

# Challenges for Estimating Policy Preferences: Announcing an Open Access Archive of Political Documents

KENNETH BENOIT, THOMAS BRÄUNINGER and  
MARC DEBUS

*We provide a comparative perspective on the contributions of the special issue with regard to their applied methods and findings. In addition, we discuss problems that arise when using 'wrong' or at least 'incorrect' versions of election manifestos by presenting replications of estimated policy positions of German parties. We show that the latter can result in biased estimates that may affect the outcome of theoretical models. On the basis of those findings, we present the idea of the open access archive polidoc.net to build up a common database for political texts.*

## INTRODUCTION

As shown through the studies in this special issue, the analysis of political texts provides considerable leverage in the estimation of the policy preferences of political actors. The leverage offered by textual analysis is not limited to political parties, furthermore, but also extends to political actors within parties or legislatures – precisely the level of politics where preferences are difficult or impossible to measure using other methods. Different problems call for different kinds of text analysis, each with its own advantages and drawbacks, but the variety of measurement challenges presented in this issue illustrates the ways that different approaches can produce reliable and valid estimates, compared with other methods and reinforced by the similarity of their findings compared to one another. Despite the complex and changing character of the German federal party system, with its rich set of intra-party differences, the articles presented in this issue reinforce each other with the similarity of their findings. Overall, the results demonstrate convincingly that the analysis of textual content, whether manual, partly, or fully automated, provides an effective, efficient, and replicable means for estimating the policy positions of political actors. The articles gathered in this special issue have used these methods to study very different questions pertaining to German political parties or individual party members in different arenas of party competition. One study found evidence of important vertical differences between positions of parties at the state and federal levels, by finding that the programmatic profiles of state parties tend to reflect the state-level differences in the social and economic composition of the electorate. Another study used textual analysis to draw attention to party cohesion and party discipline by identifying differences between the programmatic positions of individual Bundestag members, showing that they were not only

distinctive but also different from those of their party cores. A third study highlighted the assertiveness of parliamentary parties and governments in the legislative process by using the wording of legislative proposals to map the positions of parties and the federal government.

No matter how advanced our techniques for estimating political preferences from texts, of course, our ability to use them will depend on the availability and the quality of political documents to be analysed. These documents must be carefully selected, furthermore, based on our knowledge that they contain sincere information about their authors' political preferences. Even when we may not understand the precise nature of the textual representation of a political actor's preferences, all valid methods of a textual analysis rely on the assumptions that these preferences are contained in text in some fashion. No amount of technical or statistical wizardry, in other words, will extract valid policy preferences from the Free Democratic Party (FDP)'s phone directory or from the Bundestag's Bavarian cookbook.

While these pitfalls in textual selection may seem like easy traps to avoid, less obvious may be the problem of how to deal with subtle differences in texts that make them difficult to compare. Whatever method is used – human interpretative coding or any of the recent computational techniques – the quality of the results will also depend on whether the texts containing information on actors' preferences can be meaningfully compared. Proksch and Slapin's contribution to this volume critically highlights this aspect when comparing the election manifestos of the federal parties to the selection of speeches and other documents that have been used by the 'Comparative Manifesto Project' (CMP)<sup>1</sup> to measure the ideological position of German political parties, and this is something we return to briefly below. We can simply put more trust in inferences drawn from texts from similar contexts than from a comparison of political texts drafted by different types of actors for different purposes and addressing different audiences.

Finally, in addition to choosing valid, comparable texts, we must also pay careful attention to practical issues related to checking for errors in our textual data. Not only do such errors tend to creep easily and frequently into textual data, but they also tend to be much more difficult to detect than errors in the typical numerical dataset, especially when dealing with large quantities of text or texts in different languages. In their contribution, König and Luig stress the importance of working with corrected party programmes that have been adjusted with the original document, because some parts or words of the original text could be missing due to errors in scans or transcriptions. As we demonstrate below, computational methods, most often based on word or phrase frequencies, may be sensitive to misspellings. Fortunately, methods do exist to detect and fix these data-related problems, as long as we are aware of them. Detecting and fixing such problems, furthermore, will be greatly enhanced when the full texts used for analysis are easily accessible by the political science community. Indeed, for all three types of challenges we have outlined in the analysis of political text, the free and easy accessibility of texts used in political analysis is vital for both replicability and extendability – key objectives for further disciplinary progress in using political texts as a fundamental source of information to learn more about political actors and their policies. Part of our message in this conclusion to the special issue, therefore, is to introduce the idea of an open access archive, *polidoc.net*, that allows researchers

to access, archive, and disseminate political texts of all kinds, notably electoral manifestos, government declarations and the like.

In the following section, we review the findings and implications of the various contributions in this issue on party politics in Germany and of content analytical approaches to measuring policy positions. The third section illustrates the importance of text selection and pre-processing for virtually any type of text analysis when comparing the results of different sources of text information: different versions of one election manifesto, different word pre-processing rules and the like. On the basis of those findings, the fourth section presents the idea of the open access archive *polidoc.net* to build up a common database for political texts. The final section discusses incentives for further studies on patterns of German party competition and its measurement with various methods of text analysis.

#### ESTIMATING POLICY POSITIONS OF POLITICAL ACTORS IN GERMANY: IS THERE CONGRUENCE IN THE RESULTS?

The initial question raised in the introduction of this special issue of German politics was how German parties behaved programmatically with regard to changes in the preferences inside the electorate, the number of relevant parties in the party system and different patterns in coalition politics. It was argued that the identification of ideological moves or – to be more specific – changes in the positions of parties in various policy areas requires the analysis of political texts as the most reliable and unbiased source of preferences of political actors. Compared to analysis of roll call votes or survey data of party elites or party supporters, texts drafted by the parties (or their intra-party groups) seem to be the best choice to identify their respective preferences, in particular if the research question deals with changes of the positions of political actors over time.<sup>2</sup> Each contribution therefore made use of pre-existing data and refined it with new statistical approaches or referred to recent techniques of estimating policy positions of political actors. Wüst used data from the Euromanifesto dataset, which is based on a CMP-style coding of election manifestos for the European Parliament.<sup>3</sup> He found that parties change their programmatic profile by strengthening European policy issues in their election manifestos for the European Parliament, but that the distance between parties and voters on a European policy dimension seems to play a minor role for most voting decisions. Linhart and Shikano incorporate two leading theories of voting behaviour – the proximity and directional model<sup>4</sup> – into the analysis of party competition in Germany and developed a new method to estimate the positions and intensities of German federal parties on economic and social policy dimensions. Their results not only provide confidence intervals for the estimated party positions, but also make clear that a one-dimensional account is inappropriate for the analysis of German party competition. Bernauer and Bräuninger and Müller assumed on the basis of theoretical considerations that two dimensions – one based on left/right-wing economics and another based on social and moral liberalism versus conservatism – are required to analyse patterns of party competition in Germany. When using the results for coalition theories, Pappi and Seher even recommend distinguishing four to five large policy domains, each characterised by its own policy dimension and party configuration. Pappi and Seher then apply scaling techniques to estimate latent dimensions in the

text from relative word frequencies, in particular the ‘Wordscores’<sup>5</sup> and ‘Wordfish’<sup>6</sup> techniques. König and Luig apply a dictionary-based method (G-LIS) to extract the policy positions of parties and governments by analysing election manifestos and government declarations with regard to the legislative context, that is each proposal from the 4th to the 16th Bundestag terms. They derive the dimensionality of the policy space in Germany on the basis of a Bayesian factor analysis and identify two structuring dimensions in one of 14 policy areas (‘labour and social policy’ domain): the first dealing with economic or ‘resource’ issues, and the second defined by differences on ‘value’ issues. Their novel Bayesian approach not only allows empirical estimation of the policy space by allowing the importance of the issues to emerge from the data, but also makes it possible to estimate confidence intervals for the estimated quantities.

While the different approaches presented in this volume have indicated general agreement that German political competition is two-dimensional, there is less consensus as to exactly how the German parties should be located on these dimensions and how their programmatic positions have changed over time. All contributors agree that the Party of Democratic Socialism (PDS) as well as its successor party, the ‘Linke’, mark the leftmost border of a general left/right dimension or of a conflict line that specifically deals with economic and welfare issues. The Social Democrats and the Greens are located slightly to the left of centre, while the Christian Democrats are placed at the centre-right. In addition, König and Luig, Pappi and Seher as well as Proksch and Slapin identify a move of the Social Democrats towards the right during the 1990s on an economic left/right dimension (König/Luig and Pappi/Seher) as well as on the overall left/right dimension (Proksch/Slapin). Furthermore, the Greens also moved from their leftist positions in the 1980s and early 1990s towards more centrist policy goals. This is in line with the development of German party competition described in the introductory essay and also supported by Linhart and Shikano’s analysis of the CMP data. These patterns of change in the programmatic profile of German political parties on the federal and the state level may have affected coalition formation. One may argue that the Green party became a potential coalition partner at the federal level when it adopted moderate policy positions on economic issues and unequivocal support for the general political system. As an analysis of coalition formation in the German states has shown, using the preferences of state parties provides better results than applying the positions of federal parties. The data on the position change of German parties on the federal and state level also allows us to analyse whether parties change their programmatic profile for tactical motivations. Following the studies by McDonald and Budge<sup>7</sup> and by Adams *et al.*,<sup>8</sup> one may argue that the parties change their positions with regard to the median voter or the preferences of their likely voters, respectively. While this argument is already tested on the national level in a comparative way, Bräuninger finds evidence that parties on the sub-national level behave in a similar way.<sup>9</sup> None of these findings would be possible without applying new approaches of text analysis to the full text of sub-national election programmes. The data on the policy positions of German parties on the sub-national level and on the policy preferences of single MPs provide useful information for deeper-level analyses of legislative politics and the representation of constituency interests in parliament. One could, for instance, look at the impact of the electorate’s policy preferences and the macro-economic performance at the constituency level on

the policy position of the legislator that represents the respective constituency in the federal parliament.

There is only one major disagreement in the placements of German parties: the ‘Wordfish’ technique (employed by Proksch and Slapin as well as by Pappi and Seher) places the Liberals on the right-wing end of the social policy dimension, while analyses using the Wordscores technique indicate that the FDP in state election manifestos adopted a much more leftist policy position on this dimension (Müller), while the analysis of parliamentary speeches by Bernauer and Bräuninger indicates that even members of the FDP’s left-wing intra-party faction belong to the conservative spectrum of the social policy dimension or at least hold conservative positions similar to Christian Democratic MPs. Findings such as these at the intra-party level underscore the contribution of computerised text analysis: It enables researchers to relax the (often-criticised) assumption that parties act as unitary actors. By focusing on state election manifestos or the speeches of individual politicians, researchers can use estimated positions of state parties or MPs to analyse legislative decision-making, in particular in bicameral political systems such as the German one. Furthermore, text analysis can deliver new insights into decision-making inside political parties by looking at the preferences of – more or less – organised intra-party factions or even single party members. These types of intra-party questions would be difficult or impossible to investigate using party-level data such as manifestos.

Pappi and Seher argue that an application of the Wordfish technique tends to recover manifest policy differences between parties, whereas categories established by the Manifesto Research Group (MRG) capture ideological overtones which are built into traditional party images, in particular in societal policy. According to Pappi and Seher, this results in the progressive policy position of German Liberals on the social policy dimension when applying MRG/CMP data. One may also argue that Wordscores could produce biased estimates: because of the – potentially wrong – FDP’s placement on the social policy dimension by a number of expert surveys<sup>10</sup> and their use as reference scores in the estimation process, the ‘true’ and thereby more conservative policy position of the Free Democrats on social issues cannot be estimated. Since Bernauer and Bräuninger also use the Wordscores technique to extract policy positions of speeches, but get different results from Müller (who analyses election manifestos on the state level), one may wonder whether the reason for the different results on the social policy position of the FDP at the state level and its MPs in the Bundestag is caused by the applied method of text analysis. One different reason could be that the Free Democrats use a very different vocabulary in their written documents that is complicated to capture for computational text analysis. Proksch and Slapin raise this point in their contribution to this special issue when discussing the relative frequency of the word ‘liberal’ in the FDP’s election manifestos.

While we can hardly purport to pass final judgement on the accuracy of one text scaling method versus another, the FDP example does demonstrate that different methods may yield different results, and suggests that these differences may be attributable to how different techniques handle positioning in higher dimensional spaces. The contributions by Proksch and Slapin and König and Luig make an additional point: they stress the quality of political documents that were used for estimating the preferences of German parties for each federal election. This aspect sets an incentive to look

at what happens when ‘wrong’ or ‘incorrect’ documents are used as proxies to extract policy positions of political actors. We discuss this aspect and the idea of creating an open access database of political documents in the following sections.

#### ESTIMATING POLICY POSITIONS FROM POLITICAL TEXTS: WHY SELECTION AND QUALITY OF TEXTS IS SO IMPORTANT

The selection of an appropriate set of texts is a crucial aspect of any type of text analysis, but probably even more so for methods that treat texts as objective data rather than as objects of subjective interpretation.<sup>11</sup> The same presumably holds for the quality and nature of the source data, that is readability, spelling and pre-processing of political texts to be analysed. Human coders find it relatively easy to detect misspelled words, deleted pages, ‘junk’ text (such as line numbers, repeated chapter titles, etc.), but their coding of text may be unreliable. Computerised methods for processing texts, on the other hand, will treat good text and garbage text with the same impartial, unknowing, and perfect reliability. Given the rapidly advancing spread and sophistication of automated text analysis techniques, we draw attention to the potential for automated methods to be affected by undetected errors in text data, illustrating the dangers using three examples from the analysis of German party competition. Our examples use the Wordscores scaling technique of Laver, Benoit and Garry (LBG)<sup>12</sup> to illustrate potential differences caused by selection effects from changing reference documents and scores. We have used Wordscores because of its simplicity and because of its explicit identification of training documents, although our point is a very general one and not intended to be restricted to any particular one of the text analysis methodologies used in this volume.<sup>13</sup>

According to the CMP codebook, the Christian Democratic Union (CDU)’s 1962 election text coded by the project was entitled ‘Düsseldorf Declaration of the CDU’ (‘Düsseldorfer Erklärung der CDU’).<sup>14</sup> This is in fact the correct title of the election manifesto adopted in 1965 at the CDU national convention in Düsseldorf. However, the CMP dataset coded not only the short Düsseldorf Declaration as the Christian Democratic election manifesto, but also comments made by a number of party officials on the content of the ‘Düsseldorf Declaration’ that were added to the declaration in the brochure published after the party convention. For this reason, Proksch and Slapin decided to exclude the 1965 federal election from their analysis in their contribution in this volume. While the Düsseldorf Declaration contained only 599 words, the document coded by the Comparative Manifesto Project was 16,011 words long because it also included extra comments of the party officials. It is quite possible, then, that the 1965 position of the CDU/CSU<sup>15</sup> estimated by the CMP might be biased because of the inclusion of speeches and comments on the real manifesto. To show that the use of the ‘extended’ version of the CDU manifesto in 1965 results in biased (or at least very different) estimates, we apply Wordscores to estimate the economic policy position of German parties for the federal election in 1965. Reference texts are the election manifestos of CDU/CSU, Social Democratic Party (SPD) and FDP in 1990, scored with data from the Laver and Hunt expert survey. As Table 1 indicates, when using the long version of the CDU election manifesto of 1965, i.e. the one that includes comments of CDU politicians on the programmatic document, one would conclude that the Christian Democrats

TABLE 1  
ECONOMIC POLICY POSITIONS OF MAJOR GERMAN PARTIES IN 1965 WITH TWO  
DIFFERENT VERSIONS OF THE CDU/CSU ELECTION MANIFESTO

	Score	95% Confidence interval	Total words scored	Share of scored words (%)
CDU/CSU (long)	7.7	[6.5; 9.0]	13,161	82.2
CDU/CSU (short)	13.8	[8.4; 19.1]	559	89.4
SPD	9.2	[8.1; 10.3]	17,140	73.4
FDP	18.3	[16.2; 20.4]	4,964	72.0

*Source:* Own calculations; reference texts are the federal election manifestos of CDU/CSU, SPD and FDP in 1990; reference scores are the Laver and Hunt estimates of German parties on the 'increase taxes vs. cut spending' dimension.

adopted an economic policy position similar to that of the SPD, which would be a remarkable result, in particular when taking into account that the parties mentioned last formed a coalition government one year after the election. However, when using only the 599-word 'Düsseldorf declaration' as the CDU election manifesto without the speeches, the position of the Christian Democrats is clearly more uncertain and ranges between 8.4 and 19.1, so that no one could conclude that the CDU adopted an economic policy position close to the Social Democrats. Thus, the CMP dataset not only has to be handled with care because of missing information on measurement error,<sup>16</sup> but also because of the inclusion of partly or fully misclassified documents.

A further example comes from the estimation of policy positions of German state parties from 1990 until 2008. In a study on government formation in the German states, Bräuninger and Debus used the Wordscores approach to estimate the programmatic heterogeneity of party coalitions on an economic and a social policy dimension.<sup>17</sup> Application of the Wordscores technique, described in the contributions by Müller and by Bernauer and Bräuninger in this issue, requires the analyst to carefully select reference texts and to identify reference scores to anchor these reference texts. Bräuninger and Debus used the election manifestos of the federal parties in 1990 and 2002 as their reference texts and scored them with positions taken from the expert surveys by Laver and Hunt (for the 1990 election manifestos)<sup>18</sup> and by Benoit and Laver (for the 2002 manifestos).<sup>19</sup> In 1990, the first election after unification, West German and East German Greens ran separate lists and, as they had postponed their unification until after the election, separate electoral platforms. The Greens in East Germany, competing under the name 'Alliance 90/Greens' ('Bündnis 90/Grüne'), won 6.2 per cent of the vote in five Eastern Länder and thereby captured eight seats in the Bundestag.<sup>20</sup> The West German Green Party, however, won only 4.8 per cent of the vote—short of the 5 per cent legal minimum required to win seats – and thus failed to gain any parliamentary representation.<sup>21</sup> Because the 1990 election manifesto of the West German Greens was not available at the time of their study, Bräuninger and Debus used the manifesto of the East German Greens, which was much shorter than the manifesto of the Western section of the party. Using the East German Green Party manifesto along with reference scores from the Laver and Hunt study certainly is admittedly a second-best solution, as the party simply did not exist when Laver and Hunt gathered their data in 1989.

Furthermore, the East German Greens were seen as clearly more moderate compared to the Greens in the West, which then were dominated by the fundamentalist left wing of the party.<sup>22</sup> By examining the election manifesto of the West German Greens (available on the website of the party foundation<sup>23</sup>), we can see what difference it makes. Table 2 shows the results of the estimated economic policy positions of the parties in the state of Hessen. We selected this state because its elections were held a few months after the federal elections of 1990 and 2002. As the results reveal, the only clear differences that emerge in the case of the Green party lie in its estimated economic policy position. When choosing the election manifesto of the West German Greens as a reference text, the economic policy positions of the Green party in Hessen in 1991 and in 2003 significantly shifts to the right if compared with the estimated position when using the election manifesto of the East German Greens as a reference text. In addition, the economic policy position of the 'red-green' coalition government that formed after the 1991 state election has a significantly more liberal position on the economic left/right axis when using the 1990 manifesto of the Eastern German Greens as one reference text. As this example clearly demonstrates, even when two parties appear to be the same, differences in the texts used as the basis for scaling may yield substantially different results, further underscoring the importance of carefully selecting texts for analysis.

One final example shows the importance of using 'correct' versions of election manifestos when estimating their policy positions. In their foundational article on the Wordscores method, LBG use German federal election manifestos from 1990 and 1994 to demonstrate the applicability of their fully language-blind method.<sup>24</sup> The character-encoded data file of the election manifestos they used was provided by the CMP project. This file, however, had been altered to make it more useable for the type of hand-coded content analysis used by the CMP project. Not only did

TABLE 2  
ESTIMATING POLICY POSITIONS HESSEN STATE PARTIES IN 1991 AND 2003 ON AN  
ECONOMIC POLICY DIMENSION WITH ONE DIFFERENT REFERENCE TEXT

	1991		2003	
	1990 Election Manifesto of East German Greens used as reference text	1990 Election Manifesto of West German Greens used as reference text	1990 Election Manifesto of East German Greens used as reference text	1990 Election Manifesto of West German Greens used as reference text
CDU	15.2 [14.4; 16.0]	13.9 [13.1; 14.7]	16.5 [16.1; 16.9]	16.4 [16.0; 16.7]
SPD	13.0 [12.5; 13.5]	12.0 [11.5; 12.5]	13.4 [13.0; 13.7]	13.4 [13.1; 13.8]
FDP	17.2 [16.7; 17.8]	16.8 [16.3; 17.4]	19.1 [18.8; 19.5]	19.0 [18.6; 19.3]
Bündnis 90/Greens	<b>12.1 [11.8; 12.4]</b>	<b>10.1 [9.7; 10.4]</b>	<b>13.3 [12.9; 13.7]</b>	<b>12.8 [12.4; 13.1]</b>
Coalition agreement	<b>14.8 [14.4; 15.2]</b>	<b>13.9 [13.5; 14.3]</b>	16.1 [15.7; 16.5]	16.2 [15.8; 16.6]

Note: Estimates that are significantly different are shown in bold text.

Sources: T. Bräuninger and M. Debus, 'Der Einfluss von Koalitionsaussagen, programmatischen Standpunkten und der Bundespolitik auf die Regierungsbildung in den deutschen Ländern', *Politische Vierteljahresschrift* 49/2 (2008), pp.317–19; reference texts are the federal election manifestos of CDU/CSU, SPD, FDP and East (West) German Greens in 1990, and of CDU/CSU, SPD, FDP, Bündnis 90/Greens and PDS in 2002.

all of the data files include running line numbers, but they also included duplicated text – (quasi-)sentences that were duplicated as many times as they were classified to one of the 56 thematic categories of the CMP coding scheme. Thus, some sentences and therefore words have an incorrectly high frequency, which should bias methods like Wordscores or Wordfish<sup>25</sup> that rely on relative word frequencies for generating results.

Applying the Wordscores method to the CMP versions of German political texts from 1990 and 1994 and the original versions of these documents results in clear differences of the estimated policy positions as in the example before (see Table 3). The results presented in Table 3 are replications of the ones presented in the Laver, Benoit and Garry article. The first column provides the estimated position of German parties on the economic policy position in 1994 based on the uncorrected reference texts from 1990, while we used the ‘cleaned’ manifestos as reference texts to get the policy positions reported in the second column. Although the differences between the two estimations are generally not drastic – or statistically distinguishable when taking into account the standard errors – the two extreme parties’ positions are both significantly changed when the cleaned texts are used. The FDP position shifts to the centre by three points, and the PDS estimate appears as more centrist when using the erroneous texts: the economic position of the FDP is underestimated by nearly the three points (from the 1–20 point scale) and the Communist PDS is estimated as being more than two points further to the extreme left. While some of these differences arise from the rescaling method that LBG used to place scaled texts on the same metric as those of the reference texts, these differences nonetheless illustrate the importance of avoiding junk text that clearly has no substantive political information that would yield leverage on estimating party policy positions.<sup>26</sup>

Small differences in the estimated policy positions may be substantial when policy positions are used as explanatory variables, as they typically are.<sup>27</sup> When applying De Swaan’s theory of minimal ideological range coalitions<sup>28</sup> to the above results, then a coalition between CDU/CSU and SPD is expected for 1994 when using party policy positions presented in the first column of Table 3 (and assuming that coalition formation in Germany is determined only by economic policy positions). The programmatic distance between the two largest German parties reaches a score of 3.0 according to the results provided by LBG,<sup>29</sup> while the programmatic heterogeneity reaches 3.5 between the Christian parties and the Liberals. Using the estimates based on the

TABLE 3  
ESTIMATING THE ECONOMIC POLICY POSITIONS OF GERMAN FEDERAL PARTIES IN 1994  
WITH CORRECTED REFERENCE TEXTS FROM 1990

	Reference texts used by LBG	Corrected reference texts
CDU/CSU	13.7 [12.9; 14.5]	13.2 [12.6; 13.8]
SPD	10.7 [9.9; 11.4]	10.7 [10.1; 11.2]
FDP	17.2 [16.8; 17.66]	14.1 [13.7; 14.4]
Bündnis 90/Greens	7.5 [7.0; 8.0]	7.1 [6.7; 7.5]
PDS	4.0 [3.0; 5.0]	1.3 [0.6; 2.0]

Source: M. Laver, K. Benoit and J. Garry, ‘Extracting Policy Positions from Political Texts Using Words as Data’, *American Political Science Review* 97/2 (2003), pp.311–31; Comments: reference texts are the federal election manifestos of CDU/CSU, SPD, FDP and the Eastern German Bündnis 90/Grüne in 1990.

'corrected' texts, however, the policy distance between Christian Democrats and the FDP is clearly smaller (0.9 units) than the distance between CDU/CSU and Social Democrats (2.5 units). Eventually, it was a Christian Democrat/FDP government that was formed under Chancellor Helmut Kohl. The example also alerts us to the fact that even language-blind methods may require language-fluent human analysts to detect such sources of error.

#### THE OPEN SOURCE ARCHIVE POLIDOC.NET

The above examples demonstrate how important it is to put great effort into carefully selecting political texts, carefully considering how they were generated and for what purpose, as well as paying careful attention to more quotidian matters such as detecting errors from text preparation, transcription, and encoding. These are all vital facets of efforts to produce replicable, valid, and comparable estimates of the 'true' preferences of political actors when using political texts as a source of information. While most of our focus here has been on the special version of these issues applied to computational, automated text analysis, they apply quite generally to any method of text analysis.

A key issue for text analysts concerns the availability of political texts for both original analysis and replication. Although the methods discussed by the contributors to this collection can be (and actually are) applied to different types of political texts, party manifestos are the class of political text studied the most extensively in political science. Manifestos are normally written a few months before each election and cover the policy positions of the respective party on all relevant issues. Parties use these documents to announce their policy positions to voters, as well as to mark their starting positions for the process of government formation following an election. Coalition negotiations often result in another important policy document: the coalition policy agreement. In that document, the parties involved in the government formation process formulate their policy goals for the forthcoming legislative period.<sup>30</sup> A large number of studies in political science have focused on these topics and analysed the congruence between election manifestos, government declarations and coalition agreements.<sup>31</sup> Despite being the most commonly studied kind of political text, however, it is still extremely difficult in practice simply to obtain original manifesto texts, either those used in widely cited articles, books, and datasets or yet unanalysed manifestos that might contain new and interesting information about party policy. Furthermore, once the hurdle of acquiring texts has been overcome, would-be analysts may also face the additional time-consuming tasks of scanning, encoding, cleaning, and pre-processing texts before they are ready for analysis. As the example above illustrates with the widely cited LBG using texts from the CMP, errors inherited from different levels of this text processing procedure may be quite subtle and difficult to detect.

To minimise these problems, we have established an electronic archive of political texts, *polidoc.net*, which maintains digital versions of election manifestos, coalition agreements, government declarations and various other documents of political actors from developed democracies. The aim of the repository is to provide political texts to facilitate scholarly research in this area but also to encourage researchers to share their documents with others. We invite everybody to participate in this project so that the political science community can easily make use of documents already

collected. The archive has a website available at <http://www.polidoc.net>, and is maintained by the authors. All those political documents that we have used in our previous studies are online, so that our findings can be replicated. Other manifestos and speeches of politicians were collected originally by researchers named in the codebook available on the webpage. At the time of writing, there are more than 800 political documents from 12 Western European and North American countries available. In the case of Germany, for instance, we also provide around 400 election manifestos, coalition agreements and government programmes from the German Länder.<sup>32</sup> All documents are provided as ASCII data files using a fixed set of pre-processing rules (e.g. page numbers removed) and therefore easy to import into word processing or text analytical software. After receiving the password, users can download the documents from the online archive, as required, by country and election year. A description and codebook provides more detailed information about the sources of the respective documents. This handbook will also be made available at the homepage of the online archive. As some students of party politics are interested in analysing political documents in a more qualitative manner, e.g. print style of the document, whether pictures are used and if so what message they intend to send, and so on, we also intend to provide documents in photograph-style format (e.g., in pdf file format).

#### STEPS FOR FUTURE RESEARCH

One obvious next step in the development of the open access archive is to increase the country sample and to widen the time period the respective programmes reflect. This would allow for more intensive empirical tests of theoretical models of party system change, party competition, coalition politics and legislative decision-making in a comparative perspective. Changes in the programmatic profile of parties from Eastern European democracies, for instance, have not received great attention in the scholarly literature. While the expert surveys by Benoit and Laver as well as by Marks and Hooghe cover Eastern Europe,<sup>33</sup> the limited time points at which these surveys collected data make it difficult to analyse the movement of party policy positions over time. The recent version of the CMP dataset<sup>34</sup> includes the relevant party groups from the Eastern European democracies, but still codes – with the exception of the bipolar categories such as ‘Military: positive’ and ‘Military: negative’ – saliencies rather than positions that were mentioned in the election manifestos. A further incentive to continue the agenda of improving methods for extracting policy preferences from political text is the promise of unlocking information from historical texts, since unlike experts, texts never become forgetful, tired or just plain fed up at being asked to complete repeated questionnaires. Using roll call votes, research on the United States Congress<sup>35</sup> and the British Parliament<sup>36</sup> has yielded very important insights into the development of (intra-)party competition and parliamentary decision-making. In the case of Germany, studies of the voting behaviour in previous parliaments has also shown promising results when using data from parliamentary archives: recent studies focus on the German national assembly during the failed revolution in 1848 and 1849,<sup>37</sup> on the Reichstag of the Weimar Republic between 1920 and 1932<sup>38</sup> and of the Bundesrat of the Federal Republic of Germany since 1949.<sup>39</sup> The last-mentioned studies provide promising results on the dimensionality of parliaments

and the position of each single actor (an MP or a state) on the extracted dimensions. In systems with party discipline, however, MPs may not always behave sincerely during roll call votes;<sup>40</sup> in addition, roll call votes are not always representative of overall legislative activity where political preferences might play a role. Because it is widely thought that what is said (or written) may be more sincere and less subject to party discipline than voting, the analysis of historical texts such as party documents, election manifestos, legislative debates, and party convention speeches offer a rich source of historical information on party preferences for which few other methods hold any promise of being able to extract useful or indeed any, information.<sup>41</sup> Finally, analysing the speeches of MPs or the documents of intra-party groups will help to relax the often criticised unitary actor assumption and to analyse intraparty heterogeneity more adequately.<sup>42</sup> Related to this objective, a more detailed analysis of multi-level systems requires information on the preferences of political actors on the respective sub-national units, in particular in those states where political actors on the sub-national level have strong competencies.<sup>43</sup> A number of studies analyse patterns of party competition at the regional level in one state, but none to date has established a comparative dataset on political preferences of different regional parties.<sup>44</sup>

In this collection we have demonstrated various approaches to estimating political preferences from texts, and have outlined issues both theoretical and practical that will have to be addressed if future progress is to be made in applied, computational text analysis. Apart from methodological issues, furthermore, one of the most significant practical challenges that text analysts face lies simply in obtaining access to texts in a format that are ready for use. In order to reduce the barriers posed by this practical challenge, we have established the freely accessible archive *polidoc.net* to archive and disseminate textual data, and to establish a network for exchanging and accessing data that has been checked for errors and assured for quality. We hope that the effort will be supported and make it more likely that future generations of researchers, students, and analysts of all stripes will be able to focus more on producing valid and reliable findings from empirical studies of textual data rather than struggling unnecessarily to obtain and process texts.

#### ACKNOWLEDGEMENTS

We thank all contributors to this volume for their comments on an earlier version of this paper.

#### NOTES

1. Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara and Eric Tanenbaum (eds.), *Mapping Policy Preferences. Estimates for Parties, Electors and Governments 1945–1998* (Oxford: Oxford University Press: 2001); Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Ian Budge and Michael McDonald (eds.), *Mapping Policy Preferences II: Estimates for Parties, Electors and Governments in Eastern Europe, European Union and OECD 1990–2003* (Oxford: Oxford University Press, 2006).
2. K. Benoit and M. Laver, *Party Policy in Modern Democracies* (London: Routledge, 2006), pp.64–5.
3. A. Wüst and A. Volkens, *Euromanifesto Coding Instructions* (Mannheim: Mannheimer Zentrum für Europäische Sozialforschung Working Paper No. 64, 2003).

4. See A. Downs, *An Economic Theory of Democracy* (New York: Harper, 1957); S. Rabinowitz and E. Macdonald, 'A Directional Theory of Issue Voting', *American Political Science Review* 83/1 (1989), pp.93–121.
5. M. Laver, K. Benoit and J. Garry, 'Extracting Policy Positions from Political Texts Using Words as Data', *American Political Science Review* 97/2 (2003), pp.311–31.
6. J. Slapin and S.-O. Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts', *American Journal of Political Science* 52/3 (2008), pp.705–22.
7. M. McDonald and Ian Budge, *Elections, Parties, Democracy: Conferring the Median Mandate* (Oxford: Oxford University Press, 2005).
8. J. Adams, M. Clark, L. Ezrow and G. Glasgow, 'Understanding Change and Stability in Party Ideologies: Do Parties Respond to Public Opinion or to Past Election Results?', *British Journal of Political Science* 34/4 (2004), pp.589–610.
9. T. Bräuninger, 'Responsivität und strategische Adaption im Parteienwettbewerb in den deutschen Bundesländern', in C. Henning, E. Linhart and S. Shikano (eds), *Parteienwettbewerb, Wählerverhalten und Koalitionsbildung* (Baden-Baden: Nomos, 2009), pp.27–46.
10. M. Laver and W. Hunt, *Policy and Party Competition* (New York: Routledge, 1992); K. Benoit and M. Laver, *Party Policy in Modern Democracies* (London: Routledge, 2006); P. Warwick, *Policy Horizons and Parliamentary Government* (Houndmills: Palgrave Macmillan, 2006).
11. Laver *et al.*, 'Extracting Policy Positions from Political Texts Using Words as Data', p. 314.
12. *Ibid.*
13. It has also been shown that Wordscores produces extremely similar results to Proksch and Slapin's Wordfish procedure when the Wordfish extremes are selected as reference texts. See Sven-Oliver Proksch and Jonathan B. Slapin, 'Position-Taking in European Parliament Speeches', *British Journal of Political Science* 48/1 (2009), pp.58–79.
14. Budge *et al.*, *Mapping Policy Preferences*; Klingemann *et al.*, *Mapping Policy Preferences II*. We would like to thank Andrea Volkens for detailed information on the coding of the 1965 German election manifestos.
15. The Christian Social Union (CSU) is the CDU's sister party in Bavaria.
16. See K. Benoit, M. Laver and S. Mikhaylov, 'Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions', *American Journal of Political Science* 53/2 (2009), pp.495–513; S. Mikhaylov, M. Laver and K. Benoit, 'Coder Reliability and Misclassification in Comparative Manifesto Project Codings', Paper presented at the 66th MPSA Annual National Conference, Palmer House Hilton Hotel and Towers, 3–6 April 2008.
17. T. Bräuninger and M. Debus, 'Der Einfluss von Koalitionsaussagen, programmatischen Standpunkten und der Bundespolitik auf die Regierungsbildung in den deutschen Ländern', *Politische Vierteljahresschrift* 49/2 (2008), pp.317–19.
18. Laver and Hunt, *Policy and Party Competition*.
19. Benoit and Laver, *Party Policy in Modern Democracies*.
20. O. Niedermayer, 'Die Entwicklung des deutschen Parteiensystems bis nach der Bundestagswahl 2002', in O. Niedermayer (ed.), *Die Parteien nach der Bundestagswahl 2002* (Opladen: Leske und Budrich, 2003), p.16.
21. For the first Bundestag election after unification, West and East Germany were considered as two electoral areas. To receive parliamentary representation, a party had to pass the 5 per cent threshold or win three seats directly in any region.
22. See, e.g., M. Klein and J.W. Falter, *Der lange Weg der Grünen* (München: Beck, 2003), pp.46–47.
23. See the pdf files on the web page of the Heinrich-Boell-Foundation at [http://www.boell.de/alt/de/13\\_archiv/2297.html](http://www.boell.de/alt/de/13_archiv/2297.html).
24. Laver *et al.*, 'Extracting Policy Positions from Political Texts Using Words as Data'.
25. See *ibid.*; Slapin and Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts'.
26. The effect is caused by the LBG approach to rescaling the estimated party scores using a normalisation procedure that disproportionately affects parties at the extremes of the scale. This rescaling procedure is designed to mitigate the shrinking effect caused by centrist scores assigned to words found in all reference texts – in this case, junk text such as line numbers or repeated phrases found in common across texts. For a fuller discussion of the rescaling of 'virgin' text scores, see K. Benoit and M. Laver, 'Compared to What? A Comment on "A Robust Transformation Procedure for Interpreting Political Text"' by Martin and Vanberg', *Political Analysis* 16/1 (2008), pp.101–11.
27. Benoit *et al.*, 'Treating Words as Data with Error'.
28. A. de Swaan, *Coalition Theories and Cabinet Formation* (Amsterdam: Elsevier, 1973); see M. Laver and N. Schofield, *Multiparty Government. The Politics of Coalition in Europe* (Ann Arbor: The

- Michigan University Press, 1998), pp.97–103; L. Martin and R. Stevenson, 'Government Formation in Parliamentary Democracies', *American Journal of Political Science* 45/2 (2001), pp.33–50.
29. Laver *et al.*, 'Extracting Policy Positions from Political Texts Using Words as Data', p.326.
  30. See Wolfgang C. Müller and Kaare Strøm, 'Coalition Agreements and Cabinet Governance', in Kaare Strøm, Wolfgang C. Müller and Torbjörn Bergman, *Cabinets and Coalition Bargaining: The Democratic Life Cycle in Western Europe* (Oxford: Oxford University Press, 2008), pp.159–99.
  31. I. Budge and M. Laver, 'The Policy Basis of Government Coalitions: A Comparative Investigation', *British Journal of Political Science* 23/4 (1993), pp.499–519; A. Timmermans, 'Conflicts, Agreements, and Coalition Governance', *Acta Politica* 33/4 (1998), pp.409–32; A. Timmermans, 'Standing Apart and Sitting Together: Enforcing Coalition Agreements in Multiparty Systems', *European Journal of Political Research* 45/2 (2006), pp.263–83; P. Warwick, 'Coalition Policy in Parliamentary Democracies. Who Gets How Much and Why', *Comparative Political Studies* 34/10 (2001), pp.1212–36; M.D. McDonald and I. Budge, *Elections, Parties, Democracy: Conferring the Median Mandate* (Oxford: Oxford University Press, 2005); M. Debus, 'Office and Policy Payoffs in Coalition Governments', *Party Politics* 14/5 (2008), pp.515–38.
  32. Some of these documents were made accessible by the Central Archive in Cologne (ZA), were the full text of the manifestos analysed by the Comparative Manifestos Project is available. Other manifestos and speeches of politicians have been collected originally by researchers named on the codebook available at the webpage. A large number of people helped us in collecting and editing those documents. We received documents from and are grateful to GESIS-ZA Cologne, Wolfgang Güttel, Marcelo Jenny, Gail McElroy, Paul Mitchell, Arco Timmermans, and Andrea Volkens. Julian Bernauer, Martin Brunner, Thomas Däubler, Christina Eder, Gunnar Flämig, Andreas Grosser, Katrin Kirschmann, Laszlo Kovats, Jochen Müller and Christine Spannagel helped us editing the documents and maintaining the website.
  33. Benoit and Laver, *Party Policy in Modern Democracies*; L. Hooghe, G. Marks and C.J. Wilson, 'Does Left/Right Structure Party Positions on European Integration?', *Comparative Political Studies* 35/8 (2006), pp.965–89.
  34. Klingemann *et al.*, *Mapping Policy Preferences II*.
  35. K.T. Poole and H. Rosenthal, 'A Spatial Model for Legislative Roll Call Analysis', *American Journal of Political Science* 29/2 (1985), pp.357–84; K.T. Poole and H. Rosenthal, *Congress: A Political-Economic History of Roll Call Voting* (New York/Oxford: Oxford University Press, 1997); J. Clinton, S. Jackman and D. Rivers, 'The Statistical Analysis of Roll Call Data', *American Political Science Review* 98/2 (2004), pp.355–70.
  36. C. Schonhardt-Bailey, 'Ideology, Party and Interests in the British Parliament of 1841–1847', *British Journal of Political Science* 33/2 (2004), pp.581–605.
  37. H. Best and E.J. Zimmermann, 'Dimensionen politischer Konflikte: Die Analyse von namentlichen Abstimmungen in Parlamenten mit dem Verfahren der Mokken-Skalierung', in H. Best and H. Thome (eds.), *Neue Methoden der Analyse historischer Daten* (St. Katharinen: Scripta Mercaturae, 1991), pp.41–79.
  38. M. Debus and M.E. Hansen, 'Die Dimensionalität der Reichstage in der Weimarer Republik von 1920 bis 1932', Mannheim Centre for European Social Research, University of Vienna: Unpublished Manuscript.
  39. S. Shikano, 'The Dimensionality of German Federal States' Policy Preferences in the Bundesrat', *German Politics* 17/3 (2008), pp.340–52.
  40. Clinton *et al.*, 'The Statistical Analysis of Roll Call Data'.
  41. M. Laver and K. Benoit, 'Locating TDs in Policy Spaces: Wordscoring Dáil Speeches', *Irish Political Studies* 17/1 (2002), pp.59–73; D. Giannetti and M. Laver, 'Policy Positions and Jobs in the Government', *European Journal of Political Research* 44/1 (2005), pp.91–120.
  42. Daniela Giannetti and Michael Laver, 'Intra-Party Politics and Coalition Governments in Parliamentary Democracies', in Daniela Giannetti and Kenneth Benoit (eds.), *Intra-Party Politics and Coalition Governments* (London: Routledge, 2008), pp.3–24; Marc Debus and Thomas Bräuninger, 'Intraparty Fractions and Coalition Bargaining in Germany', in Daniela Giannetti and Kenneth Benoit (eds.), *Intra-Party Politics and Coalition Governments* (London: Routledge, 2008), pp.124–45.
  43. See G. Marks, L. Hooghe and A.H. Schakel, 'Patterns of Regional Authority', *Regional and Federal Studies* 18/2–3 (2008), pp.167–81.
  44. Bräuninger and Debus, 'Der Einfluss von Koalitionsaussagen, programmatischen Standpunkten und der Bundespolitik auf die Regierungsbildung in den deutschen Ländern'; L. Libbrecht, B. Maddens, W. Swenden and E. Fabre, 'Issue Salience in Regional Party Manifestos in Spain', *European Journal of Political Research* 48/1 (2009), pp.58–79; I. Stefuriuc, 'Government Formation in Multi-Level Settings: Spanish Regional Coalitions and the Quest for Vertical Congruence', *Party Politics* 15/1 (2009), pp.93–115.