Policy Domains and the Speakers' Gender in ParlaMint-DK 4.1

Costanza Navarretta
University of Copenhagen, Denmark
costanza@hum.ku.dk

Dorte Haltrup Hansen
University of Copenhagen, Denmark
dorteh@hum.ku.dk

Bart Jongejan
University of Copenhagen, Denmark
bartj@hum.ku.dk

Abstract

In this paper, we describe the ParlaMint-DK 4.1 corpus, which consists of the Danish parliament speeches from 2014 to 2022 annotated with 20 general policy domains mapped to the codebook of the Comparative Agendas Project. The policy domains were added to the speeches semi-automatically using the agenda titles under which the speeches occurred. In the paper, we also account for how some of the linguistic annotations of the corpus were improved using the Text Tonsorium and present some of our previous studies on parliament data. We also describe novel investigations, based on the policy domain annotations in ParlaMint-DK aimed at determining which domains are most frequently addressed in the speeches and the frequency by which policy areas are debated by female and male politicians during the various governments covered by the corpus.

1 Introduction

In this paper, we present the latest version of the ParlaMint-DK corpus, ParlaMint 4.1, focusing on the policy domain annotation, which it contains. ParlaMint-DK is the Danish part of the corpora of parliamentary speeches that were collected and annotated under the CLARIN's flagship project ParlaMint (Erjavec, Kopp, Ljubešic, et al., 2024)[1]. ParlaMint-DK 4.1 differs from the other ParlaMint corpora, because it contains the annotations of policy domains. These annotations are mapped to the categories in the Comparative Agendas Project[2]'s codebook. The annotation of general policy areas or domains in parliamentary debates has been addressed by political scientists for a long time, since these enable the analysis and comparison of how specific topics are dealt with by different parties or political wings, nationally or internationally, in various periods of time (Baumgartner et al., 2011; Ivanusch, 2024; Merz et al., 2016a; Ristilä & Elo, 2023; Yu et al., 2023; Zirn et al., 2016).

The policy domain annotations were first added to some of the Danish speeches in a pilot study described in (Hansen et al., 2019). The speeches used in this work covered the period from October 2009 to June 2017 and were released as the Danish Parliament corpus with subject annotations v.2 (Hansen & Navarretta, 2021) in 2021. This corpus, as it is also the case for ParlaMint-DK, was downloaded from the Danish Parliament (*Folketinget*)'s website[3] and is available in CSV format. The corpus was used for training and testing classification algorithms to identify the speeches' main policy domain automatically. Hansen et al. (2019) report a 0.8 F1-score when identifying 18 domains in a balanced data set. In other experiments, classifiers were trained to identify both primary and secondary policy domains in speeches that are annotated with two domains (Navarretta & Hansen, 2022). The policy domain annotations of the Parliament corpus v.2 with subject annotations were also used in other studies, for example, to investigate how *Immigration* and *Environment* have been handled by different left- and right-wing parties in the covered period (Navarretta & Hansen, 2023; Navarretta et al., 2022).

[1]The corpora are available at (Erjavec, Kopp, Ogrodniczuk, Osenova, Agerri, et al., 2024; Erjavec, Kopp, Ogrodniczuk, Osenova, Agirrezabal, et al., 2024)

[2]https://www.comparativeagendas.net/

[3]ftp://oda.ft.dk

Recently, we coded the policy domains of the speeches from June 2017 to June 2022 and added these to the speeches in the ParlaMint-DK corpus. In this paper, we use these policy domain annotations to determine which areas were most frequently discussed in the Danish parliament, and to find which policy domains were more often addressed by female and male politicians, and whether the frequency of speeches about the various policy domains by the two genders changes over time. ParlaMint-DK 4.1 also contains improved linguistic annotations compared to version 4.0, and in the present paper we also describe these improvements and how they were achieved.

The paper is organized as follows. In section 2, we present studies aimed at classifying policy areas in political discourse, and in section 3 we account for the classification system which we adopted, and the annotation method which we applied. In section 4, we describe how the linguistic annotations of ParlaMint-DK were improved while in section 5, we discuss the distribution of the main policy domains and the co-occurring policy areas in the corpus. In section 6, we account for which policy areas were most frequently addressed by female and male politicians during the governments covered by ParlaMint-DK. Finally, section 7 contains a short conclusion and discussion of future work.

2 Background work

Various classifications of policy domains have been proposed to cover different types of data. The classification systems most often used are the ones created by the Comparative Manifesto Project[4] (Merz et al., 2016b), and by the Comparative Agendas Project[5] (Baumgartner et al., 2011).

The Comparative Manifesto Project classification is fine-grained and comprises more than 550 categories used to annotate so-called quasi-sentences[6] in party election manifestos from many countries. The annotations distinguish positive and negative quasi-sentences and are produced manually.

The scheme used in the Comparative Agendas Project (CAP) builds on the classification applied in the Policy Agendas Project[7], whose aim was to structure the US policy data. The CAP scheme is a modified version of this classification to cover not only the policy activities of the US data, but also the policy domains of other countries (Baumgartner et al., 2011). The CAP classification scheme comprises 21 main domain categories and 192 subcategories. Danish researchers in political science from the University of Aarhus have manually annotated political data from 1953 to 2007 in the Danish Policy Agendas Project[8] using an adapted version of the CAP scheme. We have been inspired by their work.

3 The classification of Policy Areas in ParlaMint-DK

The classification scheme of policy domains which we have used in ParlaMint-DK 4.1, consists of 20 classes. 19 of these correspond to the areas of responsibilities in the Danish Parliament (spokesmanships) in the covered period, while the latter class, *Other*, was added to cover government operations. Table 1 shows the 20 policy domain classes, the corresponding areas of responsibility in the Danish parliament, the corresponding CAP codes and CAP areas.

3.1 The annotation method

The policy domain annotations were semi-automatically added to the speeches in The Danish Parliament Corpus (2009-2017) extracting them from the titles of the agenda items of the meetings. The method was described in (Hansen et al., 2019) where the first pilot annotations were presented. The method consists of the following steps: 1) extraction of the agenda titles 2) normalization, e.g., "Third reading of bill N: XYZ" becomes "XYZ", 3) manual annotation of the agenda titles with one or two policy areas, and 4) automatic assignment of the policy area(s) to each speech in the meeting covered by the relevant agenda titles. For example, for the title *Tax on saturated fat in food*, the domain *Agriculture*, which comprises the food subcategory, was assigned as the primary policy domain, while *Economy*, which comprises *Tax*,

[4]https://manifesto-project.wzb.eu/

[5]https://www.comparativeagendas.net/

[6]Quasi-sentences correspond to sentences in the majority of cases, but they can also indicate entities such as titles of films and books.

[7]https://liberalarts.utexas.edu/government/news/feature-archive/the-policy-agendas-project.php

[8]http://www.agendasetting.dk/.

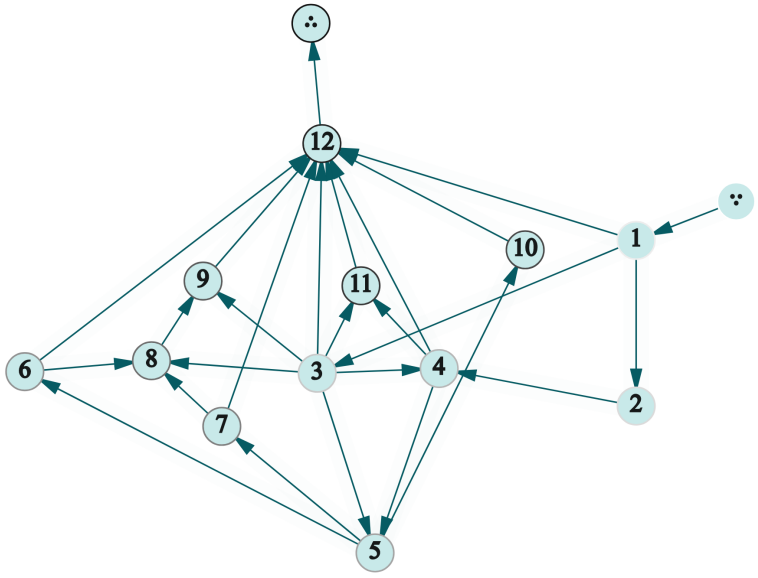
Policy Domain	Area of Responsibility	CAP no.	CAP Areas
Economy	Finance, Fiscal Affairs	1	Domestic Macroeconomic Issues
Health Care	Psychiatry, Health	3	Health
Agriculture	Animal Welfare, Fisheries, Food, Agriculture Consumer Policy	4 1525	Agriculture Consumer Policy
Labour	Labour market	5	Labour and Employment
Education	Higher Education and Research Education	6	Education
Environment	Environment	7	Environment
Energy	Energy Climate	8 705	Energy Air and Noise Pollution, Climate Change and Climate Policies
Immigration	Immigration and Integration, Alien Affairs, Naturalization	9	Immigration and Refugee Issues
Infrastructure	Transportation IT, Media	10 17	Transportation Space, Science, Technology and Communications
Justice	Legal Affairs Constitutional Matters	12 20	Law, Crime, and Family Issues Government Issues
Social Affairs	Children, Family, Social Services, Senior Citizens Gender Equality	13 2	Social Welfare Civil Rights, Minority Issues, and Civil Liberties
Housing	Housing	14	Community Development & Housing Issues
Local and Regional Affairs	Rural Districts and Islands Municipal Affairs	4 2001	Community Development & Housing Issues Local Government Issues
Business	Trade and Industry	15	Industrial and Commercial Policy
Defence	Defence	16	Defence
Foreign Affairs	Foreign Affairs, Development, Cooperation	19	International Affairs and Foreign Aid
European Integration	EU	1910	International Affairs and Foreign Aid
Territories	Faroe Islands, Greenland	2105	Dependencies and Territorial Issues
Culture	Cultural Affairs Ecclesiastical Affairs Sport	23 210 1526	Cultural Policy Issues The Danish National Church Sport and Gambling
Other	-	2000	Government Operations

Table 1: Policy domains, and corresponding responsibility areas, CAP numbers, and CAP areas in ParlaMint-DK

was chosen as the second domain. More difficult cases require knowledge of the content of specific bills. For example, the title of the agenda *1. Behandling af B 47 om Amager Fælled* (1. Treatment of B 47 about Amager Fælled) refers to a bill proposing to sell a green area in Copenhagen as building area. The speeches referring to this agenda item were classified with the main domain *Local and Regional Affairs* and the secondary domain *Housing*. Later, when the parliament discussed the legality of selling this area with respect to the Environment legislation, the speeches about *Amager Fælled* were annotated with the policy domain *Justice*.

5000 speeches that were coded with two policy areas were reviewed by two annotators independently. The two annotators did not find any errors in the assignment of the two policy areas, but in some cases they disagreed on which of the two annotated areas should be considered the primary (Navarretta & Hansen, 2022).

The extended annotations of policy domains covering speeches from 2017 to 2022 were performed according to the same methodology as in (Hansen et al., 2019), but, in the ParlaMint-DK annotations of policy domains, we decided to use an extra domain *Other*, which covers speeches about government operations and other issues. To add all the annotations to the ParlaMint-DK speeches, we first created a TEI taxonomy over the policy domains, and then the policy domain annotations were added to each



ID	Tool	ID	Tool
⋱	input (TEI P5)	7	Anno-splitter (PoS tags)
1	TEI tokenizer (token & sentence boundaries)	8	PoS translator (UD → CST)
2	Sentence extractor (sentence boundaries)	9	CSTlemma (replaces UD lemmas)
3	Token extractor (defines token ID's)	10	Anno-splitter (syntax)
4	TEI-segmenter (defined as token ID ranges)	11	CSTner (named entities)
5	udpipe (PoS, lemmas, morphology, syntax)	12	TEI annotator (aggregates)
6	Anno-splitter (morphological features)	∴	output (TEI P5)

Figure 1: The Text Tonsorium workflow for the annotation of the Danish ParlaMint dataset.

speech as an *@ana* attribute in the *u*-element[9]. Since we had assigned a unique unifier to each parliament speech, and the unifier is based on the exact time and date of each speech, this step was trivial.

4 The linguistic annotations and Text Tonsorium

The linguistic annotations were performed as in the previous versions of ParlaMint-DK through Text Tonsorium[10] (TT) (Jongejan et al., 2021). TT is a workflow management system that not only executes linguistic annotation workflows, but can also compose workflows by combining different Natural Language Processing tools. The linguistic annotations of ParlaMint-DK were made using ten different tools in a workflow comprising twelve steps, see figure 1.

Evaluating the linguistic annotations of ParlaMint-DK 4.0, we found some systematic lemma annotation errors. We corrected them by taking morphology and word form into account when mapping between the Universal tag set output by UD-pipe and the CST tag set used by CSTlemma (Jongejan, 2016). Already for the first published version of ParlaMint-DK we decided not to use the UD-pipe software for delivering lemma annotations because, as a lemmatizer, UD-pipe performs worse than CSTlemma. Also, we required that the application of a lemmatization rule was conditioned on the word class of the input word, while UD-pipe often applies lemmatization rules that are not meant for the word classes

[9]This was done after consulting the head of the ParlaMint project, Tomaž Erjavec.

[10]Via CLARIN-DK website: https://dkclarin.dk/clarin.dk/ or directly: https://cst.dk/texton/, Source code: https://github.com/kuhumcst/texton

assigned by the UD-pipe software. Previously, the mapping from the UD-tag set onto the CST tag set was performed by CSTlemma itself, using a simple lookup table. Now, the mapping task is delegated to a separate tool, the PoS translator (tool 8 in figure 1). The tool combines information from tokenization, PoS-tagging and morphological analysis to provide the correct information to CSTlemma.

The NER annotations were also improved in this version of the annotated corpus. This improvement especially dealt with the abbreviations of parties' and organizations' names. The refined mapping between tag sets had the desired effect and the frequent, systematic lemmatization errors and many NER errors were gone.

The TT is also available via the CLARIN Language Switchboard. We are currently working on additions to its toolbox that can be interesting to users outside CLARIN-DK. We are integrating the Stanford CoreNLP tools to handle Chinese, English, French, German, Hungarian, Italian, and Spanish texts even better than we already do. We are also continuously extending the TT's capabilities and improving its user interface. A recent example of the latter is the graphical representation of workflows. A TT workflow is structured as an Acyclic Directed Graph (DAG) with potentially a few tens of nodes. In the new graphical representation, each tool is represented by a numbered circle. Data streams are drawn as arrows between these circles, with the arrow head indicating the direction towards the output. Input and output are represented as circles with ·.· resp. .·. symbols (see figure 1 for an example). The TT consists of a toolbox and software that manages the tools and the metadata about the tools. The toolbox is in part filled with 3rd-party tools (UD-Pipe, CoreNLP, Tesseract and many others) and in part with our own tools. All software in the TT is free, open source and not depending on frameworks, DBMS'es, virtualization or new technologies with short expected lifespans. Although automated, the TT is not a black box; it can be opened and inspected. There is also an option to export a workflow as a list of command lines, so users with technical skills can recreate workflows even after the TT has gone out of service.

Compared to other workflow managers, the management of the tools by the TT is more advanced, since the TT computes workflows while other workflow managers depend on human workflow architects[11]

5 ParlaMint-DK 4.1

ParlaMint-DK 4.1 comprises the transcriptions of 398,610 speeches. The transcriptions were produced by the *Office of the Parliament Hansard*. They state that the transcribed speeches are reported literally, but with small editions whose aim is to adapt the speeches into a colloquial and syntactically coherent written language ensuring that the intentions of the speaker are clear. Factual errors and minor speech errors are corrected in the transcribed speeches. Moreover, punctuation marks have been added to the transcriptions, and spoken language characteristics, such as pauses, speech marks, hesitations, and self corrections have been left out (Hansen et al., 2018). One characteristic of the Danish parliament is that politicians must follow specific rules during the debates. The rules are spelled out in *The Standing Orders of the Danish Parliament*, which the Speaker (the chairman of the Parliament), the Chair, henceforth, enforces during the debates. According to these rules, expressions of approval or disapproval during the debates are considered disorderly, and the Chair can stop a politician who is judged to say something improper.

5.1 A quantitative analysis of the policy domains

In our analysis of policy domains in ParlaMint-DK 4.1, we removed the utterances of the Chair. These utterances have the same policy domain annotation as the speeches that are chaired in that section, if they occur under an agenda point with that policy domain. These utterances should not be considered when analyzing the content of policy domains since the Chair does not address the domains and only chairs the meetings, e.g. introducing the various speakers or enforcing the rules for the debates. The utterances by the Chair and speeches which do not address a policy domain[12] can be though interesting in studies

[11] Due to the complexity of the task, it would have been hard and time-consuming to implement the TT in main-stream programming languages such as Java, Python, or C++. We have used an internally developed programming language, Bracmat (https://github.com/BartJongejan/Bracmat), itself written in C, for all functionality not related to the handling of internet protocols. For the latter, we used Java and PHP.

[12] These speeches have no domain annotation.

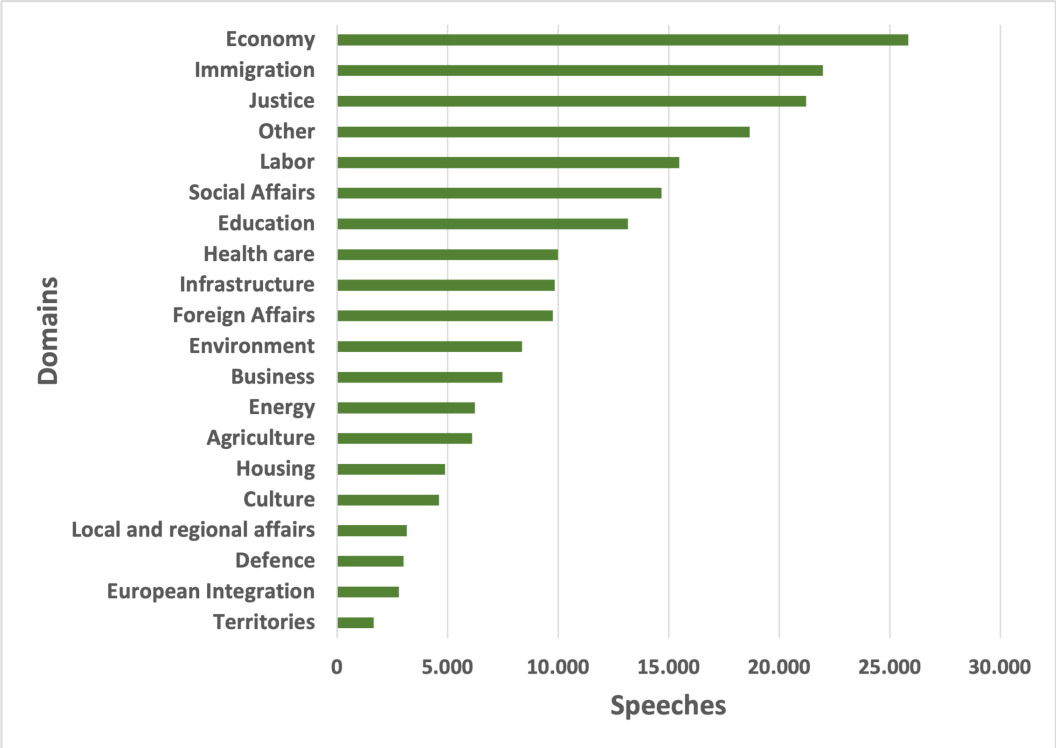


Figure 2: The distribution of policy domains.

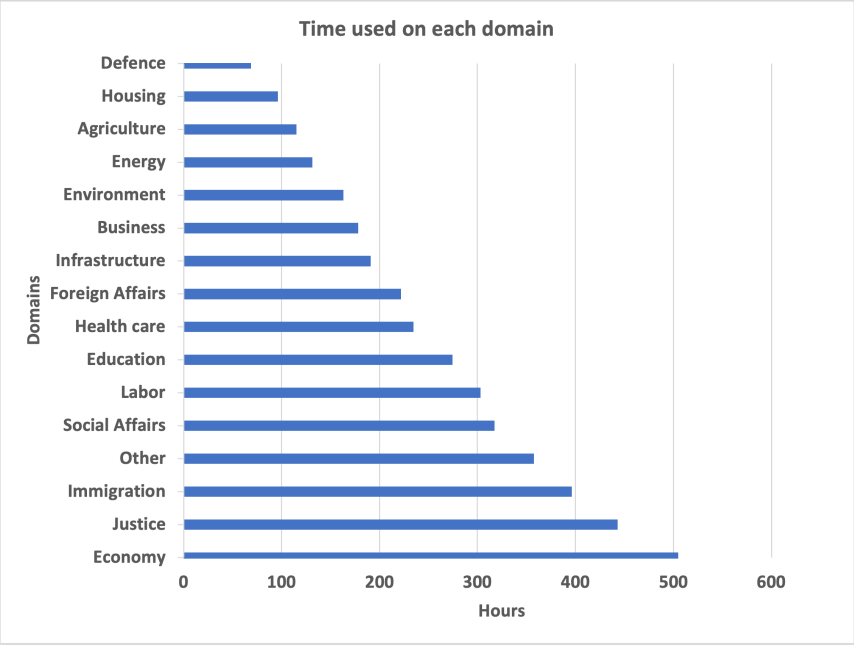


Figure 3: The Duration of speeches about policy domains.

addressing for instance the debates' dynamics. After removing the utterances of the Chair, we obtained 208,881 speeches with policy domains.

The distribution of the most frequently debated policy domains in ParlaMint-DK is given in figure 2. The policy domain that is most frequently addressed in the speeches is *Economy*, followed by *Immigra-*

tion, *Justice*, and *Labor*. The prominence of the *Economy* domain is not surprising, while the frequency of speeches about *Immigration* indicates the importance that this topic has had in Danish politics the past ten years. In one of our previous studies (Navarretta et al., 2022), we investigated how *Immigration* was addressed by seven Danish main parties not only in the speeches, but also in their manifestos.

The policy domain that is debated for the longest period of time is *Economy*, which relates to it being the most frequent speech domain. The second domain with respect to the duration of the speeches, which address it, is *Justice*, followed by *Immigration*, *Social Affairs* and *Labor*.

Approximately 18% of the speeches in ParlaMint-DK (38,425) are annotated with two policy domains. In Figure 4, the most frequently co-occurring domains in the corpus are displayed. They are the follow-

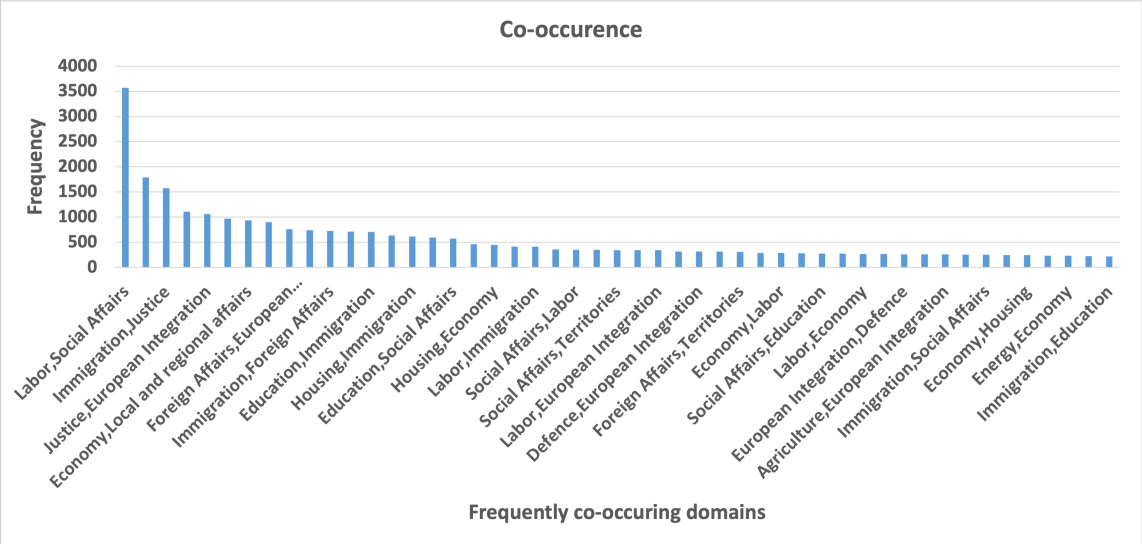


Figure 4: The Distribution of co-occurring policy domains.

ing: a) *Labor* and *Social Affairs*, b) *Immigration* and *Justice*, c) *Justice* and *European Integration*, d) *Economy* and *Regional affairs*.

6 Investigation of the policy domains most often addressed by women and men

6.1 Related studies

In this study, we used the annotations of policy domains in ParlaMint-DK 4.1 to investigate which areas were more often addressed by male vs. female politicians, and whether the most frequently debated domains by each of the two groups change over time. Gender differences in political speeches have been addressed in studies focusing on various aspects. For example, Paxton et al. (2007) analyze a number of articles describing the political participation of women in different countries. They conclude that even if all studies show that the number of female parliament members is low, the figures vary from country to country. The female parliament members are most numerous in the Scandinavian countries, while they are fewest in the Middle East. In 2005 their percentage was 38% in Scandinavia and 8% in the Middle East.

Dahllöf (2012) addresses the automatic identification of gender, age (young vs. old) and political affiliation (left or right wing) of politicians in the initial words of selected transcriptions of the Swedish parliament debates from 2003 to 2010. Words characterizing each class were used as features, and a support vector machine was applied to the data for classification. The accuracy for gender identification on balanced train and test sets was around 0.6 and gender identification was better for right-wing than for left-wing politicians. Mandravickaitė and Krilavičius (2017) compare the most frequent words in the transcripts of parliamentary speeches by female and male members of the Lithuanian parliament, and they apply hierarchical clustering to samples obtained on the basis of similarity measures. They find

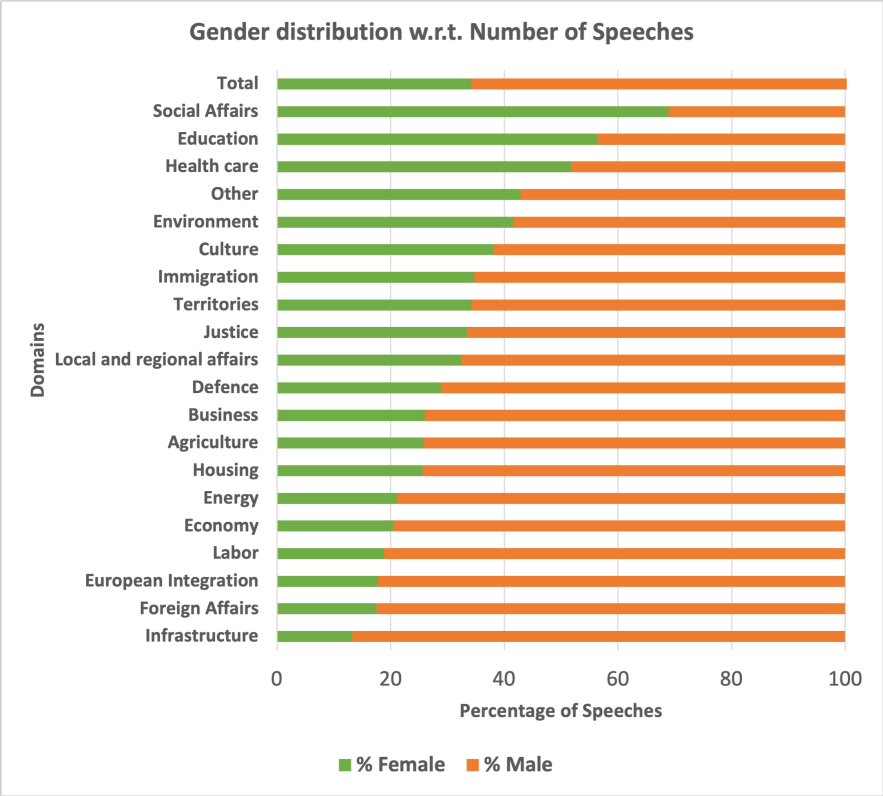


Figure 5: The distribution of speeches by female and male politicians.

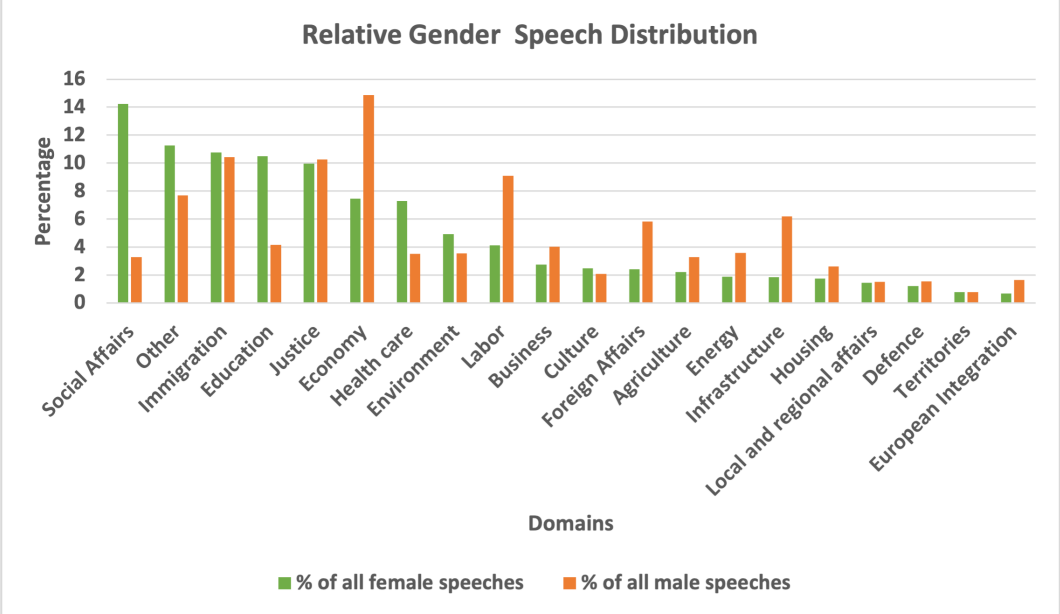


Figure 6: The relative distribution of speeches by female and male politicians.

differences in both the lexical items used by female and male politicians and in the stylometric figures for the speeches of the two groups.

Hansen et al. (2018) address gender differences in the Danish Parliament Corpus (2009-2017). They

find statistically significant differences in the number and duration of speeches by male and female politicians, and their results show that the role of politicians in their party influences their participation in the debates. Hansen et al. (2018) also find that female ministers talk more in periods when the prime minister is a female than they do when the prime minister is a male. This indicates that female politicians are inspired by the presence of a female head of government. The positive influence of female ministers on the active participation of female parliament members in the British debates is underlined in a study by Blumenau (2021).

Bäck and Debus (2019) study the Swedish parliamentary debates from 2002 to 2010 and find that female politicians take the floor less often than male politicians, and they talk less about "harder" policy areas such as energy, finance, macro economy, foreign affairs, and national security than their male colleagues.

Analyzing 200 debates of the UK's House of Commons between 1997 and 2006, Hargrave and Langengen (2021) find that female politicians refer more often to their personal experience, discuss policies in a more concrete way, and are less antagonistic than men in their speeches. Hargrave and Blumenau (2022) also find that the debating styles of the female members of the British parliament have changed over time (1997-2019) and that women have adopted stylistic traits, which traditionally have been considered masculine. The authors conclude that the division between female and male stylistic traits for politicians does not longer correspond to the reality and builds upon old stereotypes.

6.2 Our analyses

In the following, we investigate the frequency of speeches about the various policy domains addressed by female and male politicians in ParlaMint-Dk 4.1 and, inspired by Hargrave and Blumenau (2022), we also look at whether the frequency figures change over time. In the period covered by our corpus, the percentage of women in the Danish parliament is of 38.5%. More specifically, the percentage is 39.1% after the elections in 2011 and 2019, and 37.4% after the elections in 2015. The percentage of speeches by women in the same period is slightly lower (32.3%), which could be due to the fact that the majority of ministers have been men. Figure 5 shows the absolute frequency of speeches by women and men about each policy domain in ParlaMint-DK 4.1. The policy domains that are addressed by female politicians more frequently than by male politicians are *Social Affairs*, *Education*, and *Health Care*. The speeches regarding *Environment*, *Culture* and *Immigration* are also frequent. The policy domains that are debated by women less often than by men are *Labor*, *European Integration*, *Foreign Affairs* and *Infrastructure*.

According to the frequency numbers in figure 5, the women in the Danish parliament, similarly to the women in the Swedish parliament (Bäck & Debus, 2019), talk more often about "soft" policy areas than men. This was also found in the Danish speeches from 2009 to 2017 by Hansen et al. (2018). They note that women are often responsible for these areas in their parties and are therefore also assigned them as ministers when they are part of the government.

In figure 6, we show the frequency of speeches in each policy domain by female and male politicians relative to the number of speeches made by each group, respectively. According to figure 6, female politicians talk frequently about *Social Affairs*, *Immigration*, *Education* and *Justice*, while male politicians talk most often about *Economy*, *Immigration*, *Justice*, *Labor*, and *Infrastructure*. The figure also shows that *Immigration* and *Justice* are prominent topics in the speeches of both groups.

Figure 7 shows the policy domains that have frequently been discussed by female and male politicians during the two governments under the social-democrat Helle Thorning-Schmidt (2013-2014). She was the first female prime minister in Denmark. In her first government 34% of the ministers were female, while in her second government the percentage of female ministers was 47%. Under Helle Thorning-Schmidt's first government, the policy domain which is most frequently addressed by female politicians is *Education* followed by *Justice*, *Social Affairs*, and *Economy*. In the same period, the domains that are addressed by male politicians most frequently are *Economy*, *Justice*, and *Culture*. In Helle Thorning-Schmidt's second government, women speak frequently about *Justice*, *Social Affairs*, and *Education* while the domains most often addressed by men are *Economy*, *Labor* and *Justice*.

Figure 8 shows the frequency of speeches made by female and male politicians during the two govern-

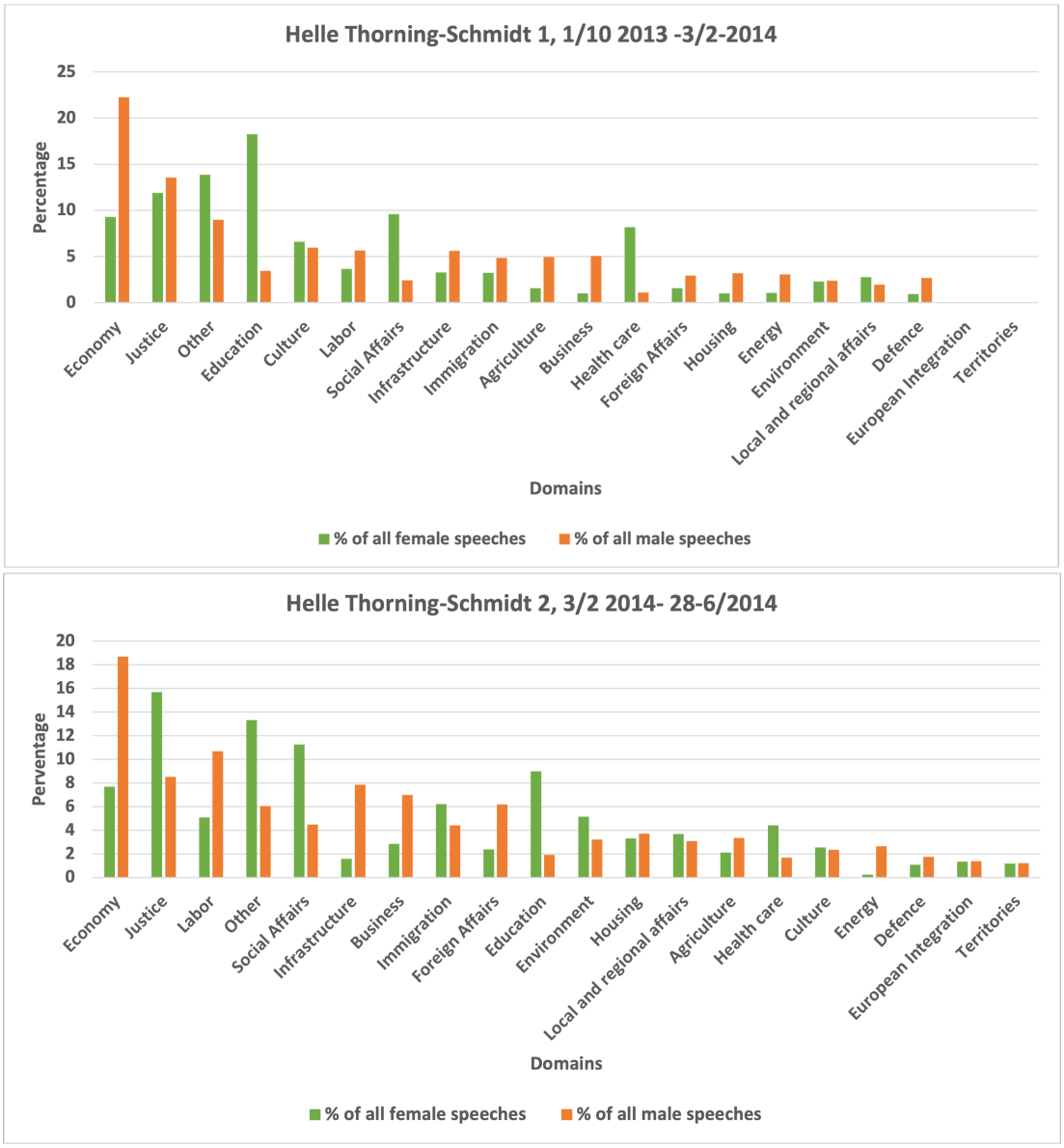


Figure 7: The relative distribution of speeches by female and male politicians under the first and second government under Helle Thorning-Schmidt.

ments under Lars Løkke Rasmussen (2014-2019), who was, at the time, the leader of the Liberal Party. In the first of these periods, the percentage of female ministers was 29% while in the second period the percentage was 41%. Figure 8 shows that women under the first Lars Løkke Rasmussen government talk most frequently about *Immigration*, while the second most frequently addressed area is *Justice*. The high number of speeches on *Immigration* is a consequence of the immigration crisis in Europe in September 2015, as also discussed in Navarretta et al. (2022). Male politicians also speak more often about *Immigration* in this period than earlier, but the most frequently addressed speech by them is still *Economy*. In the third government under Lars Løkke Rasmussen, women more often talk about *Social Affairs*, *Health Care*, and *Immigration* than men, while the domain most often addressed by men is *Economy* followed by *Labor* and *Justice*. These are the same domains that were most often debated by male politicians under Helle Thorning-Schmidt's second government. The difference between the number of speeches about *Immigration* by women and men in this period is remarkable and this can be due to the fact that

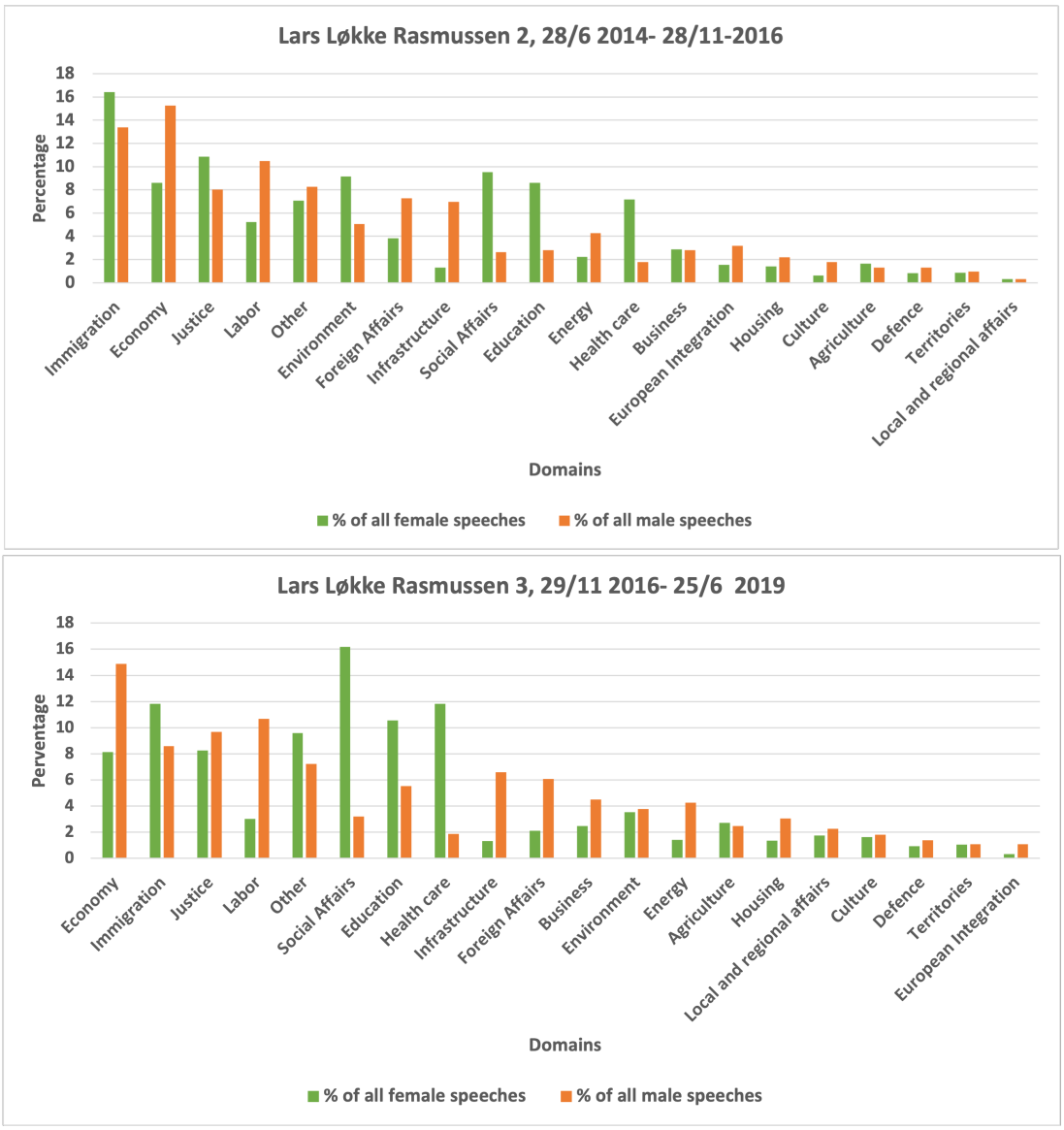


Figure 8: The distribution of speeches by female and male politicians during the second and third government under Lars Løkke Rasmussen.

the minister of immigration and integration in this period was a woman, Inger Støjberg.

In figure 9 the frequency of speeches by female and male politicians during the first government of the social democrat Mette Frederiksen (2019-2022) is shown. 29% of the ministers were women during this government. The policy domains most often addressed by female politicians between 2019 and 2022 are *Social Affairs*, *Education*, and *Immigration* while men most often spoke about *Immigration*, *Economy*, and *Justice*. Figure 9 clearly indicates that *Immigration* is a policy domain that continues to be a central topic in the debates. In this period, men talk more about *Health care* than women, and this can be due to the fact that the minister of health during the COVID crisis was a man, Magnus Heunicke.

The analysis of the frequency of speeches about various policy domains addressed by women and men in the Danish parliament shows that the frequency vary over time. Not surprisingly, it can be influenced by external events such as the immigration crisis in 2015 and the COVID-19 pandemic. In general, however, female politicians address ”soft areas” more often than their male colleagues. Another factor to

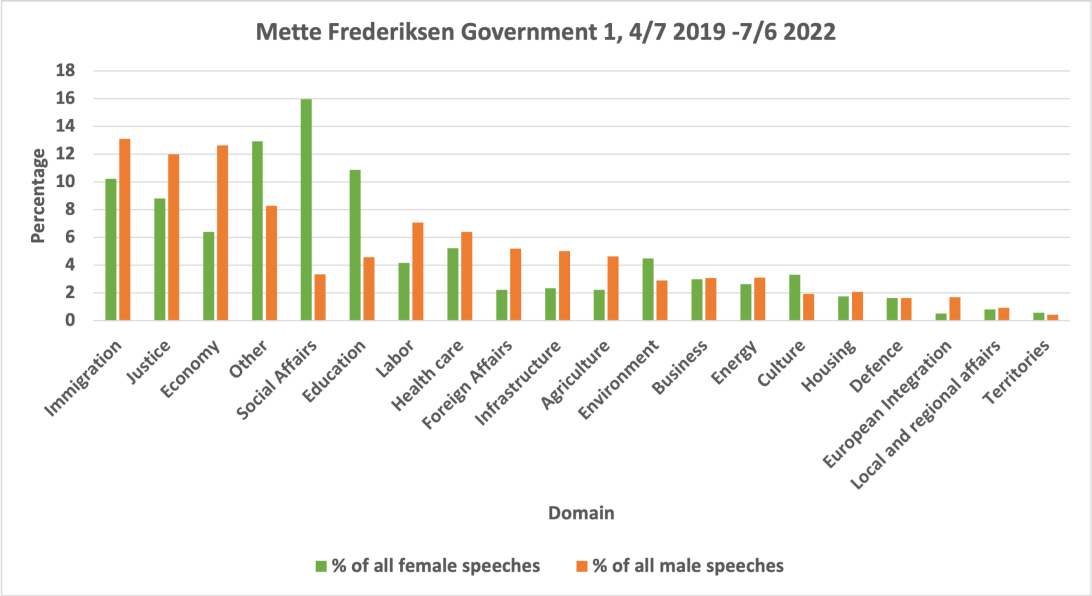


Figure 9: The relative distribution of speeches by female and male politicians during the first Mette Frederiksen's government.

take into consideration is the gender of the minister for a given domain since ministers often talk more than other politicians about that domain. It is also interesting that women often talked slightly more about *Environment* than men while men talked much more about *Energy*, even though the two domains are strongly related. The fact that female politicians often talk about "soft" areas also reflects the job situation in Denmark since there are more women employed as social workers, nurses, and educators than men.

7 Conclusions and future work

In the paper, we have presented ParlaMint-DK 4.1 with policy domain annotations and improved lemma and NER annotations. We have also presented a quantitative analysis of the most frequently addressed policy domains in the corpus and of the most frequently co-occurring domains in it. Quantitative analyses of the most frequently addressed policy domains in the speeches made by female and male politicians under the five governments covered in the corpus are also presented. Our analyses confirm that Danish female politicians talk about "soft" domains such as *Social Affairs*, *Education*, and *Health care* more often than their male colleagues do, and men talk about "hard" areas such as *Economy* and *Labor* much more frequently than women. However, events such as the immigration crisis in 2015 and the COVID-19 pandemic from 2020 can change the relevance of specific policy areas. The gender distribution of the speeches on policy domains also reflects the gender of the ministers for those areas.

In the future, we will investigate how well the policy domain annotations can be used for automatically annotating new speeches, and to determine how different parties have dealt with the same domain over time.

References

Bäck, H., & Debus, M. (2019). When do women speak? a comparative analysis of the role of gender in legislative debates. *Political Studies*, *67*(3), 576–596.

Baumgartner, F. R., Jones, B. D., & Wilkerson, J. (2011). Comparative Studies of Policy Dynamics. *Comparative Political Studies*, *44*(8), 947–972. https://doi.org/10.1177/0010414011405160

Blumenau, J. (2021). The effects of female leadership on women's voice in political debate. *British Journal of Political Science*, *51*(2), 750–771.

Dahllöf, M. (2012). Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches — A comparative study of classifiability. *Literary and linguistic computing*, *27*(2), 139–153.

Erjavec, T., Kopp, M., Ljubešic, N., Kuzman, T., Rayson, P., Osenova, P., Ogrodniczuk, M., Çöltekin, Ç., Koržinek, D., Meden, K., Skubic, J., Rupnik, P., Agnoloni, T., Aires, J., Barkarson, S., Bartolini, R., Bel, N., Pérez, M. C., Darġis, R., … Fišer, D. (2024). Parlamint ii: Advancing comparable parliamentary corpora across europe. *Language Resources and Evaluation*. https://doi.org/https://link.springer.com/article/10.1007/s10579-024-09798-w

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agerri, R., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M. d. M., … Fišer, D. (2024). Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.1 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1911

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M. d. M., Calzada Pérez, M., … Fišer, D. (2024). Multilingual comparable corpora of parliamentary debates ParlaMint 4.1 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1912

Hansen, D. H., & Navarretta, C. (2021). The Danish Parliament Corpus 2009 - 2017, v2, w. subject annotation [CLARIN-DK-UCPH Centre Repository]. http://hdl.handle.net/20.500.12115/44

Hansen, D., Navarretta, C., & Offersgaard, L. (2018). A Pilot Gender Study of the Danish Parliament Corpus. *Proceedings of the ParlaClarin workshop at the Eleventh International Conference on Language Resources and Evaluation*, 67–72.

Hansen, D., Navarretta, C., Offersgaard, L., & Wedekind, J. (2019). Towards the Automatic Classification of Speech Subjects in the Danish Parliament Corpus [https://cst.dk/DHN2019/DHN2019.html]. *CEUR Workshop Proceedings*, *2364*, 166–174.

Hargrave, L., & Blumenau, J. (2022). No Longer Conforming to Stereotypes? Gender, Political Style and Parliamentary Debate in the UK. *British Journal of Political Science*, *52*(4), 1584–1601. https://doi.org/10.1017/S0007123421000648

Hargrave, L., & Langengen, T. (2021). The Gendered Debate: Do Men and Women Communicate Differently in the House of Commons? *Politics & Gender*, *17*(4), 580–606. https://doi.org/10.1017/S1743923X20000100

Ivanusch, C. (2024). Where do parties talk about what? party issue salience across communication channels. *West European Politics*, *0*(0), 1–27. https://doi.org/10.1080/01402382.2024.2322234

Jongejan, B. (2016). Implementation of a Workflow Management System for Non-Expert Users. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 101–108.

Jongejan, B., Hansen, D., & Navarretta, C. (2021). Enhancing CLARIN-DK Resources While Building the Danish ParlaMint Corpus. *CLARIN Annual Conference 2021 Proceedings*, 70–73.

Mandravickaitė, J., & Krilavičius, T. (2017, April). Stylometric analysis of parliamentary speeches: Gender dimension. In T. Erjavec, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger, & R. Yangarber (Eds.), *Proceedings of the 6th workshop on Balto-Slavic natural language processing* (pp. 102–107). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1416

Merz, N., Regel, S., & Lewandowski, J. (2016a). The manifesto corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, *3*(2), 2053168016643346. https://doi.org/10.1177/2053168016643346

Merz, N., Regel, S., & Lewandowski, J. (2016b). The manifesto corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, *3*(2), 2053168016643346. https://doi.org/10.1177/2053168016643346

Navarretta, C., Haltrup Hansen, D., & Jongejan, B. (2022, June). Immigration in the manifestos and parliament speeches of Danish left and right wing parties between 2009 and 2020. In D. Fišer, M. Eskevich, J. Lenardič, & F. de Jong (Eds.), *Proceedings of the LREC22 Workshop ParlaCLARIN III* (pp. 71–80). ELRA. https://aclanthology.org/2022.parlaclarin-1.11

Navarretta, C., & Hansen, D. H. (2023, September). According to BERTopic, what do Danish parties debate on when they address energy and environment? In C. Klamm, G. Lapesa, V. Gold, T. Gessler, & S. P. Ponzetto (Eds.), *Proceedings of the 3rd KONVENS Workshop on Computational Linguistics for the Political and Social Sciences* (pp. 59–68). Association for Computational Lingustics. https://aclanthology.org/2023.cpss-1.6

Navarretta, C., & Hansen, D. H. (2022). The Subject Annotations of the Danish Parliament Corpus (2009-2017) - Evaluated with Automatic Multi-label Classification. *Proceedings of LREC 2022*.

Paxton, P., Kunovich, S., & Hughes, M. M. (2007). Gender in Politics. *Annual Review of Sociology*, *33*(1), 263–284. https://doi.org/10.1146/annurev.soc.33.040406.131651

Ristilä, A., & Elo, K. (2023). Observing political and societal changes in Finnish parliamentary speech data, 1980–2010, with topic modelling. *Parliaments, Estates and Representation*, 1–28.

Yu, H.-C., Rehbein, I., & Ponzetto, S. P. (2023). Policy domain prediction from party manifestos with adapters and knowledge enhanced transformers. *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, 229–244.

Zirn, C., Glavas, G., Nanni, F., Eichorst, J., & Stuckenschmidt, H. (2016). Classifying topics and detecting topic shifts in political manifestos. *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016)*, 88–93.