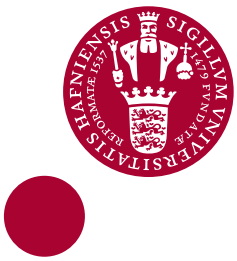

Identifying Parties in Manifestos and Parliament Speeches

Navarretta, Costanza; Hansen, Dorte Haltrup

Published in:
Creating, Using and Linking of Parliamentary Corpora with Other Types of Political Discourse (ParlaCLARIN II)

Publication date:
2020

Document version
Publisher's PDF, also known as Version of record

Identifying Parties in Manifestos and Parliament Speeches

Costanza Navarretta, Dorte Haltrup Hansen
Centre for Language Technology, Department of Nordic Studies and Linguistics
University of Copenhagen Emil Holms Kanal 2
2300 Copenhagen S, DK
{costanza,dorteh}@hum.ku.dk

Abstract

This paper addresses differences in the word use of two left-winged and two right-winged Danish parties, and how these differences, which reflect some of the basic stances of the parties, can be used to automatically identify the party of politicians from their speeches. In the first study, the most frequent and characteristic lemmas in the manifestos of the political parties as well as their language complexity are analysed. The analysis shows inter alia that the most frequently occurring lemmas in the manifestos reflect either the ideology or the position of the parties towards specific subjects, confirming for Danish preceding studies of English and German manifestos. Successively, we scaled our analysis applying NLP methods to the transcribed speeches by members of the same parties in the Parliament (Hansards) and trained machine learning algorithms in order to determine to what extent it is possible to predict the party of the politicians from the speeches. The speeches are a subset of the Danish Parliament corpus 2009–2017. The best results of the classification experiments gave a weighted F1-score of 0.57. These results are significantly better than the results obtained by the majority classifier (weighted F1-score = 0.11) and by chance results. They show that the party of the politicians can be distinguished from their speeches in nearly 60% of the cases, even if they debate about the same subjects and thus often use the same terminology. In the future, we will include the subject of the speeches in the prediction experiments.

Keywords: Parliament Speeches, Machine learning, Corpus analysis

1. Introduction

This paper concerns the relation between political parties' stances and the words the parties use as well as applying natural language processing methods and classification algorithms in order to identify the party of Parliament members from their speeches. The language of politicians has been analysed by researchers from various disciplines such as linguistics, rhetoric and political sciences. Moreover, the digital availability of parliament debates, party manifestos and other political data has extended this research to other fields such as computational linguistics and computer science, while political science researchers are using NLP methods and tools in order to test their theories about political opinions and investigate new aspects of political discourse taking advantage of big data technologies.

Being able to distinguish the party of politicians when they talk about important issues such as economy, culture and immigration investigating whether politicians follow their party's positions in practice is one of the long term-aims of the present research. On the short term, it is interesting to find out to which extent politicians use party specific terminology when they speak in the parliament reflecting eventual differences in their parties manifestos. Therefore, we extracted and analysed frequent lemmas in the political manifestos of four political parties in Denmark applying NLP techniques on the manifestos as a way to present differences and similarities in the positions of the four parties. Successively, we scale the study up applying NLP methods to the transcriptions of the parliament speeches of members of the same parties and training classifiers on the resultig data in order to determine the party of politicians from their speeches. We also investigate which features and algorithms perform best on this task. To our best knowledge, this is the first work, at least for Danish, in which NLP techniques are applied to Parliament speeches in order to automatically predict the party of the speakers.

The paper is organised as follows. Firstly, we discuss related research in section 2., secondly we describe the manifestos that we used in our qualitative analysis and the Hansards which were the data in our machine learning experiments (section 3.). Thirdly, we present the qualitative analysis of the manifestos (section 4.) and in section 5. we account for the prediction experiments and their evaluation. Finally, in section 6. we conclude and present future work.

2. Related Studies

The past decades researchers from different disciplines have addressed political discourse taking advantage of the digital availability of political texts and speeches and of NLP tools for processing them. On the one hand, large collection of political data have been collected and/or annotated, e.g. the collection of Hansards from different countries [1] , among many (Alexander and Davies, 2015; Hansen et al., 2018) and outside Europe, e.g. the Canadian bilingual Hansards (Germann, 2001) and the New Zealand's Hansards[2]. Moreover, party manifestos have been collected and annotated in the Comparative Manifesto Project (Merz et al., 2016), and projects associated with the Comparative Agendas Project[3] have manually classified political speeches into domain specific classes. On the other hand, researchers have used raw or annotated data in or der to determine the policy preferences of a number of political parties e.g. (Zirn, 2014; Zirn et al., 2016) and applied

[1] A list of these corpora is under https://www.clarin.eu/resource-families/parliamentary-corpora.

[2] https://www.parliament.nz/en/pb/hansard-debates/.

[3] https://www.comparativeagendas.net/in.

Proceedings of ParlaCLARIN II Workshop , pages 51–57
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

sentiment analysis techniques to Hansards in different languages in order to extract the politicians' stances towards particular subjects, e.g. (Onyimadu et al., 2014; Abercrombie and Batista-Navarro, 2018). Similarly, (Schumacher et al., 2019) have applied sentiment analysis to a collection of speeches by Danish and Dutch politicians at party congresses in order to determine over time the number of positive and negative words used by the different parties.

Some researchers propose to count the number of metaphors used by politicians in their speeches for identifying their political ideology (Landtsheer, 2009). The metaphors used in speeches are also studied in order to distinguish the language of male and female politicians in Italy(Ahrens, 2009). Other researchers have used more simple linguistic data such as word scores obtained from political texts in order to determine the political positions of parties over specific dimensions such as economy and culture in other political texts. For example, Laver et al. (2003) determine word scores from British and German parties' manifestos in order to classify political positions in different manifestos produced by the same political parties. They also found that the words used in manifestos cannot be used to classify political speeches in e.g. parliament since the language in manifestos and in parliament speeches is quite different. Slapin and Proksch (2008) use word occurrences in order to estimate political positions in German manifestos. Diermeier et al. (2012) apply support vector machines to the speeches of conservative and liberal politicians in the US Senate in order to find the words that characterize each group mostly. They conclude that cultural terms are more distinctive than economic ones when differentiating the two groups. We follow this line of research of using word- and sentence-based scores for distinguish political discourse by different parties in our analysis of the manifestos.

More recently, the use of word embedding for analysing political speeches has been addressed (Denny and Spirling, 2018) as a better way to determine the semantics of political speeches than word scores since word embeddings account for the context in which words appear. Rheault and Cochrane (2020) apply therefore word embeddings in order to determine the ideology positions of the left and right wing over time in large British, Canadian and U.S.A. parliament corpora. They assess their results against various indicators from e.g. party manifestos, surveys and roll-call votes.

Differing from the preceding studies, in our classification experiments we use NLP techniques applied to Parliament speeches in order to predict the party of the speakers. We do not consider aspects such as sentence length and punctuation marks, since the speeches were converted to written texts by the Parliament language department and punctuation marks and sentence decision are not part of the original speeches. We apply machine learning on our data, but not deep learning since we do not have large-scale parliament data from many decades as it is the case for e.g. in the study by Rheault and Cochrane (2020). Moreover, the political situation in Denmark is different from that of countries such as U.S.A. and Great Britain were there is a clear ideologicall difference between left and right wing parties. In Denmark, the distinction between left and right parties are often not very strong and parties from the left and the right have common positions on some subjects. For example, the two largest left- and right-wing parties are often accused of being too similar in the Parliament. For these reasons, it is interesting to investigate whether different parties can be in fact distinguished from the speeches of their members.

3. The data

The party manifestos (*principprogrammer*) and the Hansards we address concern the following four Danish parties:

- Dansk Folkeparti (DF, The Danish People's Party) a nationalist party which supports right-wing governments,
- Venstre (V, The Liberal Party) which is the largest right-wing liberal party
- Socialdemokratiet (S, The Social Democratic Party), the largest centre-left party supported by most left-wing parties
- Enhedslisten (Ø, The Red-Green Alliance), the most left-oriented party in the Danish Parliament.

The last twenty years, the Danish prime ministers have belonged to The Liberal Party or The Social Democratic Party. On the contrary, The Danish People's Party and The Red-Green Alliance have never been part of a Government, but they have been very active in the media and in the Parliament debates.

3.1. The manifestos

The party manifestos are interesting since they describe in general terms the ideology of a party and therefore they have been investigated in many projects, e.g. (Merz et al., 2016; Zirn et al., 2016; Laver et al., 2003). In this work, we downloaded the currently valid manifestos from the four parties' homepages. They were published between 2002 – 2017 since parties change their manifestos with varying frequency. In table 1 the length and the publishing date of the four manifestos are given . The oldest and shortest manifesto is from The Danish People's Party. The second oldest manifesto, The Liberal Party's one, is also the longest manifesto, while the length of the two most recent manifestos, The Social Democratic Party's and The Red-Green Alliance's ones, are slightly shorter than that of The Liberal Party.

Party manifesto	Tokens	Year
The Danish People's Party	1132	2002
The Red-Green Alliance	8015	2014
The Social Democratic Party	8835	2017
The Liberal Party	9241	2006

Table 1: Length and year of the manifestos

3.2. The Hansards

The dataset of our second study is part of the Danish Parliament Corpus 2009–2017. It consists of the Hansards of the sittings in the Chamber of the Danish Parliament. The corpus is available as a collection through the Danish CLARIN research infrastructure [4] which is part of the European Research Infrastructure for Language Resources and Technology, CLARIN [5]. The corpus consists of xml-files, each covering the Hansards of a parliamentary year which runs from October to June. The xml-files contain metadata providing information about the meetings, the speeches, the name of the speakers, their role (member, minister, chairman), their party and the timing of the speeches as well as the speeches' text. The Hansards contain the exact transcripts of the speeches with the exception of some editing, transforming the spoken speeches into syntactically coherent written texts. In the Hansards factual errors and slips of the tongue are for example corrected and spoken language characteristics such as filled pauses and retractions are not recorded. A more comprehensive description of the corpus is in (Hansen et al., 2018).

The Danish Parliament Corpus consists of approx. 41 million running words and 182,192 speeches. For this work we used a subset of the corpus also used in a preceding study act to the automatic classification of speeches in general domains (Hansen et al., 2019). In this study we only include the speeches by ordinary Parliament members excluding speeches by ministers since these only belonged to the two parties, The Liberal Party and The Social Democratic Party.

4. An analysis of the manifestos

The manifestos were tokenized, PoS-tagged and lemmatized (Jongejan and Damianis, 2009) with the Centre for Language' tools for processing Danish available at the Danish CLARIN infrastructure[6]. In table 2, we report for each manifesto: the number of running words, the number of lemmas, the number of lemmas which only occur in the specific manifesto (unique lemmas), their percentage with respect to the number of lemmas in the manifesto, and finally the manifestos' LIX-score. The LIX-score was originally proposed by (Björnsson, 1968) as a readability score and is often used in the Nordic countries. However, it is also one of the features that has been found useful to characterize the authors of texts, e.g. (Pennebaker et al., 2007). The LIX-score is calculated as $LIX = \frac{W}{S} + \frac{LW \cdot 100}{W}$, where W is the number of words, S is the number of sentences, and LW is the number of long words, that is words that consists of more than 6 letters. The LIX-score formular is similar to e.g. the Flesch-Kincaid Grade Reading Level and other readability scores (see e.g. (Zhou et al., 2017) for a comparison of various readability scores), which do not include text external evidence such as the frequency of words, or syntax, e.g. information on subordinate clauses. We only use it as an extra factor in the comparison of the four manifestos. Since the manifestos are written by professionals, the LIX-score to some extent reflects the chosen complexity of the manifesto texts with respect to the target audience. Unfortunately, this score cannot be used as a feature for analyzing the speeches from the parliament, since sentence length, delimited by punctuation marks, is not a natural property of spoken language. The Social Democratic Party's manifesto has the lowest LIX score, followed by The Danish People's Party's manifesto. The highest LIX score is that of The Red-Green Alliance's manifesto. Not surprisingly, the manifesto of The Danish People's Party contains the lowest number of unique lemmas since it is the shortest one, while the difference between the number of unique lemmas in the manifestos of The Social Democratic Party and The Liberal Party and their length are not related. In fact, the former manifesto contains approx. 400 tokens less than the second, but has relatively fewer unique lemmas.

The five most frequent adjectival, verbal and nominal lemmas and their relative frequency with respect to the lemma's class in each manifesto were extracted and they are shown in table 3 while table 4 shows the three most frequent unique lemmas in each manifesto. Auxiliary verbs were not included in table 3 .

It is not surprising that some of the frequent lemmas in the table are common to more manifestos. However, many of the frequently occurring lemmas and most of the unique frequent lemmas reflect clearly the political stance of the party. This is especially the case for the manifesto of the most right- and left-winged parties. More specifically, The Danish People's Party's manifesto contains many times the adjective *Danish*[7] and *free*, the substantive *democracy* and *country* and the verb *secure*, while the most frequent unique lemmas for this party are *christianity*, *cultural heritage* and *health care* reflecting the main stance of the party: the defense of the Danish culture, religion, and democracy against the influence of non christian immigrants as well as the need for keeping the Danish welfare system. The Red-Green Alliance's manifesto on the other hand contains many occurrences of the lemmas *socialist*, *capitalist*, *capitalism*, *create*, *work*, and *movement* which point towards the party's ideology aiming towards a socialist state and against capitalism. Similarly, the most frequently occurring lemmas in The Liberal Party's manifestos are partly common to those of the other right-winged party and partly characteristic of their liberal ideology, e.g. *free*, *freedom*, *possibility*, *secure*. Moreover, their most frequent unique lemmas are *liberal* and *liberalism* and *police* which reflect their liberal ideology and their intention to secure a strong policy as middle against criminality, one of the themes in the party's manifesto. Finally, the manifesto of the social democrats contains many lemmas common to the manifestos of the other parties, while the most frequent unique lemmas show their general plan of ensuring a social model and integrating the legal immigrants in the Danish society. This reflects the position of the party in the parliament (center-left) and the fact that the social democrats' attitude towards e.g. immigrants the past years has become more similar to that of the right-winged parties.

[4] https://clarin.dk.

[5] https://clarin.eu.

[6] https://clarindk/toolchains-wizard.jsp.

[7] The occurrences of the adjective in the party's name were removed from the frequency numbers.

Party	Token	Lemma	UniqLemma	% Uniq	LIX
Danish People's Party	1132	389	83	21.3	47.02
Red-Green Alliance	8015	1294	514	39.7	50.05
Social Democratic Party	8835	1286	469	36.5	39.22
Liberal Party	9241	1668	825	49.5	49.45

Table 2: Number of tokens, lemmas, unique lemmas and LIX of manifestos

Danish People's Party					
%	ADJ	%	VERB	%	NOUN
20.56	dansk (Danish)	3.55	ønske(wish)	6.91	land (country)
4.67	stor (big)	2.37	sikre (secure)	3.62	folk (people)
3.74	høj (high)	2.37	følge (follow)	2.30	folkestyre (democracy)
3.74	fri (free)	21.78	udvikle (develop)	1.97	borger (citizen)
2.80	offentlig (public)	1.18	værdsætte (value)	1.64	udvikling (development)
Red-Green Alliance					
%	ADJ	%	VERB	%	NOUN
4.34	socialistisk (socialist)	3.04	skabe (create)	4.46	menneske (human)
4.34	al (all)	3.04	arbejde (work)	3.22	samfund (society)
3.72	demokratisk (democratic)	1.13	leve (live)	3.10	Kapitalisme (capitalism)
3.22	stor (big)	1.13	se (see)	1.49	land (country)
2.48	økonomisk (economic)	1.13	stå (stand)	1.49	arbejde (work)
Social Democratic Party					
%	ADJ	%	VERB	%	NOUN
8.19	god (good)	2.78	skabe (create)	3.37	verden (world)
5.42	al (all)	2.71	gøre (do)	2.93	land (country)
4.50	mange (many)	2.08	sikre (secure)	2.60	menneske (human)
3.34	stor (big)	1.39	tro (believe)	2.48	fælleskab (community)
2.77	social (social)	1.18	gå (go)	1.66	mulighed (possibility)
Liberal Party					
%	ADJ	%	VERB	%	NOUN
5.48	fri (free)	4.43	sikre (secure)	2.06	menneske (human)
5.02	offentlig (public)	1.33	give (give)	1.83	borger (citizen)
4.46	god (good)	1.14	ønske (wish)	1.78	mulighed (possibility)
4.00	al (all)	1.01	udvikle (develop)	1.78	frihed (freedom)
3.81	enkelt (few)	1.01	skabe (create)	1.69	samfund (society)

Table 3: Most frequent lemmas and % of same in the word class

Partys manifesto	1.unique	2.unique	3.unique
Danish People's Party	kristendom (christianity)	kulturarv (cultural heritage)	sundhedspleje (health care)
Red-Green Alliance	socialistisk (socialist)	kapitalistik (capitalist)	bevægelse (movement)
Social Democratic Party	sammenhængskraft (cohesion)	samfundsmodel (society's mode)l	integration (integration)
Liberal Party	liberal (liberal)	frisind (tolerance/liberalism)	politi (police)

Table 4: Most frequent unique lemmas in the parties' manifestos

A manual analysis of all the unique lemmas in the manifestos shows also that while the manifesto of the Danish People's Party addresses the general themes which are connected with the party's ideology, the manifestos of the other three parties, and especially of the social democrats and liberals, also address general political domains such as the environment, the economy and the education policy. Concluding, the analysis of the most frequently occurring lemmas in the four Danish parties' manifestos show that manifestos' lemma frequencies are a useful feature for extracting the political stance of the political parties confirming the importance of word-related scores investigated in party manifestos in other countries, e.g. (Laver et al., 2003; Slapin and Proksch, 2008).

5. The Prediction experiments

As noticed by (Laver et al., 2003), the language of party manifesto is different from that used in political speeches and therefore it cannot be used as a reference language for making predictions in political debates. However, we hypotesize that the words used by politicians of various parties during Parliament debates differ to some extent since they should reflect the different stances of their party on specific issues and make use of words preferred by their political group. Therefore, the main aim of our second study is to determine to what extent it is possible to automatically predict the party of Parliament members from their speeches in the parliament chamber applying various language models built on their words and lemmas. Furthermore, we want to evaluate the performance of several NLP methods and algorithms on this task.

First, we extracted all the speeches uttered by members of the four parties whose Manifestos were analysed in the previous sections. Then, we removed the speeches which were produced by ministers and the Speaker in order to have a uniform corpus of speeches by ordinary Parliament members, since the speeches of ministers are generally longer while the Speaker only chairs the debates without participating actively in them. We also removed from the data the speeches which contained less than 7 words, getting a dataset of 15911 speeches and 3,145,226 tokens. The number of speeches and the number of tokens per party in the resulting datasets are in table 5. The experiments were run

Party	Number	Tokens
Danish People's Party	3864	785,785
Red-Green Alliance	3711	732,422
Social Democratic Party	4255	858,880
Liberal Party	4081	768,139

Table 5: Speeches and tokens per party

using the scikit-learn library in python. The transcriptions of the speeches were tokenized and lemmatized using the Centre for Language's tools available in the Danish infrastructure, CLARIN.DK. The data were transformed in csv-format so that every line contained a speech, the lemmas of the speech, and the party of the speaker. Punctuation marks were removed from the speeches. The module's algorithms which were tested are K-nearest Neighbors (KNN) multinomial Naive Bayes (NB), Multi-layer Perceptron classifier (MLP) with a lbgs solver, Support Vector Machine (SVM) with a rbf kernel, and Logistic Regression with the lbgs solver (LR). The dataset was randomly divided in a training set, 60% of the data, a testing set (20% of the data) and an evaluation set (the remaining 20% of the data). The baseline is provided by a majority classifier, and the results are reported in terms of precision (P), recall (R) and weighted F1-score (F1). Speeches of a politicians couls occur both in the training and test data. The algorithms were trained on the following datasets: a dataset consisting of bag-of-words (BOW), BOW of the speeches' lemmas (BOWL), the term frequency-inverse document frequency (tf*idf) extracted from the words (TFIDF) of the speeches and from their lemmas (TFIDFL). The tf*idf measure was developed in the field of Information Retrieval (Salton and McGill, 1986) in order to determine how central a word is to a document in a collection of documents and is also often used in NLP.

Stoplists consisting of the most frequent tokens (n>2500) and of the least frequent tokens (n<10) were applied when pre-processing the data. The removal of the least frequent tokens resulted in the deletion of most of the wrongly tokenised elements and numbers. The most frequently occurring lemmas on the other hand consisted of words like *tak* (thanks), *minister* (minister) and *lovforslag* (law bill) which often occur in the speeches from all parties, and therefore are not particularly characteristics of one of them. Table 6 shows the results of the baseline and of the three best performing algorithms, that is Naive Bayes , Support vector machine and Logistic Regression on the various language models. The results of all classifiers are significantly bet-

Algorithm	Data	P	R	F1
Majority		0.07	0.27	0.11
Multinom. Naive Bayes	BOW	0.57	0.57	0.57
	BOWL	0.52	0.52	0.51
	TFIDF	0.57	0.46	0.44
	TFIDFL	0.52	0.47	0.44
Support Vector Machine	BOW	0.52	0.52	0.52
	BOWL	0.47	0.46	0.46
	TFIDF	0.57	0.57	0.57
	TFIDFL	0.55	0.55	0.55
Logistic Regression	BOW	0.52	0.52	0.52
	BOWL	0.49	0.49	0.49
	TFIDF	0.57	0.56	0.56
	TFIDFL	0.53	0.53	0.53

Table 6: Results of predictions experiments

ter than those obtained by the majority classifier or those that can be obtained by chance (0.25). The best results, a weighted F1-score of 0.57, were obtained with the multinomial Naive Bayes trained on bag of words and the support vector classifier trained on tf*idf over words. The second best results were obtained by Logistic regression trained on the tf*idf over words (F1-score 0.56). The results are very promising since some of the speeches are short and the parliament members discuss the same law bills, and therefore they often use the same terminology. Moreover, speakers' individual characteristics in the form of e.g. number of disfluencies and self corrections were removed from the speeches. Therefore, the differences between the various speeches are not caused by these factors. Instead the differences in word use by different parties' members can be explained by party specific terminology and by party specific interests in various subjects . Both aspects should be investigated further in future studies.

Figure 1 is the normalized confusion matrix obtained with the support vector machine's tf*idf model. The diagonal of the confusion matrix shows the proportion of speeches which were correctly classified, while the other slots show the speeches which were wrongly attributed to another party. The confusion matrix shows that the model predicts

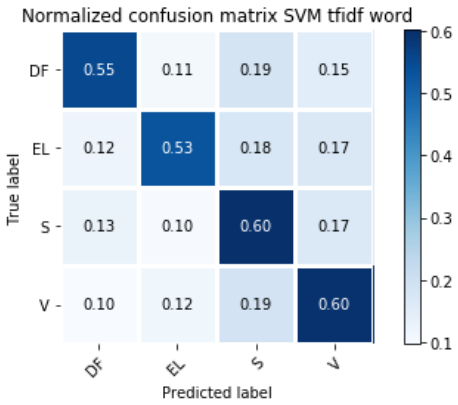

Figure 1: Confusion matrix

speeches from the four parties with F1-scores between 0.53 and 0.60, and the best scores were obtained on speeches by The Liberal Party's (V) and Social Democracy (S)'s members and the worse result was achieved in the identification of speeches by The Red-Green Alliance (EL). Not surprisingly, the best results regard the speeches of the two parties with the highest number of speeches. Similarly, the F1-score for The Red-Green Alliance's speeches is the lowest one since the speeches from the latter party are the least numerous. Moreover, the fact that the language model predicts in 0.19 of the cases that the speeches from The Danish People's Party are uttered by members of The Social Democratic Party confirms the qualitative analysis of the two parties' manifestos which indicated that The Danish People's Party and The Social Democratic Party use often the same terminology in a number of subjects (section 4.). The confusion matrix also shows that the speeches of the two parties which are less frequently confused are those from the most left- and most right-winged party. Furthermore, the matrix shows that the speeches of Social Democrats and Liberals are also often attributed to the other party (19% and 17% of the cases). Interestingly, the best performing algorithms give similar Precision and Recall scores (the same in our tables since we rounded the results up to two decimal digits). This shows that the false negatives and false positives are often the same, indicating again that the members of the Parliament talk about the same subjects and have some common terminology in approximately 40% of the cases even if they have different ideologies.

6. Conclusions and future work

In the paper, we described work act to a) present an analysis of the content of the manifestos of two left- and two right-winged Danish based on the most frequent and specific lemmas occurring in them, b) determine to what extent the words used in the parliament debates by members of the four parties can be used to train models that can distinguish the party of the speech producers c) test the performance of various features and classifiers on this task.

The analyses of the frequency of content lemmas in the manifestos indicate similarity and differences between the four parties ' programs, confirming that parties from both the left and right wing have similar positions on a number of subjects. The analyses also confirm previous research that successfully use word-based scores from party manifestos in order to distinguish the party's positions towards specific subjects (Laver et al., 2003; Slapin and Proksch, 2008). The results of our prediction experiments involving various language models based on NLP-technologies show that the best results are achieved by a support vector machine trained on a td*idf vector (F1-score$= 0.57$) obtained from the speeches' words and by the multinomial Naive Bayes trained on bag of words. These results are striking since the politicians discuss the same subjects in the Parliament Chamber. The confusion matrix for the best performing language model also confirms that the speeches of some parties (Social Democratic Party and Liberal Party as well as Social Democratic Party and Danish People Party) are more similar to each other than the speeches by members of other parties (Danish People's Party and Red-Green Alliance) with respect to lexical choice. In the future, we will include in the study the subjects of the speeches and other factors such as the age and gender of the parliament members. Moreover, the speeches of more parties and covering a longer period of time will be used in prediction experiments. Finally, our study could be extended to the Hansards of more parliaments and the words used by left-wings and right-wings politicians in different countries could be compared.

7. Acknowledgements

This work is done in CLARIN.DK.

8. Bibliographical References

Abercrombie, G. and Batista-Navarro, R. T. (2018). A sentiment-labelled corpus of hansard parliamentary debate speeches. In Darja Fiser, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Ahrens, K. (2009). *Politics, Gender and Conceptual Metaphors*. PALGRAVE MACMILLAN.

Alexander, M. and Davies, M. (2015). Hansard corpus 1803-2005. Available online at http://www.hansard-corpus.org.

Björnsson, C.-H. (1968). *Läserbahed*. Liber, Stockholm.

Denny, M. J. and Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189.

Diermeier, D., Godbout, J.-F., Yu, B., and Kaufmann, S. (2012). Language and Ideology in Congress. *British Journal of Political Science*, 42(1):31–55.

Germann, U. (2001). Aligned hansards of the 36th parliament of canada. https://www.isi.edu/natural-language/download/hansard/.

Hansen, D. H., Navarretta, C., and Offersgaard, L. (2018). A Pilot Gender Study of the Danish Parliament Corpus. In Daria Fiser, et al., editors, *Proceedings of LREC 2018 Workshop ParlaCLARIN*.

Hansen, D. H., Navarretta, C., Offersgaard, L., and Wedekind, J. (2019). Towards the Automatic Classification of Speech Subjects in the Danish Parliament Corpus. In Costanza Navarretta, et al., editors, *DHN 2019 Digital Humanities in the Nordic Countries Proceedings*, volume 2364, pages 166–174.

Jongejan, B. and Damianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-,in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th AnnualMeeting of the ACL and the 4th International Joint Conference on Natural LanguageProcessing of the AFNLP*, pages 145–153, Singapore. ACL.

Landtsheer, C. D., (2009). *Collecting Political Meaning from the Count of Metaphor*, chapter 5, pages 59–78. Springer.

Laver, M., Benoit, K., and Garry, J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review*, 97(2):311–331.

Merz, N., Regel, S., and Lewandowski, J. (2016). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research and Politics*, pages 1–8, April-June.

Onyimadu, O., Nakata, K., Wilson, T., Macken, D., and Liu, K. (2014). Towards Sentiment Analysis on Parliamentary Debates in Hansard. *Semantic Technology. JIST 2013. Lecture Notes in Computer Science*, 8388.

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., and Booth, R. (2007). *The development and psychometric properties of LIWC2007*. LIWC Inc, Austin Texas.

Rheault, L. and Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1):112–133.

Salton, G. and McGill, M. (1986). *Introduction to modern information retrieval*. McGraw-Hill.

Schumacher, G., Hansen, D., van der Velden, M. A., and Kunst, S. (2019). A new dataset of Dutch and Danish party congress speeches. *Research and Politics*, 1-7.

Slapin, J. B. and Proksch, S.-O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722.

Zhou, S., Jeong, H., and Green, P. A. (2017). How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Transactions on Professional Communication*, 60(1):97–111.

Zirn, C., Glavas, G., Nanni, F., Eichorst, J., and Stuckenschmidt, H. (2016). Classifying topics and detecting topic shifts in political manifestos. In *Proceedings of the International Conference on the Advances in Computational Analy-sis of Political Text (PolText 2016)*, pages 88–93, Dubrovnik, Croatia, July.

Zirn, C. (2014). Analyzing Positions and Topics in Political Discussions of the German Bundestag. In *Proceedings of the ACL Student Research Workshop*, pages 26–33.