4 Assessing Similarity in Manifestos
An Overview and a New Measure

Finding and evaluating similarities and differences is the core of the comparative method. In one way or another, this basic principle is applied in many studies, regardless of whether policy moves, party strategies, or other phenomena are studied.

The comparative principle is also common in linguistics and computer science. Computer-assisted methods are used at the intersection of these disciplines to compare texts, detect patterns, and extract information from them. Therefore, the study of (political) documents represents an exciting connection to those disciplines.

In this work, I examine the central thesis that established parties can influence the electoral success of new parties by making strategic changes to their election programs. Such changes may affect the position of the party or the issue salience. I combine two different measures to integrate these two aspects into my models. Based on innovative methods of computer-assisted text analysis, cosine similarity scores capture the degree of similarity between two documents. This allows tracking changes in the parties' issue salience. In addition, I resort to the RILE, which is based on manual content analysis, to identify changes in the parties' positions in the left-right dimension.

The use of a new method like cosine similarity must be justified, its quality validated. This is what I will deal with in this chapter. Therefore, natural language processing and machine learning techniques are applied to election programs of political parties. By comparing the results with those obtained with the help of content analysis methods from political science, it will be shown that computer-assisted procedures are suitable instruments for dealing with political science questions.

Accordingly, this chapter deals with the question: How do different position or salience measurements of election programs behave compared to text analytic measurements? A novel simulation experiment was developed to compare the different position and salience measurements to answer this question.

In the experiment, synthetic data are generated from existing election programs that exhibit predetermined known properties. Based on these synthetic manifestos, it can be shown that changes in position and salience can be detected by text analysis at least as well as by established methods.

In order to present these results, the first step will be to examine the characteristics and content of election programs and how they have been analyzed in political science so far. For this purpose, the criticism of previous methods will be addressed.

In a second step, I will discuss which computer-assisted text analysis methods are available today for the automated analysis of text content and the principles on which they are based.

Finally, the simulation to compare the existing indices is presented in detail, and the results of the experiments are discussed.

4.1 The Content of Election Programs

The study of party programs has occupied political science for decades. As Budge and Bara present in a brief historical outline, the systematic content analysis of party programs can be traced back to the work of David Robertson in the early 1970s (Budge and Bara, 2001, p. 6).

Robertson used the method of manual content analysis, which in turn has a long tradition whose earliest roots lie ”in theological studies in the late 1600s” (Krippendorff, 2004, p. 3). He was among the first to analyze party manifestos to gain insights into policy positions. This way, he laid the foundation for a new branch of research.

While this early study was still limited to the election manifestos of the Conservatives and Labour party (Robertson, 1976, p. 72), it is thanks to the Comparative Manifesto Project that party manifestos are analyzed over long periods and in many countries.

The Comparative Manifesto Project was created to extract party policy positions from the documents that parties themselves create: the election programs. One goal was to test Down’s economic theory based on more than voter surveys (Budge and Bara, 2001, p. 6-7).

The core of the CMP is the collection and content analysis of the election programs of political parties. Currently, the dataset covers 56 countries, 753 elections, and 1154 parties. In total, 4582 manifestos were coded. More than 2000 of them are as raw or annotated documents available for download (Volkens et al., 2020). These are better known as the Manifesto corpus.

As part of the project, the election programs are broken down into so-called quasi-sentences and assigned to one of 56 content policy categories through a coder. A further category is used to record non-assignable quasi-sentences. These data allow long time series to be created and compared between different countries. Furthermore, indices can be created to capture the content of party programs along different ideological dimensions. The best-known use of CMP data is to measure the content of left and right issues in a party program. The corresponding index is widely known as RILE and considered the ”crowning achievement” (Budge and Klingemann, 2001, p. 19) of the project.

This data has led to an explosion in manifesto-based research studies, dealing with numerous questions of party competition. Furthermore, research projects have been inspired that expand the selection of documents (e.g., Regional and Local manifesto project) and the coding scheme (cf. Gemenis, 2013, p. 4).

Political science owes a multitude of insights to these data. A key finding of the manifesto project is that election programs differ in their emphasis on individual policy issues, i.e., the amount of text devoted to a particular topic. At the same time, direct confrontation is rarely found. This finding is generally regarded as the most important confirmation of the saliency theory.

While the comparative manifesto projects focus on the policy content of party platforms, scholars started to emphasize the content of election programs beyond policy positions. A notable example is the Austrian National Election Study. Dolezal et al. (2018) analyze Austrian election manifestos to shed light on so far neglected content of election manifestos like ”references to the past (party records), promises about the future (pledges) and attacks on competitors (negative campaigning) as well as the degree of personalization” (p. 3). Other authors approach platforms in a similar way: For example, election pledges (Mansergh and Thomson, 2007; Thomson et al., 2017) as well the temporal focus of statements (Müller, 2022) are examined.

These studies have in common that they go beyond the traditional category scheme of the Comparative Manifesto Project and perform their own manual or automated content analysis on the documents.

This development puts an old branch of research into a new light: With the availability of manifesto corpus documents and increasing computing power and new software packages, a ”back to the roots” movement can be observed in more recent research. Nowadays, interest in raw texts is growing again, as computer-assisted content analysis methods enable cost-effective (re-)analyses.

So, despite these terrific achievements of the CMP, new research techniques question the ”gold standard” (Pennings, 2011) status of the data. It is, therefore, worth taking a critical look at both the CMP data and the alternative approaches. Therefore, the following section summarizes the CMP and the criticism of the data and its use.

Subsequently, alternative, computer-based text analysis methods and the simulation experiment to compare both approaches are presented and discussed.

4.1.1 The CMP and CMP-based Measurements

The most extensive and systematic criticism of the CMP data comes from Gemenis (2013). In his article, he differentiates between problems of ”(1) theoretical underpinnings of the coding scheme; (2) document selection; (3) coding reliability; and (4) scaling” (Gemenis, 2013, p. 4). The upcoming discussion follows this classification and supplements the individual points of criticism where necessary.

The Theory Behind the Coding Scheme

The CMP's category system was developed explicitly based on the assumption that parties behave according to the saliency theory. From that point of view, most issues are valence issues, i.e., parties consider only one position on these issues to be occupiable: they advocate education, peace, environmental protection, and economic development; they are opponents of inequality, injustice, or high inflation. As Budge put it: ”Long digressions on the growth of unemployment are presumably saying it is a bad thing and the party would do something to counter it. Is any party going to say explicitly that it is in favour of unemployment?” (Budge, 2001b, p. 219).

Political competition can thus be described as selective emphasis: Parties choose issues they credibly represent and emphasize them in their election program. Topics that other parties have successfully occupied are ignored wherever possible. A decidedly contrary stance, the articulation of contradiction, will be encountered only very rarely: ”A party might, however, say very little about unemployment and expatiate greatly on the evils of inflation, implying that all other considerations should be subordinated to fighting this problem. These tricks of party rhetoric are no doubt familiar to every reader. They do not leave much room for parties to line up for or against each issue.

What party wants to appeal for votes by extolling either unemployment or inflation – or supporting war against peace?” (Budge, 2001b, p. 219).

The category scheme of the CMP reflects the consequences of the assumptions of this theory: According to the self-description of the project, a ”salience coding” is performed. Accordingly, most of the categories are formulated in such a way that the naming of the respective topic is collected, ”whether they seem to have a direct policy content or not” (Budge, 2001b, p. 219).

This coding is described by Budge as ”one-positional” and justified by the nature of the texts considered: ”Coding-categories are inductively derived – basically formed by grouping related sentences in the text – and so they reflect the textual practice of only endorsing the ’obvious’ position on each issue – against unemployment, inflation and high taxes, for extending services, etc. Hence the codings directly reflect party assumptions that there is only one tenable position on each issue” (Budge, 2001b, p. 220).

Deviating from this, dichotomous positive or negative statements were collected for twelve topic areas: ”Scepticism on the part of certain members of the Manifesto Research Group at the very beginning of the coding operation resulted in ’pro-con’ codings being put in for certain issue areas where confrontation between parties was thought most likely” (Budge, 2001a, p. 78). However, from the point of view of the CMP, these categories essentially confirmed the salience assumptions and were mainly used as validity checks.

The problem with this theoretical basis is that the saliency theory is much less secure than assumed by Budge. Gemenis (2013) names various empirical and theoretical studies that question the validity of saliency theory and thus the appropriateness of the CMP category scheme.

For instance, Laver pointed out that there ”are issues deemed highly salient by people with radically different substantive policy positions. They include issues involving: the redistribution of resources in an unequal society, which generates a fundamental conflict of interest between rich and poor; a range of potent ’moral’ issues such as abortion, capital punishment and euthanasia; issues generating conflicts of interest between religious, linguistic, ethnic or other social groups; and so on” (Laver, 2001, p. 74). For such issues, it is simply inappropriate to assume a single reasonable position and base the coding scheme on this from the outset.

For genuine valence issues like the environment, empirical findings have also shown that even the choice of an ineligible position does not have to detract from success (cf. Gemenis, 2013, p. 13). Furthermore, even the classification as a valence or position issue is not constant over time (Gemenis, 2013, p. 6; Franzmann and Kaiser, 2006, p. 170). It has also been shown that attacks on opposing parties are not frequent but still do occur in party manifestos (Dolezal et al., 2018, p. 9).

Document Selection

A second important criticism of the CMP data is the use of so-called proxy documents (Gemenis, 2012). These documents are analyzed instead of election programs wherever they were not available. Proxy documents are, for example, newspaper articles, interviews, or speeches. This concerns a significant number of observations (Gemenis, 2012, p. 596-597).

The problem with these documents conceptually is that they were not always published directly by the party and thus contain less the self-representation of the party's policy position than potentially inaccurate perceptions from outside or, in the case of speeches, the possibly distorted presentation of individual politicians. Furthermore, it is questioned whether the CMP coding scheme captures ”accurately the policy content of proxy documents” (Benoit et al., 2012, p. 605).

Empirically it was shown that proxy ”documents can introduce measurement error in addition to the error introduced into the CMP by other means” (Gemenis, 2012, p. 601). Several solutions to this problem have been proposed, including the replacement of proxy documents by the correct election programs, the exclusion of these data from the analysis, and the use of alternative scales (Gemenis, 2012, p. 601-602; Benoit et al., 2012, p. 608). A separate section is devoted to the latter proposal.

Content Analysis

The quality assurance of manual content analysis is of central importance for the usability of the resulting data. Therefore, it must be ensured that all coders assign the same text component to the same categories. This is called reliability. Common measures of reliability require that the same coder either produces the same results at different times (stability, or intra-coder reliability) or that different coders produce the same result (reproducibility, or inter-coder reliability) (Krippendorff, 2004, p. 214-216). The correspondence between the two coders is measured with the Holsti coefficient or Krippendorf's alpha (Krippendorff, 2004, p. 221-243).

The Comparative Manifesto Project is criticized because only one coder processed all election programs at a single time. Accordingly, established reliability measures cannot be given. Instead, extensive coder training has been provided to ensure the reliability of the measurements. Although this is one of the commonly used steps of manual content analysis, it cannot replace a check using the results of other coders.

That these concerns are more than mere speculation is shown by an experiment of Mikhaylov et al. (2012). Using former coders of the CMP project, the study shows a considerable lack of reliability: ”Our examination of coder disagreement using experimental recoding of core CMP documents clearly indicates that the CMP coding process is highly prone to misclassification and stochastic coding errors. Bearing in mind that the minimum standard conventionally deemed acceptable for the reliability coefficients reported in Table 2 is 0.8, the coefficients we find are worryingly low, almost all in the range [0.3, 0.5]” (Mikhaylov et al., 2012, p. 90).

It can be assumed that there is a high amount of noise in the data, which is based on wrong assignments of quasi-sentences to categories. This noise is adding bias to the CMP estimates, ultimately leading to ”bias of estimated causal effects when CMP quantities, especially Rile, are used as covariates in regression models”(Mikhaylov et al., 2012, p. 90).

The coding scheme and the coding process would have to be fundamentally revised to solve this problem, which is unlikely or impossible due to the high costs involved. However, one possible way out is computer-assisted automatic coding and more robust scaling techniques.

The Right-Left index

The previous sections have dealt with the basic principles of data collection of the CMP. However, in passing, it has already been mentioned that these problems also affect scaling based on this data. This applies in particular to the standard left-right scale of the project, the RILE: ”Aggregation of misclassified categories to coarser scales - such as the Rile scale of left-right policy - does not eliminate this problem” (Mikhaylov et al., 2012, p. 90).

This is very important because the RILE index is ”by far the most common way to use the manifesto dataset (arguably for 80-90 percent of users of the data)” (Mölder, 2016, p. 38). The importance of RILE was emphasized not least by the project leaders themselves: ”The crowning achievement of the Manifesto Research Project has been to measure party policy change in a variety of countries over an extended time period along the Left-Right dimension" (Budge and Klingemann, 2001, p. 19).

Created by Laver and Budge in the context of their work about party policy and government coalitions (Laver and Budge, 1992), the RILE became the standard left-right scale of the CMP. Moreover, indeed, there are good reasons for the popularity of this scale. The data is readily available, but more importantly, "the rich time series produced by MRG/CMP, covering a 50 year period for many democracies" (Budge and Pennings, 2007a, p. 123) is outstanding in the field.

Furthermore, the basic construction of the index is easy to understand and reproduce: The RILE is based on the identification of thirteen right and an equal number of left categories. Their observed relative frequency is summed separately for the left and right categories. Subsequently, the sum of all left categories is subtracted from the sum of the right categories. The result is a value between -100 (the manifesto is left) and +100 (all quasi sentences of the considered categories in the manifesto are right).

The CMP group considers the measurement results obtained in this way to have a good face validity (Budge and Bara, 2001, p. 14). In order to prove that, line plots of the party policy movements in many different countries have been published (Budge and Klingemann, 2001, p. 19-50), as well as comparisons with an expert survey, have been conducted (Budge and Pennings, 2007b, p. 136).

Unfortunately, not everyone could be convinced this way. As a result, the right-left index has been criticized both conceptually and empirically.

The most fundamental criticism of RILE results from the nature of political competition, often portrayed as inherently multidimensional (Adams et al., 2005; Albright, 2010; Benoit and Laver, 2012). Besides that "there is no 'one true' dimensionality for any given policy space" (Benoit and Laver, 2006, p. 110), the analysis of one dimension for a given research interest can be justified. However, this does not necessarily have to be the left-right dimension, even though it has proven its great importance in many contexts.

Another point of criticism focuses on RILE's assumptions about the nature of the left-right dimension: "For the index it has been assumed that the left-right dimension is meaningfully invariant across time and space" (Mölder, 2016, p. 40). However, research results on the change in values in Western European societies (Inglehart, 1977) clearly show that there is a change in the meaning of right and left.

Nor does Western European conceptualization work in the context of Central and Eastern Europe (Mölder, 2016, p. 40). As Benoit and Laver put it: ”However, our results also suggest quite strongly that the substantive meaning of left and right is a poor international traveler” (Benoit and Laver, 2006, p. 152).

As Jahn pointed out, at least some common ground must exist to be able to speak meaningfully of right and left (Jahn, 2011, p. 5). At the same time, this does not exclude that parties are also ”able to ‘modernize’ the left-right semantic by integrating new issues within their ideology” (Jahn, 2014, p. 299). Unfortunately, these differences in meaning and importance between countries and across time are not taken into account by the RILE index.

Attempts to solve these problems with inductive methods such as the vanilla method (Gabel and Huber, 2000) or the FK index (Franzmann and Kaiser, 2006) have contributed significantly to the understanding of time- and country-dependent differences, but they suffer from the fact that they are challenging to interpret (Mölder, 2016, p. 46; Jahn, 2011, p. 4). For this reason, Jahn (2011) combines a deductive core (LR core) with inductively gained complementary issues (LR plus).

In empirical terms, the criticism of RILE is even more pronounced. To begin with, Mölder showed that the issues grouped as left or right have hardly any common inner context (Mölder, 2016), which is a core assumption of summated rating scale construction.

Furthermore, changes in the RILE not only occur because the number of quasi-sentences devoted to the left or right change, but also because all excluded sentences change. Suppose a party decides to give more weight to an issue that is not left or right. In that case, the RILE subsequently portrays the party as more centrist: ”To take a very simple example, imagine a document from a left-wing party with a total (N) of 100 sentences, in which 50 sentences were coded left (L) and zero coded right (R). The Rile score is (R-L)/N = -0.5. Now imagine that 50 sentences are added to the manifesto, consisting of uncodable rhetoric singing the praises of the party leader and trashing the other parties. The Rile score is now -0.33 and the party appears to have moved to the center” (Benoit et al., 2012, p. 606).

Kim and Fording (1998) tried to correct this problem ”by dividing the difference between the left and right components, not by the total number of quasi-sentences in the manifesto, but by the total number of quasi-sentences included in the L–R scale” (Gemenis, 2013, p. 13). But unfortunately, this adjusted scale tends to force ”scores toward the extremes” (Benoit et al., 2012, p. 607; Lowe et al., 2011).

This is a severe problem because it means that a more left RILE score could be the result of a higher number of quasi-sentences referring to left topics, or a reduction of sentences referring to right topics, or a reduction of the number of topics that are neither right nor left, or a combination of all these sources. In the worst case, a change in the RILE score is thus a pure measurement construct without a corresponding basis for a party's change in position.

As a consequence of these shortcomings, ”implausible results for left-right scores based on CMP data for party systems as diverse as Austria, Belgium, Denmark, France, Germany, Italy and The Netherlands” (Franzmann and Kaiser, 2006, p. 164) and significant differences compared to expert surveys (Benoit and Laver, 2006) have been reported. This has raised doubts about the proposed face validity (Jahn et al., 2018b; Pelizzo, 2003) and usability of the index: ”The locations and the corresponding differences between parties as assumed by the index [...] capture only a marginal amount of variance that is present in the political positions of parties according to the manifesto dataset. Therefore, it is questionable whether such a measure is suitable for evaluating the political differences between parties” (Mölder, 2016, p. 45).

Conclusion

The points of criticism of the CMP and RILE outlined in the previous four sections illustrate the significant problems in the valid and reliable measurement of party positions.

Based on this discussion, it should be noted that the CMP is still one of the best datasets available for comparative political science. Therefore, the criticism expressed should not mean a complete turning away from the CMP data, but a reflected use instead of the ”earlier suggestions to accept the CMP data 'as is'” (Gemenis, 2012, p.602) or to declare them the ”gold standard” (Pennings, 2011).

This can, for example, consist of using ”CMP's codings but not its policy scale” (Benoit et al., 2012, p. 608). As already mentioned, this path has been followed several times and has produced a series of indices that were intended to remedy the weaknesses of RILE. These include the vanilla approach (Gabel and Huber, 2000), the FK index (Franzmann and Kaiser, 2006), the LR index (Jahn, 2011) and the logit scaling method (Lowe et al., 2011).

These proposals have given rise to lively debates that intensively discuss the strengths and weaknesses of the respective approaches. Common to all alternatives mentioned is that they are based on the CMP-category scheme and thus on the saliency theory. Hence they share specific problems of the RILE, like misclassification and ”implicitly positional and censored” (Gemenis, 2013, p. 5) categories. Thus, these approaches only address the fundamental problems to a limited extent. Concerning the use in empirical studies, it should be noted that none of the approaches has so far come close to the popularity of RILE.

In order to avoid the discussed weaknesses of the CMP in principle, we have to go back to the original documents. However, due to the high cost of manual content analysis, computer-assisted text analysis methods were suggested as an alternative. The advantages and disadvantages of this approach are discussed in the next section.

4.1.2 Computer-Assisted Text Analysis in Political Science

The previous section reported the problems of determining a party position based on the CMP data in detail. Very similar analyses exist on the problems with expert interviews (Benoit and Laver, 2007b; Laver and Garry, 2000) and other data sources. In the end, all procedures and data sources have ”serious methodological and practical problems” (Laver et al., 2003, p. 311).

The baseline of this debate is that party programs are the most reliable source for party positions: ”Even though party manifestos are not written to inform citizens about a party’s position on a Left-Right dimension, but rather to accommodate strategic challenges in order to win an election (Laver 2001), they can be used to deduce a party’s underlying ideological position” (Jahn, 2011, p. 2). They are ”concrete by-products of strategic political activity” (Laver et al., 2003, p. 311) and can be ”analyzed, reanalyzed and reanalyzed again without becoming jaded or uncooperative” (Laver et al., 2003, p. 311).

These advantages raise the question of how valid and reliable party positions can be extracted from political texts without the need for cost- and time-intensive manual content analysis. Two answers have been given: The first one is the ”direct attempt to reproduce the hand-coding of texts, using computer algorithms to match texts to coding dictionaries” (Laver et al., 2003, p. 312). As one of the earliest representatives, Laver and Garry (2000) should be mentioned here. This approach is promising, but unfortunately, it cannot do without human coders developing and testing the dictionaries. However, recent breakthroughs in machine learning suggest that there will be significant progress in this area in the future. The second answer is more radical because it touches the structure of the texts themselves, treating ”words unequivocally as data” (Laver et al., 2003, p. 312). Of course, this refers to the ”Wordscore” (Laver et al., 2003) and the ”Wordfish” (Slapin and Proksch, 2008) approaches.

Despite significant differences in the procedure, both share several assumptions and procedural fundamentals. The common basis of both approaches was an important inspiration and source for the cosine method proposed here for measuring party strategy.

Wordscore and Wordfish have shown that policy positions of political parties can be measured using bag-of-words approaches, thus laying the foundation for further developments in this field of research. Therefore, the following section explains the principles of bag-of-words (or vector space) models and how the cosine approach works.

Fundamentals of Bag-of-Words Models

The Wordscore approach was the first to establish the bag-of-words model in political science. Slapin and Proksch (2008) took up this model and developed their own method for determining party positions from manifestos. Scholars of political science using bag-of-words approaches assume ”that relative word usage of parties provide information about their placement in a policy space” (Slapin and Proksch, 2008, p. 708).

The bag-of-words model was initially developed at the interface between linguistics and computer science in the field of information retrieval. However, it is of great importance today in many natural language processing tasks.

The core assumption of bag-of-words approaches is that the frequency of words in a document is sufficient to extract relevant information. In contrast, the order of words and sentences in the document can be ignored: ”Automated text analysis methods usually treat documents as a vector containing the count of each word type within the document, disregarding the order in which the words appear. This ‘bag-of-words’ assumption reduces the dimension of natural language text, representing each document as a single vector with length equal to the number of unique words in the text” (Lucas et al., 2015, p. 257).

Due to its initially seemingly simple form of document representation, this approach is often met with skepticism: ”Critics of word frequency-based approaches are quick to point out that such algorithms are ignorant of sentence structure and context. For instance, the expressions “We are against lowering taxes, and for tax increases” and “We are for lowering taxes, and against tax increases” use the exact same words with the same frequencies, even though the meaning is reversed. A word frequency approach used on only these statements, however, will provide identical estimates. While this may indeed be cause for concern for short statements, we believe that this is not problematic for the analysis of long texts such as election manifestos" (Proksch and Slapin, 2009, p. 324).

In addition to the word order, possible problems due to the changing meaning of words are often pointed out. Especially if the intention is to create long time series, it can become a problem that the meaning of words changes over time: "For Wordscores, the difficulty is that the political lexicon changes over time" (Benoit and Laver, 2007a, p. 132). If, on the other hand, only two consecutive election dates are compared, the impact of the language change is negligible: "We are in effect assuming that party manifestos in country c at election t are valid points of reference for the analysis of party manifestos at election t + 1 in the same country. Now this assumption is unlikely to be 100 % correct, since the meaning and usage of words in party manifestos change over time, even over the time period between two elections in one country. But we argue not only that it is likely to be substantially correct, in the sense that word usage does not change very much over this period, but also that there is no better context for interpreting the policy positions of a set of party manifestos at election t + 1 than the equivalent set of party manifestos at election t" (Laver et al., 2003, p. 314).

Furthermore, lexical ambiguity can be a problem. Scholars of lexical semantics have developed concepts to capture differences in the relationship between words and their meanings: "Synonyms are words with the same meaning (or very similar meaning): Car and automobile are synonyms. Homonyms are words that are written the same way, but are (historically or conceptually) really two different words with different meanings which seem unrelated. Examples are suit ("lawsuit" and "set of garments") and bunk(sic!) ("river bank" and "financial institution")" (Manning and Schütze, 1999, p. 110).

While these may seem like big problems at first, practice shows that, in reality, they are comparatively small problems, especially when texts of the same genre and time are compared. Political texts as means of communication are carefully written to ensure that their meaning is as unambiguous as possible. Again, especially long texts, like party manifestos, are less susceptible to this kind of problem. Even semantic errors rarely occur, so they have little influence or even out in longer texts.

In essence, the "bag-of-words" approach is therefore considered to have a good performance: "An ongoing surprise and disappointment is that structurally simple representations produced without linguistic or domain knowledge have been as effective as any others" (Lewis, 1998, p. 6).

In order to determine policy positions or party ideology based on the bag-of-words approach, some further assumptions are necessary. First of all, the construct to be measured should be considered as a latent variable: "This means that ideology is not something that the researcher can directly observe, rather it must be indirectly estimated based upon observable actions taken by parties and their members. The observable action we are most concerned with here is the writing of election manifestos" (Proksch and Slapin, 2009, p. 324).

Building on the distinction between ideal policy positions and stated policy positions (Laver, 2001), a more fine-grained operationalization presents the writing of a manifesto as "a stochastic text generation process" (Benoit et al., 2009, p. 497), in which ultimately three different policy positions can be differentiated.

First of all, there is a true (or ideal) policy position, which is "fundamentally unobservable even, arguably, to the author" (Benoit et al., 2009, p. 498). The true position must be distinguished from the "intended message" about the position.

The intended message can be the honest attempt to formulate one's true position or the strategic communication of another position to be taken for one's own. This intended message "exists only in the brain of the author and is also fundamentally unobservable" (Benoit et al., 2009, p. 498).

In order to communicate this intended message, the author produces the observable text, the stated position. Even if the intended message is the same, each new attempt to formulate it will differ. A text can therefore be understood as the result of a random experiment. A true value exists, but every single run of the experiment produces a slightly different result.

When trying to put a message into words, the authors are not entirely free. The number of synonyms is limited. The rules of grammar allow only certain phrases, words have fixed meanings, and therefore there is only a finite number of ways to formulate a particular meaning through them.

For this reason, it is reasonable to assume that "the language used by political parties expresses political ideology. Ideology manifests itself in the word choice of politicians when writing party documents. More specifically, Wordfish assumes that parties' relative word usage within party documents conveys information about their positions in a policy space (Proksch and Slapin, 2009, p. 324).

This is the same assumption that guides other methods like probabilistic topic models as well: ”Topic models ... are based on the idea that documents are mixtures of topics, where a topic ... is a probability distribution over words” (Steyvers and Griffiths, 2007, p. 427).

In linguistics, this assumption is known as the ”distributional hypothesis”: ”This hypothesis is often stated in terms like ‘words which are similar in meaning occur in similar contexts’ (Rubenstein & Goodenough 1965); ‘words with similar meanings will occur with similar neighbors if enough text material is available’ (Schütze & Pedersen 1995); ‘a representation that captures much of how words are used in natural context will capture much of what we mean by meaning’ (Landauer & Dumais 1997); and ‘words that occur in the same contexts tend to have similar meanings’ (Pantel 2005), to quote a few representative examples. The general idea behind the distributional hypothesis seems straightforward. There is a correlation between distributional similarity and meaning similarity, which allows us to utilize the former in order to estimate the latter” (Sahlgren, 2008, p. 34).

Since different topics are described with different words (even agreement is signaled differently than disagreement), it seems linguistically justified to conclude from the distribution of words to latent constructs such as policy position. So in the next section, I will show how the bag-of-words approach is applied in practice.

Elements, Vectors and Matrices

In order to extract information from texts, they are represented as a vector whose individual elements represent the frequency of unique words in the text. These vectors are combined into the document-feature matrix (DFM) to compare several texts.

The DFM is a table whose individual rows correspond to a unit of investigation (mainly a document), while the columns stand for a feature (mainly unique words). The observed frequency of each feature is entered in the individual cells.

The resulting document-feature matrices are usually characterized by a high number of individual features, whereby many cells remain unoccupied so that we speak of a sparse matrix.

Document-feature matrices can be manipulated in many ways. The most important is the stemming of individual words and the removal of so-called stopwords that carry little meaning but are grammatically necessary. Furthermore, the values of the matrix can be inverted (term frequency-inverse document frequency or tf/idf) (cf. Manning and Schütze, 1999, p. 543) to reflect the importance of words or otherwise weighted in order to meet the requirements of the research project.

Based on this data, very different statistical inference methods can be applied. In the following section, one of these methods, the vector space model, is presented.

4.1.3 Cosine Similarity and the Vector Space Model

Party programs are generally seen as "encyclopedic statements of the parties' positions" (Slapin and Proksch, 2008, p. 709), from which information on left-right positioning of parties can be obtained (Jahn, 2011). In addition, Pelizzo emphasizes that election programs and measurements based on them, such as the RILE, "indicate parties' direction, that is how (and how much) parties move to adjust to changing political conditions and to remain competitive" (Pelizzo, 2003, p. 67).

This section explains how statistical methods can measure party positions using bag-of-words approaches. More specifically, the measurement developed here is intended to capture the strategy of an established party vis-à-vis a new party, whether or not these changes occur on issues of the classical right-left dimension. This is necessary because there is a particular intersection between new and niche parties.

The cosine similarity approach presented here has certain parallels with the well-known Wordscore method of Laver et al. (2003). Conceptually, the main difference is that no external reference texts are used as content validation of the measured dimension. Instead, a pairwise measurement of party programs is used, whereby the selection of these party programs allows conclusions to be drawn about which party is developing in which direction.

In technical terms, there are further differences. Wordscore uses reference texts to locate the texts to be analyzed closer to one pole or the other with respect to their correspondence of the observed word frequencies with the frequencies of the reference texts. For this purpose, conditional probabilities are calculated for each word (Laver et al., 2003, p. 317).

Wordfish, on the other hand, estimates regression parameters based on the assumption that words are used according to the Poisson distribution (Slapin and Proksch, 2008, p. 709-710). Both procedures have in common that they try to map several texts on one dimension, whereas in this study, a pairwise comparison of party manifestos is intended.

A further difference is that both methods are proprietary developments in political science. This is surprising because one of the most basic analysis methods for bag-of-words approaches, the vector space model (or vector similarity model), has never been applied in political science. However, it ”is one of the most widely used models for ad-hoc retrieval, mainly because of its conceptual simplicity and the appeal of the underlying metaphor of using spatial proximity for semantic proximity (Manning and Schütze, 1999, p. 539). This directly addresses the theory of party competition, which was presented earlier.

The basic idea of the vector space model is to equate spatial and content proximity of documents. The documents represented as vectors are thus mapped in a multidimensional space. The cosine of the included angle of both documents is a measure for the similarity of the content. Thus, from the frequency of words in different documents, the similarity of these documents is inferred. Documents that have a higher degree of correspondence between their terms are thus considered to be more similar.

This method has been successfully used for search queries. The calculation of the vector similarity between the search query, on the one hand, and the available documents, on the other hand, has proven that relevant documents can be found: ”The most relevant documents for a query are expected to be those represented by the vectors closest to the query, that is, documents that use similar words to the query. Rather than considering the magnitude of the vectors, closeness is often calculated by just looking at angles and choosing documents that enclose the smallest angle with the query vector” (Manning and Schütze, 1999, p. 539).

To illustrate this principle, consider the following example: A researcher wants to know which parties have a similar attitude towards environmental topics. Therefore, a highly simplified dictionary is being developed that consists only of the terms ”pollution” and ”sustainability” to address this question.

In the first manifesto, document A, the term pollution (called feature i) is observed twice. The term sustainability (or feature j), on the other hand, is observed four times. The corresponding vector is called A(2,4). In document B, feature i occurs four times, but feature j occurs only three times. Therefore, the vector is called B(4,3). Both vectors can be represented in a two-dimensional coordinate system (Figure 4.1).

The cosine of the included angle determines the difference in the direction of both vectors. If a third document C would be added, where feature i occurs four times and feature j two times, the angle between documents A and C can be determined additionally. Since the angle is larger, it is clear that documents B and C are more similar to each other than documents A and C or an A and B. Documents B and C use the word pollution equally often. However, the emphasis on sustainability differs by one reference. Document A uses the word pollution half as often but emphasizes sustainability. Thus, it can be concluded that parties B and C have a more similar attitude towards environmental topics than parties A and C do.

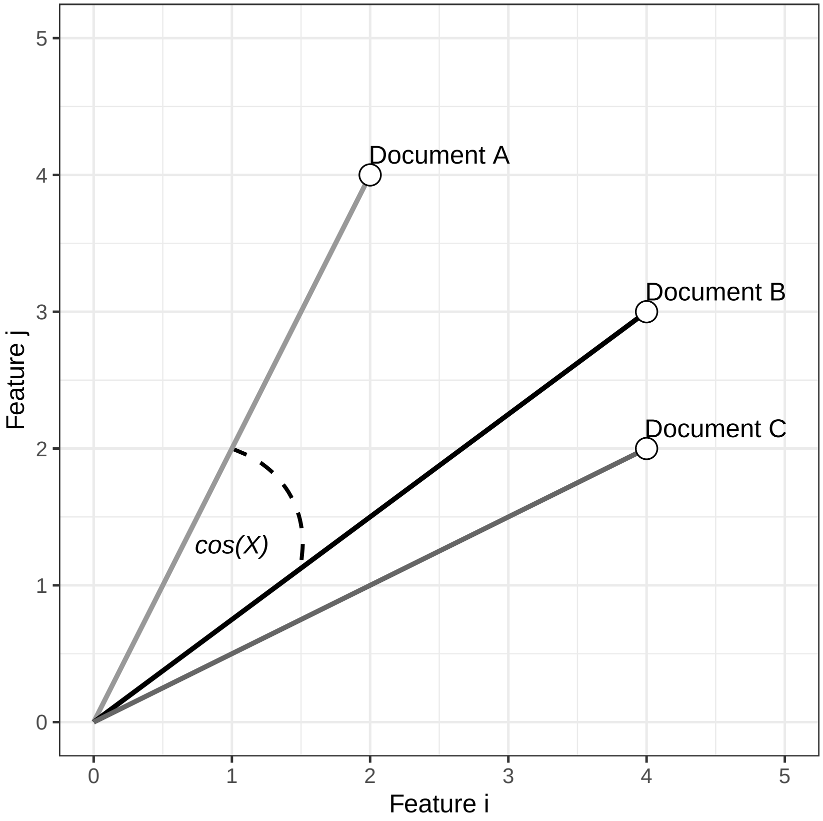

Figure 4.1: Cosine Similarity of Three Documents in a Two-dimensional Vector Space

This fact can not only be read off graphically but also calculated as cosine similarity according to the following equation:

$$\cos x \quad \frac{{}^{n}_{i1}\, A_i \cdot B_i}{\sqrt{{}^{n}_{i1} A_i^2} \cdot \sqrt{{}^{n}_{i1} B_i^2}} \tag{1}$$

Figure 4.2: Equation to Calculate Cosine Similarity of Documents A and B

In the formula, the ”inner product” of the two vectors in the numerator is divided by the vector magnitude in the denominator. This makes cosine similarity insensitive to different document lengths: ”To compensate for the effect of document length, the standard way of quantifying the similarity between two documents d1 and d2 is to compute the cosine similarity of their vector representations V (d1) and V (d2) where the numerator represents the dot product (also known as the inner product) of the vectors V (d1) and V (d2), and the denominator is the product of their Euclidean lengths” (Manning et al., 2008, p. 111).

To return to the above example: According to the formula, the cosine similarity between document A and B is 0.89, between document A and C 0.8, and between document B and C 0.98 if only the features of the ”environment dictionary” are used.

The cosine similarity approach has several advantages that make it suitable for analyzing political texts:

First, the approach follows the principle of parsimony. Whether parties behave as saliency theory assumes or not is not presupposed but is the subject of empirical research. This corresponds to the demand in the literature.

Furthermore, cosine similarity considers the number of different features, their frequency and the length of the compared documents. This is a significant advantage over other measurements of vector similarity like the Jaccard index. For example, the Jaccard index covers only the vocabulary overlap but is blind to the frequency of words and the length of the manifesto.

Thus, the cosine approach also meets Slapin and Proksch’s (2008) requirements, who pointed out that parties sometimes write manifestos of above-average length and measurements, therefore, have to correct this [p. 706].

Even the weighting of words according to the probability of their occurrence, which is essential for Wordfish, can also be easily taken into account in the vector similarity approach by weighting the document-feature matrix.

In order to avoid potential problems caused by the change in political vocabulary and at the same time ensure the interpretability of the index, the chosen documents play a decisive role.

Election programs of new and established political parties are compared so that the measurement of similarity indicates the direction of policy change. These strategies are determined based on two measurements of three party programs:

First, the similarity between the election program of the established party at election t $_0$ and the reference election program of the new party at the same election is determined. Then the election program of the established party at the subsequent election t_1 is compared with the reference election program of the new party at t_0.

From the comparison of both measurements, it can be concluded whether the established party has brought its election program closer to that of the new party or not.

By using election manifestos, it is ensured that texts are sufficiently long to obtain reliable measurements. Among other things, this ensures that individual problematic terms have only a negligible effect on the measured value. Furthermore, removing terms that do not make sense ensures that similarity between texts is not based on grammatically necessary words that have no meaning. Third, the comparison of election programs is limited to a maximum of two consecutive election dates so that a stable political vocabulary can be assumed.

Last but not least, the cosine approach also allows for an ideological calibration that goes beyond the use of election programs as reference texts, much like it is known from other text analysis methods like Wordscore or the study from Proksch and Slapin (2006), who ”examine party positions in two dimensions (economic and social)” [p. 540] by parsing ”the reference texts into economic and social sections and then estimate positions using the respective sections only” (Slapin and Proksch, 2008, p. 707).

So by selecting document sections that are assigned to a specific dimension, for example, the left-right or the green-growth dimension (Jahn, 2016), it is possible to focus more specifically on aspects of content that are of interest.

Here, too, it is important to note that, on the one hand, as many different terms as possible should be included to cover the phenomenon in its entire range, and, on the other hand, that the selection of reference texts should not be too comprehensive. In the first case, there is the danger of not capturing essential elements of the dimension; in the second case, the measurement would no longer discriminate between concepts (cf. Laver et al., 2003, p. 315).

While cosine similarity has proven its usefulness in diverse natural language processing tasks, the evidence is still missing that political texts can be analyzed as well. I will try to provide this proof in the next section.

4.2 Synthetic Manifestos – Assessing Measurement Properties

The previous sections discussed why previous measurements of party position have been met with criticism. Then, based on the information retrieval literature, a new approach was presented that can be used to explore parties' strategies.

The crucial question now is how well the presented computer-assisted approach can measure differences in party programs. In other words: How valid are the measurements?

Assessing the measurement quality of different indices is a great challenge. While reliability can be determined relatively easily by repeating the same measurement on the same data, this is not the case for validity.

In the literature, different validity measurements have been discussed (Adcock and Collier, 2001). The so-called external validation, i.e., comparing one measurement with the result of another measurement from a different independent data source, is considered the ideal solution. Different methods for assessing external validity are conceivable for scholars of party policies.

One possible method is the comparison to expert surveys. The problem with that method is that experts, due to the definition of the term, really know their subject matter well, i.e., they consider the election manifestos and the corresponding research. Accordingly, this measurement is not independent of other measurements.

Another particularly frequently used method is the comparison with a ”gold standard” like the RILE: ”Computers can easily count words in an electronic text. But how do we know that these are really telling us what we want to know about policy? An easy way is to compare the estimates these generate with the previously validated ones from the Manifesto data” (Budge and Bara, 2001, p. 2).

Problems with this method arise when established indices like RILE have validity problems. Furthermore, if the same data is used for both scaling methods, the independence requirement is violated again.

So, as long as it is ultimately unclear what ideological content a document has, it is challenging to establish the validity of a new measurement beyond doubt. To address this problem, a simulation experiment was conducted here.

The simulation aims at better understanding the behavior of different position measurements by generating synthetic manifesto data. Synthetic manifestos have the advantage that the researcher can determine the ideological content of the party programs under study. Thus it is then also clear which measurement correctly represents the latent variable.

The theoretical basis of the simulation is the assumption that different topics are usually formulated with different terms, better known as the distribution hypothesis. The connection between words and topics is probabilistic, i.e., there is also the situation that the same terms are used to describe different things. However, this is not the rule.

From the analysis of the frequencies of terms, it is possible to conclude the underlying topics. As already discussed, this is the basis of numerous methods, for instance, topic modeling procedures, such as the Latent Dirichlet Allocation, as well as classification procedures for texts, as they are already used in political science.

As a basis of the analytical process, I assume that party programs are random selections from the universe of all possible propositions that express a particular policy position. On the sentence level, this corresponds to Slapin and Proksch's (2008) approach which assumes on the word level that a concretely observed word comes from a random selection of all possible terms.

The starting point of the simulation developed here are 33 German party programs with 67989 quasi-sentences encoded by the Manifesto Project. An average party program was constructed based on these codes: it contains as many sentences on the respective policy area as the average of the 33 party programs. The same is true for the length of this manifesto. This average party program is then processed, resulting in synthetic manifestos that differ from the average party program by exchanging a defined set of randomly selected sentences.

The next step is to compare the average manifesto with the party program changed by 10, 20, or even 1000 sentences, so the impact of a marginal sentence can be determined. The comparison is based on established measurements such as RILE or log RILE, but also on measurements of text similarity such as cosine similarity or Jaccard similarity. In order to ensure that the results are robust, this process is repeated ten times, with the sentences of the average party program being randomly selected each time. This repeated measurement thus ensures that no measurement artifacts are produced.

Of course, the result of this comparison process is determined by the population of the fed in sentences. The repeated generation of the average manifesto as well as the selection of sentences to be exchanged corresponds to a random sample from a population or, in terms of probability theory, it can be described as the basic urn model, where a ball (or sentence, in this case) is drawn from an urn (or here the entirety of all sentences) and then put back before the process is repeated.

Hence, the drawn sentences represent the entirety of all sentences in the same sense, as a random sample represents the population it is sampled from. Since the average manifesto refers to the same population as the entirety of the sentences, there would be no changes in the respective measurement values if the exchanged sentences come from the population of all sentences.

If, on the other hand, the population changes in advance, e.g., by making only left-wing or right-wing party programs the population, the replacing sentences change accordingly. Thus, as the number of replaced sentences increases, the following picture emerges: The measured values move closer and closer to the right or left population, i.e., they move away from the average manifesto with which the simulation was started.

The choice of different populations or selection bases thus corresponds to different test cases for the measurements used. This allows the advantages and disadvantages of the individual indices to be exemplified in terms of their ability to measure different issues.

Therefore, two experiments were performed. The first experiment changed the average manifesto by feeding in exclusively left sentences. This experiment aims to test the standard RILE scale where it should be strongest: When measuring the left-right dimension in election programs. The comparison with other measurements shows the extent to which they can detect positions on the left-right axis.

In the second experiment, the selection of sentences to be fed in was changed: Instead of left sentences, sentences are now fed that are assigned to the green-growth dimension. This second experiment reverses the first experiment: Instead of feeding sentences belonging to the left-right dimension, only sentences are added which do not belong to it. A valid measurement of a party's left-right position should not react to such changes.

Taken together, both experiments provide information about the specificity and sensitivity of the measurements under investigation. In the following sections, the results of these two experiments are presented.

4.2.1 The Left-Right Experiment

The main line of conflict in the political competition continues to be the left-right dimension. Accordingly, it is crucial to validly capture changes in party position in this dimension. In order to assess the measurement quality, the content analytical measurements of the Manifesto Project need to be compared to text analytical measurements.

To ensure that the contents of party programs are known, synthetic manifestos are constructed. Therefore, the experiment uses the annotated sentences from the corpus of the Comparative Manifesto Project.

The starting point of the simulation is the construction of an ”average manifesto”. This average election program contains all 56 topics of the CMP in frequency as they are contained in all available German election programs. Deviating from this general rule, the number of sentences carrying left or right topics were determined to be precisely the same on the left and the right side. This was done because the RILE has a well-known tendency towards the center, which should be excluded from the measurement here.

In order to generate the average program, the first step was to analyze how frequently each issue occurs in all German election programs. In the next step, quasi sentences were randomly selected from the corpus and arranged to correspond to the calculated averages topic frequencies. This average party program was then replaced sentence by sentence with quasi-sentences devoted to left issues to simulate the growing importance of this dimension. The sentences to feed in were randomly selected from the collection of all sentences assigned to a category classified as left in RILE.

This simulation process was repeated ten times. This means that ten different average manifestos were constructed from randomly selected sentences based on given average frequencies. From each of these ten texts, 1000 sentences were randomly selected and replaced by a random sentence taken from those associated with the left position. Thus, this test series consists of 1000 different synthetic election programs. Due to the tenfold repetition, the simulation results are based on a total of 10000 election programs (Figure 4.3 and Figure 4.4).[1]

The correlation between the absolute differences of the RILE and the cosine dissimilarity measurement with r=0.997 shows that both measurements are virtually identical.

A similar agreement is shown when the absolute RILE differences are compared with the cosine similarity scores calibrated to the left-right dimension. Here the correlation coefficient is 0.936.

Interestingly, the degree of agreement breaks down when the individual parts of the calibrated cosine similarity measurement are compared with the absolute RILE differences. While the correlation with the cosine scores calibrated with left sentences is still r=0.911, the correlation with the right sentences changes direction and goes down to r=-0.784. This is because the RILE is constructed as a summated rating scale. The calibrated right cosine score is negative because some right sentences are deleted and replaced by left sentences through the simulation process. The RILE hides this by design.

I present the raw results of the simulation to give an impression of the different ranges of values and variances, as well as the min-max standardized values, which allow a better comparison of the indices.

On the X-axis, the number of exchanged quasi-sentences is deducted. On the left Y-axis, the absolute differences of the RILE between the average manifesto and the respective synthetic manifesto are noted. On the right Y-axis, the text similarity or dissimilarity between 0 and 1 is recorded. The graph shows that the RILE correctly represents the increasing salience of the left position, so the index can be considered sensitive as long as the CMP codings are valid.

For better comparability, the measurement of text similarity between the average manifesto and the respective synthetic manifesto was inverted here so that dissimilarity is measured. As a result, both election programs' increasing degree of dissimilarity is correctly captured.

Interesting is the comparison with the calibrated measurement. Here, 1000 left, and 1000 right sentences were defined as reference texts. After the two measurements were performed, the result for the right reference text was subtracted from the left reference text.

[1] Many overlapping data points pose a challenge for the graphical representation. To avoid overplotting, I drew a random sample of 200 measurement results to be shown in the graphs. The random sample ensures that the interpretation of the results does not differ from the original data.

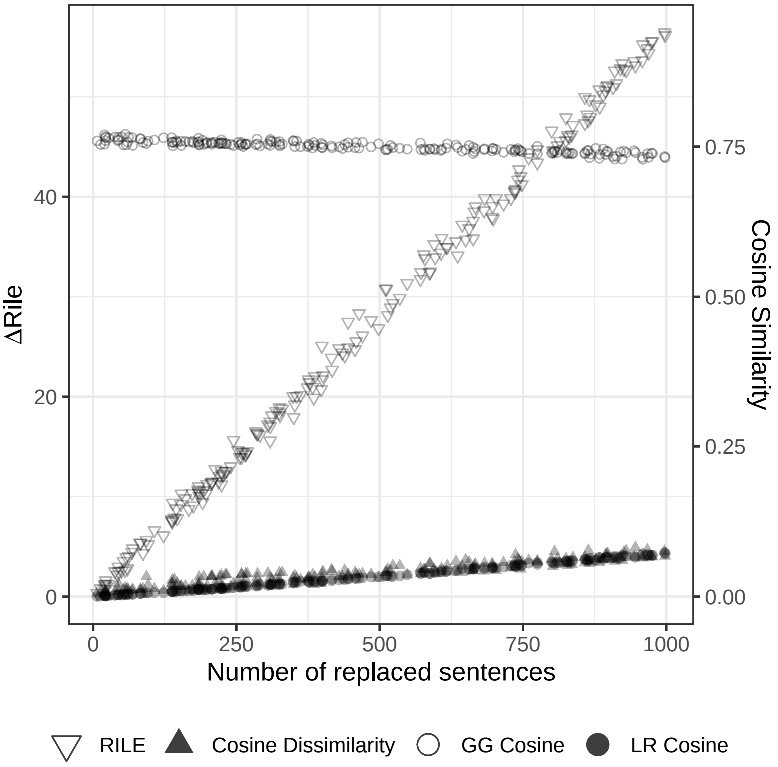

Figure 4.3: Raw Results of the Left-Right Simulation Experiment

Although the calibrated cosine method uses different reference points, the results are very similar. That makes it clear that the robust and straightforward cosine dissimilarity measurement based on the party programs can also be used to detect changes in the left-right dimension.

The calibrated green-growth measurement is based on the random selection of 1000 sentences dedicated to these issues. Surprising at first is the recorded changes of this dimension in the simulation. However, this can be easily explained: Another sentence from the average manifesto is replaced by a left sentence with each simulation run. The selection of these deleted sentences is random. Thus, sentences associated with the green-growth dimension can also be deleted during each run. Due to the random selection,

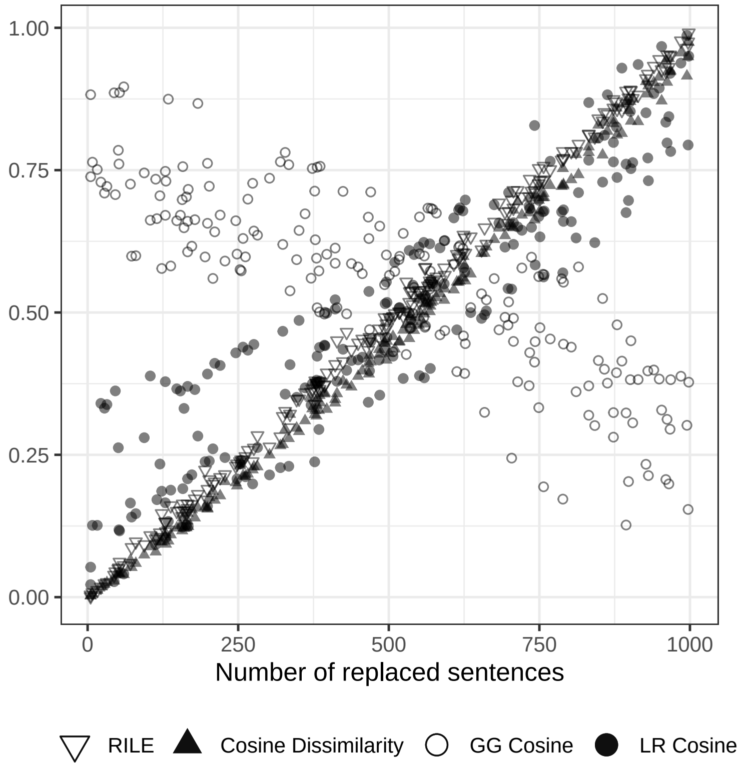

Figure 4.4: Normalized Results of the Left-Right Simulation Experiment

some deviation is possible at each measurement repetition. Certain variations occur, which are recorded here as a variance of similarity to green-growth issues. All in all, of course, the number of topics that are not left is decreasing. Correspondingly, the similarity to green-growth topics decreases overall.

This connection becomes even more apparent when the normalized values are considered (c.f. Figure 4.4). The min-max normalization standardizes the value ranges of the individual indices to values between 0 and 1. This facilitates the comparison of the individual measurements, as it shows the differences in the variance of the ten measurements and thus the size of the confidence intervals even more clearly.

RILE and cosine similarity perform equally well and are nearly indistinguishable. However, the calibrated cosine measures have higher uncertainty because there are different ways to state the same position. This uncertainty could be reduced by increasing the number of sentences used for calibration. However, this leads to conceptual arbitrariness and endogeneity problems at a certain point.

4.2.2 The Green-Growth Experiment

The green-growth experiment was conducted to test the specificity of the RILE. In the experiment, a manifesto was simulated, in which the green-growth dimension becomes more and more salient in each run. Because no left or right categories are changed, the correct measurement would show no differences.

While very good matches between the indices were found in the left-right experiment, the green-growth experiment shows completely different results.

The correlation between RILE and cosine dissimilarity is only 0.119. The comparison with the left-right calibrated cosine measurement shows that this can be attributed to the RILE. Here the correlation is only r=-0.202. The RILE shows a similar, though low correlation with the calibrated green-growth measurement of r=0.183, while the cosine dissimilarity is correlated with the calibrated green-growth measurement by r=0.974. All in all, this pattern of correlations indicates that the RILE shows considerable noise here and therefore has a low correlation with all dimensions. This becomes even clearer when the graphs are considered (Figure 4.5 and Figure 4.6).

The simulation shows that the RILE measurement deviates significantly from the ideal. The purely random removal of sentences leads to deviations of up to 2.5 points when measuring the RILE. This value is just as high as party movements that are observed in real elections and are therefore considered worthy of explanation in empirical studies (Jahn et al., 2018a).

In contrast, the calibrated cosine measurement of the right and left issues shows the desired flat slope over all simulation runs. The cosine dissimilarity measure shows a parallel course to the calibrated green-growth measurement on a much lower level.

If the normalized data are used, how similar these two measurements perform becomes even more apparent. The cone-shaped course of the cosine dissimilarity measurement is also interesting here. The increasing number of exchanged sentences explains this: The number of unique words is related to

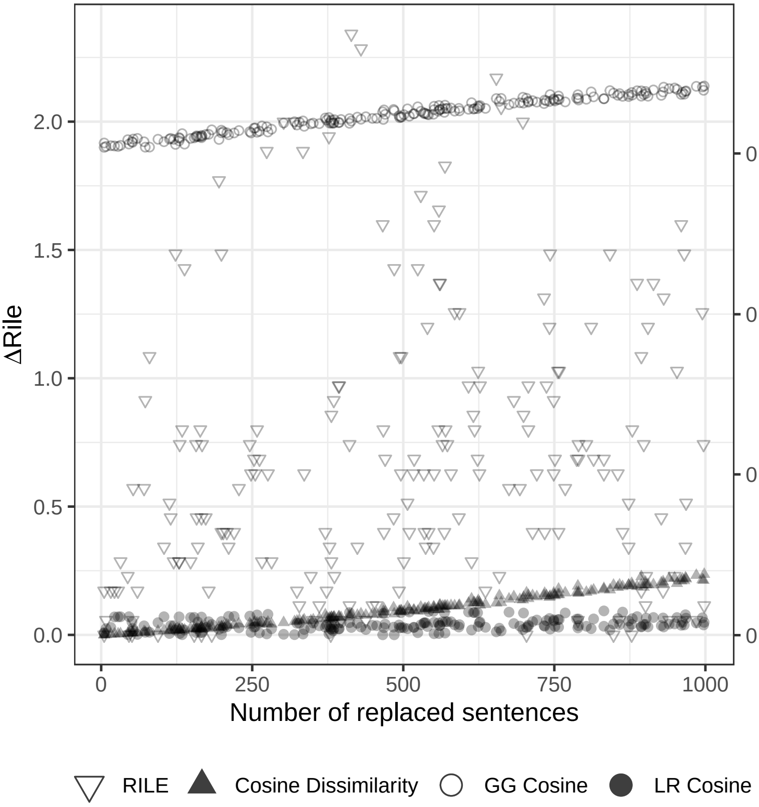

Figure 4.5: Raw Results of the Green-Growth Simulation Experiment

the text length. When 1000 sentences are exchanged, the number of features is greater than when only ten sentences are exchanged. Accordingly, the probability of unique words that do not occur in the other measurement increases with a higher number of replaced sentences.

4.2.3 Conclusion

When the results of both experiments are combined, it can be said that computer-assisted text analysis methods perform similarly well, if not better than methods based on manual content analysis. Despite their respective ad-

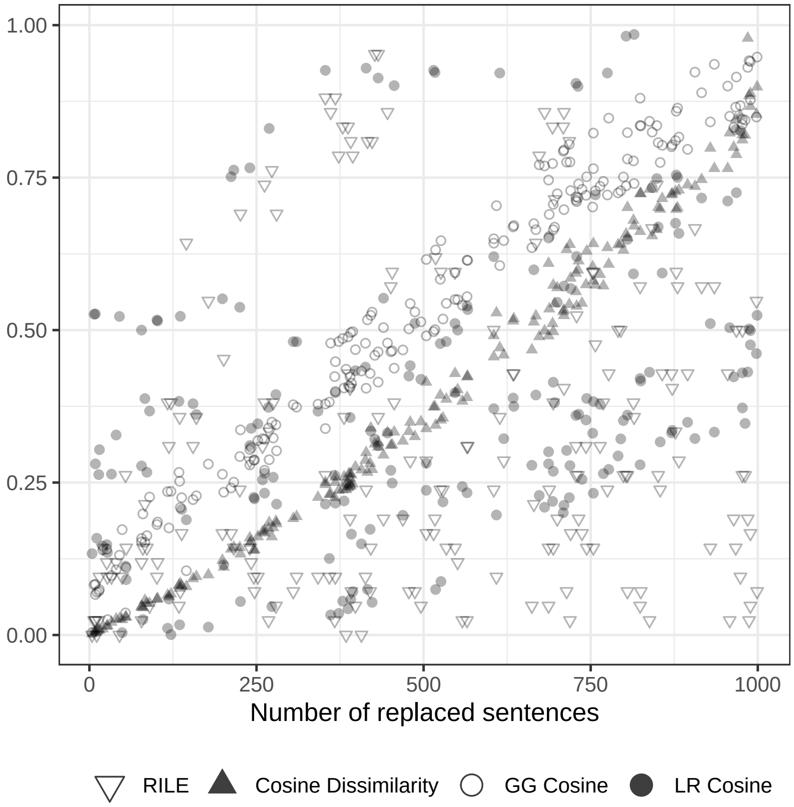

Figure 4.6: Normalized Results of the Green-Growth Simulation Experiment

vantages and disadvantages, both methods are suitable for capturing changes in party programs.

If the sensitivity to changes in the issue salience is the key issue, cosine similarity between the party programs can be used without hesitation. However, if specific issues or dimensions of political competition are to be addressed, a calibrated measurement is preferable.

The RILE can be used as long as it is clear that the main line of conflict is the left-right dimension. If there are concerns about this, for example, because niche parties advance issues for which there are indications that their ideological profile is different, text-analytical measurements are at an advantage.

Based on this experiment, I argue that a combination of both measurements is reasonable: While the RILE captures changes in position on the left-right dimension, changes in issue salience are accounted for by the cosine similarity scores. Thus, if both measurements are integrated into one model, it can be assumed that they complement each other well. I deal with such modeling issues in greater detail in the next chapter.