DS4DH-II Final Project: Bundestag Protocol Corpus

Analyzing Political Discourse over 6 Legislative Periods

1998-2021

Open Discourse Project

- *“The German Bundestag is the parliamentary heart of democracy, with public deliberations mandated by the constitution.”* [Open Discourse Website]
- Since 1949, a stenographic report documenting every word spoken in each plenary session has been produced.
- Protocols are available in Plain-Text, XML, PDF format
- Goal: Make these transcripts **widely available** and ready for digital research, by **unifying all formats into a coherent representation**, and aggregate all available metadata.

Dataset

- We will use a subset of the dataset covering the 6 legislative periods from 1998-2021.
- Added faction names.

	id	session	electoralTerm	firstName	lastName	politicianId	speechContent	factionId	documentUrl	positionShort	positionLong	date	factionName
583702	588546	11	14	reinhard	loske	11003176	Herr Präsident! Liebe Kolleginnen und Kollegen...	3	https://dip21.bundestag.de/dip21/btp/14/14011.pdf	Member of Parliament		1998-12-03	Bündnis 90/Die Grünen
584009	589046	14	14	Norbert	Wieczorek	11002502	Ja, bitte.\n\n	23	https://dip21.bundestag.de/dip21/btp/14/14014.pdf	Member of Parliament		1998-12-10	Sozialdemokratische Partei Deutschlands
584190	588157	10	14	petra	bläss	11000189	Eine zweite Zusatzfrage.\n\n	-1	https://dip21.bundestag.de/dip21/btp/14/14010.pdf	Presidium of Parliament	vizepräsidentin	1998-12-02	not found
584871	589161	14	14	Hans-Peter	Friedrich	-1	Herr\nPräsident! Meine Damen und Herren! Ich g...	4	https://dip21.bundestag.de/dip21/btp/14/14014.pdf	Member of Parliament		1998-12-10	Christlich Demokratische Union Deutschlands/Ch...
585348	589230	14	14	Regina	Schmidt-Zadel	11002026	Frau Präsidentin!\nMeine Damen und Herren!	23	https://dip21.bundestag.de/dip21/btp/14/14014.pdf	Member of Parliament		1998-12-10	Sozialdemokratische Partei Deutschlands

Legislative Periods

Bundestag	Periode	Mandate	CDU/CSU	SPD	FDP	Grüne[1]	PDS/Linke	DP	AfD	Sonstige
14. Bundestag	1998–2002	669	245	298	43	47	36	–	–	–
15. Bundestag	2002–2005	603	248	251	47	55	2	–	–	–
16. Bundestag	2005–2009	614	226	222	61	51	54	–	–	–
17. Bundestag	2009–2013	622	239	146	93	68	76	–	–	–
18. Bundestag	2013–2017	631	311	193	–	63	64	–	–	–
19. Bundestag	2017–2021	709	246	153	80	67	69	–	92	2

Task I: Data Cleaning & Preprocessing

Protocols capture everything, not only speeches but also interjections.

Task: Remove or flag everything that is not a proper speech.
Task: Analyze the data carefully and remove everything else that is not the content of the speech itself.

Task II: Exploration & Descriptive Statistics

Get to know the dataset and the metadata.

Task: Analyze & visualize basic text properties (length, etc.)
Task: Analyze & visualize metadata (speaker distributions, etc.)
Task: Combine different features to gain meaningful insights!

Task III: Content Exploration

Explore the content of the covered period with Topic Modeling.

Task: Perform topic modeling to explore recurring topics or themes in the dataset.
Task: Use topic distributions to identify topics that become more or less prominent over the 20 years.
Tasks: Combine Topics with metadata and determine if and how topics are connected to certain parties.

Task IV: Speaker Networks

Use Graph Analysis to explore the connections between speakers.

Task: Create a speaker network by connecting all speakers who spoke at the same debate.
Task: Identify the most prominent members by analyzing the speaker network.

Task V: Sentiment Analysis

Political parties try to evoke emotions to amplify their messages.

Task: Use the provided sentiment dictionary (or another tool) to create sentiment scores for speeches. Aggregate scores over parties and periods, and analyze and visualize your findings.
Tasks: Find out if there are **significant** differences in sentiment scores between parties.

Task VI: Text classification as analytical tool

Enrich your analysis of the difference between factions using text classification.

Task: Train logistic regression model to perform party classification
Task: Analyze the trained model and determine how it differentiates between the parties. Visualize your findings!

Task VII: Go your own way

After completing Tasks I-VI and gaining familiarity with the dataset, it's time to define and answer your research question. If you do not find your own RQ, do some research to find inspiration

Task: Define a clear and concise research question.
Task: Identify an appropriate methodological approach to address your question.
Task: Present your research question and findings coherently and structured.

Feel free to extend the dataset!

Task VII: Places that might inspire you

- NLP + Computational Social Sciences Workshops Proceedings: https://aclanthology.org/venues/nlpcss/
- Journal of Computational Social Science https://link.springer.com/journal/42001
- Journal of Cultural Analytics: https://culturalanalytics.org
- Journal of Computational Literary Studies https://jcls.io
- The Stone and the Shell -- Ted Underwood's blog: https://tedunderwood.com
- … many more!

Report

- Summarize all the steps your project in a short report (7-10 pages).
- Tell us **what you did** and **why you did it**, but not *how you implemented it*!
- For each task feel free to add visualization if you find them important to explain your findings.
- If necessary, you can also use tables to summarize your data.
- Your report can be written in German, or English.
- Use the template provided in WueCampus.

- Hand in your report, your code, and any additional data you used via WueCampus. Zip everything together into single archive.

Sources

- WueCampus Dataset Download
- WueCampus Sentiment Dictionary Download

- Open Discourse Project Website: https://opendiscourse.de/
- Open Discourse Project Code: https://github.com/open-discourse/open-discourse
- Open Discourse Project Data Repository: https://doi.org/10.7910/DVN/FIKIBO
- Sentiment Dictionary:
  R. Remus, U. Quasthoff & G. Heyer: SentiWS - a Publicly Available German-language Resource for Sentiment Analysis. In: Proceedings of the 7th International Language Ressources and Evaluation (LREC'10), 2010