VALIDATING WORDSCORES FOR USE IN VOTING ADVISE APPLICATIONS

Sebastianus Cornelis Jacobus Bruinsma, BSc

EUROPEAN STUDIES

Dr. Kostas Gemenis (Universiteit Twente) Prof. Dr. Oliver Treib (WWU Münster) Dr. Andreas Warntjen (Universiteit Twente)

JULY 2014

UNIVERSITEIT TWENTE.



Westfälische Wilhelms-Universität Münster

ABSTRACT

Since 2006, voting advice applications have shown a significant increase in use in Europe. One of the main factors in a successful VAA is its ability to accurately locate parties in policy space. *Wordscores*, a quantitative text analysis scaling method, has been proposed as a method that can be used to accurately estimate positions on *a priori* specified dimensions. This study tries to establish if *Wordscores* is able to perform this function in the context of voting advice applications. To do so, I apply Wordscores on parties' 2009 Euromanifestos and compare the estimates against a commonly accepted benchmark: the party positions of the 2009 EU Profiler voting advice application. The analysis showed that in countries as varied as the Netherlands, Germany, Slovakia, Finland, & Poland, Wordscores was unable to match the benchmark' estimates. The concordance correlations coefficients between wordscores and EU Profiler estimates are too low for the method to be able to be judged valid.

TABLE OF CONTENTS

1	lr	ntroduction	. 4
	1.1	Research Question	. 5
2	L	iterature	. 6
	2.1	About Voting Advice Applications	. 6
	2.2	Party Placement In Practise	. 8
	2.3	Party Placement In Voting Advice Applications	11
	2.4	What Is Wordscores?	13
	2.5	Faults In The Method	17
3	S	tudy Design	20
4	D	Data	22
	4.1	Benchmarks	23
	4.2	Sample	25
	4.3	Dimensions	25
5	Ρ	re-Processing Of Documents	26
6	A	nalysis In Other Countries	29
7	C	Conclusion	31
8	R	ecommendations	32
9	R	eferences	33
A	ppe	endix I – List of Parties	38
A	ppe	endix II – Remarks on the data file	42
A	PPE	NDIX III – Wording of questions in benchmarks	43
A	PPE	NDIX IV – Results For The Netherlands	45
A	ppe	endix V – Number of words for 2004 reference manifestos in the netherlands	47
A	PPE	NDIX VII - Spearman's Correlations For The Left-Right Dimension In The Netherlands	48
A	ppe	endix VIII - Results for the Netherlands, Germany, Slovakia, Finland & Poland	49
A d	PPE ocu	NDIX IX – Word count and unique words for Germany, Slovakia, Finland, and Poland ments	51

1 INTRODUCTION

Voting Advice Applications (VAAs) are online applications that aim to give the prospective voter a voting advice, by matching their political preferences with the positions of parties or candidates. VAAs are built around a database containing the parties or candidates' positions on a number of issues. A graphical interface allows voters to express their own views on the same issues by indicating their position on a scale, or either choosing if they agree or disagree with a statement. By calculating the proximity of the voter to the position of the parties, the VAA can present the voters their position in the political landscape, sometimes accompanied with a voting recommendation.

Over the last years, the popularity of VAAs has shown a spectacular rise. For example, during the 2010 general elections in the Netherlands, Stemwijzer provided 4.2 million users with a voting advice, roughly 45% of the electorate¹. Fivaz et al. (2010) explain this popularity of VAA's because they take away the costs voters make to collect the information to make a decision what to vote. According to Garzia (2010, pp. 18–19) VAA's already *procure* the relevant information on any major policy issue, *analyse* the data and compare it to the position of the voter, and *assist* the voter in giving him a voting advice. The benefits of this are that VAA's lead to a higher voter turnout, a fact often claimed by the designers of the VAA's, and supported by findings in the Netherlands, Switzerland, and Finland (Carlson & Strandberg, 2005; Fivaz et al., 2010)

There are five aspects in the design of a VAA: issue selection, issue formulation, party positioning, voter placement, and presentation of results. Any changes in one of these five aspects changes the VAA and with it, its outcome (Fossen, Anderson, & Tiemeijer, 2012; Louwerse & Rosema, 2013; Walgrave, Nuytemans, & Pepermans, 2009). This research focuses on the third aspect: the positioning of the parties. The way a party is positioned has consequences for the party and the voter: voters could be advised to vote for a different party than the one closest to their interests, and parties can "miss" voters who would otherwise have been advised to vote for them. Non-accurate placings of parties can thus lower the reliability, and usefulness, of a VAA. Current methods of party positioning are based on the ideas of party self-placement, sometimes combined with expert opinions. Nevertheless, while being accurate enough to use, these methods are not flawless. Expert opinions are sometimes difficult to verify, and idiosyncratic interpretation of the questions influences both its reliability and validity (Krouwel & van Elfrinkhof, 2013). In addition, experts often have difficulty placing new parties, parties that have been shifting position, or cases where many parties

¹ See http://www.stemwijzer.nl/Nieuws2/Recordaantal-bezoekers-voor-StemWijzer

are involved (Marks, Hooghe, Steenbergen, & Bakker, 2007, p. 27). Party placements, while simple to carry out, suffer from strategic use of political parties, the sometimes-blunt refusals by parties to place their party at all (Gemenis, 2012, pp. 4–5), or from parties 'making up' positions on issues they have not yet developed a position on (Krouwel & van Elfrinkhof, 2013, pp. 4–5)

This research looks into a new way to place parties in VAA's by using automated content analysis. Krouwel & Elfrinkhof (2013) have shown that these kinds of analyses can produce party positions that carry high reliability, verifiability, and cost and time effectiveness, but on the other hand lack validity and comparability. This research will test if these disadvantages are indeed problematic for the use in VAA's. To do so, it will make use of one often-used method for automated content analysis, the *Wordscores* method developed by Laver, Benoit, & Garry (2003). The results of this method will be compared with established benchmarks, and the results will either deem the method useable for VAA's, or not.

Until now, automated content analysis has not yet been considered for use in VAA's (however, see a paper presented by Baudewyns et al. (2011)). This research will try to give a first assessment if it can be of any use. To do so, it will test if Wordscores is able to position the parties on for the 2009 European Parliament elections on three different dimensions (left-right, EU integration, and EU Foreign and Security Policy) in the same way as a series of benchmarks does. If so, this proves that the Wordscores method is able to place parties in a correct way, suitable for VAA's. Wordscores is a good choice to start with, for despite the fact that Wordscores is a popular method for party positioning, as yet, to our best knowledge, no authors have used it in VAAs – a gap this research aims to close. Results outside VAA's seem promising, as the method has been used to a fair degree of success by several authors (Benoit & Laver, 2003; Hug & Schulz, 2007; Proksch & Slapin, 2009; Ruedin, 2013). The simplicity of the programme, its relative ease-of-use and cost-effectiveness, make this programme worthwhile for research and a validation can have consequences for the future design of VAA's.

1.1 Research Question

In this research, we focus on the positions of political parties taking part in the 2009 European Parliament elections, on three different dimensions, left-right, EU integration, and the Common and Foreign Security Policy (CFSP). We want to know whether the score Wordscores assigns to these

parties on the dimensions, is the same as those of well-established benchmarks. Thus, our main research question is:

Can Wordscores be used to estimate parties' positions on policy issues and dimensions of interest to VAAs in a valid way?

2 LITERATURE

2.1 About Voting Advice Applications

Voting advice applications are based on the theoretical model of issue voting (Andreadis, 2013; Mendez, 2012). This model lies, together with party loyalty, at the basis of models trying to predict voting behaviour. There are three different variations of this model. The first is the directional model of voting, which is based on the idea that the more voter and candidate agree on certain issues or positions, the higher the change is that the voter will vote for that candidate. The second is the spatial, or proximity, model, which assumes that an individual will choose that candidate whose political stances are closest (not necessarily the same) to their own. The third model, the salience model, sees candidates as possessing their own issues and the voter decision based upon how important, or salient, that opinion is to the voter (Borre, 2001; Iversen, 1994; Meier & Campbell, 1979; Mendez, 2012; Pierce, 1997; Tomz & Van Houweling, 2007). Altogether, this means that voters will choose the candidate that stands closest to them, voices their ideas the most clearly, and whose issues he considers most important. VAA's try to tap into these variations by asking not only for a voter's position on an issue, but also how salient he finds it. Based on this information, a VAA tries to show a voter which party 'fits' him best.

VAA's do so in a multitude of different ways. According to Garzia & Marschall (Garzia & Marschall, 2012, pp. 207–209) there are nine aspects that characterize a specific VAA: the organization in charge, the range of parties involved, the quantity and quality of the issues involved, the development of these issues, the identification of the party's positions on these issues, the answer pattern, the weighing of the issues, the calculation method, and the illustration of the voter's proximity to the parties. These aspects can characterize how detailed a VAA is, how politically independent it is, and

how accurate its predictions are. Here, we are primarily interested in the first part of this process and then especially in which issues to take and how to score parties on them.

Deciding which issues to select has a considerable influence on the workings of the VAA. Most important, the issues should tap into dimensions that can be used to place all parties accurately. Moreover, the issues themselves should be balanced: in other words, there should not be a single party, which "owns" a majority of the issues, as it would make the VAA more favourable to them. Indeed, a study by Walgrave, Nuytemans, & Pepermans (2009) in Belgium shows that voters are provided with a different voting advice based on the issues selected. The issues selected tap into either a low-dimensional or high-dimensional idea of party space. In the low dimensional model, one or two dimensions are decided *a priori*, to best represent the party space. Popular dimensions in the European context are the left-right dimension or the pro-European integration/anti-European integration dimension. The selected issues are than seen as "pointing" towards either a left or right position. In the *high dimensional* model, specific issues point to specific dimensions. For instance, a question on the opening up of the borders points to an immigration dimension. As such, a VAA's has as much dimensions as it has issues (Mendez, 2014).

When the decision is made which issues to select, the parties need to be assigned a position with regard to this issue. Nevertheless, getting a party's *true* position is difficult, as parties tend to behave in a strategic way in order to gain the largest amount of votes. Such a *Downsian* perspective sees party space as a market in which parties try to obtain the largest share. A party might thus consider it wise to take another position than its true position in order to obtain a larger share (Downs, 1957)². Moreover, until now we have considered that parties are unitary actors. However, in reality, it may well be that the party elite has a different perception of the position of the party than the party members. If we then want to determine the position of a party, we need to get "inside" the party and see how the different standpoints have evolved, for example, through a party congress. However, most scholars, for sake of clarity, tend to add all the different positions within a party together and view the *composite* position as the position we are interested in, ignoring the way in which this position has come about.³

² Of course, parties may lie about their position, for example to obtain a more popular position in a VAA. However, most VAA's, such Stemwijzer and Kieskompas, have ways to filter these untrue statements out and correct them.

³ Benoit & Laver (2006) also consider the differences between a party in government and a party in opposition, or a party in a coalition or a party ruling alone. Parties, they argue, can change positions overnight in such cases

2.2 Party Placement in Practise

Considering we view the party as a single unit for analysis, how do we practically go about measuring its position? First, we need to decide how to measure the positions. Are we doing this second-hand, i.e. with help of evaluations of these positions by politicians or voters, or are we doing so directly, with help of political manifestos, interviews with party leaders, or speeches made at party meetings (Hooghe et al., 2010). Second, we need to decide if we are measuring high-dimensional or low-dimensional and if we want to specify these dimensions *a priori* or want to "discover" them later. Last, we need to think of the statistical properties of our placement (Benoit & Laver, 2006, pp. 90–92). How precise, accurate, reliable, and with what levels of uncertainty do we want to have our positions? Of course, while it would be preferable to have a VAA that is as precise as possible, has high accuracy and reliability and low levels of uncertainty, such benefits come at the price of time and resources, which are dependent on the organization designing the VAA (see Garzia & Marschall's list above).

Out of these three considerations, seven ways to place political parties are identified by political scientists (Benoit & Laver, 2006; Krouwel & van Elfrinkhof, 2013; Mair, 2001):

- I. Expert surveys, in which experts rank parties on certain dimensions
- II. Mass or voter surveys, in which respondents place the parties on what they think is their position in the spectrum
- III. Elite studies, such as studies amongst politicians or high-ranking government officials
- IV. Secondary reading, using historical sources such as newspapers, studies and other media
- V. Analysis of roll-call votes
- VI. A priori judgements, in which parties are placed on the basis of their origins or basic beliefs
- VII. Analysis of political contents, often manifestos, either manual or automatic

and party positions may well be different. Nevertheless, they also conclude that very little work has been done on the subject, which is the reason we will not go further into it here. All of these approaches have their advantages and disadvantages and each approach pays for its strength at the cost of a weakness in the other (Volkens, 2007, p. 118). We will handle all six of the above methods here, both focus mostly on the automated forms of content analysis.

Expert surveys have the advantage that they use experts knowledgeable on the subject, who are able to integrate all their knowledge into establishing their opinion on the position, and that these type of surveys can be deployed quickly, inexpensively, at relative ease across a wide range of countries(Benoit & Laver, 2006, pp. 114–116), and that they can be directly quantified into a type of score (Marks, 2007). Their disadvantages are that different experts can have different opinions of what parties are, which time-period they need to take into account, and if parties are to be judged on the promises or merits (Volkens, 2007). These disadvantages withstanding, expert studies are still the preferred approach for correctly identifying party positions (Benoit & Laver, 2007b; Klemmensen, Hobolt, & Hansen, 2007; Laver et al., 2003). Many studies view them as "authoritative" and as a benchmark against which other methods can be compared (Ruedin, 2013, p. 85), despite their shortcomings. This is not only because of its advantages and because of proofed usefulness, but also because other methods have not yet obtained the same level of acceptance and reliability in the field. For this study, we will subscribe to the consensus and use the expert studies as a benchmark to test our case against, while keeping the critiques in mind.

Voter surveys share the same disadvantages as expert surveys, but lose both the cost-effectiveness part, as a large sample is needed to obtain valid results, and the political knowledge part, which is in any case smaller than those of the experts (Krouwel & van Elfrinkhof, 2013). Moreover, non-response is higher, especially among those less politically engaged, leading to skewed results (Latcheva, 2011).

Elite studies have the advantage that they allow parties themselves to state their own position, which should be close to their real position, and that a relative small number of respondents are needed. However, not all parties are unitary actors and the selection which part of the elite is surveyed and what their own strategic behaviour is, can lead to different positions then the "real" one – and thus leads to "insincere" positions (Benoit & Laver, 2006, p. 93). Moreover, the timing of the survey (close to the elections, before or after) and the image the party wants to present influence the quality of the data (Debus, 2009, pp. 287–288; Krouwel & van Elfrinkhof, 2013).

Secondary reading involves the use of other studies and newspaper articles to establish the position of a political party. For example, changing descriptions of a party in the newspapers can be used to chart the movement of a party during the elections. Nevertheless, as the results are based upon the

9

researcher's subjective judgment, it is hard to incorporate this method in a statistical framework, which makes it unsuitable for many purposes (Hakhverdian, 2009, p. 725).

Roll-call analysis analyses what parties actually voted for, instead of what they promise to vote. Advantages of this method are the large availability of data, the cost-effective way of analysis, and the possibility to plot party positions over time. Nevertheless, due to differing parliamentary systems and differing political cultures, cross-national comparison will be difficult. In addition, "new" parties have had less possibility to vote than older established parties, leading to a lower level of certainty about their position, and roll-call voting may be used strategically by parties in the political process (Debus, 2009; Hug, 2009).

A priori judgments are rude, often ordinal, measures of party placement, based upon were a researcher feels that a party is positioned, taking all aspects in account. This is mostly based upon what is considered as "common knowledge". As with secondary reading however, the measure is extremely subjective, and therefore unsuitable for many types of analysis.

Content analysis uses political documents as published speeches, conference papers, and party manifestos to establish the position of a political party. In most research manifestos are used, as there is a high degree of availability of documents, and that the position in the party manifesto is usually the mean of the opinions of all the intra-party groups (Hug, 2009, p. 289). Content analysis can be divided in manual or automatic methods. The difference between them is that manual methods are based on the information in the sentences or the manifestoes themselves. In other words: they analyse what the manifestoes or sentences "tell". From this perspective, a sentence can have a negative or positive meaning, and can be analysed as such. Automated methods are based on the idea of the relative frequency of the words used in a manifesto. Instead of using the information that sentences "tell", party positions are derived from the different issues parties emphasize, and thus the different words they use. Even on similar issues, such as the European Union, parties use different words to frame their positions. The underlying assumption here is that of saliency: the more often a word, or expression, associated with a particular position is used (making it more salient), the more close the party is said to stand to this position. For example, even on the same issue, right-wing parties tend to speak more of "illegal immigrants" where left-wing parties speak of "asylum seekers". The choice of words thus tells something about the party itself.

Manual content analysis, of which the Comparative Manifesto Project (CMP) is the main representative, has as its main advantage that over time, the project has built an impressive database

of party positions, making the CMP an often-used tool in comparative politics. However, various scholars have pointed out several theoretical and methodological flaws in the method. Gemenis (2013) divides these in four aspects: the theoretical underpinnings of the method are not validated, the selection of the documents is haphazard and contains not only manifestos but also other party documents (see also the investigation of Hansen (2008) of the documents used in Denmark), the reliability of the coding is lacking as each document is only coded once by a single voter (see also Benoit, Laver, & Mikhaylov (2009)), and the method of scaling especially left-right issues is questionable. Besides, the different coding categories have not been updated since the eighties, thus ignoring any change in the world, and making it increasingly difficult for coders to categorize new events (Krouwel & van Elfrinkhof, 2013, p. 10). Another form of hand-coded analysis is the manual coding of manifestoes using checklists. Here, instead of sentences (or quasi-sentences, which more often corresponds to a small portion of a sentence), the whole manifesto is coded. Using a checklist, the coder is given several questions, and is then asked to give the manifesto's position on it. Again, the added up scores of this make the parties', or rather the manifestos', position on the issue.

The *automated forms of analysis* have the advantage over hand-coded forms that they are highly reliable, can be easily verified as both the input and the algorithm used are known, and are able to analyse their input quick. Moreover, given the same data, the method if fully replicable for anyone wanting to confirm the estimates it makes (Trechsel & Mair, 2011, p. 2). The automated forms now used all stem from the semi-manual "dictionary approach" such as used by Laver & Garry (Laver & Garry, 2000). Two of the most used methods are the Wordscores method (Laver et al., 2003), which compares known reference texts with unknown virgin texts, and the Wordfish method (Slapin & Proksch, 2008), which uses a scaling method instead of the reference texts. All of these methods however, lack validity as the context of words and sentences analysed is hardly taken into account (Krouwel & van Elfrinkhof, 2013, pp. 11–12), which makes that automated methods always need to be validated in order to be of any use to the researcher (Grimmer & Stewart, 2013).

2.3 Party Placement in Voting Advice Applications

Not all of the methods described thus far have been used to position parties in VAA's, and sometimes VAA's have developed alternative ways to position parties. Broadly speaking, VAA's use four methods to position parties: the above-mentioned expert studies, self-positioning, the so-called "Kieskompas"-

method in which self-positioning is later corrected by experts, and the Delphi method, which uses consensus across a panel (Garzia & Marschall, 2012; Garzia, 2010; Gemenis & van Ham, 2014).

The most used of these is *self-positioning*, which is broadly comparable to elite studies. VAA's such as Stemwijzer, VoteMatch, and Wahl-O-Mat use this method, which in its most simple form consists of a questionnaire sent to parties in which can are asked to place themselves on various issues and justify this claim. However, not only are parties not always keen to do so, as with elite studies, parties may have ulterior motives to place themselves on a certain position (Gemenis & van Ham, 2014). For example, Wagner & Ruusuvirta (2011) describe a case in Finland, where a candidate had chosen the middle category for all issues, thus hoping to catch the largest amount of voters.

The "Kieskompas" method tries to circumvent such problems by combining the idea of selfplacement with expert opinions and analysis of party manifestos. The questionnaire is still sent to the parties, but concurrently, a team of experts place the parties based on their election manifestos and statements. In cases where both disagree, parties are sent the text snippets that made the experts to place them otherwise, and are asked to reconsider their position or justify the difference. In the end, either the experts or the parties change their claim and adjust their position (Krouwel & van Elfrinkhof, 2013; Krouwel, 2012, pp. 194–197). Despite these solutions, however, the Kieskompas method still suffers the same problems as self-placement as the parties can still refuse to take part in the study. Moreover, it is unknown how the expert groups reach their consensus and which factors played a role in the ultimate decision.

To combat the second problem, Gemenis (2012) proposes to use the *Delphi-method*, in which experts work independently of each other and are asked to position the political parties and justify their thoughts. A moderator then collects this information, feeds them, anonymously, back into the panel, and asks them to update their estimations. Once consensus is reached, the results are aggregated en used for the VAA. The advantages of the method are that the experts work more structured, but also that the parties are not needed, thus solving the problem of strategic behaviour. However, as the method asks for considerable estimation and updating, the process can take a long time, making the process considerably more expensive and time-consuming.

The use of *roll-call behaviour* for VAA's is relatively rare. Current examples are Stemmentracker (The Netherlands) and KohoVolit.eu (Finland). Nevertheless, roll-call behaviour has an "excessive" concentration on the past (Škop, 2010), is not able to contain new parties, and cannot respond to new issues. Manual content analysis, such as the Comparative Manifesto Project (CMP), suffers from the

same problems, as the CMP has a pre-defined coding scheme. Other forms of content analysis, such as automated content analysis, can circumvent this problem, but suffer from a lack of validity. Automated content analysis also is rarely used, for most part because the positions are still deemed non-valid enough to place parties exactly (Volkens, 2007).

Automated methods are still untried in VAA's. Nevertheless, the methods have the possibility to beat the current methods on the fields of verifiability and reliability. Besides, the downsides of the method, its lack of validity, may also not be as large a problem as thought. In various studies, Wordscores has proved to be a reliable estimator to measure valid party positions (Benoit & Laver, 2003; Hug & Schulz, 2007; Proksch & Slapin, 2006; Ruedin, 2013). Moreover, the method allows us to calculate measurements of uncertainty, making it possible to give an indication of how precise the measurement of the party's position is.

2.4 What Is Wordscores?

Unlike other methods, Wordscores is not dependable on either predefined dictionaries or placing by human coders. By treating words as "data", and assuming that the relative frequencies of specific words provide manifestations of underlying political positions (Klemmensen et al., 2007, p. 748), the programme is able to place parties on any "a priori" specified dimension.

To commence analysis we first need a set of reference texts that each have a position on a dimension that we can assume with some confidence (for example, when they are obtained through expert coding). We can call this position of reference text *r* on dimension *d* the *a priori* position of the text, A_{rd} . In addition, as we are looking at the frequency of words used, we can call the frequency of each different word *w* used in the texts F_{wr} . Once we have calculated F_{wr} for all reference texts, the probability that an occurrence of word *w* implies that we are reading text *r* is⁴:

$$P_{wr} = \frac{F_{wr}}{\sum_{r} F_{wr}}$$

We can then use this probability to calculate the position of any given word on the specified dimension, giving the word a score:

⁴ I use the notation as used originally by Benoit, Garry, and Laver (2003). Different notations are used by Lowe (2008) and Martin & Vanberg (2008a, 2008b)

$$S_{wd} = \sum_{r} (P_{wr} \times A_{rd})$$

Thus, we have calculated the Wordscores for all the words used in the reference texts. Now we can use these to calculate the position of the texts we want to know the position of (the so-called *virgin texts*)⁵. To do so, we first calculate the relative frequency of each virgin text word, as a proportion of the total number of words used in the virgin text, just as we did with the reference text. This frequency (F_{wv}) can then be used to calculate the total score of the virgin text (S_{vd}). The score is then in essence the mean score of all the scored words, weighted by their frequency:

$$S_{vd} = \sum_{w} (F_{wv} \times S_{wd})$$

This score S_{vd} is what can then be used for further analysis and is called the *raw score*. However, while these raw scores can be used to study relative movement between parties, for any substantive interpretation of results, we need to transform the raw scores back to the original metric, as this allows us to compare the raw scores of the virgin texts with the assigned scores of the reference texts and thus spot party movement (Laver et al., 2003; Martin & Vanberg, 2008a)

In our case, we are not interested in party movement, but we are interested in placing the parties back on the original metric. This because we rescaled all benchmarks to the same 0-1 scale. In order to be able to compare the obtained raw score to the 2009 benchmark, we want our raw scores to be on the same metric as these benchmarks. As the metrics of the reference texts (and their assigned scores) are the same as these of the benchmarks however, it is enough to make sure the scores are placed on the original metric, without the need for any further transformation.

In their original paper, Laver et al. (2003) suggest the following transformation:

$$S_{vd}^{*} = (S_{vd} - S_{-vd}) \left(\frac{SD_{rd}}{SD_{vd}}\right) + S_{-vd}$$

Here, S_{vd}^* is the transformed score, S_{vd} the raw score, S_{-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. This metric

⁵ Indeed, we can only score words in the virgin documents that also occur in the reference document. Thus, the more the set of reference documents differs from the set of virgin texts in terms of word use, the less useful the method is.

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 this 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 "but", and as they occur frequently in *all* documents, they get centrist scores. As such, the distances between the virgin texts are thus shrunken, and all texts bounce towards the mean. LBG 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, increasing polarisation between parties, or joint movement to the sides of a set of parties, is hard, if impossible, to discern as the mean of the virgin texts is always close to that of the reference scores due to shrinkage, while LBG's transformation makes sure that the variance is the same as for the referenced documents (Lowe, 2008, pp. 359–360).

Martin & Vanberg (2008a, pp. 95–97, 2008b) agree with the above criticism and note several more shortcomings of this method. First, as the transformation uses the SD of the virgin text raw scores it depends on the set virgin texts themselves. This makes the score non-robust with regard to the virgin texts, and any difference in the set automatically leads to a difference in the scores. This way, a researcher may obtain different results – and reach different conclusions – solely based on which texts are included or excluded. Second, they claim that the transformed scores and the virgin scores are *not* on the same metric, as the method fails to recover the accurate relative distance ratios between the texts. This is due to shrinkage, as we pointed out above. To combat these problems, they provide a new transformation based on the idea of relative distance ratios:

relative distance ratio for text
$$S_i = \frac{S_i - S_{R1}}{(S_{R1} - S_{R2})}$$

Here, 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, 2008a, p. 97). They then use these ratios to construct a new transformation:

$$S_{vd}^* = \left((S_{vd} - S_{r1}) \frac{A_{R2} - A_{R1}}{S_{R2} - S_{R1}} \right) + A_{R1}$$

Here, S_{vd}^* is the transformed score, S_{vd} the raw score, A_{R2} and A_{R1} are the assigned scores to reference texts R_1 and R_2 (where R1 is located to the left of R2), and S_{R2} and S_{R1} are the reference text's raw scores. In their article, Martin & Vanberg use two reference texts, or "anchor texts" (one located to the left and one located to the right of the text analysed) to analyse a single text. As seen in the equation above, both the assigned scores for the reference texts are recovered, and the text under analysis is thus placed on the original metric. However, as soon as more than two reference texts are used – which is strongly suggested by Laver et al. (2003) – not all the original exogenous scores of the reference texts cannot be recovered exactly, as only two texts can be used to define the metric⁶. The authors thus suggest a change to the transformation:

$$S_{vd}^{*} = \left((S_{vd} - S_{min}) \frac{A_{max} - A_{min}}{S_{max} - S_{min}} \right) + A_{min}$$

Here A_{min} and A_{max} 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 because 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, 2008a).

To conclude, while the transformation by LBG depends on the virgin texts and is indifferent to the composition of the reference texts, the transformation by MV depends on the reference texts and is indifferent to the composition of the set of virgin texts (Lowe, 2008, p. 360). Moreover, LBG assume that the variances of both the set of reference texts and virgin texts are the same, while MV's transformation does not do so (Benoit & Laver, 2008, p. 110). *In this paper, we use both scoring 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.*

⁶ In the Wordscores suite for STATA, when using MV scaling, the programme automatically uses the two most extreme reference texts R_1 and R_n , where $A_1 < ... < A_n$ (Benoit & Laver, 2008, p. 105).

2.5 Faults in the Method

Problems with Wordscores are given in a paper by Lowe (2008) and in a series of papers between Benoit & Laver (2007a) and Budge & Pennings (2007a, 2007b). Lowe (2008) criticizes the method primarily on its heavy dependence on reference texts. He views Wordscores as an approximation to correspondence analysis and goes on to treat the method as a statistical ideal point model for words. Doing this, he identifies six conditions that Wordscores needs to fulfil in order to ensure consistent and unbiased estimation of the parameters of the ideal point model:

- 1. The Wordscores of the virgin texts need to be equally spaced and extend over the whole range of wordscores for the reference texts
- 2. The Wordscores of the virgin texts need to be spaced relative to the informativeness term (all texts are thus informative)
- 3. The scores of the reference texts need to be equally spaced and extend past each wordscore of the virgin texts in both direction
- 4. The Wordscores 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, condition 5 & 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 (Lowe, 2008, p. 369). 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 & 2 suffer when there is not enough overlap between word distributions between the reference documents, but with many documents available, the conditions might be well approximated. With regard to condition 1, we tried to include manifestos of the same parties in the batch of virgin texts and the batch of reference texts. This way, the texts more or less extend over the whole range. Condition 2, however, suffers from the fact that some documents are small, and thus contain very little to no information. This thus creates a large bias in the estimates, negatively influence the validity of the new document scores. Conditions 3 & 4 are similar to 1 & 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, p. 369), which makes the conditions even more plausible. Lastly, Lowe (2008, p. 369) considers that condition 1 & 3 can never hold simultaneously, as this would require an infinite data set – and thus concludes that bias in wordscores is inevitable.

Budge & Pennings (2007a, 2007b) also criticize the heavy dependence on reference texts and compare the method against Comparative Manifesto Project estimates. Their findings are unfavourable as they find Wordscores *flattening* out the scores compared to the CMP estimates. However, in a response, Benoit & Laver (2007a) criticize these findings as Budge & Pennings applied the Wordscores method in the wrong way, as well, as used the CMP as the gold standard while, the authors argue, it is flawed in several places. These latter findings are supported by Klemmensen et al. (2007), who use Danish party manifestos to compare Wordscore to the CMP and find that the CMP only outperforms Wordscores slightly. Moreover, in another article, Hansen (2008) criticizes this specific CMP dataset of the Danish party manifestoes for having large variations in the number of coded sentences, having large variations in length, and for sometimes being no manifesto at all. As such, using the CMP as the gold standard would be questionable in the first place.

Summarizing the above, the main threats to Wordscores are its proneness to generate unreliable scores when dealing with a small amount, or short, documents, its reliance on well-chosen reference texts and its clustering of results towards the mean. We could circumvent these problems by using large corpora of text, using accurate reference scores for our documents, using reference scores at the extremes of a well-defined scale, using equally spread scores, and by using long texts (Klemmensen et al., 2007; Lowe, 2008)

In our case, this is problematic. First, while LBG state that the method is word blind, we do have to know something about the documents under analysis and the political situation in the country to be able to use the method successfully (Klemmensen et al., 2007, p. 754). For example, Benoit and Lowe (2013) show that while analysing speeches of MP's in parliament, it is necessary to know the content of the text to decide when the MP is talking about the issue or whether he goes off-topic. Party manifestos also contain a good deal of off-topic material, such as biographies of party members or profiles of voters, which add no to very little information. Besides, the researcher should know which parties can be relied on to hold extreme positions, and if the reference and virgin documents are of the same kind. While we could argue that anyone establishing a VAA should be able to know

something about the country, the idea of using the method in a European context implies that a multitude of countries and languages are used. As we are unaware of the political context and language of some of the countries analysed, this poses a challenge. We thus have to "trust" that the documents we analyse are indeed of the same nature and that the extreme texts indeed to define the extremes. Second, as we analyse our set of documents per country, per dimension, the resulting corpora of texts is small (at an average of 10 manifestos). In addition, the length of the documents varies, with especially the extreme right or left parties having remarkably short manifestos. Moreover, not all of the 2004 manifestos were digitized and thus obtainable, leading to some gaps in the corpus. Third, in many cases the reference scores are not equally spread over the dimension analysed. Often, scores are clustered either to the extreme dimensions, or in the middle. In addition, scores for extreme reference texts are often lacking in the data set.

3 STUDY DESIGN



The study design used is shown in the diagram below.

In step 1, we will use position estimates to score the reference texts on pre-defined dimensions. These scored reference texts will then be fed into the Wordscores algorithm in step 2, and the words in the reference documents will then be scored. In step 3, we will feed Wordscores the virgin manifestos, which are then scored, using the word scores calculated. The results of this analysis are the *raw data scores*. We do so using the STATA statistical software package using a command line⁷ written by Kenneth Benoit.

Benoit & Laver (2008, pp. 103–106) describe four ways to analyse these: use no transformation at all, use relative distance ratios, use the original method by Laver, Benoit & Garry or use the method by Martin & Vanberg. Of these four methods, we only use the latter two methods. This is because we do not want to observe the *relative movement* of the positions of the manifestos, but are only

⁷ See http://www.tcd.ie/Political_Science/wordscores/software.html

interested in the positions of the 2009 manifestos themselves. In VAA's, the difference in position of a party manifesto with its predecessor is not important – it is its position that counts. As such, the distance and the relative distance are not of interest. The Martin & Vanberg (MV) and Laver, Benoit, and Garry (LBG) scoring, however, give us scores that we can compare to expert scores. No one of the methods is the most suitable. Both give different results and are based on a different idea. To see if there is any substantial difference between them and if one of them is more suitable for use in VAAs, we will include both in this study. After the transformation, for both the MV and LBG scores (step 5), we calculate the concordance (step 6) between the benchmarks and the Wordscores.

By calculating the concordance, we assess to what degree the party positions in the benchmarks and the party positions as given by Wordscores are the same. When the concordance is perfect, this would mean that both scores are exactly the same. To calculate the concordance, we can rank-ordered methods (such as Spearman's Rho, Kendall's tau *a*, or Kendall's tau *b*) or continuous methods (such as Pearson's *r*). Problematic with the Spearman's Rho or Kendall's a, however, is that because they assign ranks to values, they cannot handle so-called ties in the data set. In other words, when two parties occupy the same position (which happens frequently), the method does not function correctly. Kendall's b overcomes this problem by making adjustments for these ties. Nevertheless, these non-parametric tests assume rank order, while the party positions are in fact continuous data. Nevertheless, in order to use Pearson's *r*, we need to assume bivariate normality, which is not satisfied by the skewed (or non-normal) nature of the data (Gemenis & van Ham, 2014).

In order to overcome these problems, we use the concordance correlation coefficient (Lin, Hedayat, Sinha, & Yang, 2002; Lin, 1989, 1992, 2000; Nickerson, 1997). This coefficient is defined as:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

Where μ_x and μ_y are the means for the two positions σ_x^2 and σ_y^2 are the corresponding variances, and ρ Pearson's correlation coefficient. A ρ_c of 0 would denote absence of concordance, while a ρ_c of 1 denotes perfect concordance. Lin's CCC seems to overcome most problems mentioned above, and has been proven to work for as little as 10 pairs of data (Lin, 1989). Following McBride (2005, p. 6), the following strength-of-agreement criterion to assess the agreement is used:

Almost perfect:	> 0.99
Substantial	> 0.95–0.99
Moderate	0.90–0.95
Poor	< 0.90

Though this criterion might seem strict, we use it here as we are measuring the same quantity (that is, the position of the party), and thus should expect a high concordance. Nevertheless, as an extra assessment for these benchmarks, we compare them against the correlations between the benchmarks themselves in paragraph 4.1.

4 DATA

Four types of data are used in this study: the reference documents, the virgin documents, the benchmarks for the reference texts, and the benchmarks to compare the scored documents against. In this case, the reference texts are party political manifestos from the 2004 European Parliament elections, and the virgin text are party political manifestos from the 2009 European Parliament elections. As we use documents for the same type of elections merely five years apart we can be sure that the lexicon has changed little. In addition, as the European policy space is quite diverse, we can expect the reference texts to span a considerable amount of the observed dimensions. Moreover, electoral manifestos usually contain enough words, in comparison with other material such as promotional flyers or pamphlets. Last, confident benchmarks to score the documents on are available from previous expert studies.⁸ As such, we think we are safe in our selection of reference texts (Laver et al., 2003, pp. 214–215).

⁸ For both the 2004 and 2009 manifestos, the database of the DFG-funded Euromanifestos Project is used (http://www.mzes.uni-mannheim.de/projekte/manifestos/manifestos.html). These manifestos are available in either .txt or .pdf format.

4.1 Benchmarks

We use benchmarks to validate if Wordscores is able to position parties correctly. While we agree that no benchmark is fully correct, or is able to show the true position of a political party, the methods used in the benchmarks below are comparable to those used in the current positioning of VAA's. As such, these benchmarks are already deemed sufficient enough for use in VAA's. Question is now if Wordscores is in any way able to position the parties in the same way as they do.

For the 2004 documents, the following benchmarks will be used: the 2003 Benoit & Laver expert survey, the 2004 Euromanifestos Project, and the 2002 Chapel Hill Expert Survey. The 2003 Benoit & Laver (Benoit & Laver, 2006)⁹ expert study scores parties from 24 European countries, including former communist countries, on several dimensions. A broad selection of experts from each country was asked to position the parties. This work was carried out mostly in 2003, only a year before the 2004 EP elections. Given this, we deem the estimates of the positions of parties this study makes valid enough to act as a reference. The 2004 Euromanifestos Project¹⁰ (Schmitt & Wüst, 2012) is even more specific. Using a coding scheme based on that of the CMP, all the party manifestos of the 2004 election have been coded. As these coding are done in expert teams, this gives us reliable positions of the manifestos themselves on various issues. Both variables used here specifically asked for the coder's idea of the position of the manifestos. The 2002 Chapel Hill Expert Survey (Hooghe et al., 2010) ¹¹ estimates the positions of parties in 24 European countries, using the same approach as Benoit & Laver. For the same reasons as described there, we consider this dataset useful to establish the positions of the parties.

For the 2009 benchmarks against which we compare the wordscores estimates, we use the party position data from the 2009 EU profiler (Treschel, 2010)¹², the Euromanifestos-Project 2009 (Braun, Mikhaylov, & Schmitt, 2010)¹³, and the 2010 Chapel Hill Expert Study (Bakker et al., 2012)¹⁴. The first of these contains positions actually used to score the parties in the EU Profiler VAA. As we want to validate Wordscores for use in VAA's, comparing the Wordscores positions and the EU Profiler

⁹ http://www.tcd.ie/Political_Science/ppmd/

¹⁰https://dbk.gesis.org/dbksearch/sdesc2.asp?no=4457&db=e&doi=10.4232/1.4457

¹¹ http://www.unc.edu/~hooghe/data_pp.php

¹² http://dvn.eudo.eu/dvn/dv/euprofiler

¹³ https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5057&db=e&doi=10.4232/1.10204

¹⁴ http://www.unc.edu/~hooghe/data_pp.php

positions gives us a good idea how close Wordscores is to positions actually used in VAA's. The Euromanifestos-Project 2009 is a continuation of the 2004 Euromanifestos Project and uses the same coding procedure. The same goes for the 2010 Chapel Hill Expert Study, which is a continuation of its 2002 and 2006 predecessors. An overview of the exact wording of the questions is given in Appendix III.

Before commencing the analysis, we calculate the correlations between the benchmarks. On the leftright dimension, the benchmarks seem to agree on most fronts, with CHES2010 and EUP2009 showing the lowest correlation at $\rho = 0.670$. Also for the EU Integration 2009 and EU Foreign benchmarks the values show moderate to high relation. The same cannot be said, however, of the relation between EUP2009 and CHES2010 on the Foreign & Security dimension, were correlations are low (r = 0.517). Most importantly however, all of the benchmarks (except CHES2002 and BL2003) score 'poor' when we use the criteria described above. As such, it would be partly unfair on Wordscores to expect it to perform better. Thus, when assessing Wordscores, we will use both the criteria by McBride, and the correlations we calculated here. For example, if the CCC between Wordscores estimates and the EMP2009 on the left-right dimension is 0.8, we can say that it fails McBride's criteria, which compares worse to EMP2009 than the EUP2009, but better than CHES2010.

Table 1 Correlations between benchmarks								
			Spearman's Rho	Pearson's r	Lin's CCC			
	EUP2009	EMP2009	0.848	0.8435	0.830			
	CHES2010	EMP2009	0.777	0.782	0.747			
Left-Right	CHES2010	EUP2009	0.651	0.670	0.653			
	CHES2002	BL2003	0.928	0.9652	0.962			
	EMP2004	BL2003	0.872	0.913	0.907			
	EMP2004	CHES2002	0.874	0.894	0.871			
	CHES2002	EMP2004	0.478	0.660	0.636			
EU Integration	EMP2009	CHES2010	0.731	0.737	0.728			
	EMP2009	EUP2009	0.770	0.822	0.810			
	EUP2009	CHES2010	0.750	0.710	0.700			
CFSP	EUP2009	CHES2010	0.517	0.519	0.417			

4.2 Sample

We analyse five EU member states: the Netherlands (NL), Germany (DE), Poland (PL), Finland (FI), and Slovakia (SK). We do so for several reasons. First, as we strive for maximum variation, we include countries from both Eastern en Western Europe. Countries as the Netherlands and Finland have a long tradition of party democracy, while Poland and Slovakia have a history of communist rule. Second, for all countries, both the virgin and reference manifestos are available for a width variety of parties. Third, also for all the parties in these countries, we have estimates for the positions of their reference texts, as well as benchmarks with which to compare their virgin texts (see Appendix I for a full list of parties analysed). Note that it is not important that the manifestos of some parties are only available in 2004 or 2009, as we are not comparing the parties across time.

4.3 Dimensions

As Wordscores uses an a priori approach, we need to define the dimensions we want to the manifestos to be scored on beforehand. As we are interested in applying Wordscores to VAA's, we want to include in these dimensions the two different ways in which VAA's use dimensions. They do this in a so-called *high dimensional* or *low dimensional* way (Louwerse & Rosema, 2013; Mendez, 2012). In low dimensional matching, there parties are scored on general or broad dimensions, such as the above-mentioned left-right dimension or EU integration. In high dimensional matching, the parties are scored on many different and specific issues, such as support for the Euro, the Common Agricultural Policy or the admission of Turkey into the European Union. We choose two dimensions belonging to the low dimensions, and one belonging to the high dimensions:

- I. A general left-right dimension (Left-Right)
- II. A dimension tapping in to general support for EU integration (EU Integration)
- III. A dimension displaying the position of the party leadership on EU foreign and security policy (Foreign)

We choose the first two because as a low dimension, they are present in almost all VAAs. To give Wordscores any credit therefore, it should be able to place parties on this dimension. We choose the last one because we want to test if Wordscores is able to estimate a party's position on any specific issue. This issue could be any of the 20-30 issues VAAs use. In this case, we opted for the party's position on the EU foreign and security policy as an example of such a dimension. The first two

dimensions are available in all the above-mentioned data sources except 2003 Benoit & Laver, while the third dimension is only available in the 2002 and 2010 CHES. See Appendix III for the wording of the items in the different data sources.

5 PRE-PROCESSING OF DOCUMENTS

Grimmer & Stewart (2013, p. 292) argue that pre-processing of documents is able to increase the validity of Wordscores. To test if this is the case, we make five different "sets" of methods:

- I. A raw version that is, a version directly translated from the .pdf file into a UTF-8 text file
- II. A *transformed* version the above, but lowercase, and cleaned of currencies, numbers, and other special characters, using the Transformer¹⁵ tool
- III. A stemmed version the above, but stemmed using JFreq¹⁶
- IV. A *cleared* version as in II, but cleared of so-called "stop words". Debus (2009) uses a rather random number of the 20 most frequently occurring words, we opt for the removal of stop words by use of a prepared list¹⁷
- V. A stemmed and cleared version including both stemming and clearing

Of these five, the raw version is expected to give the worst results. This as the documents converted from .pdf to .txt contain a high number of strange characters and signs which are filtered out in the transformed version. With stemming, we significantly reduce the number of words in the document. For example, *family* and *familiar* are both stemmed to become *famil*. The total number of unique words per document thus decreases this way, as will the discrimination between texts as the number of words being the same in the set of documents will be more alike. With the clearing, we remove stop words such as *and*, *the*, or *but*, which do not add information to the documents and are found

¹⁵ Holmes M., UVic HCMC (2005-2008) 'Transformer 2.0.0.0 – A Unicode batch search/replace/scripting tool', URL http://www.tapor.uvic.ca/~mholmes/transformer/

¹⁶ Lowe W. (2011) 'JFreq: Count words, quickly'. Java software version 0.5.4, URL

http://www.conjugateprior.org/software/jfreq/

¹⁷ Downloaded from http://snowball.tartarus.org/

amongst all the documents in large numbers. This leads to the documents being clustered towards the mean as words with high probability will give the documents a more centrist position.

Appendix V shows the differences in words and unique words for reference documents from the Netherlands. As we see, the overall number of "words" drops when transformed, and even further so when cleared. When stemmed, the number of words does not drop compared to the transformed documents, but the number of unique words do. The same goes for the difference between the cleared documents and the stemmed and cleared documents. For example, the CU-SGP documents contain 12,591 words in the "normal" version, with 3,276 unique words. This drops to 12,504 and 3,166 respectively when they are transformed. When stemmed, the total number of words stays the same, but the number of unique words drops to 2,784. When cleared (but not stemmed) the number of words drops to 8,174, indicating a high number of "stop words" adding no to little information to the document. When cleared (and stemmed) the number of words stays the same, but the number of and stemmed) the number of words stays the same, but the number of an addition of "stop words" adding no to little information to the document. When cleared (and stemmed) the number of words stays the same, but the number of unique words drops again.

Table 2. Means and SD's for different types of text								
ТҮРЕ	N	MEAN	STD. DEVIATION					
Normal LBG	9	-0.13	0.1					
Normal MV	9	-0.09	0.07					
Trans LBG	9	0.48	0.07					
Trans MV	9	0.07	0.05					
Cleared LBG	9	0.42	0.06					
Cleared MV	9	0.07	0.05					
Stemmed LBG	9	0.55	0.07					
Stemmed MV	9	0.11	0.08					
Stemmed & Cleared LBG	9	0.48	0.06					
Stemmed & Cleared MV	9	0.09	0.07					
Total	9	0.21	0.07					

We run an analysis for the Netherlands. scoring the 2009 documents, based upon a set of 2004 reference documents. Following LBG's guidelines on selecting reference texts, we excluded the LPF, ET, D66 and CDA from the reference set, as they contained both a low number of words and a low number of unique words. This left us with five texts, which were relatively equally spread over all three dimensions of interest, had known positions, and had a significant number of unique words. The results for the analysis. with the calculated concordance correlations, are shown in Appendix IV.

For comparison amongst the documents we only consider the left-right dimension, as this gives us the highest number of values. The normal documents differ significantly from the other sets of documents in either the LBG (F(4) = 45, p < 0.0005, Wilk's Λ = 0.0715) or MV transformation (F(4) = 45, p < 0.0005, Wilk's Λ = 0.4284). The other four groups, however, do differ significantly in the LBG (F(3) =36, p > 0.0005, Wilk's Λ = 0.6126) method, but not in the MV transformation (F(3) =36, p > 0.0005, Wilk's Λ = 0.9043)¹⁸ (see also the Spearman's correlations in Appendix VII)

For further analysis, we will thus drop the normal documents and continue using the Transformed documents. We do so as these documents involve the least amount of work, and have (apart from the normal documents) the highest number of unique words.

¹⁸ Pillai's Trace, Lawley-Hotelling Trace, and Roy's largest root gave slightly different values, but led to the same conclusions

6 ANALYSIS IN OTHER COUNTRIES

Appendix VIII shows the result of the analysis, and Appendix IX the word count for the documents. In none of the countries, concordance correlations higher than 0.899 (under the lowest 0.9 mark) are reported.

For the Netherlands, VVD, CU-SGP, GroenLinks, SP, PvdA were chosen as reference texts. CDA was excluded as the document continued to yield errors during the estimation, and D66, ET and LPF were excluded because of the low amount of words. Overall, however, the number of words was high (average 9725). The highest concordance scores were reached on the left-right dimension with LBG scoring, with 0.580 the highest overall score. Both the MV scoring and the other dimensions scored around 0, with occasional small negative values.

For Germany, B90/Die Grünen, SPD, CDU, CSU, FDP and Die Linke were chosen as reference documents. Both FDP and Die Linke had relatively long documents (10,105 and 19,434 respectively), while other documents were around 3000 words each. Both CSU and CDU documents were included however, to accommodate the right-wing words, and thus to stretch the dimensions. The best results were reached on the left-right dimension with LBG-scaling, with values between 0.665 and 0.899. MV values were significantly lower, between 0.164 and 0.251. Values for integration and foreign were clustered around 0, with 0.633 on the Foreign scale as an outlier. When compared how the benchmark compare to each other, the results are even more promising. When compared with EUP2009, the results (0.802, 0.825, and 0.899) only differ marginally from the CCC between EUP and the EMP (0.830) and are substantially better than the CCC between EUP and CHES (0.653). The same goes for CHES 2010, whose CCC's (0.757, 0.768, 0.755) are in all cases higher than those between CHES and EMP or EUP (0.747 and 0.653). The EMP is the exception, as two of the scores (0.665, 0.685) are worse (though only marginally) than those between the benchmarks (0.830 and 0.747), and one is marginally better (0.766).

For Slovakia, SMK-MKP, SMER, SDKU, LS-HZDS, and KDH were included, as these were the only parties from which reliable positions in all benchmarks could be obtained, and whose documents were longer than 1000 words. As such, documents from ANO, KSS, and SF were not included. The average length of the documents was 2,073, with values ranging between 1,111 and 3,211. Concordance correlations clustered around 0, with values ranging between -0.295 and 0.267.

For Finland, SDP, RKP-SFP, KOK, KESK, VAS, VIHR, KD, were chosen to act as reference documents. While not necessary to include for means of stretching the dimensions, the decision to include all the parties was due to the low number of words in the documents (average 1,503), with the document from VIHR only counting 273 words. Including all documents would lead to a larger word universe, and thus increase discriminating power. Moreover, especially in the shortest documents, the number of unique words was high, pointing to a low amount of word overlap. Both these characteristics can be explained by the structure of the Finnish language and its high amount of agglutination. This in turn, makes that differently modified words are treated as separate words by Wordscores, leading to the high amount of unique words. All concordance values are negative, ranging between -0.461 and -0.102., and spread out equally between all three dimensions.

For Poland, documents from PO, PSL, PiS, SLD-UP, UW, were used to assure stretching of the dimensions and because from them all benchmarks had valid positions. Manifestos were relatively short, between 308 and 1,283 words (average 799). The highest scores were reached on the left-right dimension with LBG scoring (0.184), while the MV scoring yielded a lower maximum (0.023). Scores on the integration and foreign dimension also clustered around 0, with 0.114 on the integration dimension being the highest score.

Overall, concordance scores were highest on the left-right dimension with LBG scoring. Both the integration and Foreign dimensions obtained low scores, as did all the dimensions when MV scoring was applied.

7 CONCLUSION

For none of the dimensions Wordscores has been able to assign a score which reached a higher correlation with established benchmarks than 0.899, still under the lowest level of 0.95. Also, with the CCC of the benchmarks considered, only the left-right dimension in Germany showed positive results. Therefore, we consider the scores Wordscores gives to parties as *not the same* as those provided by the benchmarks. As such, the method fails our test to be able to serve as an automated way of scoring parties for use in voting advice applications. Of the two scorings available, especially the MV scoring led to low scorings. Most successful was the well-established left-right dimension, but also here scores are too low to deem the method valid.

Reasons for the failure of the Wordscores method can be divided into the nature of reference documents, language issues, and the method itself. With regard to the first, Laver, Benoit, and Garry warn in their original article for the dangers of selecting the wrong reference documents. Nevertheless, in order to establish the positions of the parties, a series of documents is necessary in any case. Thus, while for example, the documents for Poland were *short*, they had to be included to construct a sizeable word universe. Also, the number of documents per country, and per dimension was low, with in no case more than 10 documents involved. For the second reason, the fact that some languages as Finnish have a low amount of words, with tenses altering the word itself, makes for a smaller word universe and thus lower discriminating power. Most important, however, is that the original purpose of Wordscores seems never to have been the correct placing of parties on a scale. In almost all articles, the positions. Even with the transformations, the results are in most cases clustered, and while it is possible to distinguish relative positions of the parties with regard to each other, the absolute positions differ much from the positions assigned to them in the benchmarks.

8 **RECOMMENDATIONS**

For future research using the Wordscores method applied to VAAs, the following is recommended. First, that the number of countries analysed is extended and that a larger amount of party documents are used. While in most countries the results were low overall, in Germany, the well-known left-right axis performed relatively well, and similar results might be obtained in other countries, especially where the dimension is well visible in the countries' party space. Also, as we used 2004 election documents and found that these were over the whole shorter than the 2009 documents, an analysis with 2009 and 2014 documents might show better results. Not only are there more documents (digitally) available, it is also likely that the length of (especially) documents in the new Eastern European member states has increased, allowing for better analysis.

Second, it is better not to use the *wordfreq* function included in Kenneth Benoit's Wordscores package for STATA, but instead use either the *wordfreqj* function included in the same package, or use JFreq application designed by Will Lowe, as the *wordfreq* function has the habit of producing unrecognizable characters, especially when using letters with diacritical marks.

Third, a perhaps wider (thus involving more countries) research into which type of texts to use (normal, transformed, etc.) might show different results, as the conclusions here are solely based on results from the Netherlands. Given the differences between languages, it is highly plausible that every language has a preferred type of text which should be used.

Last, as the reference texts are the main weakness of the Wordscores method, other methods, such as *Wordfish*, which does not depend on reference documents, can be used for future use. An investigation similar to the one conducted here could show if this method is able to place parties in a more correct way.

9 REFERENCES

- Andreadis, I. (2013, September). Voting Advice Applications: a successful nexus between informatics and political science. In *Proceedings of the 6th Balkan Conference in Informatics* (pp. 251-258). ACM.
- Bakker, R., De Vries, C., Edwards, E., Hooghe, L., Jolly, S., Marks, G., ... Vachudova, M. A. (2012).
 Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999-2010. *Party Politics*, 1–15.
- Baudewyns, P., Bol, D., Dumont, P., Rezsohazy, I., and Teuber, F. (2011). Positionner les partis dans l'espace politique par l'analyse du contenu de leurs programmes : évaluation d'alternatives récentes (SAV et codage assisté par ordinateur). Paper presented at the *Quatrieme Congres international du réseau francophone des Associations de Science politique*, Brussels, Belgium.
- Benoit, K., & Laver, M. (2003). Estimating Irish party policy positions using computer wordscoring: the 2002 election – a research note. *Irish Political Studies*, 18(1), 97–107.
- Benoit, K., & Laver, M. (2006). *Party Policy in Modern Democracies*. London: Routledge.
- Benoit, K., & Laver, M. (2007a). Benchmarks for text analysis: A response to Budge and Pennings. *Electoral Studies*, *26*(1), 130–135.
- Benoit, K., & Laver, M. (2007b). Estimating party policy positions: Comparing expert surveys and hand-coded content analysis. *Electoral Studies*, *26*(1), 90–107.
- Benoit, K., & Laver, M. (2008). Compared to What? A Comment on "A Robust Transformation Procedure for Interpreting Political Text" by Martin and Vanberg. *Political Analysis*, 16(1), 101– 111.
- Benoit, K., Laver, M., & Mikhaylov, S. (2009). Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions. *American Journal of Political Science*, *53*(2), 495–513.
- Borre, O. (2001). Issue Voting: An Introduction. Aarhus: Aarhus University Press.
- Braun, D., Mikhaylov, S., & Schmitt, H. (2010). Manifesto Study Documentation 2009. Retrieved from GESIS Data Archive website: https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5057&db=e&doi=10.4232/1.10204
- Budge, I., & Pennings, P. (2007a). Do they work? Validating computerised word frequency estimates against policy series. *Electoral Studies*, *26*(1), 121–129.

- Budge, I., & Pennings, P. (2007b). Missing the message and shooting the messenger: Benoit and Laver's "response." *Electoral Studies*, *26*(1), 136–141.
- Carlson, T., & Strandberg, K. (2005). The 2004 European parliament election on the web : Finnish actor strategies and voter responses. *Information Polity*, *10*, 189–204.
- Debus, M. (2009). Analysing Party Politics in Germany with New Approaches for Estimating Policy Preferences of Political Actors. *German Politics*, *18*(3), 281–300.
- Downs, A. (1957). An Economic Theory of Democracy. New York, NY: Harper.
- Fivaz, J., Pianzola, J., & Ladner, A. (2010). More than toys: a first assessment of voting advice applications' impact on the electoral decision of voters (Working Paper No. 48). Retrieved from National Center of Competence in Research website: http://www.nccrdemocracy.uzh.ch/publications/workingpaper/pdf/WP_48.pdf
- Fossen, T., Anderson, J., & Tiemeijer, W. (2012). Wijzer stemmen? StemWijzer, Kieskompas en het voorgeprogrammeerde electoraat. In C. van 't Hof, J. Timmer, & R. van Est (Eds.), *Voorgeprogrammeerd: hoe Internet ons leven leidt* (pp. 163–183). Den Haag: Boom.
- Garzia, D. (2010). The Effects of VAAs on Users' Voting Behaviour: An Overview. In L. Cedroni & D. Garzia (Eds.), *Voting Advice Applications in Europe The State of the Art* (pp. 13–33). Napoli: CIVIS s.n.c/Scriptaweb.
- Garzia, D., & Marschall, S. (2012). Voting Advice Applications under review: the state of research. International Journal of Electronic Governance, 5(3/4), 203–222.
- Gemenis, K. (2012). A new approach for estimating parties' positions in voting advice applications.
 Paper prepared for the "Interdisciplinary Perspectives on Voting Advice Applications" Workshop, 23–24 November 2012, Cyprus University of Technology, Cyprus.
- Gemenis, K. (2013). Estimating parties' policy positions through voting advice applications: Some methodological considerations. *Acta Politica*, *48*(3), 268–295.
- Gemenis, K., & van Ham, C. (2014). Comparing methods for estimating parties' positions in Voting Advice Applications. In D. Garzia & S. Marschall (Eds.), *Matching Voters with Parties and Candidates*. Colchester: ECPR Press.
- 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.

- Hakhverdian, A. (2009). Capturing Government Policy on the Left-Right Scale: Evidence from the United Kingdom, 1956-2006. *Political Studies*, *57*(4), 720–745.
- 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.
- Hooghe, L., Bakker, R., Brigevich, A., De Vries, C., Edwards, E., Marks, G., ... Vachudova, M. A.
 (2010). Reliability and Validity of Measuring Party Positions: The Chapel Hill Expert Surveys of
 2002 and 2006. *European Journal of Political Research*, 4(5), 684–703.
- Hug, S. (2009). Selection Effects in Roll Call Votes. British Journal of Political Science, 40(01), 225.
- Hug, S., & Schulz, T. (2007). Left-Right Positions of Political Parties in Switzerland. *Party Politics*, *13*(3), 305–330.
- Iversen, T. (1994). Political Leadership and Representation in West European Democracies: A Test of Three Models of Voting. *American Journal of Political Science*, *38*(1), 45–74.
- 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.
- Krouwel, A. P. M. (2012). Party Transformations in European Democracies. Albany: State University of New York Press.
- Krouwel, A. P. M., & van Elfrinkhof, A. (2013). Combining strengths of methods of party positioning to counter their weaknesses: the development of a new methodology to calibrate parties on issues and ideological dimensions. *Quality & Quantity*, 48(3), 1455-1472.
- Latcheva, R. (2011). Cognitive interviewing and factor-analytic techniques: a mixed method approach to validity of survey items measuring national identity. *Quality & Quantity*, *45*(6), 1175–1199.
- Laver, M., Benoit, K., & Garry, J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review*, *97*(02), 311–331.
- Laver, M., & Garry, J. (2000). Estimating Policy Positions from Political Texts. *American Journal of Political Science*, 44(3), 619–634.
- Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1), 255–268.
- Lin, L. I.-K. (1992). Assay Validation Using the Concordance Correlation Coefficient. *Biometrics*, 48(2), 599–604.

- Lin, L. I.-K. (2000). A Note on the Concordance Correlation Coefficient. *Biometrics*, *56*(March), 324–325.
- Lin, L. I.-K., Hedayat, A. S., Sinha, B., & Yang, M. (2002). Statistical Methods in Assessing Agreement: Models, Issues, and Tools. *Journal of the American Statistical Association*, *97*(457), 257–270.
- Louwerse, T., & Rosema, M. (2013). The design effects of voting advice applications: Comparing methods of calculating matches. *Acta Politica* advance online publication 18 October 2013, 1–27.

Lowe, W. (2008). Understanding Wordscores. Political Analysis, 16(4), 356–371.

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.

- Mair, P. (2001). Searching for the Positions of Political Actors. In M. Laver (Ed.), *Estimating the Policy Position of Political Actors* (pp. 10–30). London: Routledge.
- Marks, G. (2007). Introduction: Triangulation and the square-root law. *Electoral Studies*, 26(1), 1–10.
- Marks, G., Hooghe, L., Steenbergen, M. R., & Bakker, R. (2007). Crossvalidating Data on Party Positioning on European Integration. *Electoral Studies*, *26*(1), 23–38.
- Martin, L. W., & Vanberg, G. (2008a). A Robust Transformation Procedure for Interpreting Political Text. *Political Analysis*, *16*(1), 93–100.
- Martin, L. W., & Vanberg, G. (2008b). Reply to Benoit and Laver. *Political Analysis*, 16(1), 112–114.
- McBride, G. B. (2005). A Proposal for Strength-of-Agreement Criteria for Lin's Concordance Correlation Coefficient. Hamilton: National Institute of Water & Atmospheric Research.
- Meier, K. J., & Campbell, J. E. (1979). Issue Voting: An Empirical Examination of Individually Necessary and Jointly Sufficient Conditions. *American Politics Research*, *7*, 21–50.
- Mendez, F. (2012). Matching voters with political parties and candidates: an empirical test of four algorithms. *International Journal of Electronic Governance*, *5*(3/4), 264.
- Mendez, F. (2014). What's behind a matching algorithm : A critical assessment of how VAAs produce voting recommendations. In D. Garzia & S. Marschall (Eds.), *Matching voters to parties and candidates* (pp. 1–18). Colchester: ECPR Press.
- Nickerson, C. A. E. (1997). A Note on "A Concordance Correlation Coefficient to Evaluate Reproducibility." *Biometrics*, *53*(4), 1503–1507.

- Pierce, R. (1997). Symposium. The Directional Theory of Issue Voting: III: Directional versus Proximity Models: Verisimilitude as the Criterion. *Journal of Theoretical Politics*, 9(1), 61–74.
- Proksch, S.-O., & Slapin, J. B. (2006). Institutions and coalition formation: The German election of 2005. *West European Politics*, *29*(3), 540–559.
- 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.
- Ruedin, D. (2013). Obtaining Party Positions on Immigration in Switzerland: Comparing Different Methods. *Swiss Political Science Review*, *19*(1), 84–105.
- Schmitt, H., & Wüst, A. M. (2012). Euromanifestos Project (EMP) 1979-2004. ZA4457 Data file Version 1.0.0. Cologne: GESIS Data Archive.
- Škop, M. (2010). Are the Voting Advice Applications (VAAs) Telling the Truth? Measuring VAAs' Quality. Case Study from the Czech Republic. In L. Cedroni & D. Garzia (Eds.), Voting Advice Applications in Europe – The State of the Art (pp 199–216). Napoli: CIVIS s.n.c/Scriptaweb.
- 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), 705–772.
- Tomz, M., & Van Houweling, R. P. (2007). *The Microfoundations of Issue Voting* (Manuscript). Retrieved from http://www.stanford.edu/~tomz/working/TomzVanHouweling-2007-08a.pdf
- Trechsel, A. H., & Mair, P. (2011). When Parties (Also) Position Themselves: An Introduction to the EU Profiler. *Journal of Information Technology & Politics*, *8*(1), 1–20.
- Treschel, R. (2010). EU-Profiler: positioning of the parties in the European Elections. Bologna: European Union Democracy Observatory, European University Institute.
- Volkens, A. (2007). Strengths and weaknesses of approaches to measuring policy positions of parties. *Electoral Studies, 26*(1), 108–120.
- Wagner, M., & Ruusuvirta, O. (2011). Matching Voters to Parties: Voting Advice Applications and Models of Party Choice. *Acta Politica*, *47*(4), 1–32.
- Walgrave, S., Nuytemans, M., & Pepermans, K. (2009). Voting Aid Applications and the Effect of Statement Selection. *West European Politics*, *32*(6), 1161–1180.

APPENDIX I – LIST OF PARTIES

COUNTRY	PARTY	PARTY NUMBER	PARTY NAME	PARTY NAME IN ENGLISH
FI	KD	902	Suomen Kristillisdemokraatit	Christian Democrats of Finland
FI	KESK	903	Suomen Keskusta	Centre Party
FI	КОК	904	Kansallinen Kokoomus	National Coalition Party
FI	КА	905	Köyhien Asialla	For the Poor
FI	PERUS	906	Perussuomalaiset	Finns Party
FI	RKP/SFP	907	Suomen ruotsalainen kansanpuolue/Svenska folkpartiet i Finland	Swedish People's Party of Finland
FI	SDP	908	Suomen Sosialidemokraattinen Puolue	Social Democratic Party of Finland
FI	SKP	910	Suomen kommunistinen puolue	Communist Party of Finland
FI	STP	911	Suomen työväenpuolue	Workers Party of Finland
FI	VAS	912	Vasemmistoliitto	Left Alliance
FI	VIHR	913	Vihreä liitto	Green League
DE	B90GRUNEN	1101	Bündnis 90/Die Grünen	Alliance 90/The Greens
DE	CDU	1102	Christlich Demokratische Union Deutschlands	Cristian Democratic Union Germany
DE	CSU	1103	Christlich-Soziale Union in Bayern e. V.	Christian Social Union in Bavaria
DE	PDSDIELINK E	1104	Partei des Demokratischen Sozialismus - DIE LINKE	Party of Democratic Socialism - The Left
DE	DKP	1105	Deutsche Kommunistische Partei	German Communist Party

DE	DVU	1106	Deutsche Volksunion	German People's Union
DE	REP	1107	Die Republikaner	The Republicans
DE	FDP	1108	Freie Demokratische Partei	Free Democratic Party
DE	FW	1109	Freie Wähler	Free Voters
DE	NPD	1110	Nationaldemokratische Partei Deutschlands	National Democratic Party of Germany
DE	Newropeans	1111	Newropeans	Newropeans
DE	SPD	1112	Sozialdemokratische Partei Deutschlands	Social Democratic Party of Germany
NL	CDA	2001	Christen-Democratisch Appčl	Christian Democratic Appeal
NL	CUSGP	2002	ChristenUnie-Staatskundig Gereformeerde Partij	Christian Union - Reformed Political Party
NL	D66	2003	Democraten '66	Democrats '66
NL	GL	2005	GroenLinks	Green Left
NL	LPF	2006	Lijst Pim Fortuyn	List Pim Fortuyn
NL	PVV	2009	Partij voor de Vrijheid	Party for Freedom
NL	PVDA	2010	Partij van de Arbeid	Labour Party
NL	SP	2012	Socialistische Partij	Socialist Party
NL	VVD	2013	Volkspartij voor Vrijheid en Democratie	People's Party for Freedom and Democracy
PL	AWSP	2101	Akcja Wyborcza Solidarnosc	Solidarity Electoral Action
PL	LPR	2102	Liga Polskich Rodzin	League of Polish Families
PL	PO	2103	Platforma Obywatelska	Civic Platform
PL	PSL	2104	Polskie Stronnictwo Ludowe	Polish People's Party

וח		2105	Porozumienie dla Przyszlosci -	Coalition Agreement for the
	PDP-CL	2105	CentroLewica	Future - CentreLeft
PL	PiS	2106	Prawo i Sprawiedliwosc	Law and Justice
PL	PR	2107	Polska Razem	Poland Together
PL	SD	2109	Stronnictwo Demokratyczne	Democratic Party
PL	SDPL	2110	Socjaldemokracja Polska	Social Democracy of Poland
PL	SLD	2111	Sojusz Lewicy Demokratycznej	Democratic Left Alliance
PL	SRP	2113	Samoobrona Rzeczpospolitej Polskiej	Self-Defence of the Republic of Poland
PL	UPR	2114	Unia Polityki Realnej	Union of Real Politics
PL	UW	2115	Unia Wolnosci	Freedom Union
SK	ANO	2401	Aliancia Nového Obcana	Alliance of the New Citizen
SK	HZD	2402	Hnutie za demokraciu	Movement for Democracy
SK	КДН	2403	Krestanskodemokratické hnutie	Christian Democratic Movement
SK	KSS	2404	Komunistická strana Slovenska	Communist Party of Slovakia
SK	LS-HZDS	2405	Ludová strana - Hnutie za demokratické Slovensko	People's Party – Movement for a Democratic Slovakia
SK	LS-LU	2406	Ludová únia	Liberal Party-People's Union
SK	Most	2407	Most–Híd	"Bridge" Party
SK	OKS	2408	Obcianska konzervatívna strana	Civic Conservative Party
SK	PSNS	2409	Pravá Slovenská národná strana	Right Slovak National Party
SK	SDA	2411	Sociálnodemokratická alternatíva	Social Democratic Alternative
SK	SDKU-DS	2412	Slovenská demokratická a krestanská únia - Demokratická	Slovak Democratic and Christian Union – Democratic

			strana	Party
SK	SDL	2413	Strana demokratickej lavice	Party of the Democratic Left
SK	SF	2414	Slobodné fórum	Free Forum
SK	SMER-SD	2415	Smer – sociálna demokracia	Direction – Social Democracy
SK	SMK-MKP	2416	Strana madarskej komunity - Magyar Közösség Pártja	Party of the Hungarian Community
SK	SNS	2417	Slovenská národná strana	Slovak National Party
SK	SaS	2418	Sloboda a Solidarita	Freedom and Solidarity

APPENDIX II – REMARKS ON THE DATA FILE

- Scores in the documents were scaled from 0 1
- Labels standardized
- The 2004 VVD manifesto is the one from the ELDR
- In BL 2003 SLD and UP were combined and score was averaged
- In BL 2003 CU and SGP were combined and score was averaged
- In BL 2003 no coding for French Parties
- In BL 2003 CSU and CDU were split with own score
- In CHES 2002 CU and SGP were combined and score was averaged
- In CHES 2010 CU and SGP were combined and score was averaged
- In CHES 2010 SLD and UP were combined and score was averaged
- In EMP 2004 SMER and SDL were combined and score was averaged
- In EUP 2009 SLS was added to SLSSKD
- For Latvia LLCS and LLS combined form score for LiCS in CHES 2002, score LLS is score LiCS in EMP 2004
- EUREKO is New Horizons in BL2003
- In Italy UDC is CCD & CDU (scores are averaged for CHES 2002
- Czech electoral alliance EMP 2004 = ULD (Union of Liberal Democrats)

APPENDIX III – WORDING OF QUESTIONS IN BENCHMARKS

	LEFT-RIGHT	EU INTEGRATION	FOREIGN AND SECURITY
2002 CHES	LRGEN = position of the party in 2002 in terms of its broad ideological stance	POSITION = overall orientation of the party leadership towards European integration in 2002	FOREIGN = position of the party leadership in 2002 on a common foreign and security policy
	Extreme Left (0) – Centre (5) – Extreme Right (10)	Strongly opposed (1) – Strongly in favour (7)	Strongly opposes (1) – Strongly favours (7) –No position (8 – coded as missing)
2010 CHES	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	EU_FOREIGN = position of the party leadership in 2010 on an EU foreign and security policy
	Left (0) – Centre (5) – Right (10)	Strongly opposed (1) – Strongly in favour (7)	Strongly opposes (1) – Strongly favours (7)
3 Benoit & Laver	Left-Right – Please locate each party on a general left-right dimension, taking all aspects of party policy into account	-	-
2003	Left (1) – Right (20)	-	-
2004 EMP	LEFT [position in Dataset: 18] placement of Euromanifesto according to the coder on a left- right scale.	EU [position in Dataset: 24] placement of Euromanifesto according to coder on a pro-anti- EU-integration scale.	_

	Left (1) – Right (10)	Pro (1) – Anti (10)	-
2009 EMP	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.	
	Left (1) – Right (10)	Pro (1) – Anti (10)	-
9 EU Profiler	Socioeconomic left- socioeconomic right	Pro EU integration-Anti EU integration	On foreign policy issues, such as the relationship with Russia, the EU should speak with one voice
2005	-2 (Left) – 2 (Right)	-2 (anti-EU integration) – 2 (pro- EU integration)	1 (completely disagree) – 5 (completely agree)

APPENDIX IV – RESULTS FOR THE NETHERLANDS

			L	eft-Right		I	ntegratior	ו		Foreigr	1
	Source	Score Scaling	EUP 2009	EMP 2009	CHES 2010	EUP 2009	EMP 2009	CHES 2010	EUP 2009	EMP 2009	CHES 2010
	CHES 2002		-0.191	0.001	-0.220	-0.121	-0.073	-0.179	-0.144	-	-0.156
	BL 2003	LBG	-0.176	0.008	-0.204	-	-	-	-	-	-
nal	EMP 2004		-0.157	0.002	-0.188	-0.138	-0.075	-0.191	-	-	-
Norr	CHES 2002		-0.164	0.001	-0.157	-0.064	-0.051	-0.089	-0.125	-	-0.055
	BL 2003	MV	-0.116	0.006	-0.136	-	-	-	-	-	-
	EMP 2004		-0.105	0.001	-0.124	-0.057	-0.043	-0.075	-	-	-
	CHES 2002		0.386	0.539	0.492	-0.134	-0.115	-0.168	-0.186	-	-0.250
σ	BL 2003	LBG	0.432	0.580	0.530	-	-	-	-	-	-
ormea	EMP 2004		0.372	0.498	0.479	-0.059	0.007	0.112	-	-	-
ansfo	CHES 2002	MV	0.021	0.033	0.022	-0.018	-0.022	-0.024	-0.037	-	-0.043
μ	BL 2003		0.141	0.159	0.181	-	-	-	-	-	-
	EMP 2004		0.073	0.089	0.099	-0.009	0.002	0.019	-	-	-
	CHES 2002		0.346	0.438	0.451	-0.037	0.002	-0.140	-0.120	-	-0.276
	BL 2003	LBG	0.392	0.485	0.490	-	-	-	-	-	-
red	EMP 2004		0.327	0.399	0.429	0.095	0.202	0.175	-	-	-
Clea	CHES 2002		0.014	0.019	0.014	-0.004	0.000	-0.016	-0.021	-	-0.041
	BL 2003	MV	0.124	0.121	0.156	-	-	-	-	-	-
	EMP 2004		0.062	0.065	0.085	0.011	0.032	0.023	-	-	-
eme	CHES 2002	LBG	0.474	0.570	0.620	-0.065	-0.089	-0.051	-0.146	-	-0.180
Ster	BL 2003		0.502	0.602	0.635	-	-	-	-	-	-

	EMP 2004		0.451	0.532	0.599	0.053	0.072	0.343	-	-	-
	CHES 2002	MV	0.020	0.027	0.022	-0.007	-0.013	-0.006	-0.025	-	-0.028
	BL 2003		0.188	0.183	0.241	-	-	-	-	-	-
	EMP 2004		0.100	0.103	0.139	0.007	0.014	0.053	-	-	-
	CHES 2002	LBG	0.404	0.473	0.540	-0.011	0.016	-0,025	-0,012	-	-0,020
eared	BL 2003		0.442	0.516	0.570	-	-	-	-	-	-
& Cle	EMP 2004		0.383	0.437	0.518	0.138	0.220	0.354	-	-	-
med	CHES 2002		0.014	0.018	0.015	-0.001	0.002	-0.003	-0,020	-	-0,029
Stem	BL 2003	MV	0.155	0.141	0.196	-	-	-	-	-	-
	EMP 2004		0.081	0.078	0.114	0.017	0.037	0.049	-	-	-

Analysis results for the Netherlands. A hyphen indicates that no values could be obtained as one of the benchmarks was missing.

APPENDIX V – NUMBER OF WORDS FOR 2004 REFERENCE MANIFESTOS IN THE NETHERLANDS

	Nor	mal	Transf	ormed	Sten	nmed	Cleared		Stemmed & Cleared	
				1						
Toyt	Total	Unique	Total	Unique	Total	Unique	Total	Unique	Total	Unique
Text	Words	Words	Words	Words	Words	Words	Words	Words	Words	Words
VVD	12864	2393	12965	2309	12965	2003	825	512	825	474
SP	14631	3379	14328	3147	14328	2750	7142	3077	7142	2714
PVDA	9079	2156	9086	2106	9086	1861	3183	1445	3183	1262
LPF	2072	832	2044	802	2044	743	859	580	859	548
GL	27972	5793	27800	5503	27800	4773	15631	5412	15631	4703
ET	1429	877	1572	656	1572	615	1132	719	1132	671
D66	5694	1545	5702	1518	5702	1327	4981	2027	4981	1794
CUSGP	12591	3276	12504	3166	12504	2784	8174	3061	8174	2683
CDA	1526	592	1522	583	1521	535	7011	2230	7011	1934

Number of Words per document and number of unique words for 2004 manifestos in the Netherlands

APPENDIX VII - SPEARMAN'S CORRELATIONS FOR THE LEFT-RIGHT DIMENSION IN THE NETHERLANDS

		Normal	Stemmed	Transformed	Cleared	Stemmed & Cleared
Normal	Correlation Coefficient	1.000	300	.367	117	433
Norma	Sig. (2-tailed)		.433	.332	.765	.244
	Ν	9	9	9	9	9
Stemmed	Correlation Coefficient	300	1.000	,717*	,967**	,967**
otominou	Sig. (2-tailed)	.433		.030	.000	.000
	Ν	9	9	9	9	9
	Correlation Coefficient	.367	,717*	1.000	,850**	.583
TIANSIOITTIEU	Sig. (2-tailed)	.332	.030		.004	.099
	Ν	9	9	9	9	9
Cleared	Correlation Coefficient	117	,967**	,850**	1.000	,900**
Clourou	Sig. (2-tailed)	.765	.000	.004		.001
	Ν	9	9	9	9	9
Stemmed &	Correlation Coefficient	433	,967**	.583	,900**	1.000
Cleared	Sig. (2-tailed)	.244	.000	.099	.001	
	Ν	9	9	9	9	9

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

APPENDIX VIII - RESULTS FOR THE NETHERLANDS, GERMANY, SLOVAKIA, FINLAND & POLAND

				Left-Righ ⁻	t	Integration			Foreign		
	Source	Trans	EUP	EMP	CHES	EUP	EMP	CHES	EUP	EMP	CHES
		Turio.	2009	2009	2010	2009	2009	2010	2009	2009	2010
	CHES 2002	LBG	0.386	0.539	0.492	-0.134	-0.115	-0.168	-0.186	-	-0.250
	BL 2003		0.432	0.580	0.530	-	_	-	-	-	-
ands ¹⁶	EMP 2004		0.372	0.498	0.479	-0.059	0.007	0.112	-	-	-
Netherl	CHES 2002	MV	0.021	0.033	0.022	-0.018	-0.022	-0.024	-0.037	-	-0.043
	BL 2003		0.141	0.159	0.181	-	-	-	-	-	-
	EMP 2004		0.073	0.089	0.099	-0.009	0.002	0.019	-	-	-
Germany ²⁰	CHES 2002	LBG	0.802	0.665	0.757	0.157	-0.114	0.498	21	-	0.633
	BL 2003		0.825	0.685	0.768	-	-	-	-	-	-
	EMP 2004		0.899	0.766	0.755	0.085	0.297	0.408	-	-	-
	CHES 2002	MV	0.230	0.172	0.210	0.098	0.188	0.245	. 22	-	0.316
	BL 2003		0.231	0.173	0.209	-	-	-	-	-	-
	EMP 2004	1	0.251	0.182	0.164	. 23	_ 24	. 25	-	-	-

¹⁹Reference documents: VVD, CU-SGP, GroenLinks, SP, PvdA

- ²² No CCC could be calculated as all the values in EUP2009 were the same (1)
- ²³ No CCC could be calculated as all the values in EMP2004MV were the same (0.7778)
- ²⁴ No CCC could be calculated as all the values in EMP2004MV were the same (0.7778)
- ²⁵ No CCC could be calculated as all the values in EMP2004MV were the same (0.7778)

²⁰ Reference documents: B90/Die Grünen, SPD, CDU, CSU, FDP, Die Linke

²¹ No CCC could be calculated as all the values in EUP2009 were the same (1)

<ia<sup>26</ia<sup>	CHES 2002	LBG	0.046	0.155	0.226	-0.524	-0.126	-0.295	0.056	-	0.007
	BL 2003		0.035	0.166	0.267	-	-	-	-	-	-
	EMP 2004		-0.035	0.096	0.216	-0.073	0.011	-0.020	-	-	-
Slova	CHES 2002		-0.000	-0.000	-0.000	-0.114	-0.020	-0.056	0.010	-	0.001
	BL 2003	MV	0.003	0.014	0.019	-	-	-	-	-	-
	EMP 2004		-0.001	0.003	0.006	-0.011	0.001	-0.003	-	-	-
	CHES 2002		-0.185	-0.435	-0.270	-0.202	-0.348	-0.201	-0.150	-	-0.241
	BL 2003	LBG	-0.186	-0.461	-0.288	-	-	-	-	-	-
nd ²⁷	EMP 2004		-0.188	-0.434	-0.270	-0.210	-0.364	-0.211	-	-	-
Finlar	CHES 2002	MV	-0.102	-0.291	-0.154	-0.133	-0.244	-0.130	-0.103	-	-0.151
	BL 2003		-0.103	-0.294	-0.156	-	-	-	-	-	-
	EMP 2004		-0.118	-0.305	-0.167	-0.137	-0.253	-0.134	-	-	-
	CHES 2002		-0.012	0.169	0.150	0.114	0.025	-0.099	29	-	0.005
	BL 2003	LBG	0.012	0.184	0.174	-	-	-	-	-	-
d ²⁸	EMP 2004		0.004	0.166	0.145	0.084	-0.006	-0.116	-	-	-
Polar	CHES 2002		-0.002	0.023	0.023	0.049	0.014	-0.041	29	-	30
	BL 2003	MV	0.001	0.021	0.022	-	-	-	-	_	-
	EMP 2004		0.000	0.020	0.018	0.025	-0.002	-0.037	-	-	-
				1		1		1			

²⁶ Reference documents: SMK-MKP, SMER, SDKU, LS-HZDS, KDH

²⁷ Reference documents: SDP, RKP-SFP, KOK, KESK, VAS, VIHR, KD

²⁸ Reference documents: PO, PSL, PiS, SLD-UP, UW

²⁹ No CCC could be calculated as all the values in EUP2009 were the same (1)

 $^{\rm 30}$ No CCC could be calculated as all the values in CHES2002 were the same (0.595)

APPENDIX IX – WORD COUNT AND UNIQUE WORDS FOR GERMANY, SLOVAKIA, FINLAND, AND POLAND DOCUMENTS

			Unique Number of	
		Total Number of Words	Words	
	SPD 2009	8,705	2,436	
	SPD 2004	3,012	1,067	
	FDP 2009	11,497	3,346	
	FDP 2004	10,105	2,721	
	DIELINKE 2009	16,126	4,770	
Gormany	DIELINKE 2004	19,434	5,271	
Germany	CSU 2009	4,875	1,609	
	CSU 2004	3,285	1,220	
	CDU 2009	5,986	1,919	
	CDU 2004	2,730	1,034	
	B90GRÜNEN 2009	40,328	8,233	
	B90GRÜNEN 2004	2,020	851	
	SNS 2009	232	171	
	SMKMKP 2009	4,904	2,142	
	SMER 2009	572	319	
	SDKUDS 2009	7,146	2,342	
Slovakia	LSHZDS 2009	6,541	2,663	
Glovakia	KDH 2009	2,216	1,156	
	KDH 2004	1,806	1,019	
	LSHZDS 2004	1,135	669	
	SDKU 2004	2,257	1,037	
	SMER 2004	2,81	1,135	

	SMKMKP 2004	3,211	1,566		
	VIHR 2009	2,647	1,561		
	VIHR 2004	273	216		
	VAS 2009	1,913	1,117		
	VAS 2004	737	490		
	SDP 2009	2,97	1,692		
	SDP 2004	1,874	1,148		
	RKPSFP 2009	995	670		
Finland	RKPSFP 2004	1,68	1,046		
	PERUS 2009	2,013	1,144		
	KOK 2009	294	239		
	KOK 2004	2,366	1,345		
	KESK 2009	4,252	2,364		
	KESK 2004	3,199	1,751		
	KD 2009	6,45	3,375		
	KD 2004	392	300		
	SRP 2009	636	398		
	SLDUP 2009	7,381	2,605		
	PIS 2009	77,372	16,868		
	PSL 2009	4,438	1,603		
	PO 2009	16,309	4,96		
Poland	UW 2004	308	202		
	SLDUP 2004	704	470		
	PIS 2004	671	442		
	PSL 2004	1,028	534		
	PO 2004	1,283	749		