Are Topics changing over time in the German Bundestag?

Can we extract main topics over time in the German Bundestag out of the available minutes of plenary proceedings? Computer-based text analysis could be a suitable method for this task, especially topic modeling.
In this research four separate corpuses containing minutes of five electoral terms each are analyzed.

Alina Benz
Master of Business Administration, M.Sc.
Supplementary Study Program Data Literacy

Seminar Introduction to Computer-Based Text Analysis
Professorship for Data Modelling &
Interdisciplinary Knowledge Generation

Introduction

The German Bundestag held its first meeting on 07.09.1949. Today in 2022 we are in the 20th electoral term since the establishment of the Federal Republic. Since then all minutes of plenary proceedings are now digitalized and made public. Lots of German history is written in these files. But which topics can we extract of these using topic modeling?

Research Question

Do the relevant topics extracted with STM change over time in the German Bundestag minutes?

Data

4388 meeting minutes in .xml format for the 20 legislative periods until the 44th meeting on 23.06.2022. [1]
The .txt files for the analysis of the legislative periods 1-18 are taken from [2]. Periods 19 and 20 has been transformed with a self created python code.

Method

Text Analysis of minutes using quanteda, and topic modeling using stm package of R in RStudio[2].
The system is a Windows PC with 8GB RAM and Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz. Because of the limited computing power four different corpuses are analyzed.

Limitations

- Computing power
- Preprocessing steps
- STM as a topic modeling method

Preprocessing

- Tokens are without German stop words, numbers, punctuations, and symbols.
- An iteratively created dictionary is removed from the tokens list containing e.g., names, activities, and functions in the German Bundestag plenums. These did not change like the check of presence, elections, and the procedure itself.
- The minimum character is set to 4.
- Tokens collocations with a list of Ministries and phrases iteratively collected with the kwic() function and the top 15 words of the TFIDF matrix.
- The document feature matrices are trimmed using relative pruning between 0.05 and 0.99
- creating the stm matrix out of the dfm matrix

Results

- Using STM extract several topics over time, but the highest γ have topics, leading to different terms than the TF-IDF word clouds. e.g., in Fig1 "saar" is a key word but in the analysis the matching topic 9 has the least γ value
- The defined relevant topics have peak points in specific points of time
- Topics regarding Treasury and Social Security are in all four clusters relevant topics.
- Cluster 1 combines topics regarding agricultural policy and war victims
- In cluster 2 topics regarding housing shortage and Education are getting more important
- in cluster 3 German Reunion and the Introduction of the EURO are extracted
- In cluster 4 Brexit, organ transplantation regulations, the financial crisis and the Covid-19 pandemic as well as the Afghanistan and Ukraine war are relevant topics.

Analysis

A first impression of topics give the top TF-IDF terms

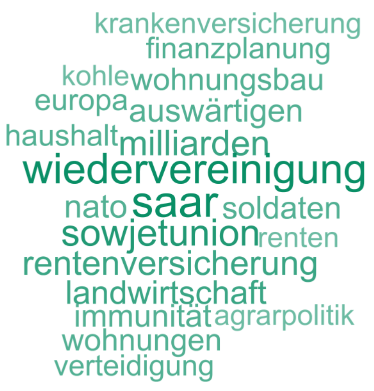

Fig1: Electoral term 1 - 5

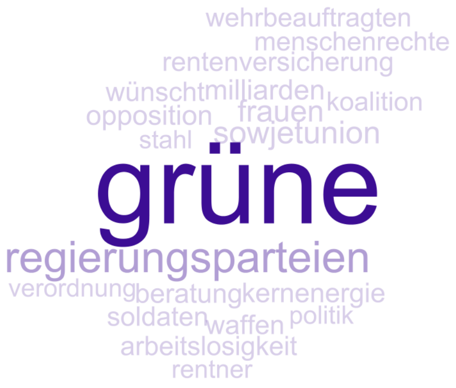

Fig2: Electoral term 6 - 10

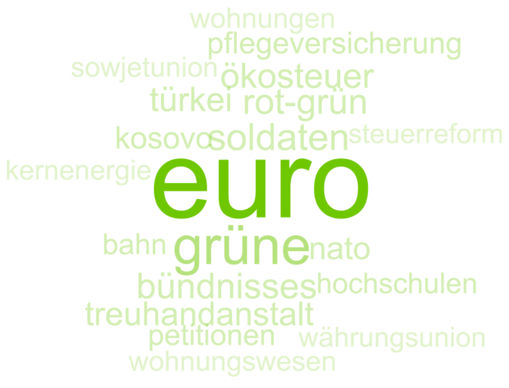

Fig3: Electoral term 11 - 15

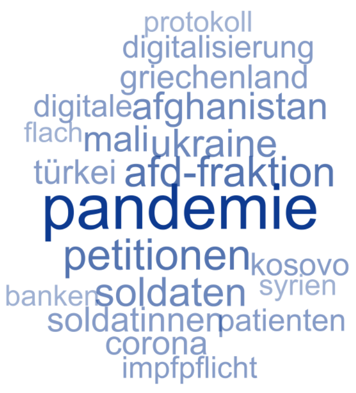

Fig4: Electoral term 16 - 20

Structural Topic Modeling with STM is used because it is possible to model which variable influences the prevalence of topics, here the legislative periods. Also, it is a non-random initialization and a robust method. Using Word-topic matrix and document Topic Matrix the models are evaluated [3].

For topic modeling, k as the number of topics has to be defined. Therefore the tradeoff between semantic coherence and exclusivity is calculated for different numbers of k.
Here the mean for key values is taken.
And the topic model is set to k=30

"A topic model with 30 topics, 1122 documents and a 80426 word dictionary."

Finding "good" topics in the final model tradeoff matrix
If a topic is exclusive and semantic coherent it is a defining topic of the corpus.

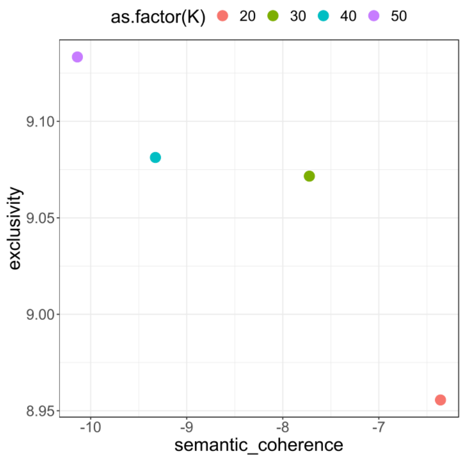

Fig5: Tradeoff Matrix Electoral term 1 - 5 per k

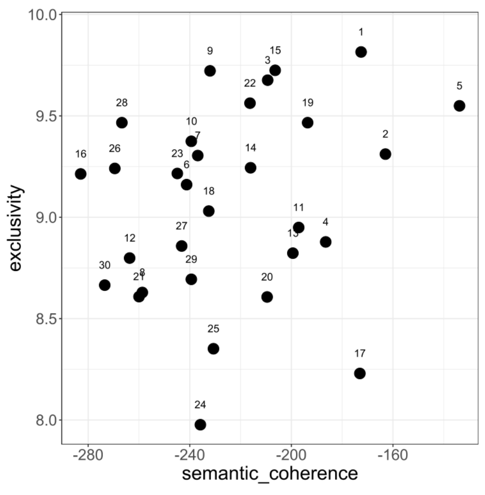

Fig6: Tradeoff Matrix Electoral term 1 - 5 (k=30)

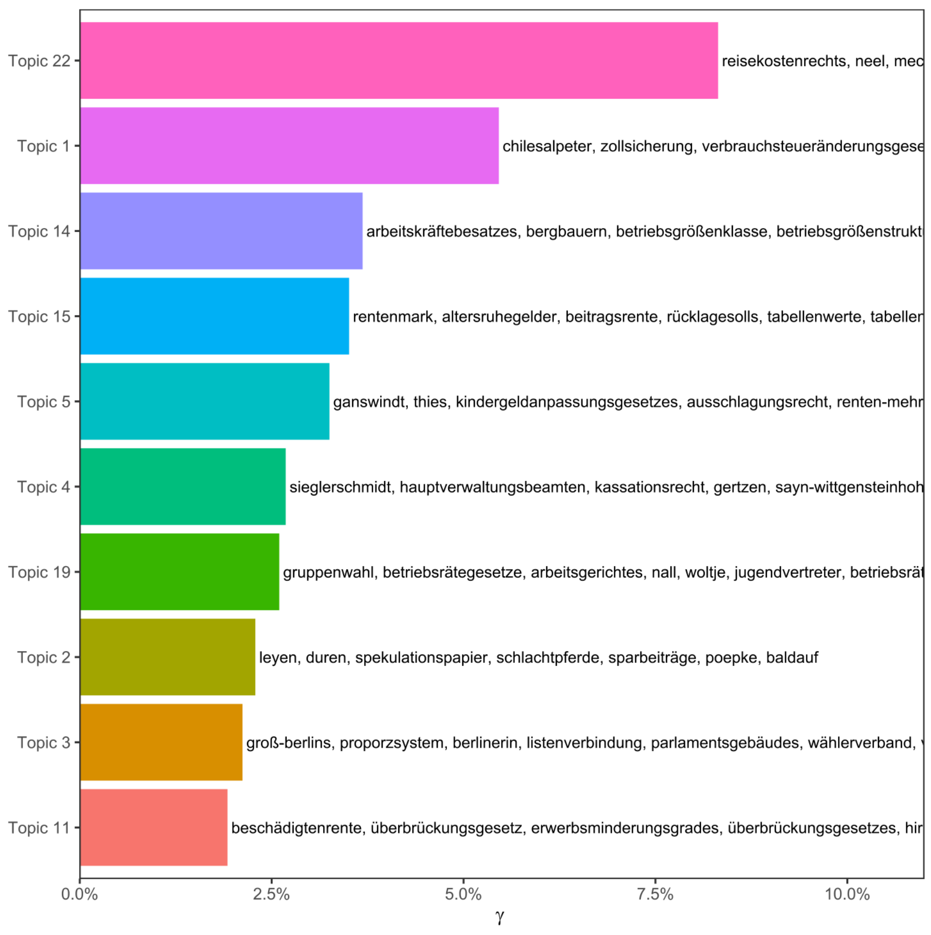

Fig7: The per-document-per-topic probabilities (γ) for relevant topics in Electoral term 1 - 5 and the key words by the lift-score

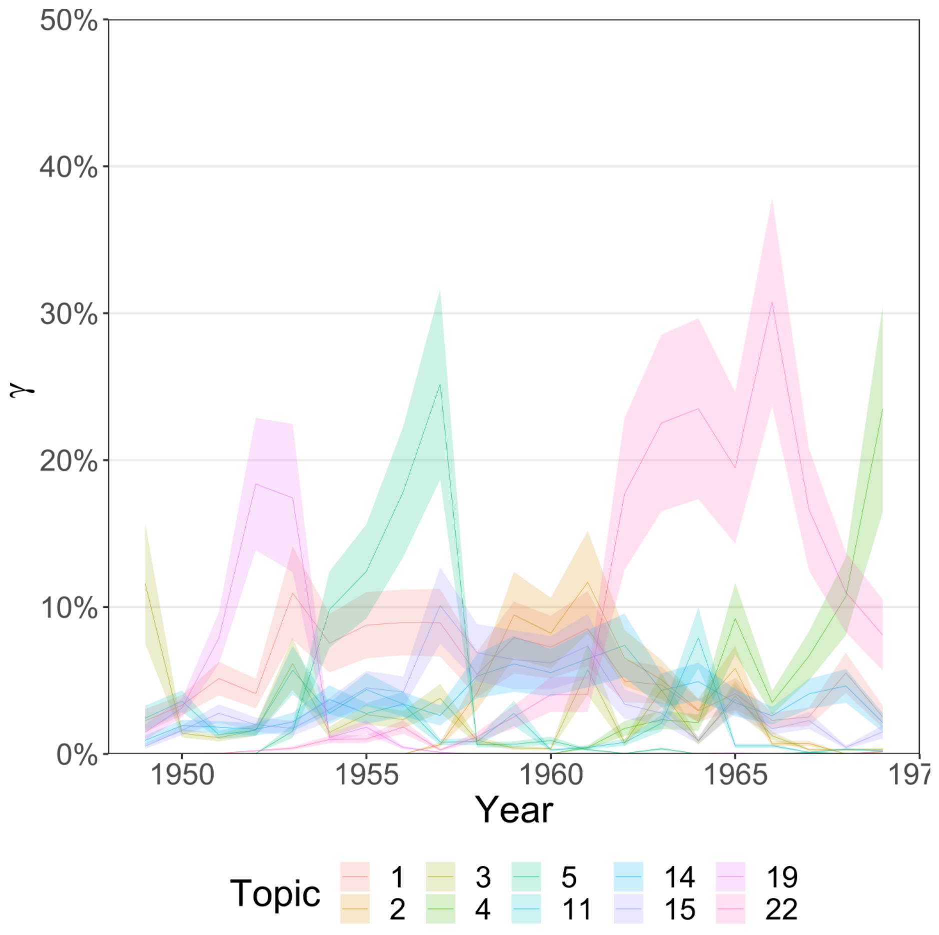

Fig8: Fluctuation of γ for relevant topics in Electoral term 1 - 5

Conclusion

In this broad overview it is difficult to tell which topics are in the clusters' focus because of the four different corpuses.
With regard to the research question, it can be confirmed through the analysis that the relevant topics extracted through topic modeling have changed over time.

Future Work

- Analysis by elected German chancellor or president
- Analysis by different Government and Opposition parties
- Topic Modeling using different algorithms, e.g., LDA

References

[1] https://www.bundestag.de/services/opendata
[2] Fobbe, Sean. (2021). Corpus der Plenarprotokolle des Deutschen Bundestages (CPP-BT) (2021-02-17) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4542662
[3] https://bookdown.org/joone/ComputationalMethods/topicmodeling.html