DORTMUND CENTER FOR DATA-BASED MEDIA ANALYSIS

technische universität dortmund

fakultät statistik

Kai-Robin Lange
Carsten Jentsch

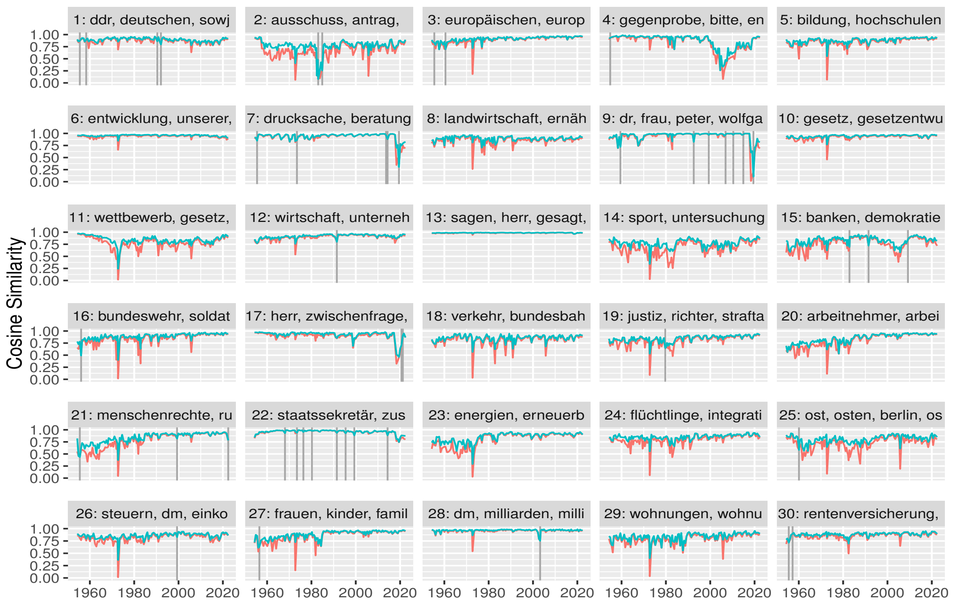
SpeakGer

A meta-data enriched speech corpus of German state and federal parliaments

Download instructions: `github.com/K-RLange/SpeakGer`

1. Motivation

Last year at CPSS: Our "Zeitenwenden" paper [2] analyzes a Bundestag corpus for **changes in German political discourse**. How to improve? We want **more data, from all federal state parliaments!** Previous data sets like [7], [5] & [1] either only use Bundestag-speeches or do not contain information about the party of the speaker!



2. The Sources

Table 1: Sources of the pdf-files containing the speeches

Parliament (English name)	Legislative period	Source
Baden-Württemberg (Baden-Wuerttemberg)	12-17 1-11	Landtag von Baden-Württemberg Württembergische Landesbibliothek
Bayern (Bavaria)	1-18	Bayrischer Landtag
Berlin	12-19 6-11 1-5	Abgeordnetenhaus Berlin Zentral- und Landesbibliothek Berlin Zentral- und Landesbibliothek Berlin
Brandenburg	8-10 1-7	Landtag Brandenburg Parlamentsspiegel
Bremen	18-20 7-17	Bremische Bürgerschaft Parlamentsspiegel
Bundestag	1-20	Deutscher Bundestag
Hamburg	20-22 6-19	Hamburgerische Bürgerschaft Parlamentsspiegel
Hessen	1-20	Hessischer Landtag
Mecklenburg-Vorpommern (Mecklenburg-Western Pomerania)	1-8	Landtag Mecklenburg-Vorpommern
Niedersachsen (Lower Saxony)	17-18 8-16	Landtag Niedersachsen Parlamentsspiegel
Nordrhein-Westfalen (North Rhine-Westphalia)	1-18	Landtag Nordrhein-Westfalen
Rheinland-Pfalz (Rhineland Palatinate)	1-18	Landtag Rheinland-Pfalz
Saarland	14-17 7-13	Landtag des Saarlandes Parlamentsspiegel
Sachsen (Saxony)	1-8	Sächsischer Landtag
Sachsen-Anhalt (Saxony-Anhalt)	6-8 1-5	Landtag von Sachsen-Anhalt Parlamentsspiegel
Schleswig-Holstein	1-20	Schleswig-Holsteiner Landtag
Thüringen (Thuringia)	4-7 1-4	Thüringer Landtag Parlamentsspiegel

3. The meta-data

From the parliaments: Dates of each parliamentary session

Wikipedia

Name Full name

Born Birth year

Consituency Constituency of the mp for that legislative period

Party Party of the mp in that legislative period

Wikidata

Last Name Last name

Born Exact birth date

Religion Religion or worldview

SexOrGender The gender or sex of the mp. Wikidata does not differentiate between the two.

Occupation Additional occupations apart from "politician"

AbgeordnetenwatchID ID on a platform for more transparent politics in Germany

The meta-data available differs based on the available information on the respective websites.

4. Getting the speeches

(i) Read text from all pdf-files (use OCR if neccessary)

(ii) Identify first and last speech

(iii) Cut off the table of contents and appendix

(iv) Look for brackets – indicators for comments

(v) Look for the word "Präsident" – indicating that the chair of the session is speaking

(vi) Look an mp's name and a colon – indicator for the start of a speech by an mp

(vii) Apply spelling correction for OCR speeches – both corrected and original speeches are published

5. Stats and descriptive analysis

Table 2: Rounded Number of speeches by party and federal state in 100. In total: 15,452,593 speeches.

State	Interjection	Chair	CDU	CSU	SPD	Linke	FDP	AfD	Die Grünen
Bayern	5926	1430	0	2637	1113	0	116	20	170
Berlin	4817	505	1395	0	2127	523	560	158	490
Brandenburg	1650	509	415	0	991	379	63	120	102
Bremen	5351	684	778	0	2280	119	334	3	486
Bundestag	21540	4997	7568	1917	7625	963	3354	285	2018
Hamburg	5670	1135	1448	0	2377	215	417	64	123
Niedersachsen	5039	1610	1951	0	1935	87	667	0	644
Saarland	2349	596	971	0	934	63	145	26	63
Sachsen	2307	853	1249	0	405	573	136	167	213
Sachsen-Anhalt	1978	658	831	0	436	432	100	171	235
Schleswig-Holstein	6499	1583	2162	0	2181	43	622	0	435
Thüringen	2671	867	1060	0	405	609	127	130	181

6. References

[1] G. Abrami, M. Bagci, L. Hammerla, and A. Mehler. German parliamentary corpus (gerparcor). In *Proceedings of the Language Resources and Evaluation Conference*, pages 1900–1906, Marseille, France, June 2022. European Language Resources Association.

[2] K.-R. Lange, J. Rieger, N. Benner, and C. Jentsch. Zeitenwenden: Detecting changes in the German political discourse. pages 47–53, 2022.

[3] K.-R. Lange, J. Rieger, and C. Jentsch. Lex2Sent: A bagging approach to unsupervised sentiment analysis, Sept. 2022.

[4] C. Rauh. Validating a sentiment dictionary for German political language—a workbench note. *Journal of Information Technology & Politics*, 15(4):319–343, Oct. 2018.

[5] C. Rauh and J. Schwalbach. The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies, Mar. 2020.

[6] J. Rieger, C. Jentsch, and J. Rahnenführer. RollingLDA: An update algorithm of Latent Dirichlet Allocation to construct consistent time series from textual data. In *Findings Proceedings of the 2021 EMNLP-Conference*, pages 2337–2347. ACL, 2021.

[7] T. Walter, C. Kirschner, S. Eger, G. Glavaš, A. Lauscher, and S. P. Ponzetto. Diachronic Analysis of German Parliamentary Proceedings: Ideological Shifts through the Lens of Political Biases, Aug. 2021.