# Towards the Automatic Classification of Speech Subjects in the Danish Parliament Corpus

Dorte Haltrup Hansen[1], Costanza Navarretta[2[0000-0002-4242-9249]], Lene Offersgaard[3] and Jürgen Wedekind[4[0000-0002-0759-6009]]

Centre for Language Technology, Department of Nordic Studies and Linguistics,
University of Copenhagen, Denmark
[1]dorteh@hum.ku.dk
[2]costanza@hum.ku.dk
[3]leneo@hum.ku.dk
[4]jwedekind@hum.ku.dk

**Abstract.** This paper addresses the semi-automatic subject area annotation of the Danish Parliament Corpus 2009-2017 in order to construct a gold standard corpus for automatic classification. The corpus consists of the transcriptions of the speeches in the Danish parliamentary meetings. In our annotation work, we mainly use subject categories proposed by Danish scholars in political sciences. The relevant subjects areas of the speeches have been manually annotated using the titles of the agendas items for the parliamentary meetings and then the subjects areas have been assigned to the corresponding speeches. Some subjects co-occur in the agendas, since they are often debated at the same time. The fact that the same speech can belong to more subject areas is further analysed. Currently, more than 29,000 speeches have been classified using the titles of the agenda items. Different evaluation strategies have been applied. We also describe automatic classification experiments on a subset of the corpus using feature extracted with NLP techniques. The best results (96% F-score) were obtained using features extracted from the agenda items. These results indicate that the gold standard corpus and agenda items can be used for automatically classify parliamentary debates with high accuracy.

**Keywords:** Parliamentary Debates, Subject Classification, Gold Standard Corpus.

## 1 Introduction

The transcriptions of parliamentary debates (Hansards) are available in many countries, and researchers from different disciplines, such as political science, linguistics and computational linguistics, have examined them in a variety of contexts. A classification of the speeches into subject areas is certainly the most basic technique for analysing their content. However, it is beneficiary for practical applications, such as search optimisation, and it is useful for more sophisticated analyses, e.g. of the tone in the debates

on immigration, a topic that can be found in the debates on *taxpaying*, *unemployment* and *foreign policy*.

In this paper, we report on the creation of a gold standard corpus consisting of the speeches from the Danish Parliament Corpus 2009-2017 classified by subject areas as well as on experiments to classify the debates in subject areas using some basic NLP methods and machine learning techniques. The corpus contains Hansards of the sittings in the Chamber of the Danish Parliament and has recently been made available as a collection through the Danish CLARIN research infrastructure [1]. The corpus consists of approx. 41 million running words and 182,192 speeches[1]. Information about the sittings, the name of speakers, their party, the time of the speeches, and the title of agenda items are provided in the corpus. However, the corpus does not contain information about the subjects of neither the speeches nor the agenda items.

The paper is organised as follows. Section 2 describes related work. In Section 3, we account for the adopted classification scheme and, in Section 4, we present the method used for constructing the gold standard corpus. The analysis and evaluation of the annotated corpus are provided in Section 5. In Section 6, we report on the automatic classification experiments and their results. The final session concludes and suggests future research.

## 2 Related Work

Political domains have been categorised according to various schemes depending on the task. The *Comparative Manifesto Project*, CMP[2] [2] and the *Comparative Agendas Project*, CAP[3] developed two domain classification systems for comparative studies.

In the Comparative Manifesto Project, party election programmes (manifestos) were annotated using 560 categories in order to determine the policy preferences of political parties. The Comparative Agendas Project classifies policy activities around the world according to 21 general categories and 192 sub-categories.

The *Danish Policy Agendas Project* at the University of Aarhus is manually annotating parliamentary activities in the Danish Parliament from 1953 and onward[4]. The data comprise e.g. policy bills, legislative hearings, parliamentary debates, and speeches by the prime minister. The project uses the CAP coding scheme. Recently, experiments with semi-automatic classification have been carried out on Danish city council agendas [3]. In the experiments, a Naive Bayes classifier was applied on a manually annotated corpus on the basis of the council agendas, and then the agendas were lemmatised and used as testing material. The best classification results on some of the data were 75%.

The CAP and CMP classifications are too complex and too broad for the scope of subjects addressed in the Danish Parliament. Moreover, the CAP scheme was originally

[1] Hansen, Dorte Haltrup, 2018, The Danish Parliament Corpus 2009 - 2017, v1, CLARIN-DK-UCPH Centre Repository, http://hdl.handle.net/20.500.12115/8.

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

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

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

proposed to describe the policy areas of the US Congress, and although it has been extended and revised to be more widely applicable, it still suffers from this bias. Some of the major categories of CAP such as *400 General Agriculture* with sub-categories comprising e.g. *403 Food Inspection and Safety* and *408 Fisheries and Fishing* describe perfectly subjects debated in the Danish Parliament, while other categories e.g. *23 Cultural Policy Issues* do not, and its sub-category *sport* is grouped under *15 Industrial and commercial policy* (*1526 Sport and Gambling*).

An alternative approach mentioned in [5, 6] is to use the names of the ministries as categories for classifying data related to German politics. Since the ministries' names and areas of responsibility can change from one election period to another [4] this is not a viable solution for our data. Therefore, Zirn [5] uses a scheme based on the responsibilities of committees to which the agenda items are assigned. Her classification scheme thus corresponds to the 22 committees of the German Parliament. Inspired by Zirn's work, we have developed a classification scheme that reflects the responsibility areas in the committees of the Danish parliament. In this way, we can also connect the domain and the spokespersons for those areas. We show that spokespersons for a particular domain are, not surprisingly, the most speaking politicians about that subject area and related subjects.

Automatic text classification of large collections of texts is a natural language processing subarea, which has developed extensively the past decades. It consists of assigning predefined classes to text documents by training machine learning algorithms on features extracted from the texts with various NLP techniques. Examples of training features are the number of words in the texts, the length of their sentences, bag of words (bow), lemmas, lemmas of particular word classes, TF*IDF values (term frequency* inverse document frequency) [8,9]. In three-fold sentiment classification of various datasets researchers have obtained between 63.9% and 98.6% accuracy depending on the data [10].

## 3 Classification Scheme

Scholars of political science in Denmark have suggested to categorise the subject areas of Danish politics into the following 23 main classes[5]: *Agriculture*, *Business*, *Culture*, *Defence*, *Economy*, *Education*, *Energy*, *Environment*, *European Integration*, *Foreign Affairs*, *Government Operations*, *Health Care*, *Housing*, *Immigration*, *Justice*, *Labour*, *Local and Regional Affairs*, *Personal Rights*, *Politics*, *Social Affairs*, *Technology*, *Territories* and *Transportation*.

We use these subject areas for our annotations and group the responsibility areas (spokesmanships) under them. The responsibility areas for 2015-17 were found on the Danish parliament website and have been used in the present work. The three categories *Government Operations*, *Politics* and *Personal Rights* have been omitted since they deal mostly with meta-content and not with specific political domains. If speeches on

[5] Mail communication with Prof. Christoffer Green-Pedersen, Political Science Department, University of Aarhus, about the CAP classification and its Danish version,

these occur, they will be categorised as *Other*. We merged the categories *Technology* and *Transport* into the category *Infrastructure*. This category also comprises IT.

In Table 1, we show the Danish specific subject areas and the corresponding CAP classes as well as the spokesmanship related to them in the Danish parliament. The latter information is based on the spokesmanships in the period 2015-17, The table shows that the Danish subject areas match the main CAP codes fairly well. Exceptions are *Local and Regional Affairs* and *Housing* which map to the same code in CAP (14) but are distinct areas in Danish politics. The same holds for *Foreign Affairs* and *European Integration*, which map to the same major subject area in CAP, but are distinguished areas in the Danish parliament. Other problematic cases are e.g.: *Consumer Policy*, which is in Denmark normally categorised under *Agriculture* together with *Food* while in CAP is categorised under (15) *General Banking, Finance, and Domestic Commerce*, and the subject area *Culture*, which in Denmark normally comprises *Sports* while the latter subject is categorised differently in CAP.

**Table 1. :** Danish subject area classification of parliamentary speeches based on spokesmanships (2015-17) and the corresponding classes in CAP.

| Chosen subject areas | Spokesmanships in the Danish parliament | Corresponding CAP subject areas |
|---|---|---|
| **Economy** | Finance | 1 Domestic Macroeconomic Issues |
| | Fiscal Affairs | 1 Domestic Macroeconomic Issues |
| **Health Care** | Psychiatry | 3 Health |
| | Health | 3 Health |
| **Agriculture** | Animal Welfare | 4 Agriculture |
| | Fisheries | 4 Agriculture |
| | Food | 4 Agriculture |
| | Agriculture | 4 Agriculture |
| | Consumer Policy | 1525 Consumer Policy |
| **Labour** | Labour market | 5 Labour and Employment |
| **Education** | Higher Education and Research | 6 Education |
| | Education | 6 Education |
| **Environment** | Environment | 7 Environment |
| **Energy** | Energy | 8 Energy |
| | Climate | 705 Air and noise pollution, climate change and climate policies |
| **Immigration** | Immigration and Integration | 9 Immigration and Refugee Issues |
| | Alien Affairs | 9 Immigration and Refugee Issues |
| | Naturalization | 9 Immigration and Refugee Issues |
| **Infrastructure** | Transportation | 10 Transportation |
| | IT | 17 Space, Science, Technology, and Communications |
| | Media | 17 Space, Science, Technology, and Communications |

| | | |
|---|---|---|
| **Justice** | Legal affairs | 12 Law, Crime, and Family Issues |
| | Constitutional Matters | 20 Government issues |
| **Social Affairs** | Children | 13 Social Welfare |
| | Family | 13 Social Welfare |
| | Disabled | 13 Social Welfare |
| | Social services | 13 Social Welfare |
| | Senior citizens | 13 Social Welfare |
| | Gender equality | 2 Civil Rights, Minority Issues, and Civil Liberties |
| **Housing** | Housing | 14 Community Development and Housing Issues |
| **Local and Regional Affairs** | Rural Districts and Islands | 14 Community Development and Housing Issues |
| | Municipal Affairs | 2001 Local Government Issues |
| **Business** | Trade and Industry | 15 Industrial and commercial policy |
| **Defence** | Defence | 16 Defence |
| **Foreign Affairs** | Foreign Affairs | 19 International Affairs and Foreign Aid |
| | Development Cooperation | 19 International Affairs and Foreign Aid |
| **European Integration** | EU | 1910 International Affairs and Foreign Aid |
| **Territories** | Faroe Islands | 2105 Dependencies and Territorial Issues |
| | Greenland | 2105 Dependencies and Territorial Issues |
| **Culture** | Cultural Affairs | 23 Cultural Policy Issues |
| | Ecclesiastical Affairs | 210 The Danish national church |
| | Sport | 1526 Sport and Gambling |

## 4 Method

As already mentioned, the Danish Parliament Corpus 2009-2017 does not contain information on subject areas or the committees responsible for them. Therefore, we use the title of the agenda items for the meetings as an indication of the subject areas of the speeches of these meetings. In total there are 182,192 speeches under 7,336 different agenda items.

We extracted the titles of the agenda items and normalized them, e.g. "First reading of bill 193: XYZ" has been normalized to XYZ. This resulted in 6,722 different agenda titles. For each title, up to three subjects from the chosen classification scheme were coded manually. For example, for the title Tax on saturated fat in food, Agriculture (comprising Food) has been chosen as the primary subject, while Economy (comprising Tax) was annotated as the secondary subject. The subject area classification of the agenda items were added automatically to the speeches in the time slots allocated to them. The process was repeated until there were more than 1,000 examples (speeches) for each of the 19 subject areas. One exception is the subject area Territories that was not assigned to so many speeches. The annotated corpus comprises currently more than

29,000 speeches. We are now using the annotations as a training and test data set for the automatic subject area classification.

## 5 Evaluation and Analysis

Of the 6,722 agenda titles, 1,079 were manually marked for subject areas by one annotator and then corrected by a second one. In 9% of the annotations, the second annotators proposed another subject area or a different ranking of the two or three subjects proposed by the first annotator. They discussed the disagreement cases and in some cases involved a third annotator, producing an agreed-upon classification. The 29,249 classified speeches contain over 615,000 tokens. Out of these speeches, 16,743 (57%) are annotated with only one subject area, 11,455 (39%) with two subject areas, and 1,051 (3.6%) with three subject areas.

As an initial evaluation of the classification, we extracted the speakers talking in each subject area in the 18,473 speeches that were classified under a single subject, and we marked the spokespersons for those areas in the period 2015-2017. We found out that the spokespersons of the subject areas and related areas are in the majority of the cases the most frequent speakers for those areas in that period. However, because politicians can be spokespersons of more than one subject area and spokespersons can change area of responsibility during the same election period, this information can only be an approximate indication that the speeches have been classified correctly.

## 6 Automatic Classification of the Speeches: First Experiments

In this section, we describe experiments for automatically categorising the parliament speeches into the given domains using supervised classification. That is, we use a training set $T = \{(s_1, c_1), \ldots, (s_n, c_m)\}$ consisting of speeches that have each been hand-labelled with the appropriate class and the task is to find a classifier and a model that are capable to map new speeches *s* to their correct class *c*. In our experiments, we have used a subset of the annotated speeches. The subset consists of 19,676 speeches, belonging to 18 classes (we excluded the class *Territories*, because of the low number of speeches). To each class belong between 900 and 1180 speeches. All the speeches in the chosen subset consist of at least 5 words. The speeches and the titles of the agenda items have been Part of Speech (PoS) tagged and lemmatised.

We have extracted the lemmas of nouns and proper nouns from the speeches and removed numbers and prepositions from the lemmas of the titles of the agenda items. The training features we have tested are the following: Bow of the agenda item titles (selected lemma types), bow of the lemmas of the speeches, TF*IDF of the lemmas of the speeches, the TF*IDF of the n-grams of the speeches' lemmas (up to trigrams), and of the characters (chars) of the lemmas (up to 4-grams), the TF*IDF of the nominal

lemmas, and information about the speakers. The latter information comprises the gender, role (minister, member) and the party of the speakers. Combinations of some of the features were also tested (see Table 2).

The Python scikit-learn package was used for the experiments. The speeches were randomised and then the data were divided in a training set (60% of the data), a test set and evaluation test (20% of the data each). We trained and tested the features on the training and test set respectively, and finally we tested the obtained models on the evaluation data. The scikit-learn multinomial Naïve Bayes and support vector machine classifier were applied. The Naïve Bayes classifier obtained the best results. This is probably because Naïve Bayes also gives good results on sparse data, and some of the speeches consist only of few words. In Table 2, we report the results obtained by this classifier on different features in terms of Precision, Recall and weighted F-score.

**Table 2.** Classification features and results

| No. | Features | Precision | Recall | F-score |
|---|---|---|---|---|
| **1** | agenda item titles – bow | 0.96 | 0.96 | 0.96 |
| **2** | all lemmas – bow | 0.75 | 0.74 | 0.74 |
| **3** | all lemmas - TF*IDF | 0.73 | 0.7 | 0.7 |
| **4** | TF*IDF n-grams (lemmas) | 0.7 | 0.69 | 0.69 |
| **5** | TF*IDF n-grams (chars) | 0.65 | 0.48 | 0.45 |
| **6** | nominal lemmas - TF*IDF | 0.73 | 0.72 | 0.72 |
| **7** | all lemmas – bow and nominal lemmas TF*IDF | 0.75 | 0.73 | 0.73 |
| **8** | agenda item titles and lemma – bow | 0.91 | 0.91 | 0.91 |
| **9** | lemma bow, TF*IDF nominal, speaker features | 0.75 | 0.74 | 0.73 |

The results show that training the classifier on the bow of the extracted agenda item titles (1) gives the best performance. This indicates that the manually classification is consistently made. When involving data from the speeches the performance drops from F-score 0.96 to F-score 0.91 for the best results (8). Furthermore, all results are significantly better than those obtained by training a majority classifier (F-score 0.06) and by chance.

A first analysis of the automatically annotated data indicates that most classification errors when using features extracted from the speeches were due to the limited length of some of the speeches, and by the fact that we did not remove from the transcriptions comments added by the transcribers concerning e.g. the poor quality of the audio file.

## 7 Discussion and Concluding Remarks

In this paper, we have presented the construction and evaluation of a subject area annotated subcorpus of the Danish Parliamentary Corpus (2009-2017). The coding scheme mainly follows the classification used by Danish scholars in politics and this classification can be mapped into the international CAP system.

The annotation of the speeches was performed by manually annotating the titles of the agenda items, and then automatically propagating their subject areas to the speeches under them. Most of the annotated speeches are classified under a single class, but in some cases, the annotator classified the speeches under two classes (39% of the speeches) or under three (3.6% of the cases). The manual evaluation and the comparison of the speeches' subject areas and the speakers' role in the parliament committees indicates that the classification is appropriate.

The automatic classification experiments, which we performed on part of the gold standard corpus show that training a multinomial Naive Bayes classifier on bow extracted from the agenda item titles results in an F-score of 0.96, which is extremely good. In a similar task on debates in municipalities, [3] obtained much poorer results. This might have been caused by the fact that in those meetings the agenda items titles were not assigned as consistently as in the case of the Danish Parliament.

The results obtained with features extracted automatically from the speeches also indicate that the parliament sessions follow quite precisely the pre-defined agendas. However, using these features as training data is only useful when the speeches have a certain length. The uneven length of the speeches is also problematic for the use of machine learning algorithms that require large amounts of data. However, linguistic features could be useful when looking for more specific topics in the general subject areas.

In the future, we will test whether we can predict the second subject area when relevant, and we will experiment with other features, and classifiers applied on more data, also selecting speeches that contain at least 50 or 100 words.

## References


1. Hansen, D. H., Navarretta, C., Offersgaard, L.: A Pilot Gender Study of the Danish Parliament Corpus. In: the proceedings of the workshop ParlaCLARIN at 11th edition of the Language Resources and Evaluation Conference, Japan (2018).
2. Budge, I., Klingemann, H. D., Volkens, A., Bara, J., Tanenbaum, E.: Mapping Policy Preferences. Estimates for Parties, Electors, and Governments 1945-1998. Oxford University Press (2001)
3. Loftis, M. W. and Mortensen, P. B.: Collaborating with the Machines: A Hybrid Method for Classifying Policy Documents. In The Political Studies Journal, https://doi.org/10.1111/psj.12245 (2018)
4. Mortensen, P. B., and Green-Pedersen, C.: Institutional Effects of Changes in Political Attention: Explaining Organizational Changes in the Top Bureaucracy. In Journal of Public Administration Research and Theory, 25(1), 165–189, https://doi.org/10.1093/jopart/muu030 . (2015)
5. Zirn, C.: Analyzing Positions and Topics in Political Discussions of the German Bundestag. In proceedings of the ACL Student Research Workshop, pp. 26 -33 (2014)
6. Zirn, C., Glavas, G., Nanni, F., Eichorst, J., & Stuckenschmidt, H.: 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, Dubrovnik, Croatia, 14–16 July (2016).

7. Spärck Jones, K.: A Statistical Interpretation of Term Specificity and Its Application in Retrieval. Journal of Documentation. 28: 11–21 (1972)
8. Allahyari M., Pouriyeh S., Assefi M., Safaei S., Trippe E.D, Gutierrez J. B., Kochut K. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. CoRR. vol. abs/1707.02919 (2017)
9. Korde V., Mahender C.N.. Text classification and classifier:A survey International Journal of Artificial Intelligence & Applications, Vol.3:2, pp. 85-99, March. (2012).
10. Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of Tricks for Efficient Text Classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 2, Short Papers, pages 427–431, Valencia, Spain, April 3-7 (2017).