration scale	STATE - placement of Euromanifesto according to coder on a state interventionism vs. free enterprise scale	LIB - placement of Euromanifesto according to coder on a libertarian-authoritarian scale.
	1=left 10=right	1 = pro 10 = anti	1=state interventionism 10=free enterprise	1=libertarian 10=authoritarian
Euromanifestos Project 2009 (Braun, Mikhaylov & Schmitt 2010)	LEFT - Left - Right	†INTEGRATION - Pro EU-Integration - Anti-EU-Integration	STATE - State Interventionism - Free Enterprise	LIBERTA - Libertarian - Authoritarian
	Coder rating on a 10-point-scale	Coder rating on a 10-point-scale	Coder rating on a 10-point-scale	Coder rating on a 10-point-scale
EU Profiler 2009 (Trechsel 2010-11) ‡	Modified Left-Right - using items 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 14, 16, 18, 19 and 20, with missing values recoded to 4 (Neutral)	Original EU Integration (Y axis), using items 12, 21, 22, 23, 24, 26 and 27	Scale composed of items 1, 2, 11, 14, 16, and 18	Scale composed of items 5, 6, 7, 8, 9, 10, 19, 20 and 25

	LR - Left-Right	EU - EU Integration	EC - Economic	SO - Social

†Denotes variables that have been reversed for subsequent analysis
‡EU Profiler data were scaled according to Gemenis (2013).

Appendix E: Content validity

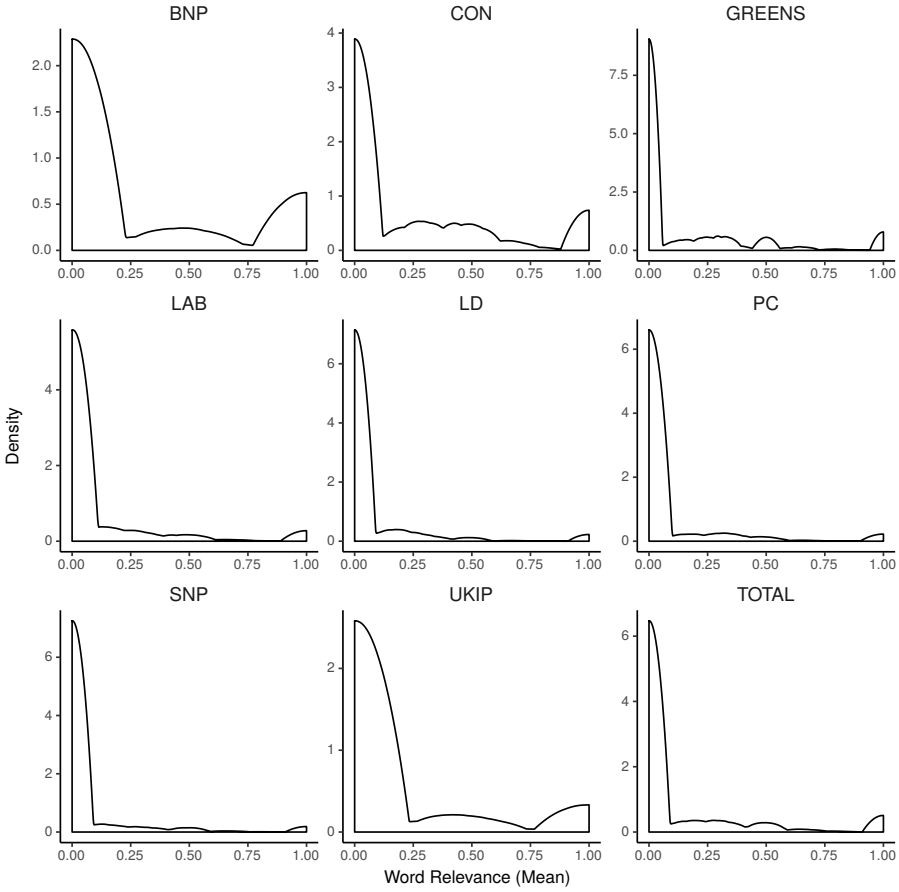
According to Carmines & Zeller (1979), content validity refers to whether the method used for measuring a latent construct represents all of its facets. If one uses multiple indicators that are scaled in a single index, then these indicators should represent all facets of the construct. Alternatively, if one uses a single indicator (for instance as done in surveys asking for a left-right placement) then this indicator has to capture all different facets of the construct. Moreover, a measure that includes facets that do not belong to the construct would be problematic in terms of content validity. As noted in the section about the previous validation attempts, the evaluation of content validity is usually of qualitative nature, so it would be difficult to see how it could be assessed in the context of *Wordscores* output. We propose a workaround this problem by conceptualizing the construct in the context of *Wordscores* as being represented by the words used in the reference texts.

When *Wordscores* places virgin texts on a dimension of interest it does so by calculating a wordscore for each of the words occurring in the reference texts. As *Wordscores* is non-discriminating and scores all words on all dimensions, treating all words as equally informative of the dimension of interest is problematic in terms of content validity. This is because we should not expect each and every word in a reference text to be associated with a dimension of interest, no matter what this dimension is. This problem of *Wordscores* is known, of course, but here we are interested in quantifying the degree of content validity in order to investigate how big of a problem it is for estimating parties' positions.

To do so, we decided to treat each of the words scored in the reference texts as an indicator of the latent concept, and evaluated whether these words relate to the latent concept/dimension of interest. To assess this, following Krippendorff (2004, 101–102) we looked at the context in which these word appear. For example, the word 'committee' can be indicative of a party's position in the dimension of EU integration when it refers to an EU committee, but not when it refers to other types of committees. We therefore hand-coded *each and every word* in the reference texts to see how many of the words used to score the virgin texts were actually used in the context of the dimension of interest. As this is a particularly time-consuming process, we restricted this analysis to British documents and the European integration dimension. Our choice of British parties should be fair for *Wordscores* given that British Euromanifestos are some of the best documents in terms of relevance for assessing parties' positions on European integration. For our hand-coding exercise we defined the context as a natural sentence that starts with a capital letter, and end with one of the following delimiters: '.', '?', '!', ';' (Däubler, Benoit, Mikhaylov & Laver 2012, 942). Items in (bullet-pointed) lists were considered as separate sentences. Each word was coded as one (1) when it was used in a context referring to European integration and zero (0) otherwise.

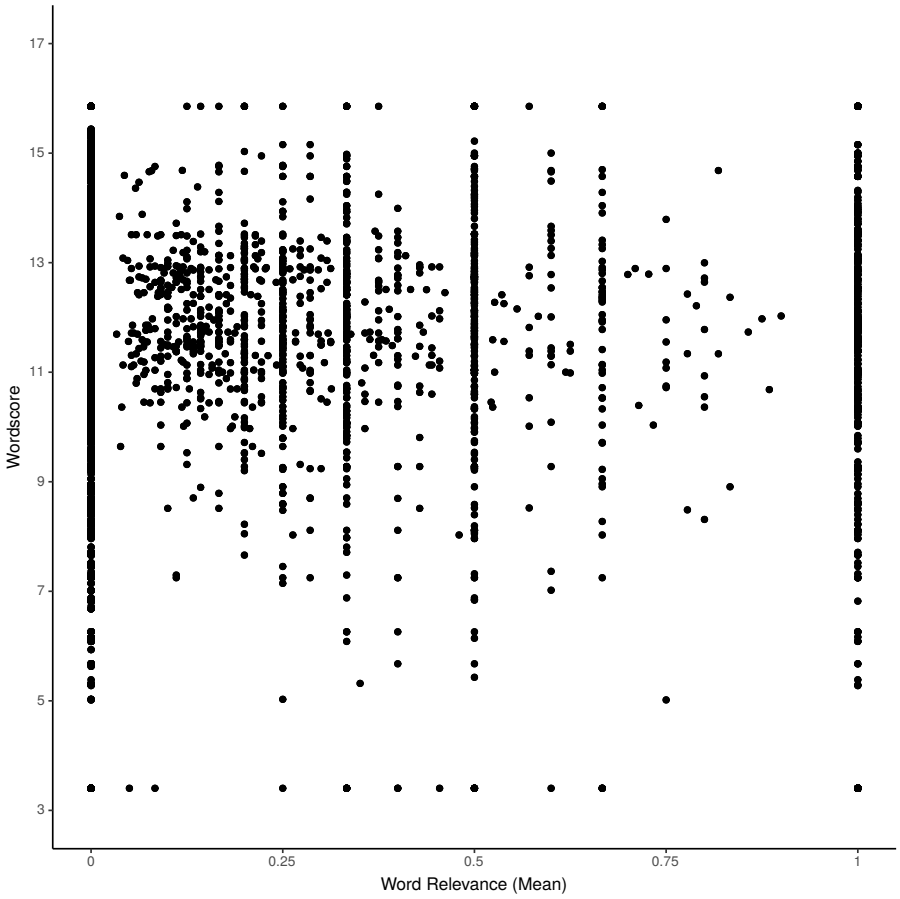
In Figure 2 we plotted the distribution of the average hand-coding evaluation for among all the words used in each virgin document of each British party. What

Figure 2: Assessing content validity against hand-coding in the European integration dimension (I).



Note: The horizontal axis refers to the rate in which words were considered by the hand-coding to be relevant.

Figure 3: Assessing content validity against hand-coding in the European integration dimension (II).



Note: The horizontal axis refers to the rate in which words were considered by the hand-coding to be relevant.

is clear from the figure is that the vast majority of words used by *Wordscores* to estimate party positions are not particularly informative if one looks at the context in which they appear. It appears that *Wordscores* uses far more noise than signal to estimate party positions. A similar pattern is shown in Figure 3 which plots the (hand-coded) word relevance against the wordscore across all British party Euromanifestos. The scatterplot shows no relationship between the two, indicating that wordscores are not particularly strong in terms of content validity.

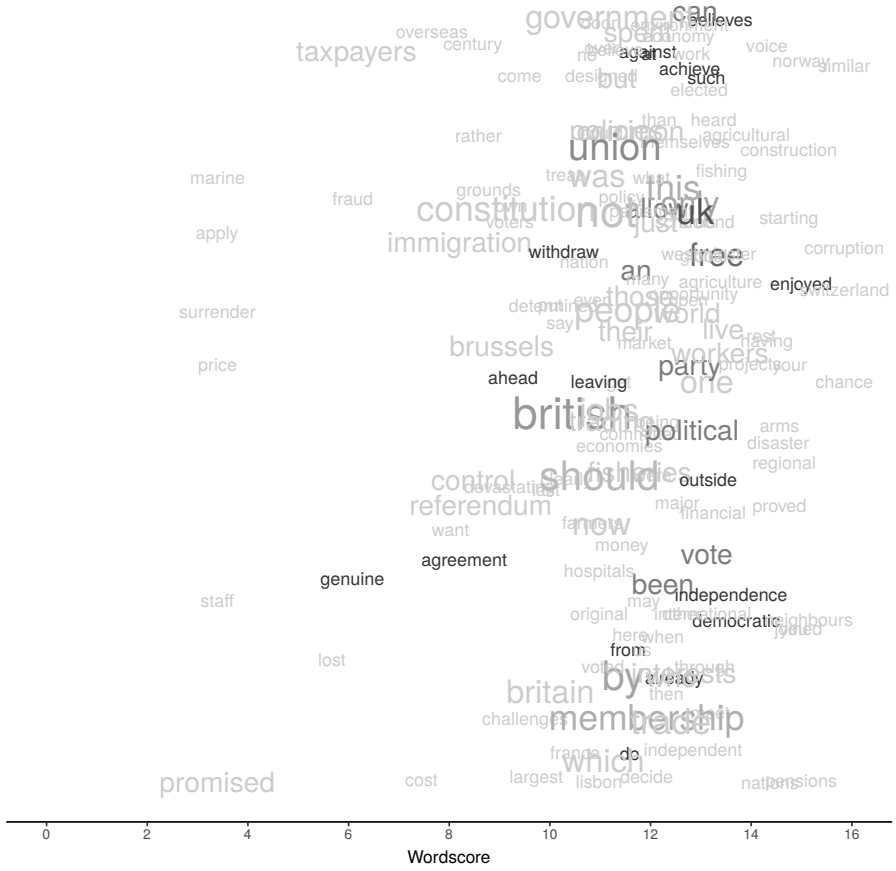
If one considers that the noise brought by the non-informative words moves party positions towards the middle of the scale, one can understand the logic behind the LBG transformation which stretches the party scores towards the end points of the scale. Although we agree that one needs to make some kind of transformation

to account for the presence of noise that leads to the centrist bias in party positions, we do not agree that such a fundamental problem in content validity present in *Wordscores* can be solved by a simple transformation of the raw scores. To give an example, we examine closely the wordscoring of the 2009 UKIP manifesto. UKIP is well-known for its extreme anti-EU stance which should leave no doubts about where the party should be placed. The *Wordscores* raw placement for UKIP is 11.5 [11.2, 11.8] and the LBG transformed one is 9.3 [5.5, 13]. In either case, the party is placed in the middle of the scale. The transformation only improves this placement by specifying that this counter-intuitive middle placement is estimated with a lot of uncertainty. *Wordscores* tells us that UKIP could be placed on either side of the scale even though one should not have much difficulty in establishing the position of the party simply by looking at the UKIP Euromanifesto.

One could argue of course, that this is a problem of the 2009 UKIP Euromanifesto being very short. However, the size of the document should only contribute to making the confidence interval around the point estimate larger. However, the problem here is that the UKIP point estimate is counter-intuitively estimated in the middle of the scale. This is not because the UKIP document is short, but because *Wordscores* is unable to accurately estimate the party position due the noise that was introduced by the scoring of non-informative words. This is clearly shown in Figure 4, where we plotted all the words scored in the UKIP 2009 Euromanifesto according to their wordscore. Most of the words scored by *Wordscores* are not informative with regards to placing UKIP on the European integration dimension and since most of the words have wordscores near the middle of the scale, the point estimate for UKIP was counter-intuitively given at 11.5 on a 20-point scale (transformed by LBG to 9.3).

The problem is therefore deeper than the uncertainty that comes with the size of the documents, and this can be established simply by looking at the cases of parties with much larger documents than UKIP. The fundamental problem lies in the content validity of *Wordscores*. The lack of content validity brought by scoring each and every word irrespective of its relevance in providing information about the dimension of interest, pushes scores towards the middle of the scale. Transforming the raw scores will pull the estimates towards the endpoints of the scale, but there is no guarantee that the estimates will be pulled to the right direction.

Figure 4: Wordscoring the UKIP 2009 Euromanifesto on the European integration dimension.



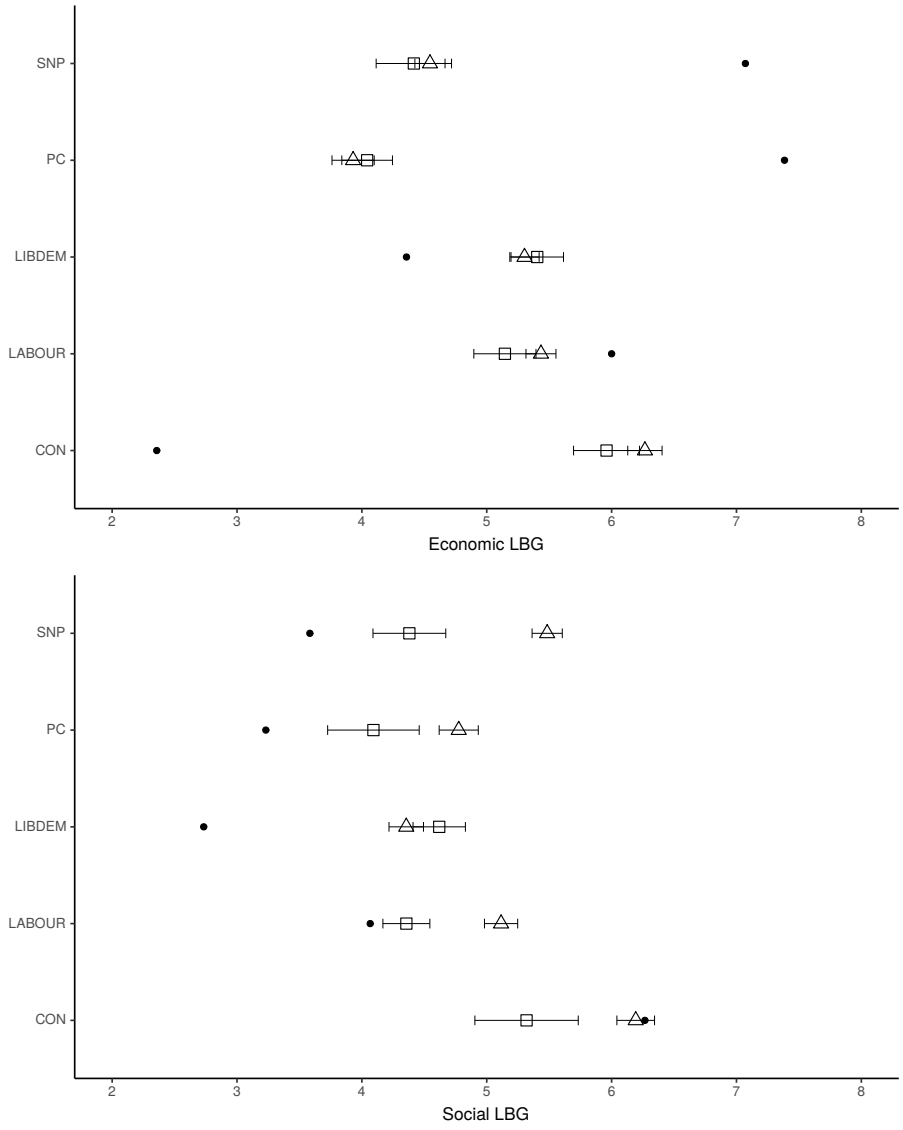
Note: Word size corresponds to the frequency of appearance in the UKIP virgin text; words that were hand-coded as being relevant at least 50% of the instances are plotted in black.

Appendix F: Document parsing

Following an approach similar to Proksch & Slapin (2006), we report an example where documents were parsed into different policy areas before estimation. The particular example regards five British parties' Euromanifestos which we manually parsed each of them into two different documents: a document containing the sections referring to economic policies, and a document containing the sections referring to social policies. Figure 5 compares the *Wordscores* estimates for the parsed versus unparsed documents against the 2010 CHES, according to the LBG transformation, while Figure 6 does the same for the MV transformation. In all cases reference scores were taken from the 2006 CHES. As evident from Figure 5, while the *Wordscores* estimates for the parsed and unparsed documents are statistically indistinguishable in socio-economic dimension, some of the parsed document estimates in the socio-cultural dimension are closer to the CHES estimates compared to the unparsed document estimates (Labour, SNP, PC), one is further away (Conservatives), and in one case there is no difference between the two (Liberal Democrats). The picture in the socio-cultural dimension is largely similar when one considers the MV transformation, but the picture is more inconclusive when one looks at the socio-economic dimension where using the parsed documents reveals some improvement in the estimates for the Conservative Party estimate and a worsening for the Labour Party estimate.

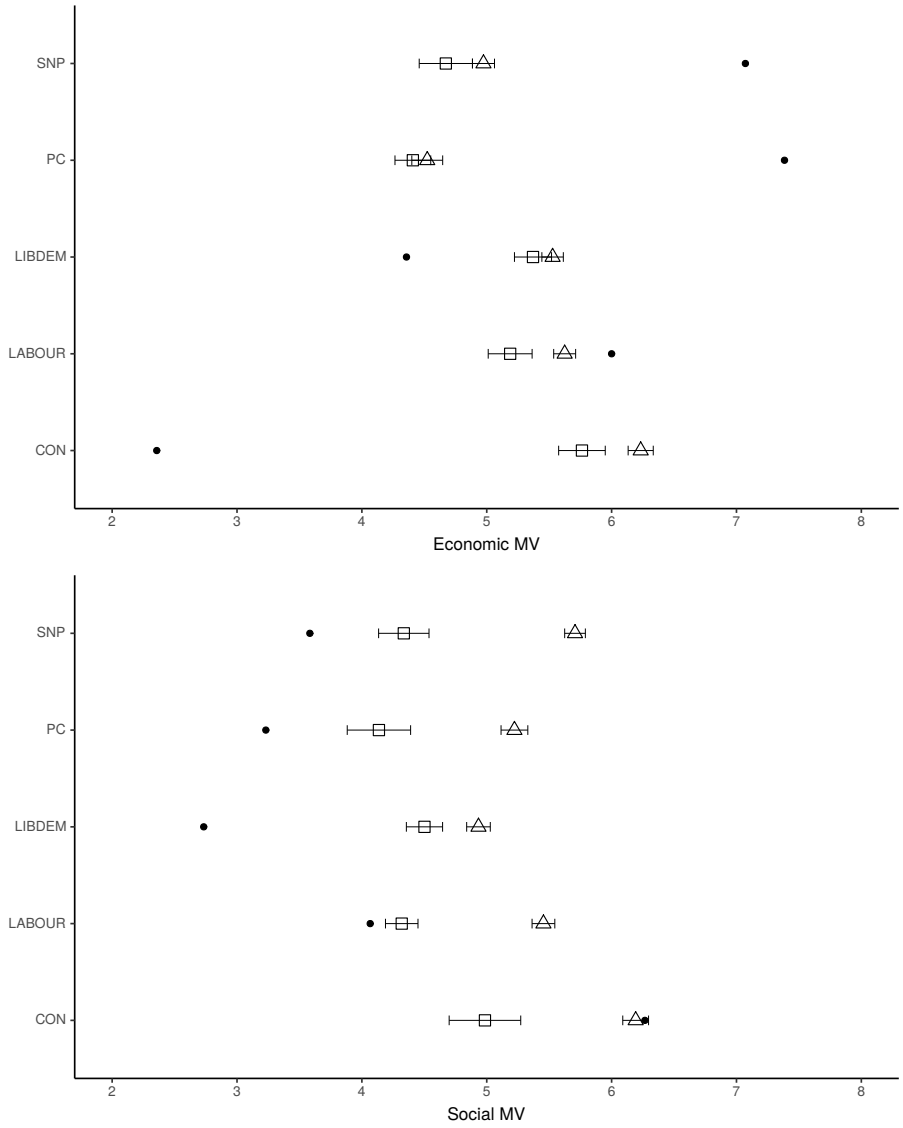
While the example using British party Euromanifestos provides some, rather limited, evidence in favour of using parsed documents for estimation with *Wordscores*, we refrain from extending the analysis to all countries in our main analysis for two reasons. Firstly, parsing documents two different policy areas is time-consuming process (as some Euromanifestos are several dozen pages long) and requires extensive skills in many European languages. Investing so much time and so many resources would make an unfair test for a method that promised to be 'simple', 'quick', 'easy', 'effortless', and 'language blind' (Laver, Benoit & Garry 2003, 312, 317, 323–326, 328). Secondly, parsing manifestos into different policy area can, by large, involve a lot of subjective choices, as it has been shown for the Manifestos Project (Däubler et al. 2012, Mikhaylov et al. 2012). We leave it to researchers to decide whether parsing documents before estimation is worth the effort when it comes to using *Wordscores*.

Figure 5: Parsed versus unparsed documents estimates on the socio-economic and socio-cultural dimensions (LBG transformation).



Note: Triangles indicate *Wordscores* estimates based on uparsed documents, squares indicate *Wordscores* estimates based on arsed documents, and dots indicate the 2006 CHES estimates.

Figure 6: Parsed versus unparsed documents estimates on the socio-economic and socio-cultural dimensions (MV transformation).



Note: Triangles indicate *Wordscores* estimates based on uparsed documents, squares indicate *Wordscores* estimates based on arsed documents, and dots indicate the 2006 CHES estimates.

References

Bakker, Ryan, Catherine de Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen & Anna Milada Vachudova. 2015. "Measuring party positions in Europe: The Chapel Hill Expert Survey trend file, 1999–2010." *Party Politics* 21:143–152.

Benoit, Kenneth & Michael Laver. 2006. *Party Policy in Modern Democracies*. Routledge.

Benoit, Kenneth & Michael Laver. 2008. "Compared to what? A comment on 'A robust transformation procedure for interpreting political text' by Martin and Vanberg." *Political Analysis* 16:101–111.

Braun, Daniela, Slava Mikhaylov & Hermann Schmitt. 2010. "Manifesto study documentation advance release." (pre-release B, 22 July 2010), `http://www.piredeu.eu/`.

Carmines, Edward G. & Richard A. Zeller. 1979. *Reliability and validity assessment*. Quantitative Applications in the Social Sciences, no. 17 Sage.

Däubler, Thomas, Kenneth Benoit, Slava Mikhaylov & Michael Laver. 2012. "Natural sentences as valid units for coded political texts." *British Journal of Political Science* 42:937–951.

Gemenis, Kostas. 2013. "Estimating parties' positions through voting advice applications: Some methodological considerations." *Acta Politica* 48:268–295.

Grimmer, Justin & Brandon M. Stewart. 2013. "Text as data: the promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* 21:267–297.

Hooghe, Liesbet, Ryan Bakker, Anna Brigevich, Catherine de Vries, Erica Edwards, Gary Marks, Jan Rovny, Marco Steenbergen & Milada Vachudova. 2010. "Reliability and validity of the 2002 and 2006 Chapel Hill expert surveys on party positioning." *European Journal of Political Research* 49:687–703.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to its Methodology*. Second ed. Sage.

Laver, Michael, Kenneth Benoit & John Garry. 2003. "Estimating the policy positions of political actors using words as data." *American Political Science Review* 97:311–331.

Lowe, Will. 2008. "Understanding Wordscores." *Political Analysis* 16:356–371.

Martin, Lanny W. & Georg Vanberg. 2008. "A robust transformation procedure for interpreting political text." *Political Analysis* 16:93–100.

Trechsel, Alexander H. 2010-11. "EU-Profiler: Positioning of the parties in the European Elections [computer file]." European University Institute [distributor], V5.