DORTMUND CENTER
FOR DATA-BASED
MEDIA ANALYSIS

DoCMA Working Paper #8

September 2021

Text mining methods for measuring the coherence of party manifestos for the German federal elections from 1990 to 2021

Carsten Jentsch, Enno Mammen, Henrik Müller, Jonas Rieger and Christof Schötz*

Cite as:
Jentsch, C., Mammen, E., Müller, H., Rieger, J. & Schötz, C. (2021). "Text mining methods for measuring the coherence of party manifestos for the German federal elections from 1990 to 2021." *DoCMA Working Paper #8, Sept. 2021.* DOI: 10.17877/de290r-22363.

Version 1.0, September 2021

*Carsten Jentsch is a professor for business and social statistics at TU Dortmund University, Henrik Müller for economic policy journalism respectively. They are members of the Dortmund Center for data-based Media Analysis (DoCMA). Enno Mammen is a professor for mathematical statistics and Christof Schötz is a postdoctoral researcher at University Heidelberg. Jonas Rieger is a researcher at TU Dortmund University and DoCMA. The authors would like to thank Eun Ryung Lee (Sungkyunkwan University, Korea) for helpful comments.

Abstract

Text mining is an active field of statistical research. In this paper we use two methods from text mining: the Poisson Reduced Rank Model (PRR, see Jentsch et al. 2020; Jentsch et al. 2021) and the Latent Dirichlet Allocation model (LDA, see Blei et al. 2003) for the statistical analysis of party manifesto texts from Germany. For the nine federal elections in Germany from 1990 to 2021, we analyze party manifestos that have been written by the parties to present their political positions and goals for the next legislative period of the German federal parliament (Bundestag). We use the models to quantify distances in the language of the manifestos and in the weight of importance the parties attribute to several political topics. The statistical analysis is purely data driven. No outside information, e.g., on the position of the parties, on the meaning of words, or on currently hot political topics, is used in fitting the statistical models. Outside information is only used when we interpret the statistical results.

Key words: Poisson reduced-rank model, Latent Dirichlet Allocation, CDU, CSU, Union, SPD, Grüne, FDP, Linke, Kenia, Jamaica, Ampel, Deutschland, R2G, coalition

1 Introduction

For analysts of political developments, the run up to the federal elections in Germany on 26 September 2021 proved an unwieldy field of research. Never before in post-war history had so many different outcomes been possible. Polls indicated that none of the traditional two-party coalitions (center-right or center-left) would be able to gain a majority sufficient to form a stable national government. As fragmentation and polarization had transformed the party system, a three-way coalition would be needed to form a stable government, an experiment without precedence at the German federal level. Among the many arithmetically possible coalitions five could be considered as prima facie workable: "Traffic Lights" (SPD, Greens, FDP), "Jamaica" (Union[1], FDP, Greens), "Red-Green-Red", or R2G (SPD, Greens, Linke), "Kenya" (Union, SPD, Greens) and "Deutschland" (Union, SPD, FDP). How well would these five hypothetical pacts fit together? Which one would stand the best chance to actually form a government? While the sympathy and antipathy of the parties' leading figures captured public attention, the real obstacles could be assumed to stem from programmatic incompatibilities and divergences. To be able to offer systematically derived answers, we have turned to a text mining approach.

In this paper we use two statistical approaches to analyze party manifestos of German parties over the period 1990 to 2021. In the Poisson Reduced Rank Model (PRR, see Jentsch et al. 2020; Jentsch et al. 2021) one assumes that there exists for each party a latent K-dimensional vector that develops in time. At each position of the text of a manifesto each word stem[2] appears with a probability that depends only on the latent position of the party at the date when the party manifesto was written and on unknown parameters. The parameters and the positions of the parties can be estimated by statistical methods. We interpret vicinity of positions of two manifestos as a similar language spoken in the two manifestos. For the choice $K = 1$ the position is a one-dimensional value. In this case one can argue that the positions are related to the location of the party on a political left/right scale: The estimated positions of the German parties appear in the order FDP (Free Democratic Party), CDU/CSU (Christian Democratic Party/ Christian Social Union), SPD (Social Democratic Party), Die Grünen (Green Party) and PDS/Die Linke (Party of Democratic Socialism/ Left Party). This is exactly the generally accepted ordering of the five parties on a political left/right scale, at least with exception of the ordering of FDP and CDU/CSU.

It is important to note that no information on the meaning of words is used in modeling PRR and in the estimation of the parameters and positions. Furthermore, positions of the parties inferred in the statistical analysis develop smoothly in time. Again, this result emerges without using that manifestos of two neighbored elections were written by the same party and without using any information on the meaning of words. The whole statistical analysis lets the data speak for themselves without any outside input.

In Latent Dirichlet Allocation models (LDA, see Blei et al. 2003) it is assumed that in

[1] We use the term „Union" to denote the national voting pact of the Christian Democratic Union (CDU) and their Bavarian sister party Christian Social Union (CSU).

[2] A word stem is one part of a word and is used to concentrate the information of a consistent lexical meaning of multiple variants of the same word, e.g. inflected variants, in one kind of expression of the word.

each manifesto a number of latent topics is addressed. A topic is given by a vector of probabilities. By this vector for each word stem a probability is given with which the word stem appears in a location of the text that belongs to this topic. Each manifesto consists of a mixture of topics. The LDA model is defined by the probabilities of word stems in the different topics and by the relative proportions of each topic in the party manifestos. Again, these values are estimated in the statistical analysis without making use of the meaning of the words. For an interpretation of the results one can look for each topic at the word stems and the text samples with the highest probability. We will see that such estimated topics make sense and can be used as a basis for a political interpretation of the data. For each topic the words with the highest estimated probabilities are related and identify a political theme like foreign policy, regional cohesion or the transition to carbon-neutral power systems. Clearly, at this point of the analysis we make use of the meaning of the word stems and attribute a political meaning to each topic generated by LDA. The relative proportion of a topic in the text body of a manifesto can be interpreted as a weight that a party assigns to this topic. If in two party manifestos the proportions of a topic are similar, this indicates that the parties attribute a comparable amount of importance to this topic. But it does not mean that they must have a similar opinion on the topic. For example, parties might agree that a certain topic (e.g. climate change) is important and has to be addressed, but they propose different solutions to tackle related challenges in their manifestos. In this sense the statistical analysis based on LDA is complementary to PRR. In PRR one analyses if two parties speak a similar language. This also cannot directly be interpreted as a similarity in opinions but it covers a different aspect than LDA. We will come back to these points when we interpret our results.

In the next section we will give a complete technical description of PRR and LDA. Section 3 contains a comprehensive summary of our statistical inference on the party manifesto data. Also a detailed description of the preprocessing steps applied to the data are given. The results of our data analysis are presented and discussed in Section 4 and Section 5 concludes.

2 Methods

To analyze the positioning and thematic compositions of party manifestos over time, we use two complementary methods. One is a Poisson reduced-rank (PRR) model to model party positions and the other is Latent Dirichlet Allocation (LDA) for more in-depth insights into the topics addressed in the manifestos.

2.1 Poisson Reduced-rank Models

Background Following Jentsch et al. 2021, we consider a Poisson model with reduced-rank, in which for individuals $\ell = 1, ..., L$, one observes Poisson random variables $Y_{j\ell}$ $(j = 1, ..., J)$ that follow a Poisson distribution with mean parameter $\mu_{j\ell}$. The matrix of mean parameters is assumed to have a reduced rank after transformation, yielding the

model

$$Y_{j\ell} \sim \text{Poisson}(\mu_{j\ell}), \quad \mu_{j\ell} = \exp\left(\alpha_j + \beta_\ell + \sum_{k=1}^{K} b_j^{(k)} f_\ell^{(k)}\right). \tag{1}$$

Here, α_j, β_ℓ, $b_j^{(k)}$, and $f_\ell^{(k)}$, as well as the dimension K, are unknown parameters. In our application, the main purpose of statistical analysis is to make inferences about the parameter $\{f_\ell^{(k)}\}_{k=1}^K$, which is interpreted as the location or position of an individual ℓ in a K-dimensional space. We use our model to develop a quantitative approach to estimate party positions over time from political manifesto texts. In this application, $Y_{j\ell}$ denotes the frequency with which a word stem j appears in a political manifesto ℓ, and we interpret $\{f_\ell^{(k)}\}_{k=1}^K$ as the position of the party at the time when manifesto ℓ was written.

Poisson reduced-rank model Model (1) is referred to as a reduced-rank model as the logarithm of the parameter matrix $(\mu_{j\ell})$ has a reduced rank after normalization, i.e., the $J \times L$ matrix with elements $\sum_{k=1}^K b_j^{(k)} f_\ell^{(k)}$ has rank $K < \min(L, J)$. There are infinitely many equivalent sets of model parameters, $\alpha_j, \beta_\ell, b_j^{(k)}$, and $f_\ell^{(k)}$ ($j = 1, \ldots, J;\ \ell = 1, \ldots, L;\ k = 1, \ldots, K$), resulting in the same Poisson parameters $\mu_{j\ell}$ of (1). To identify the model parameters, we consider the following conditions:

$$\sum_{\ell=1}^{L} \beta_\ell = 0, \quad \sum_{j=1}^{J} b_j^{(k)} = 0, \quad \sum_{\ell=1}^{L} f_\ell^{(k)} = 0 \quad (k = 1, \ldots, K), \tag{2}$$

$$\frac{1}{J}\sum_{j=1}^{J} b_j^{(k)} b_j^{(k')} = \begin{cases} \lambda_k, & k = k' \\ 0, & k \neq k' \end{cases}, \ \frac{1}{L}\sum_{\ell=1}^{L} f_\ell^{(k)} f_\ell^{(k')} = \begin{cases} 1, & k = k' \\ 0, & k \neq k' \end{cases} (k, k' = 1, \ldots, K) \tag{3}$$

with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K \geq 0$. Standard arguments from linear algebra show that (2) and (3) can always be achieved via a reparametrization. This can easily be seen for (2). For (3), consider the singular value decomposition of the matrix $(\log(\mu_{j\ell}) - \alpha_j - \beta_\ell)$. We assume that K is chosen minimal, that is, $\lambda_K > 0$. One can show that if $\lambda_1 > \lambda_2 > \ldots > \lambda_K > 0$ holds then the parametrization that fulfills (2) and (3) is uniquely defined up to the sign of $\{b_j^{(k)}\}_{k=1}^K$ and $\{f_\ell^{(k)}\}_{k=1}^K$. In the application, where $Y_{j\ell}$ counts the frequency of occurrences of the word stems j in party manifesto ℓ, the values of λ_k measure to what extent the kth dimension of the party positions $\{f_\ell^{(k)},\ \ \ell = 1, \ldots, L\}$ affect the distribution of stem counts in the party manifestos.

The identifiability conditions (2) and (3) are motivated by the following considerations. First, the distances between the party position vectors $f_\ell = (f_\ell^{(k)})_{k=1,\ldots,K}$ are invariant under orthogonal transformation, i.e., replacing f_ℓ by Qf_ℓ, where Q is an orthogonal matrix of dimension $K \times K$. Thus, the interpretation of the results in regard to the relation between the party positions is not affected. Furthermore, the conditions (2) and (3) allow for a natural quantitative interpretation of $f_\ell^{(k)}$ and $b_j^{(k)}$: Suppose the positions of two parties f_ℓ and $f_{\ell'}$ differ by $\delta = f_\ell - f_{\ell'}$; let $\delta^{(k)}$ denote the kth element of δ. Then, the expected usage of the word stem j differs by the factor $\rho_j = \sum_{k=1}^K b_j^{(k)} \delta^{(k)}$, and the overall differences can be measured using $\sum_{j=1}^J \rho_j^2$. The first identifiability condition of (3) yields $J^{-1}\sum_{j=1}^J \rho_j^2 = \sum_{k=1}^K (\lambda_k)^2 (\delta^{(k)})^2$, i.e., the weighted distance $\sum_{k=1}^K (\lambda_k)^2 (\delta^{(k)})^2$ between positions can be viewed as a quantitative measure for different word usages. In contrast, the quantity $\sum_{k=1}^K (b_j^{(k)})^2$ measures the different usages of the word stem j among

the positions ℓ as $\sum_{k=1}^{K}(b_j^{(k)})^2 = L^{-1}\sum_{\ell=1}^{L}(\sum_{k=1}^{K} b_j^{(k)} f_\ell^{(k)})^2$.

2.2 Latent Dirichlet Allocation

LDA The classical LDA (Blei et al. 2003) assumes distributions of latent topics for each text. If K denotes the total number of modeled topics, the set of topics is given by $\boldsymbol{T} = \{T_1, \ldots, T_K\}$. We define $W_n^{(m)}$ as a single token (word stem) at position n in text m. The set of possible stem tokens is given by the vocabulary $\boldsymbol{W} = \{W_1, \ldots, W_J\}$ with $J = |\boldsymbol{W}|$, the vocabulary size. Then, let

$$\boldsymbol{D}^{(m)} = \left(W_1^{(m)}, \ldots, W_{N^{(m)}}^{(m)}\right),$$

be text $m = 1, \ldots, M$, of a corpus consisting of M texts. Each text in turn consists of $N^{(m)}$ stem tokens $W_n^{(m)} \in \boldsymbol{W}$, $(n = 1, \ldots, N^{(m)})$. Topics are referred to as $T_n^{(m)} \in \boldsymbol{T}$ for the topic assignment of token $W_n^{(m)}$. Then, analogously the topic assignments of every text m are given by

$$\boldsymbol{T}^{(m)} = \left(T_1^{(m)}, \ldots, T_{N^{(m)}}^{(m)}\right).$$

When $n_k^{(mj)}(k = 1, \ldots, K; j = 1, \ldots, J)$ describes the number of assignments of stem j in text m to topic k, we can define the cumulative count of stem j in topic k over all documents by $n_k^{(\bullet j)}$ and, analogously, the cumulative count of topic k over all stems in document m by $n_k^{(m\bullet)}$, while $n_k^{(\bullet\bullet)}$ indicates the total count of assignments to topic k. Corresponding frequency vectors for a combination of texts can be determined by summing up the frequencies of the individual texts.

Using these definitions, the underlying probability model (Griffiths and Steyvers 2004) can be written as

$$\begin{aligned}
W_n^{(m)} \mid T_n^{(m)}, \boldsymbol{\phi}_k &\sim \text{Discrete}(\boldsymbol{\phi}_k),\\
\boldsymbol{\phi}_k &\sim \text{Dirichlet}(\eta),\\
T_n^{(m)} \mid \boldsymbol{\theta}_m &\sim \text{Discrete}(\boldsymbol{\theta}_m),\\
\boldsymbol{\theta}_m &\sim \text{Dirichlet}(\alpha).
\end{aligned}$$

For a given parameter set $\{K, \alpha, \eta\}$, LDA assigns one of the K topics to each stem token. Here K denotes the number of topics and α, η are parameters of a Dirichlet distribution defining the type of mixture of topics in every text and the type of mixture of words in every topic.

Estimators for topic distributions per text $\boldsymbol{\theta}_m = (\theta_{m,1}, \ldots, \theta_{m,K})^T \in (0,1)^K$ and word distributions per topic $\boldsymbol{\phi}_k = (\phi_{k,1}, \ldots, \phi_{k,J})^T \in (0,1)^J$ can be derived through the Collapsed Gibbs Sampler procedure (Griffiths and Steyvers 2004) by

$$\hat{\theta}_{m,k} = \frac{n_k^{(m\bullet)} + \alpha}{N^{(m)} + K\alpha}, \qquad \hat{\phi}_{k,v} = \frac{n_k^{(\bullet j)} + \eta}{n_k^{(\bullet\bullet)} + J\eta}.$$

LDAPrototype The Gibbs sampler in the modeling procedure of LDA is sensitive to the random initialization of topic assignments. To overcome this issue, we use the selection algorithm LDAPrototype. The method selects the LDA as prototype model of a set of LDAs that maximizes its mean pairwise similarity to all other models (Rieger et al. 2020). Thus, the LDAPrototype method increases the reliability of conclusions drawn from the resulting prototype model. The approach is implemented in the `R` package `ldaPrototype` (Rieger 2020).

3 Application to Party Manifestos

The two methods, LDA and PRR, are applied to the corpus of political manifesto text from nine German federal elections between 1990 and 2021 from six major German parties: Christian Democratic Party/Christian Social Union known for short in German as CDU/CSU or *Union*, Social Democratic Party known as *SPD*, Green Party known as *Grüne*, Party of Democratic Socialism known as PDS/Die *Linke*, Free Democratic Party known as *FDP*, and Alternative for Germany, *AfD* for short. The PDS, having its origins in East Germany, fused with the party WASG, which originated in West Germany, to become Die Linke in 2007. The AfD was founded in and published its first manifesto in 2013. The corpus consists of 3 manifestos of the AfD and 9 manifestos for each party.

Figure 1 shows a basic descriptive statistic: the length of the party manifestos (i.e., number of words before preprocessing) from 1990 to 2021. In the run up to the 2009 federal elections, word counts rose considerably, even more so in 2013. While other parties have trimmed their texts in recent years, the Greens and the Linke still publish rather extensive election programs. Even the right-wing populist AfD, that started out with a very brief manifesto in its founding year 2013 (827 words), now produces sizable papers, the 2021 edition being at par with the Social Democrats. All manifestos except AfD 2013 consist of between 5,900 and 83,000 words.

3.1 Preprocessing

We use the statistical software `R` (R Core Team 2021) with the packages `tidyverse` (Wickham et al. 2019), `tidytext` (Silge and Robinson 2016), and `SnowballC` (Bouchet-Valat 2020) to preprocess and analyse the data. Starting from raw text files extracted either by the Manifesto Project `manifestoproject.wzb.eu` (for all manifestos from the time between 1990 and 2017) or by the authors from the manifestos published by the parties (for all manifestos from 2021), we create the counting vector $(Y_{j\ell})_{j=1,\dots,J}$ for each manifesto $\ell = 1, \dots, L$ as follows: After loading the document, we split the text into words using the function `tidytex::unnest_tokens()`. By setting the arguments `strip_punct`, `strip_numeric`, `to_lower` of this function to `TRUE`, we remove punctuation and numbers, and convert each word to lower case. Then we remove certain common words – so-called stopwords – which likely do not carry much information like articles. For this we use the list provided by `tidytext::get_stopwords('de')`. We also replace umlauts and eszett by their standard transcriptions (ä → ae, ö → oe, ü → ue, ß → ss). Next words are reduced

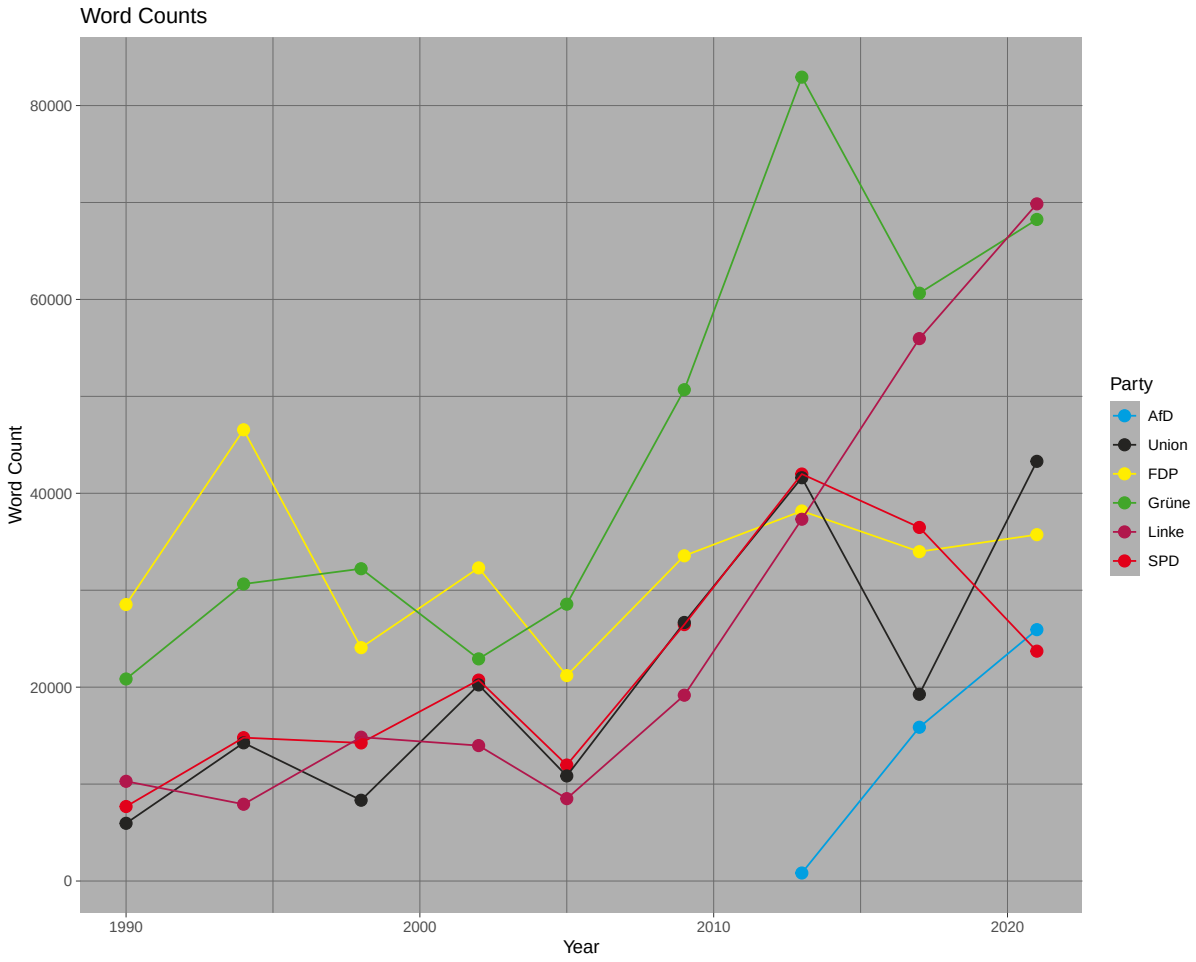

Figure 1: Raw word counts of each party manifesto.

to their stems using the stemmer `SnowballC::wordStem("german")`. We remove some further stems without semantic meaning (`"pos", "xiv", "vgl", "kap", "kapitel"`). Furthermore, we ignore stems that consist of only one symbol or contain non-alphanumeric symbols by applying the regular expression `^[:alpha:]{2,}$` on the stems. Finally, the number of occurrences of each stem is counted. If the accumulated count of a stem across all documents is below 8 the stem is ignored. For achieving this threshold occurrences of stems in AfD-manifestos are weighed by a factor of 3 as there are only 3 AfD-manifestos compared to 9 manifestos each for all other parties. For a remaining stem j the number $Y_{j\ell}$ is the number of occurrences of this stem in the document ℓ.

These preprocessing steps yield a term document matrix $(Y_{j\ell})_{j=1,\dots,J,\ell=1,\dots,L}$ with $L = 48$ manifestos and $J = 9085$ stems, i.e., 436,080 cells. The proportions of cells that contained word counts of $0, \leq 1, \leq 2, \leq 3$, or ≤ 4 were 64.4%, 80.5%, 87.1%, 90.6%, and 92.6%, respectively. The largest stem count amongst the cells was 475 (stem `"muess"` in the 2021 manifesto of *Die Linke*). The total number of stems after preprocessing was approximately 682,000.

3.2 Poisson Reduced Rank

The PRR model is applied with parameter $K = 2$ to the term document matrix $(Y_{j\ell})$ of the party manifesto corpus after preprocessing using our implementation in the package `poisrrr` available at `https://github.com/chroetz/poisrrr`. By this, we obtain estimates for the positions $(f_\ell^{(k)})_{k=1,2}$ of all manifestos $\ell = 1, \dots, 48$.

3.3 LDA

For modeling the texts using LDA, the initial structure of the dataset of complete manifestos is not useful. The text corpus consists of $L = 48$ party manifestos, which are all very long (except for the AfD manifesto in 2013) and fairly heterogeneous addressing a whole variety of topics. As Guo et al. 2021 were able to show, LDA produces better results when very long texts are split and modeled as individual texts and then - for the analysis - are combined again. Following this finding, we split texts longer than 7500 characters into text fragments after every 5000 characters "rounding" to the next word. Thus, a single party manifesto on average results in almost 45 texts, leading to $M = 2147$ texts in total, which are modeled individually by the LDA.

The process of modeling itself requires the specification of the texts and the vocabulary to be considered, as well as hyperparameters of the model. As vocabulary we consider for the models all word stems of the term document matrix (see Section 3.1) and as hyperparameters we have chosen $\alpha = \eta = 1/K$ for $K = 30$ different latent topics, i.e. we suppose that the considered party programs cover about 30 different topics. For the determination of the prototype we use the respective default values from `ldaPrototype` (Rieger 2020), i.e. in particular that our model is selected from a set of 100 standard LDAs.

4 Results: Convergence at the Center, Divergence at the Fringes

Campaign manifestos are a valuable source for research since these documents indicate the positioning of political parties on wide range of current issues. In contrast to party programs, that strive to lay the ideological foundations for longer periods of time, manifestos are frequent updates of a party's stance. In this section, we present and discuss several results obtained from using the Poisson reduced rank model and the Latent Dirichlet Allocation method to analyze the party manifesto dataset.

The results of an application of the PRR method in two dimensions to the party manifesto corpus are displayed in Figure 2, where the axes denote the two (un-specified) dimensions representing the parties' positions. For each manifesto a point in the plane is calculated. Points close to one another indicate that a similar language is used in the respective documents. Dealing with political texts, these points can be interpreted as political positions. Hence, the larger the distance between two points, the less similar are the political positions of the corresponding two parties. Note that no preliminary information about political interpretation (or any other kind) of words was used. The plot solely relies on the interplay of word counts among the manifestos without making use of their word meanings. Still a meaningful pattern emerges in Figure 2: With the passage of time depicted by the arrow symbols on the graphs, most parties are found to move from left to right when time evolves. That is, they move broadly in the same direction along the horizontal axis, arguably by picking up similar current issues; as the agenda changes, the parties respond. Top and bottom of the figure replicate the classical identification of the parties in a political left/right spectrum. These findings, however, do not apply to the AfD, which has moved in the opposite direction (to the left along the horizontal axis) to a point of maximum distance to the other parties in 2021. The peculiar AfD results might be caused by the language the right-wing party uses in its manifestos: apparently it has developed a vocabulary that differs from the other parties to an extent where two dimensions are not enough to gauge its differences with the political mainstream. The socialist party Linke has changed little with respect to the vertical dimension over the years and maintained the largest distance to the center of the spectrum. The four remaining (center) parties, in turn, form two clearly distinguishable duos, that have each moved closer to one another recently. This is particularly true for the SPD and the Greens, but also for the Christian Democrats and the Liberals.

To illustrate overall consensus and disagreement between parties, we calculate pairwise consensus scores over time. Figure 3, 4, and 5 show the degree of overall consensus (as gauged by 1 minus the Euclidean distance of the party manifestos' vocabulary, where 1 symbolizes strong agreement and 0 strong disagreement). The results underline the findings displayed in Figure 2: the best fitting party duos are SPD-Greens and Union-FDP. Their respective consensus values have increased over the past few election cycles. The results also show considerable incompatibilities: between FDP and Linke (Figure 3) as well as between Union and Linke (Figure 4) differences are great and have widened over time. What's more, Union and SPD, the long-term partners in a series of "grand" coalitions, have grown apart somewhat; in 2021, our consensus gauge for the black-red pair is close to its lowest point in more than three decades (Figure 4).

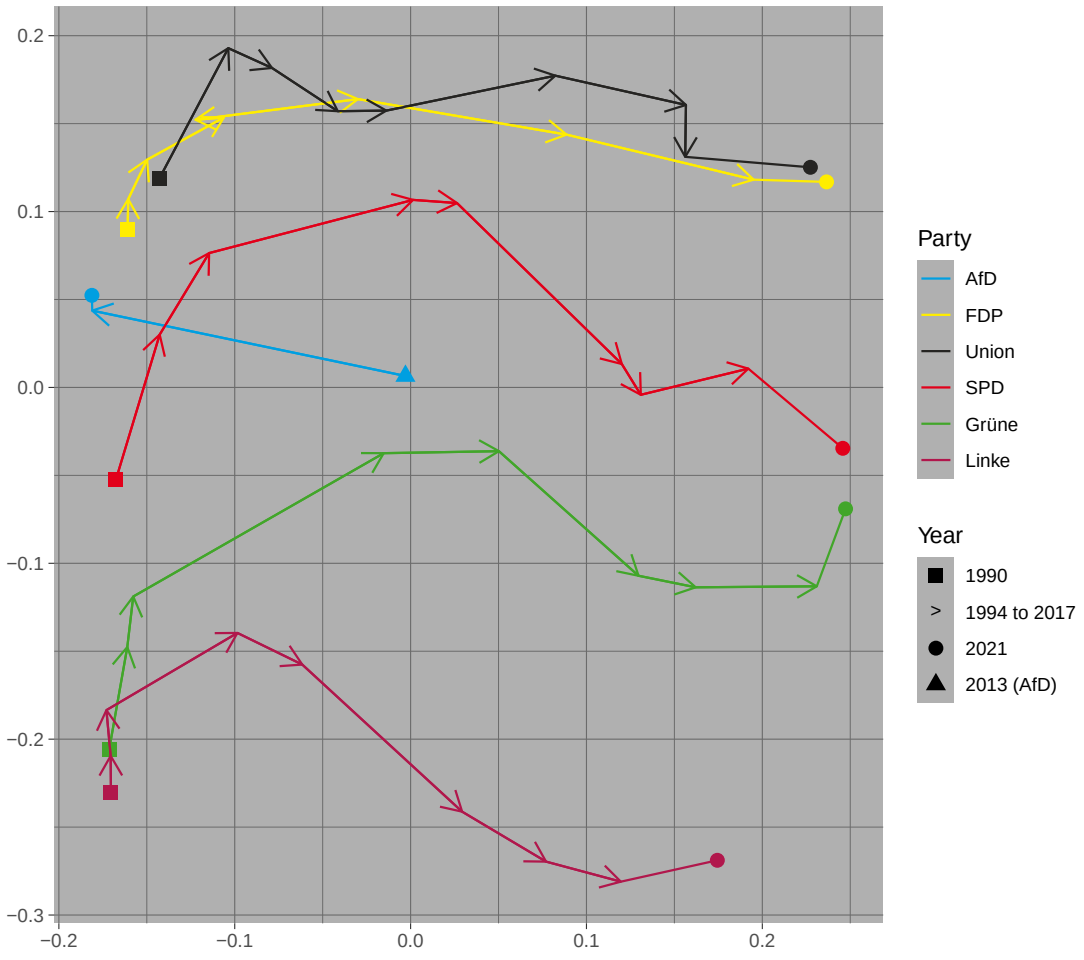

Figure 2: Party positions $(f_\ell^{(k)})_{k=1}^K$ according to a PRR model with $K = 2$.

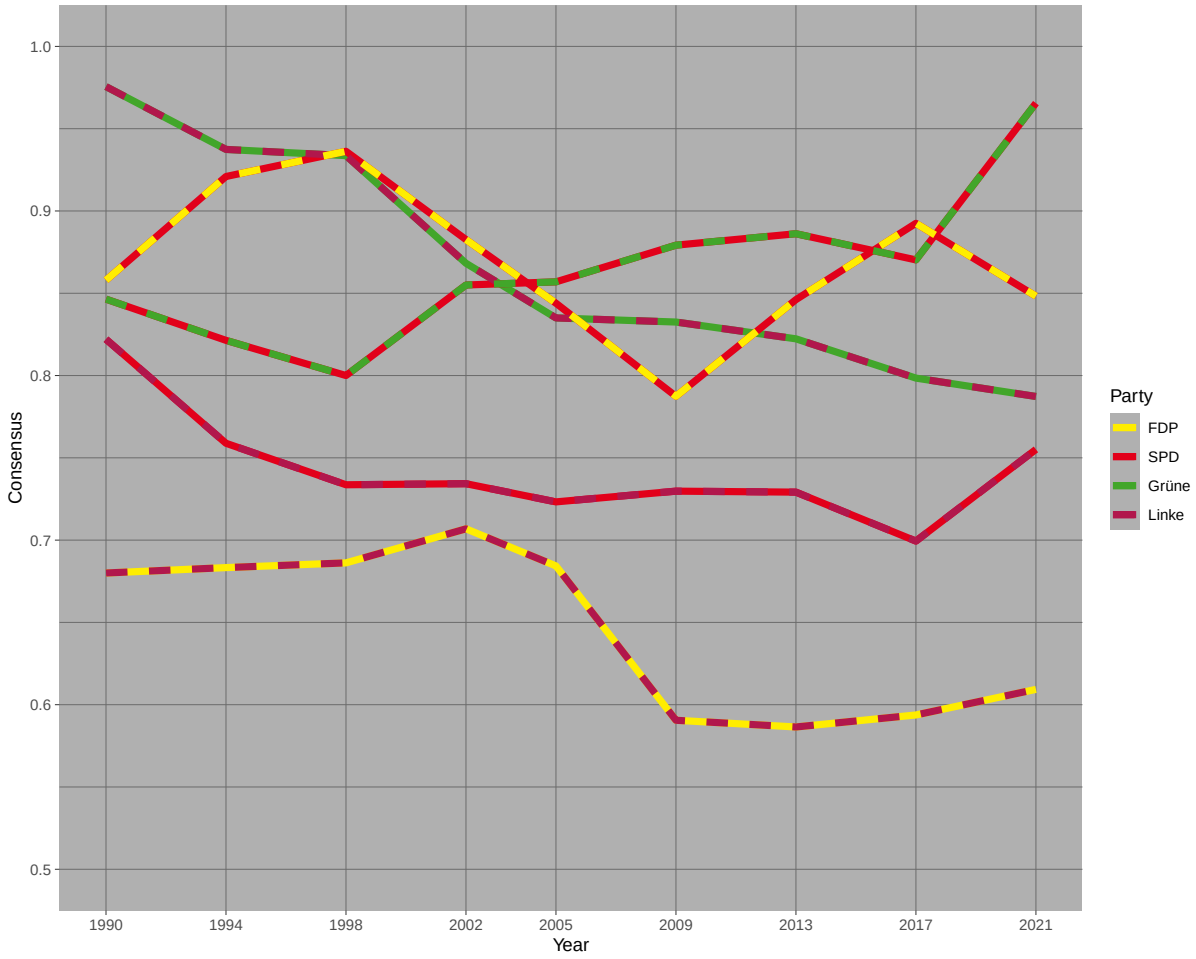

Figure 3: Consensus scores derived from PRR party positions by taking 1 minus the Euclidean distances of the respective manifestos of the same year in Figure 2. Part 1 – politically more left pairs.

Given that two-way coalitions cannot be taken for granted in Germany anymore, a crucial question concerns the pairwise compatibility of the smaller parties. The pair FDP-Greens now scores higher consensus values than ever before in our sample (Figure 4). On the other hand, Greens and Linke have diverged considerably, with consensus between the two now being at a low point since 1990. These findings suggest that a coalition involving both the FDP and the Greens should not be a mission impossible anymore, whereas any coalition involving the Linke would face considerable obstacles. Figure 5 underlines the incompatibility of the AfD with all the other parties. In this figure the values for 2013 should be interpreted with a lot of care because of the extremely short party manifesto of the AFD in this year, see Figure 1.

For the foreseeable future, it seems likely that three parties will be needed to form a coalition to win a solid parliamentary majority. Using the PRR approach we calculate the degree of linguistic consensus within the five possible three-way coalitions. (Coalitions between the Left and the Christian Democrats/Liberals have been ruled out; also, none of the other five parties is willing to engage with the AfD.) Figure 6 depicts the evolution of three-way consensus. Strikingly, the values of all possible coalitions have gone up considerably, or stayed at high levels (as is true for a "Deutschland" pact). These devel-

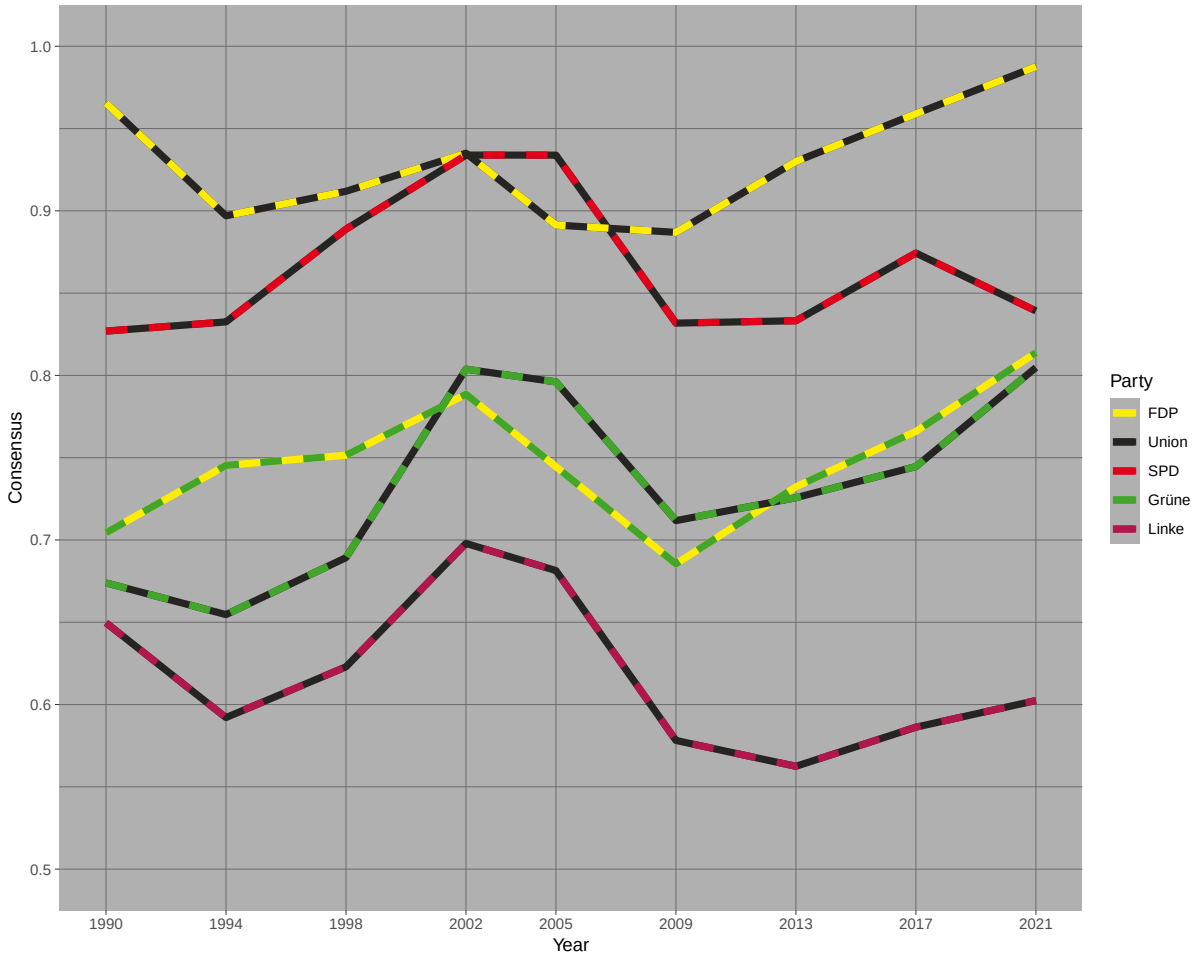

Figure 4: Consensus scores derived from PRR party positions by taking 1 minus the Euclidean distances of the respective manifestos of the same year in Figure 2. Part 2 – politically more right pairs.

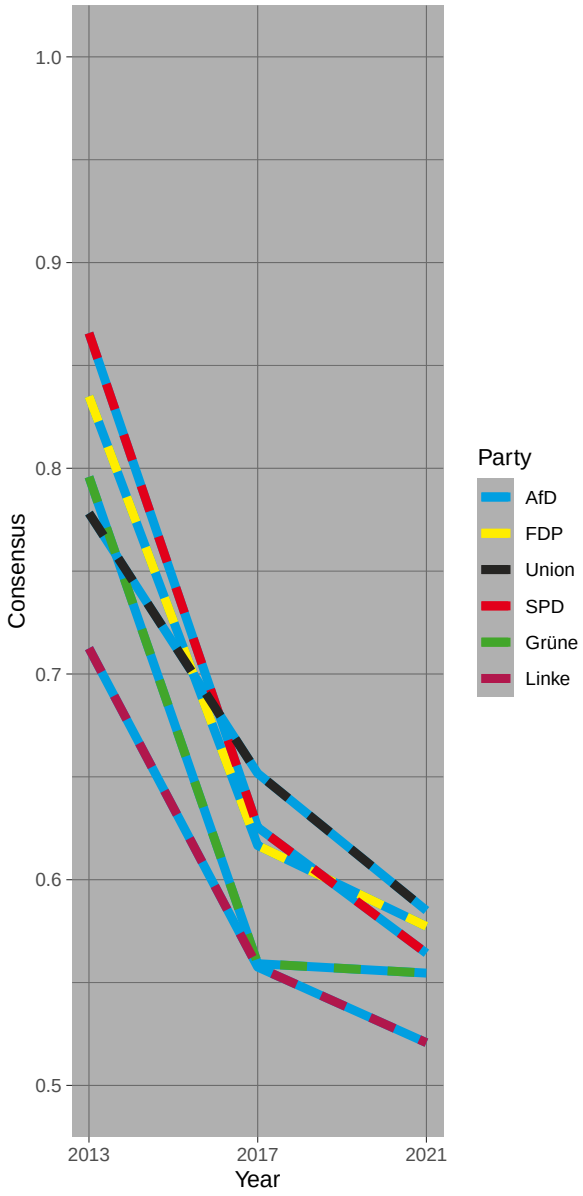

Figure 5: Consensus scores derived from PRR party positions by taking 1 minus the Euclidean distances of the respective manifestos of the same year in Figure 2. Part 3 – pairs which include the AfD.

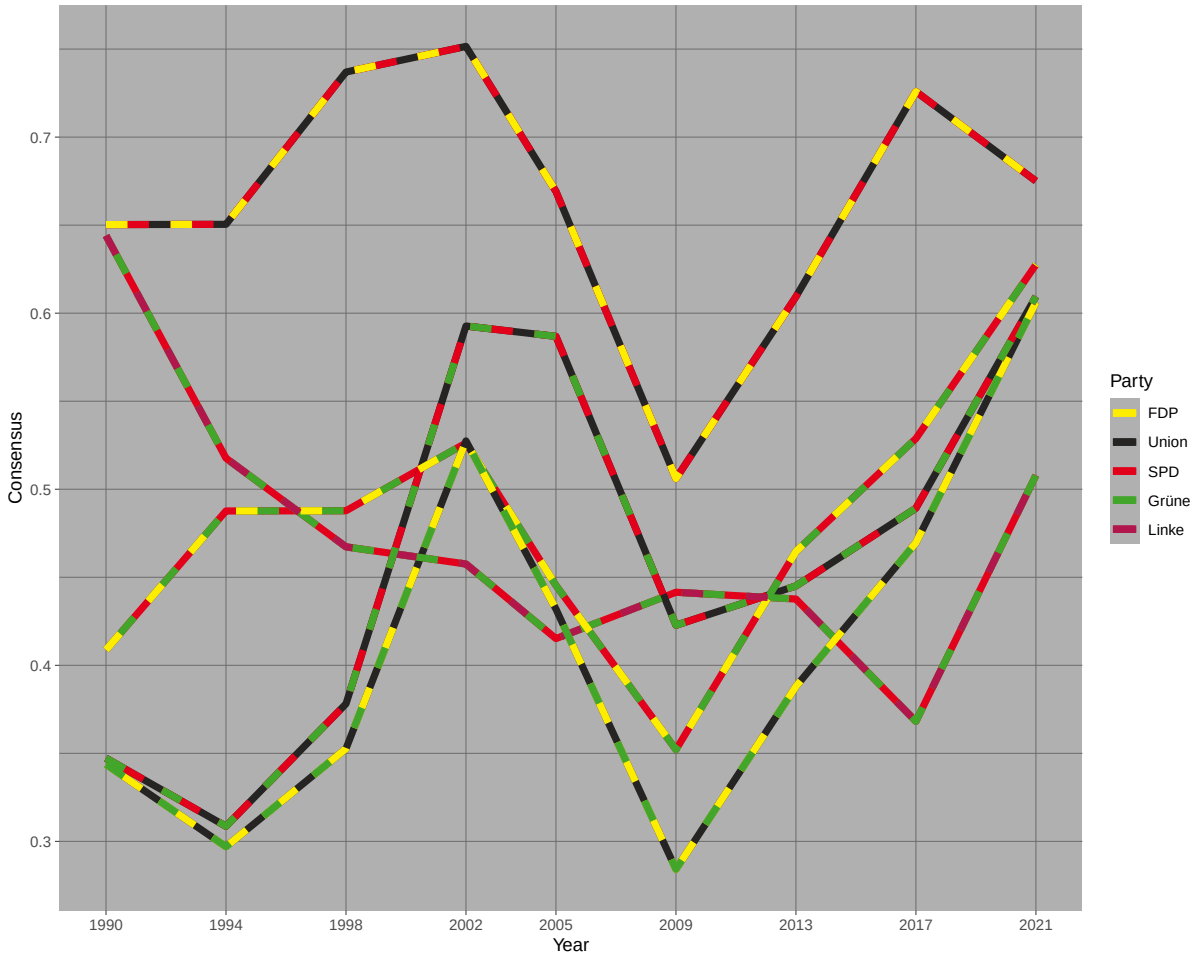

Figure 6: Consensus scores derived from PRR party positions by taking 1 minus the sum of the pairwise Euclidean distances of the respective manifestos of the same year in Figure 2.

opments could be interpreted as the parties reacting to a more fragmented and polarized political spectrum. As the fringes have grown and become more radical, particularly on the right, the four parties at the center have moved closer together. The slight decrease of consensus in a hypothetical 2021 "Deutschland" coalition can be attributed to the SPD's recent programmatic shift shown in Figure 2.

To be able to get a more detailed impression of the parties' positions on different issues, we calculated 100 LDA models for $K = 15, 20, ..., 50$ using `ldaPrototype` as selection mechanism within the same number of topics K. The results were thoroughly scrutinized by the authors, using the tables of the most characteristic terms (top words) and manifesto segments (top texts, see Section 3.3) to gauge the content. As is regularly the case with text mining techniques, their use in social sciences requires considerable interpretative understanding of the results. Since the models need to be calibrated in ways that are suitable to answer relevant research questions, the results of LDA models need to be reviewed by human researchers. Clusters of texts (topics) need to be labeled and described by at least two coders with some understanding of the issues involved. Topics should be clear-cut, that is, distinct and distinguishable. It needs to be stressed that an LDA topic may not only contain a particular issue, but also a "frame" (DiMaggio et al. 2013), i.e. a

certain spin, a problem definition, a moral judgement, possible remedies. Entman 2006 described these properties of a frame for journalistic media content, but the concept lends itself to being applied to party manifestos as well.

After reviewing the different models, we settled for a parameter setting of $K = 30$. In Figure 7 the results for the manifestos from 2021 are depicted, that is, the 15 relevant topics with the highest shares for the manifestos in 2021. For the analysis of the top 15 topics, two topics were neglected that obviously contained one-sided partisan terms. The order of the topics indicates the overall prevalence of the topics. In addition, we calculated an importance score for each cell of a topic-coalition combination in Figure 7, which is given by

$$1 - \left| \frac{\hat{\theta}_{i,k}}{\hat{\theta}_{\text{coalition},k}} - \frac{1}{|\text{coalition}|} \sum_{i \in \text{coalition}} \frac{\hat{\theta}_{i,k}}{\hat{\theta}_{\text{coalition},k}} \right|, \tag{4}$$

for $\hat{\theta}_{\text{coalition},k} := \sum_{i \in \text{coalition}} \hat{\theta}_{i,k}/|\text{caolition}|$ and coalition $\subset \{1, ..., L\}$, see Section 3.1.

The upmost one issue in Figure 7 represents a topic we labelled “Zusammenhalt” (“Regional Cohesion”); it deals with the widening gap between metropolitan and rural areas, transport policy, managing residential property markets, environmental policy, agriculture etc. “Regional Cohesion” is the most sizeable topic in the Union’s 2021 manifesto, with about a fifth of its content correspondingly attributable, as indicated by the size of the black bar in the panel on the upper left. While this topic also ranks high in the SPD’s and the Greens’ manifestos, each party tends to emphasize different aspects. Still, the wording is closely related to the effect that LDA forms a single topic containing passages from all the parties’ manifestos. In fact, we calculate a consensus of 85 per cent concerning this issue in a Kenya coalition, as the figure in the panel on the upper left indicates: Christian, Social and Liberal Democrats can be expected to be somewhat aligned on this issue, whereas a R2G pact would have considerable differences (panel on the upper right).

It needs to be stressed that the LDA results point to competing views concerning economic and social policies. Essentially three topics deal with these aspects: “Wirtschaft (Klima, sozial)” - “Economy (climate, social)” - stresses environmental and social protection (row 3), while “Wirtschaft (Wettbewerb, digital)” - “Economy (competition, digital)” - emphasizes market dynamism and innovation (row 5). In terms of the concept of the frame, the two topics offer different problem definitions and solutions, the first one being stressed mainly by the SPD and the Greens, the second one mainly by the Christian Democrats and the Liberals, as the corresponding bars indicate. A third topic (“Arbeitnehmer” or “Employees”, row 12) focuses on workers’ rights and is put forward mainly by the SPD.

In terms of “Außen- und Sicherheitspolitik” (“Foreign and Security Policies”) a leftish R2G coalition (row 4, column 5) would be the most divided. Note that the topic model probably understates these differences, since the party manifestos’ wording is rather similar, although the stances differ fundamentally, particularly with respect to NATO. Divisions concerning the views on Europe (“EU”) are even more pronounced (row 6, column 6).

In Figure 8, in addition to the importance scores considered in Figure 7, we see associated

similarity scores, which are obtained by calculating the mean pairwise cosine similarities of the stem frequency vectors of the corresponding parties involved in the coalition. Since we always consider three-party coalitions, there are three combinations of pairs, and thus the (cosine) similarity score for a topic-coalition combination is given by

$$\frac{1}{3} \sum_{\substack{r,s \in \text{coalition} \\ r \neq s}} \frac{\sum_j n_k^{(r,j)} n_k^{(s,j)}}{\sqrt{\sum_j {n_k^{(r,j)}}^2} \sqrt{\sum_j {n_k^{(s,j)}}^2}}. \tag{5}$$

From the sum of all scores for a coalition (columns) weighted by the coalition's topic frequencies, pooled importance and similarity scores can be calculated, given in Table 1. The corresponding pairwise cosine similarities can be seen in Figure 9. The table shows that with respect to the LDA results a Kenya coalition consisting of SPD, Union and Greens have the highest degree of convergence regarding the relevance of the discussed topics (importance), as well as the highest pooled agreement regarding similar word choice within these topics (similarity).

Table 1: Pooled importance and similarity scores for possible coalitions after the German federal election in 2021.

	Kenia	Deutschland	Jamaika	Ampel	R2G
Importance	0.7663	0.7251	0.7387	0.7608	0.7255
Similarity	0.5996	0.4715	0.5185	0.4950	0.5434

The differences in the importance of individual topics as well as the parties' different choice of words in the party manifestos are also reflected in the word clouds in Figure 10. The left side of the word cloud matrix shows the possible coalitions, the right side the individual parties. For each word cloud, 50 different words are shown. In contrast to the rest of the paper, these are not the word stems but, for better readability of the term, the actual word tokens from the original manifestos are displayed. For coalitions, those 50 words are determined that are used significantly more frequently in the coalition than in all other parties. Analogously, the particularly discriminatory words are also determined for parties. If $p_{i,v}$ represents the relative frequency of word v in manifesto i, then the words in the word clouds are selected according to

$$\sum_{r \in \text{coalition}} p_{r,v} \Big/ \sum_{s \notin \text{coalition}} p_{s,v}, \tag{6}$$

that is, those words are selected that realize the 50 highest scores for (6) for the specific coalition. Then, the words are colored by their discriminatory power within the coalition, which means that the words that fulfill

$$p_{i,v} \Big/ \sum_{r \in \text{coalition}} p_{r,v} > 0.42 \qquad \text{and} \qquad p_{i,v} = \max_{r \in \text{coalition}} p_{r,v} \tag{7}$$

are colored in the given party color of manifesto i. After all, the size of the representation of each word is proportional to $\sum_{r \in \text{coalition}} p_{r,v}$. For individual parties, the same formulas are used assuming one-party coalitions.

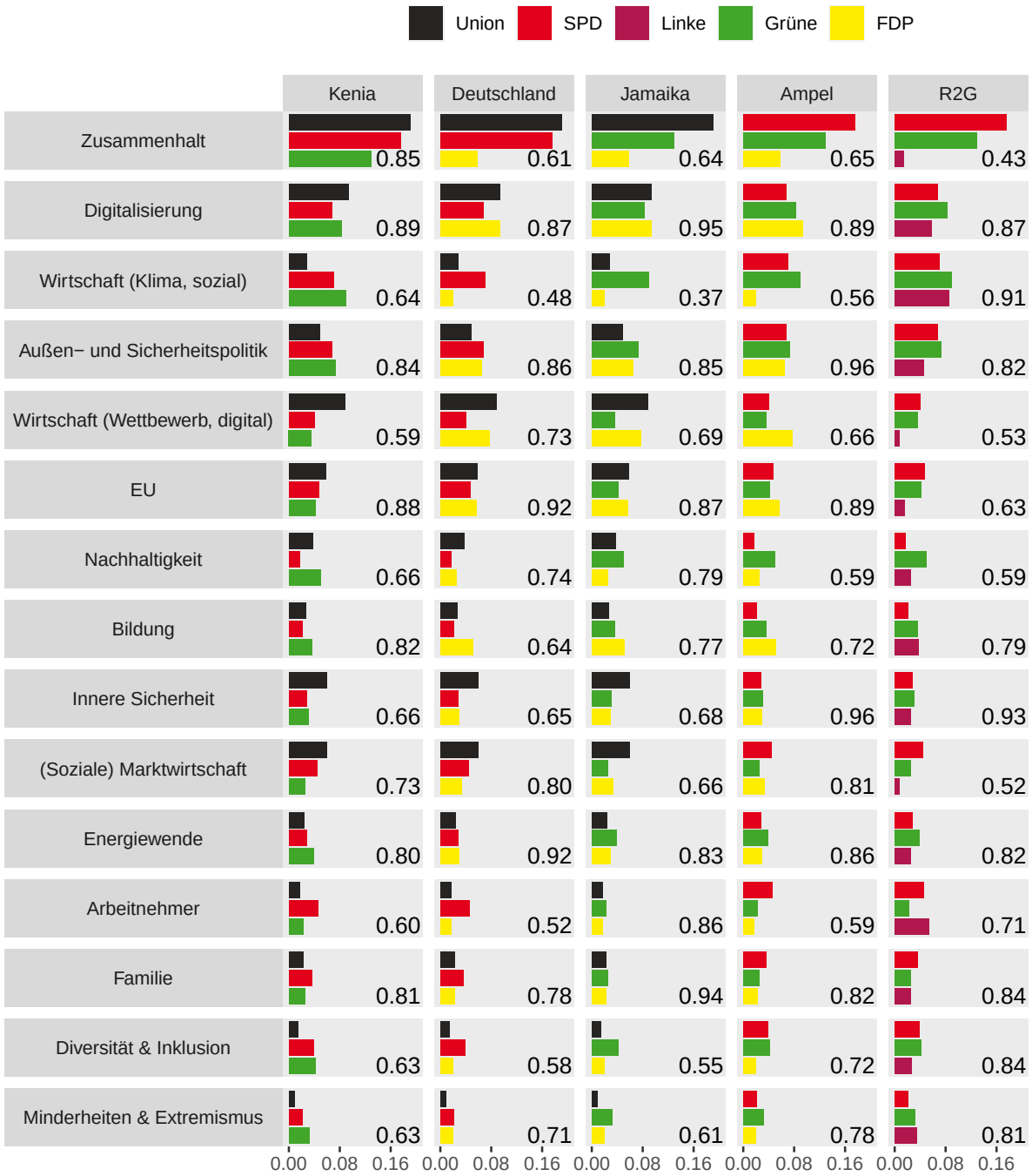

Figure 7: Share of 15 selected LDA topics for the possible parties involved in coalitions and the mean absolute difference of the standardized topic frequencies of the party manifestos in 2021.

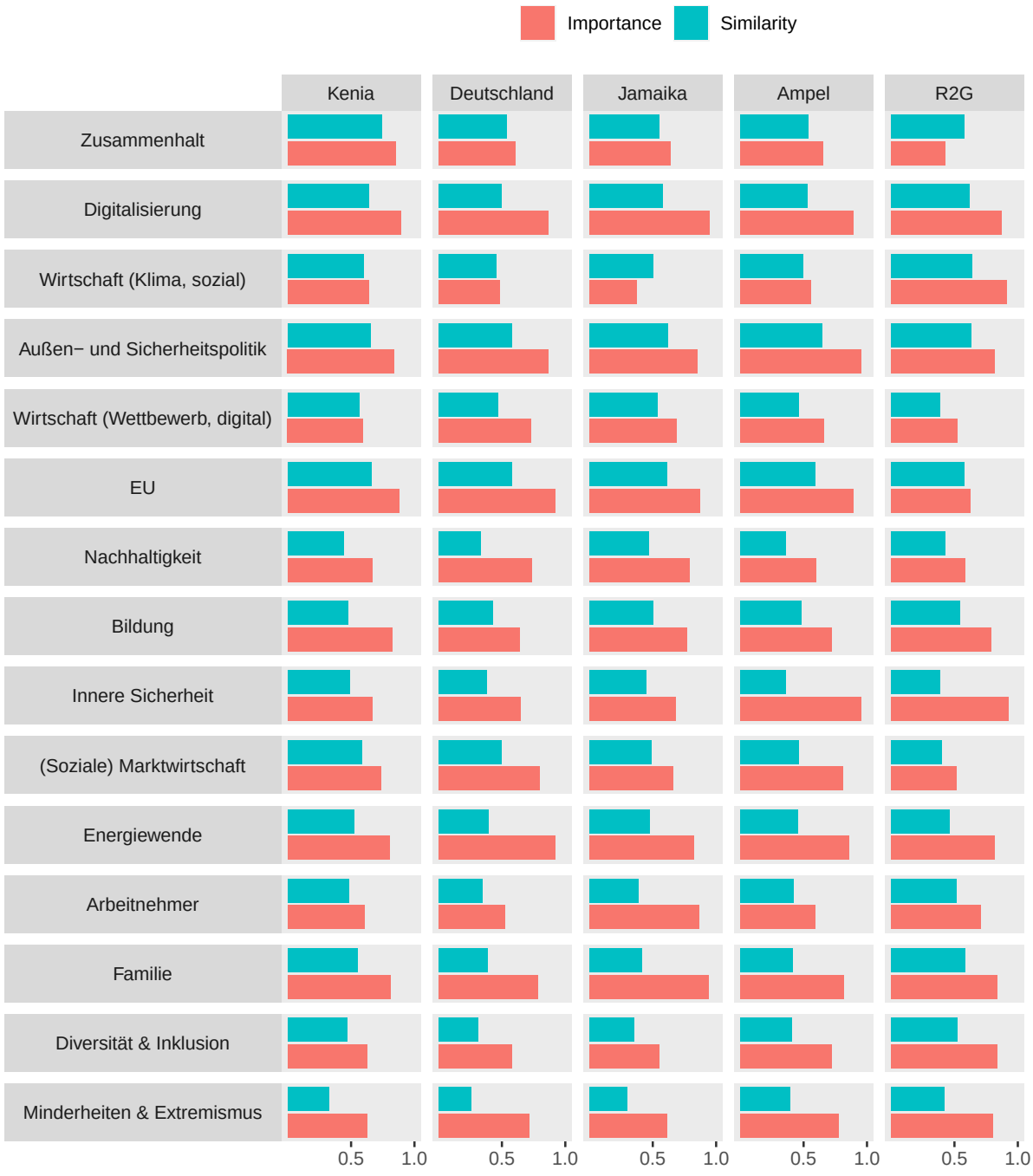

Figure 8: Comparison of importance (see Figure 7) and mean pairwise cosine similarity of stem frequency vectors stratified by party for 15 selected LDA topics and possible coalitions after the German federal election in 2021.

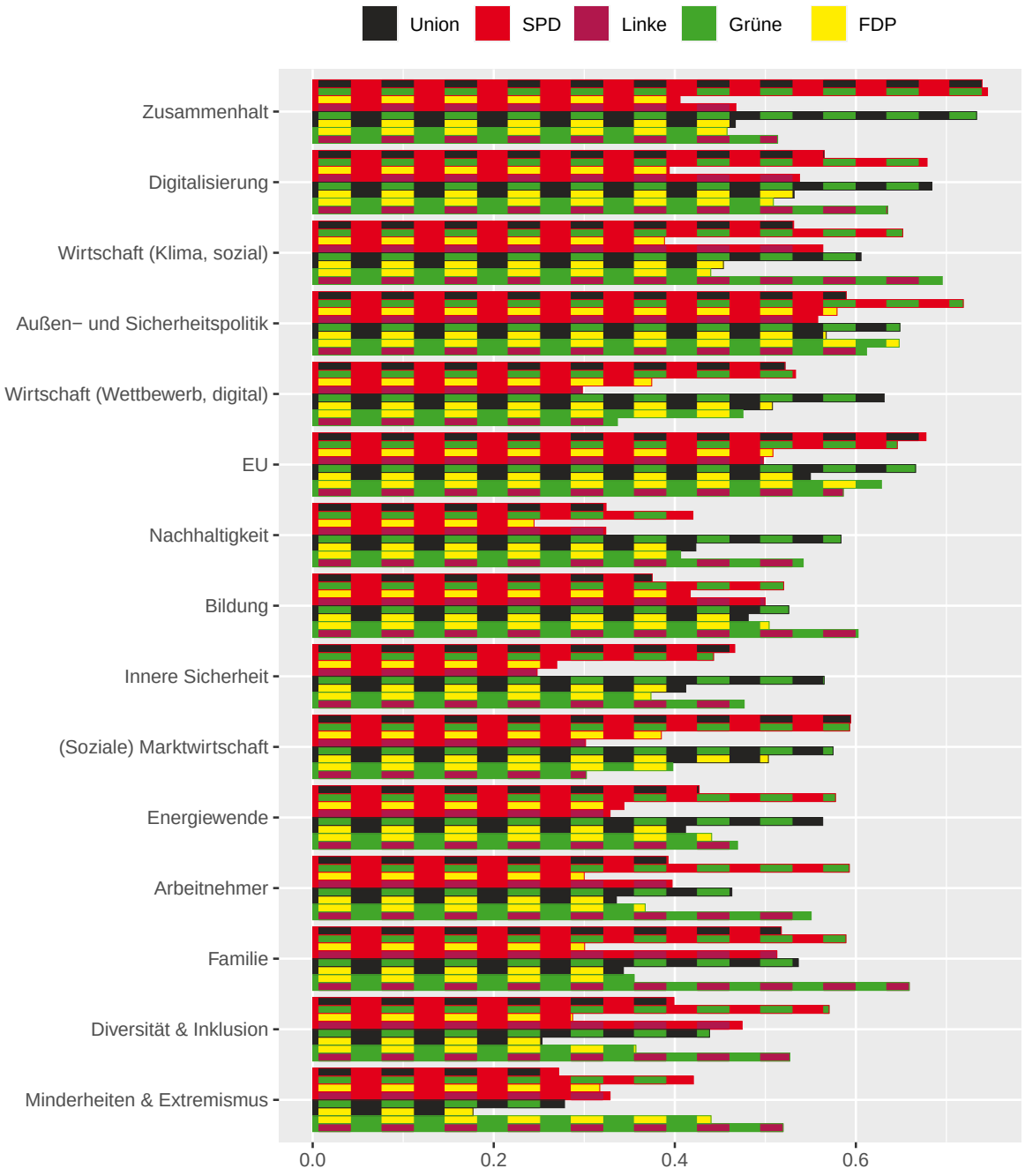

Figure 9: Pairwise cosine similarities of topic-specific stem frequency vectors for two-party coalitions as part of possible three-party coalitions after the German federal election in 2021.

Kenia:

SPD:

Ampel:

Union:

Deutschland:

Grüne:

Jamaica:

FDP:

R2G:

Linke:

Figure 10: Word clouds of the most discriminatory words between coalitions or parties with respect to their party manifestos in 2021.

5 Conclusions

We used two complementing text mining methods, the Poisson reduced rank model (PRR) and the Latent Dirichlet Allocation (LDA) for the statistical analysis of German party manifesto text data between 1990 and 2021. PRR is employed to analyze the similarity of language used in the party manifestos, which can be interpreted to reflect the respective party positions. LDA allows to study the importance of automatically identified topics addressed in the manifestos and their similarities among the corresponding parties.

Statistical methods like PRR and LDA are valuable supplements to the social scientist's tool-box. Since these methods manage with minimum assumptions, they enable an impartial cross-check of findings derived from these disciplines' more traditional methods. In this paper, we began our quest with an exploratory approach, following three broad research questions: what patterns do we find in the data? Which party combinations are the most compatible in terms of their programs? To what extent do the parties' stances vary on certain issues?

These questions matter because, with a party landscape reshaped by fragmentation and polarization, forming compromises to bolster stable governing coalitions is becoming a demanding undertaking, as three-way coalitions are bound to have a harder time to carve out common projects. Our approach enables researchers to gauge programmatic inter-party consensus and dissent over time.

Our analysis yielded three main insights:
First, as programmatic polarization has increased at the fringes of the political spectrum, particularly at the hard-right, the programs of the parties at the center have converged over the past two election cycles.
Second, consensus among the two smaller parties at the center, the Greens and the FDP, has increased. At the same time, the two bigger parties, Union and SPD, that have been long-time partners in "grand" coalitions, have diverged somewhat. The Linke, meanwhile, keeps considerable programmatic distance to the four center parties.
Third, in terms of political issues, the crucial points of conflict of a "Traffic Lights" and a "Jamaica" would be economic and social policies, where two competing approaches could be detected. For a R2G coalition the parties' stances on foreign and security policy and on Europe would be the most difficult to align.

Acknowledgements

The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993.

Bouchet-Valat, Milan (2020). *SnowballC: Snowball Stemmers Based on the C 'libstemmer' UTF-8 Library*. R package version 0.7.0. URL: https://CRAN.R-project.org/package=SnowballC.

DiMaggio, Paul, Manish Nag, and David Blei (2013). "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding". In: *Poetics* 41.6. Topic Models and the Cultural Sciences, pp. 570–606. DOI: https://doi.org/10.1016/j.poetic.2013.08.004.

Entman, Robert M. (Feb. 2006). "Framing: Toward Clarification of a Fractured Paradigm". In: *Journal of Communication* 43.4, December 1993, pp. 51–58. DOI: 10.1111/j.1460-2466.1993.tb01304.x.

Griffiths, Thomas L. and Mark Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.suppl 1, pp. 5228–5235. DOI: 10.1073/pnas.0307752101.

Guo, Chonghui, Menglin Lu, and Wei Wei (2021). "An Improved LDA Topic Modeling Method Based on Partition for Medium and Long Texts". In: *Annals of Data Science* 8.2, pp. 331–344. DOI: 10.1007/s40745-019-00218-3.

Jentsch, Carsten, Eun Ryung Lee, and Enno Mammen (2020). "Time-dependent Poisson reduced rank models for political text data analysis". In: *Comput. Statist. Data Anal.* 142, pp. 106813, 15. DOI: 10.1016/j.csda.2019.106813.

– (2021). "Poisson reduced-rank models with an application to political text data". In: *Biometrika* 108.2, pp. 455–468. DOI: 10.1093/biomet/asaa063.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rieger, Jonas (2020). "ldaPrototype: A method in R to get a Prototype of multiple Latent Dirichlet Allocations". In: *Journal of Open Source Software* 5.51, p. 2181. DOI: 10.21105/joss.02181.

Rieger, Jonas, Jörg Rahnenführer, and Carsten Jentsch (2020). "Improving Latent Dirichlet Allocation: On Reliability of the Novel Method LDAPrototype". In: *NLDB: Natural Language Processing and Information Systems*. Vol. 12089. LNCS. Springer, pp. 118–125. DOI: 10.1007/978-3-030-51310-8_11.

Silge, Julia and David Robinson (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R". In: *Journal of Open Source Software* 1.3. DOI: 10.21105/joss.00037.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani (2019). "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.