Open Political Corpora: Structuring, Searching, and Analyzing Political Text Collections with PoliCorp

Nina Smirnova and **Muhammad Ahsan Shahid** and **Philipp Mayr**
GESIS – Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8, 50667 Cologne
{nina.smirnova, ahsan.shahid, philipp.mayr}@gesis.org

Abstract

In this work, we present PoliCorp, a web portal designed to facilitate the search and analysis of political text corpora. PoliCorp provides researchers with access to rich textual data, enabling in-depth analysis of parliamentary discourse over time. The platform currently features a collection of transcripts from debates in the German parliament, spanning 76 years of proceedings. With the advanced search functionality, researchers can apply logical operations to combine or exclude search criteria, making it easier to filter through vast amounts of parliamentary debate data. The search can be customised by combining multiple fields and applying logical operators to uncover complex patterns and insights within the data. Additional data processing steps were performed to enable web-based search and incorporate extra features. A key feature that differentiates PoliCorp is its intuitive web-based interface that enables users to query processed political texts without requiring programming skills. The user-friendly platform allows for the creation of custom subcorpora via search parameters, which can be freely downloaded in JSON format for further analysis.

1 Introduction

Parliamentary debates offer a broad variety of topics for academic exploration, serving as an effective instrument for agenda-setting and influencing political power. By examining parliamentary speeches, researchers can uncover the implicit programmatic and ideological positions political parties hold. Recent studies include investigations into gender dynamics and examinations of negativity levels inferred from interjections (Ash et al., 2025) and sentiment and negativity analysis (Jenny et al., 2021). Additional efforts focus on developing automated topic modeling (Watanabe and Zhou, 2022) and discursive framing approaches within legislative settings (Reinig et al., 2024).

Parliamentary data from many countries is openly available online through governmental initiatives and open data platforms. Examples include the Open Data portal of the German parliament (Bundestag)[1], the Austrian parliament transcripts[2], and Hansard, the official record of the UK Parliament[3]. In addition to raw data, some initiatives provide preprocessed and linguistically annotated corpora, e.g., the Polish Parliamentary Corpus[4], GermaParl, a linguistically annotated corpus of German Bundestag plenary debates (Blaette, 2021, 2017), and DutchParl, which contains documents from the parliaments of the Netherlands, Flanders, and Belgium in Dutch (Marx and Schuth, 2010). Several platforms offer web-based search interfaces for parliamentary records, such as the polit-X portal[5], which includes transcripts from the German Bundestag and state parliaments; the StateParl portal[6], which covers debates from the 16 German state parliaments; and the Italian parliamentary transcripts portal[7].

Although these resources provide comprehensive collections, they often present certain limitations. Raw data typically requires significant preprocessing before it can be analyzed, and many annotated datasets are distributed as complete corpora, necessitating advanced analytical skills to extract specific information (Marx and Schuth, 2010; Blaette, 2017). Furthermore, some data sets are embedded within specialized software environments, which require a knowledge of particular programming languages for effective use (Blaette, 2021, 2020). Although certain platforms, such as polit-X, offer user-friendly online interfaces for data explo-

[1] https://www.bundestag.de/services/opendata
[2] https://www.data.gv.at/en/
[3] https://hansard.parliament.uk/
[4] https://clip.ipipan.waw.pl/PPC
[5] https://polit-x.de/
[6] https://stateparl.de/
[7] https://accademiadellacrusca.it/it/contenuti/discorsi-parlamentari/23591

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 983–992

ration, they often operate on a subscription basis, limiting open access. Overall, the lack of standardized, machine-readable annotations for parliamentary speeches limits the potential for quantitative research in this area (Wissik, 2021).

To overcome these challenges, we introduce the Pollux Political Corpora (PoliCorp) platform[8], an advanced resource that offers researchers open, structured, and searchable access to processed political corpora. The platform currently contains a collection of transcripts of Bundestag debates, spanning 76 years of parliamentary debates – from September 1949 to July 2025. The platform is designed for political scientists, interdisciplinary researchers, and others engaged in the analysis of parliamentary discourse.

Additional data processing steps were performed to enable web-based search and incorporate supplementary features. For instance, PoliCorp allows users to perform targeted searches not only within speeches but also across "interjections" and "calls to order". Calls to order can serve as indicators for analyzing incivility in parliamentary discourse, and offer a unique perspective on political polarization (Jenny et al., 2021), and are therefore of particular interest for political research. Moreover, analysis of calls to order as markers of disruptive language is a novel approach to studies of parliamentary corpora, going beyond traditional sentiment or stance analysis. Interjections constitute a key resource for understanding democratic processes, uncovering hidden power dynamics, and examining conflicts within parliament (Ilie, 2015; Truan, 2017).

A key feature that differentiates PoliCorp is its intuitive web-based interface that enables users to query processed political texts without requiring programming skills[9]. The user-friendly platform allows the creation of custom subcorpora via search parameters, which can be freely downloaded in JSON format for further analysis. The portal's content is distributed under the CLARIN PUB+BY+NC+SA license.

Background on Parliamentary Discourse

The legislative period is a specific period of time for which a parliament is elected. In Germany, it spans four years. Within each legislative period, the work of the German Bundestag is organized through plenary sessions, which follow a set procedure. The session is presided over by a member of the Bundestag Presidium, comprising the president and two secretaries. The president moderates the session, ensuring procedural compliance and managing the allocation of speaking time. Speaking time is distributed proportionally among parliamentary groups, reflecting their relative representation in the parliament. Each session usually comprises several agendas, dedicated to a specific topic.

If the speaker or an attending parliamentary member violates parliamentary order, i.e., interruptions, the president is authorized to issue a formal call to order[10]. Calls to order were analysed by (Jenny et al., 2021) from a perspective of negativity analysis. Figure 3 demonstrates an example of a call to order during a parliamentary session from the PoliCorp interface.

Interruptions or interjections during a speech may include spontaneous remarks or attempts to insert commentary by other members of parliament[11]. Interjections are widely examined in political science research. Research on interjections has largely addressed their pragmatic, rhetorical, and disruptive functions (Truan, 2017; Shenhav, 2008). Many scholarly works have further investigated interjections in relation to gendered dynamics of interjections, including gendered patterns of attacks, interruptions, and participation in legislative debates across different contexts (van Dijk and Poljak, 2025; Ash et al., 2025; Vallejo Vera and Gómez Vidal, 2022; Och, 2020; Poljak, 2022; Vallejo Vera and Gómez Vidal, 2022; Miller and Sutherland, 2023). Additional lines of inquiry have explored interjections in relation to expertise, seniority, and affiliation dynamics (Diener, 2025), cross-cultural variation in speech styles (Isosävi et al., 2025), and broader patterns of interruption behavior (Poljak, 2023).

2 Backend Implementation

Raw parliamentary speeches up to September 7, 2021, were sourced from the GermaParl corpus (Blaette, 2017), a comprehensive linguistic dataset curated by the PolMine project[12]. GermaParl covers transcripts of parliamentary debates from September 7, 1949, to September 7, 2021, and comprises of 958,100 speech contributions. Raw par-

[8] https://demo-pollux.gesis.org/

[9] A demonstration of the search functionality is available at the following link: https://youtu.be/KplgIZRVVwQ

[10] https://www.bundestag.de/services/glossar/glossar/O/ordnungsruf-869614

[11] https://de.wiktionary.org/wiki/Zwischenruf

[12] https://polmine.github.io/

liamentary speeches published after September 7, 2021, were sourced from the Bundestag Open Data project[13]. The raw GermaParl data is available for download in a GitHub repository[14]. The records from the Bundestag Open Data project were retrieved using an API.

The raw data underwent a series of processing stages, as illustrated in Figure 1. The full processed dataset, currently hosted by PoliCorp, comprises 1,035,744 speech contributions[15]. In the final stage, the processed data is indexed using Elasticsearch (version 8.12.1), facilitating efficient retrieval of user-specific information via web-interface.

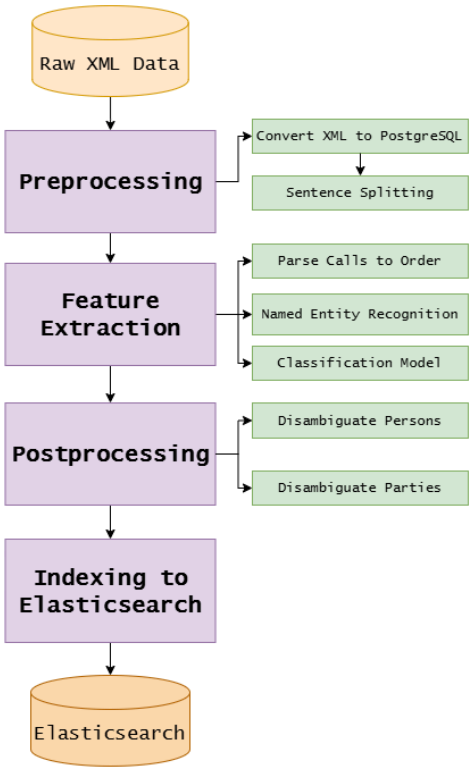

Figure 1: Data processing workflow for PoliCorp

2.1 Preprocessing

The Bundestag and GermaParl data were originally available in XML format, although with different structural schemas (see Appendix A). To address this issue, we developed two distinct conversion pipelines to map each XML structure into a unified database schema. The unified data schema comprises information about the agenda, speaker, and a corresponding speech contribution, as Figure 8 demonstrates. After the conversion process, the textual content of individual speech contributions was segmented into sentences.

2.2 Feature Extraction

Bundestag specifically marks interjections in their raw data (as Figures 11 and 12 demonstrate). Conversely, calls to order are not explicitly indicated in the data. Our objective was also to identify sentences that include calls to order. This procedure is regulated and involves the use of specific phrases. In the first step, we manually reviewed a part of the dataset containing only the speeches of the session's president. Based on this review, we developed a set of rules to identify calls to order, as illustrated in Figure 2. Following, we applied these rules to analyze only the speeches given by the session's president to detect instances of calls to order. Figure 3 shows an example of a marked call to order in the PoliCorp interface[16].

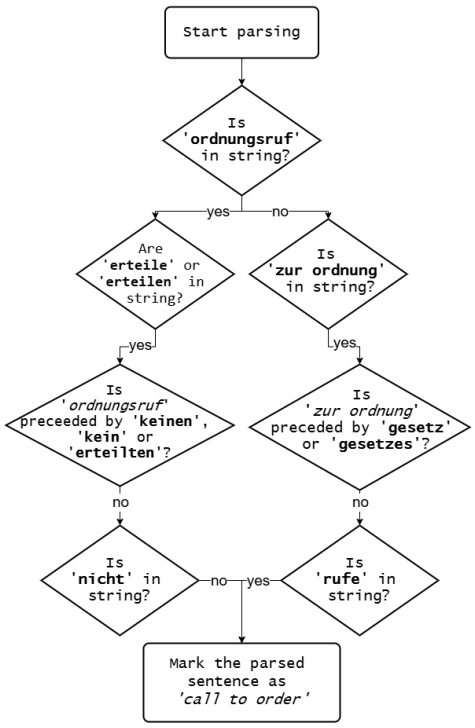

Figure 2: Rules to match calls to order

The output of the Named Entity Recognition (NER) models is integrated into PoliCorp as an experimental feature. Currently, users can see the output of two NER models, Legal German NER[17]

[13] https://www.bundestag.de/services/opendata

[14] https://github.com/PolMine/GermaParlTEI/tree/main

[15] As of September 2025.

[16] Translation of the text in Figure 3 is provided in Appendix B

[17] https://huggingface.co/flair/ner-german-legal

(Leitner et al., 2019; Akbik et al., 2018) and German NER[18] (Akbik et al., 2018). We conducted additional processing of the models' output and retrieved mentions of the German parties using pattern matching. Figure 3 shows the example of the output of the German NER model in the PoliCorp interface.

In addition, each speech contribution is annotated with a topic classification, generated using a BERT-based model[19] (Klamm et al., 2022). The classifier was applied to the full text of each speech contribution, excluding those delivered by the session president, which merely consist of procedural moderation. The underlying model was trained to differentiate between 21 distinct topics relevant to German parliamentary discourse. Contributions from the session president were instead assigned a separate category, labeled Presidency Action, thereby introducing an additional topic class. Figure 4 shows the distribution of topics in the PoliCorp collection over 21 legislative periods. The diagram illustrates the longitudinal development of the discussed topics within the Bundestag. Across all legislative periods, presidency actions and governmental affairs emerge as the most consistently debated and stable topics. In contrast, the environment topic, while being of comparatively low relevance during the early legislative periods, progressively gained relevance, becoming one of the major topics in the late periods.

2.3 Postprocessing

As the final processing step, the names of speakers and their associated political parties were disambiguated. This step was necessary due to inconsistencies in the raw data, which included spelling errors in speaker names, variations of the first names, as well as multiple variations of political party names, i.e., both abbreviated and full forms.

For the disambiguation of speaker names, a rule-based approach was employed. This approach utilized a database containing the names of all members of the German Parliament throughout its history[20]. Individuals with unique surname or surname–first name combinations were directly assigned the corresponding identifier. In cases where identical surname or surname–first name combinations appeared multiple times in the database, additional disambiguation was necessary. This was addressed by aligning the individuals with the legislative periods during which the corresponding speech contribution was delivered. If the date of a speech fell within an individual's documented parliamentary tenure, that person was considered a match. When multiple potential matches remained or no corresponding surname–first name combination could be identified in the database, the entry was kept as ambiguous, and no identifier was assigned.

Subsequently, political party names were disambiguated using pattern matching in conjunction with a comprehensive list of all German political parties and their known abbreviations across the history of the Bundestag. This process allowed for the normalization of party references into a unified format. For instance, both Freie Demokratische Partei and FDP were identified as referring to the same political entity.

3 Frontend Implementation

The frontend is developed using Express.js with Pug templates for server-side rendering. It provides an interactive web-based interface for querying, visualizing, and downloading results from parliamentary corpora. Users can perform both simple and advanced queries, with input parameters dynamically translated into Elasticsearch queries targeting the backend index. The interface includes additional informational pages (e.g., About, GitHub, FAQs, tutorial videos) to support usability. The design emphasizes responsiveness, modularity, and seamless backend integration to enable efficient and effective information retrieval for political science research.

3.1 User Interface

Figure 5 represents a basic search bar. By default, the query is executed against the full text of a speech contribution and all indexed metadata fields. Upon submission, the system returns a results list, displaying the total number of matches and an expandable metadata panel, as Figure 6 shows. Results may be reordered by relevance or chronological order.

For a more specific search, the user can proceed by selecting an advanced search option, as Figure 7 shows. The advanced interface has a set of structured input rows, each row comprising three elements: a logical operator, a field selector, and a

[18] https://huggingface.co/flair/ner-german

[19] https://huggingface.co/chkla/parlbert-topic-german

[20] https://www.bundestag.de/services/opendata

CALL TO ORDER: Herr Kollege Baumann[PER] für den Zwischenruf „Sie Hetzer!“ rufe ich Sie zur Ordnung.

INTERJECTION: (Zuruf von der AfD[PARTY] Also, hier können Leute „Nazi[ORG] sagen!)

SPEECH: – Einen Moment. Vorsicht! Die Ordnungsrufe des Präsidenten werden nicht kommentiert.

INTERJECTION: (Britta Haßelmann[PER] [BÜNDNIS 90/DIE GRÜNEN[PARTY], an die AfD[PARTY] gewandt: Herr Hilse[PER] das gilt auch für Sie! – Gegenruf des Abg. Tino Chrupalla[PER] [AfD[PARTY]: Aber auch für Sie! – Weitere Zurufe von der AfD[PARTY]

CALL TO ORDER: Dafür rufe ich Sie zur Ordnung, Herr Kollege.

INTERJECTION: (Stephan Brandner[PER] [AfD[PARTY]: Unglaublich, was hier abgeht! Unglaublich! – Zuruf: Ja, in der Tat!)

Figure 3: Example of marked call to order and NER output from the PoliCorp interface in the speech of the session's president, Wolfgang Schäuble, on October 2, 2020.

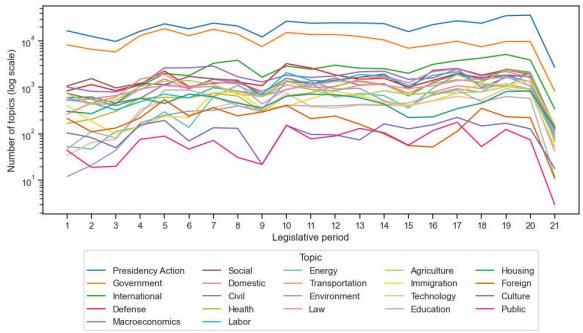

Figure 4: Distribution of topics over legislative periods.

Figure 5: Basic search

value field. Logical operator fields may be configured individually as AND, OR, or NOT, permitting execution of complex Boolean expressions. The drop-down list with the field selectors contains all searchable attributes, including but not limited to: full text, speaker, party, legislative period, topic, or date.

If a user is interested in a specific speech contribution, an expandable metadata panel can be accessed. Figure 8 displays the complete set of metadata associated with the selected speech, including general session-related information such as the legislative period, agenda and session sequence numbers, the date of the session, as well as the agenda type and a brief description. Additional metadata covers speaker-related details, including the speaker's name, party affiliation, and role or position within the Bundestag. Furthermore, the topic of the speech and a link to the corresponding source file are also provided. The web portal incorporates

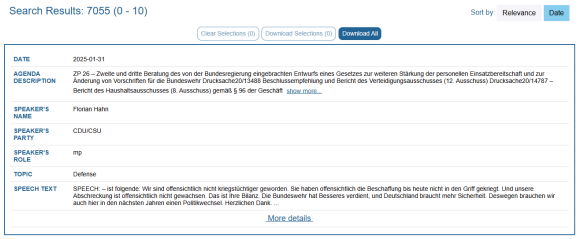

Figure 6: Search results

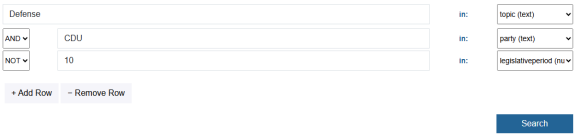

Figure 7: Advanced search

experimental features, such as the outputs of two NER models, as detailed in Section 2.2. These features are accessible via a drop-down menu labeled Speech Text. Selecting a specific NER model from the list dynamically updates the display of the speech contribution, as illustrated in Figure 3, by highlighting the recognized named entities within the text.

LEGISLATIVE PERIOD	20
SESSION NUMBER	211
AGENDA NUMBER	5
DATE	2025-01-31
AGENDA DESCRIPTION	Beratung der Unterrichtung durch die Enquete-Kommission L Drucksache20/14500
SPEAKER'S NAME	Peter Beyer
SPEAKER'S PARTY	CDU/CSU
SPEAKER'S ROLE	mp
URL TO THE SOURCE FILE	https://dserver.bundestag.de/btp/20/20211.xml
TOPIC	Defense
SPEECH TEXT	SPEECH: Frau Präsidentin! Meine sehr geehrten Damen unc Jahren lautete, 20 Jahre deutsches Engagement in Afghanis unserer Arbeit als Enquete-Kommission lag auch auf den Let

Figure 8: Available metadata

A key feature of PoliCorp is that it enables users to combine data in various ways, i.e., generating custom subcorpora and exporting them in JSON

format (Figure 9).

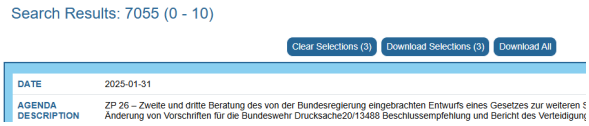

Figure 9: Download options

The data contained in the downloaded JSON file, as shown in Figure 10, fully corresponds to the information presented in the web interface. It includes all associated metadata, sentence-level segmentation of the speech contribution, and available named entity annotations. Thus, metadata comprises details about the speaker, including their identity, affiliation, and functional role. Furthermore, it encompasses information specific to the speech itself, such as the legislative period, session, and agenda numbers, date of delivery, and the speech topic.


```
"date": "2025-05-14",
"description": "(Fortsetzung der Aussprache)",
"speech_id": "4_21003_3_35",
"speechno": 35,
"sessionno": 3,
"agenda_id": "4_21003_3",
"source_file": "https://dserver.bundestag.de/btp/21/21003.xml",
"agenda_type": null,
"agenda_nr": 4,
"party_id": 27,
"speaker_id": 11004709,
"name": "Thomas Erndl",
"position": null,
"party": "CDU/CSU",
"role": "mp",
"topic": "Defense",
"text_raw": [
  {
    "seq_no": 0,
    "legal_ner": [
      {
        "end_pos": 81,
        "text": "Deutschland",
        "label": "LD",
        "start_pos": 70
      }
    ],
    "ner": [
      {
        "end_pos": 81,
        "text": "Deutschland",
        "label": "LOC",
        "start_pos": 70
      }
    ],
    "text": "Meine Damen und Herren, heute, 70 Jahre nach dem NATO-Beitritt, trägt Deuts
    "type": "speech"
  },
  {
    "seq_no": 1,
    "legal_ner": null,
    "ner": null,
    "text": "Lassen Sie uns dem mit einer starken Verteidigungspolitik gerecht werden!",
    "type": "speech"
  },
```


Figure 10: Example of JSON format

4 Portal Usage and Evaluation

Between January 5, 2025, and September 14, 2025, a total of 2,573 user queries were executed within the PoliCorp system, comprising 699 unique queries. The statistical analysis excludes queries exhibiting patterns indicative of potential malicious activity, such as those containing keywords like system, sleep, exec, bash, and similar commands. Table 1 displays the top 10 most frequently searched terms.

Statistics indicate that users typically search for data from a specific legislative period, e.g, 19 and 20. The second most popular search requests include political parties (CDU, SPD, DIE LINKE) and names of politicians (Merkel, Angela Merkel). The other common search requests relate to keywords such as Atomausstieg (nuclear phase-out), Klimawandel (climate change), or Migration (migration).

search term	frequency
19	118
20	93
cdu	89
merkel	77
spd	76
angela merkel	66
atomausstieg	55
klimawandel	52
migration	39
die linke	35

Table 1: Top 10 Most Frequently Searched Terms

The evaluation of the web portal focused on user-centered design principles. Functionality of PoliCorp was developed and refined in collaboration with a small user group with a background in political science or computer science-related disciplines, ensuring that key features align with user needs and expectations. Based on the evaluation results, certain interface and analysis features were added or removed from the portal, and the final interface layout was developed.

5 Conclusion and Future Work

In this work, we present PoliCorp, a web portal designed to facilitate the search and analysis of political text corpora. PoliCorp provides researchers with access to rich textual data, enabling in-depth analysis of parliamentary discourse over time. The platform currently contains a collection of transcripts of Bundestag debates, spanning 76 years of parliamentary debates. With the advanced search functionality, researchers can apply logical operations to combine or exclude search criteria, making it easier to filter through vast amounts of parliamentary debate data. The search can be customised by combining multiple fields and applying logical operators to uncover complex patterns and insights within the data. Selected datasets can be downloaded freely in JSON format, providing a convenient option for further analysis using com-

putational tools.

PoliCorp is a demonstration version that is currently under development. As of the current period, the platform comprises speech contributions from the German Bundestag covering the period from September 1949 to July 2025. Its content will be continuously updated with new speeches as they become available through the Bundestag Open Data portal. Ongoing work includes the implementation of a toxicity detection module, which will assign predefined toxicity labels to individual speech segments. Following, we are planning to link available transcripts with corresponding video recordings.

Future iterations will incorporate additional corpora, such as StateParl, and present them in a unified format to facilitate consistent processing and comparative analysis.

Furthermore, we are working on the integration of the analysis tools, i.e., descriptive statistics and visualization, and cross-corpus content comparison capabilities. Additionally, the available download formats will be expanded. The tool's source code will be made publicly available in a future release.

Limitations

PoliCorp represents a prototype implementation, with many features still under active development. Due to the early stage of the project, the web portal has been evaluated from a user-centered design perspective using a small group of participants. Broader evaluations, including surveys and workshops, are planned for future phases.

The platform incorporates experimental functionalities such as topic classification and NER, both of which are generated automatically and may contain inaccuracies. Users are therefore advised to interpret these outputs with caution and independently verify any critical information.

Furthermore, PoliCorp is a domain-specific resource primarily intended for users engaged in political science research.

Acknowledgements

Nina Smirnova received funding from the Deutsche Forschungsgemeinschaft (DFG) under grant number: MA 3964/7-3 (POLLUX Project).

Nina Smirnova and Philipp Mayr received additional funding from the European Union under the Horizon Europe grant OMINO (Hołyst et al., 2024) – Overcoming Multilevel Information Overload (http://ominoproject.eu) under grant number 101086321.

Muhammad Ahsan Shahid received funding from the Deutsche Forschungsgemeinschaft (DFG) under grant number: MA 3964/20-1 (OFFZIB Project).

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Elliott Ash, Johann Krümmel, and Jonathan B. Slapin. 2025. Gender and reactions to speeches in german parliamentary debates. *American Journal of Political Science*, 69(3):866–880.

Andreas Blaette. 2017. GermaParl. corpus of plenary protocols of the german bundestag.

Andreas Blaette. 2020. polmineR: Verbs and nouns for corpus analysis.

Andreas Blaette. 2021. Germaparl. download and augment the corpus of plenary protocols of the german bundestag. R package version 1.5.3.

Julius Diener. 2025. Explaining interruption behavior in parliament: The role of topic expertise, career status, and government-opposition dynamics. *Politische Vierteljahresschrift*, 66(2):303–324.

Janusz A. Hołyst, Philipp Mayr, Michael Thelwall, Ingo Frommholz, Shlomo Havlin, Alon Sela, Yoed N. Kenett, Denis Helic, Aljoša Rehar, Sebastijan R. Maček, Przemysław Kazienko, Tomasz Kajdanowicz, Przemysław Biecek, Boleslaw K. Szymanski, and Julian Sienkiewicz. 2024. Protect our environment from information overload. *Nature Human Behaviour*, 8:402–403.

Cornelia Ilie. 2015. Parliamentary discourse. In Karen Tracy, Cornelia Ilie, and Todd Sandel, editors, *The International Encyclopedia of Language and Social Interaction*. John Wiley & Sons, Inc., Malden, MA.

Johanna Isosävi, Heike Baldauf-Quilliatre, Christophe Gagne, and Eero Voutilainen. 2025. Reactions to interruptions in finnish, french and german parliamentary debates. *Journal of Language and Politics*, 24(2):301–327.

Marcelo Jenny, Martin Haselmayer, and Daniel Kapla. 2021. *Measuring incivility in parliamentary debates: validating a sentiment analysis procedure with calls to order in the Austrian Parliament*, pages 56–66.

Christopher Klamm, Ines Rehbein, and Simone Ponzetto. 2022. Frameast: A framework for second-level agenda setting in parliamentary debates through the lense of comparative agenda topics. *ParlaCLARIN III at LREC2022*.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained Named Entity Recognition in Legal Documents. In *Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019)*, pages 272–287.

Maarten Marx and Anne Schuth. 2010. DutchParl. the parliamentary documents in Dutch. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Michael G. Miller and Joseph L. Sutherland. 2023. The effect of gender on interruptions at congressional hearings. *American Political Science Review*, 117(1):103–121.

Malliga Och. 2020. Manterrupting in the german bundestag: Gendered opposition to female members of parliament? *Politics & Gender*, 16(2):388–408.

Željko Poljak. 2022. The role of gender in parliamentary attacks and incivility. *Politics and Governance*, 10(4).

Željko Poljak. 2023. Parties' attack behaviour in parliaments: Who attacks whom and when. *European Journal of Political Research*, 62(3):903–923.

Ines Reinig, Ines Rehbein, and Simone Paolo Ponzetto. 2024. How to do politics with words: Investigating speech acts in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8287–8300, Torino, Italia. ELRA and ICCL.

Shaul R. Shenhav. 2008. Showing and telling in parliamentary discourse: the case of repeated interjections to rabin's speeches in the israeli parliament. *Discourse & Society*, 19(2):223–255.

Naomi Truan. 2017. On the pragmatics of interjections in parliamentary interruptions. *Revue de sémantique et pragmatique*, 40(40):125–144.

Sebastián Vallejo Vera and Analía Gómez Vidal. 2022. The politics of interruptions: Gendered disruptions of legislative speeches. *The Journal of Politics*, 84(3):1384–1402.

Rozemarijn E van Dijk and Željko Poljak. 2025. Interrupting the interruptions. how women transform the parliamentary debate. *Parliamentary Affairs*, page gsaf040.

Kohei Watanabe and Yuan Zhou. 2022. Theory-driven analysis of large corpora: Semisupervised topic classification of the un speeches. *Social Science Computer Review*, 40(2):346–366.

Tanja Wissik. 2021. Encoding interruptions in parliamentary data: From applause to interjections and laughter. *Journal of the Text Encoding Initiative*.

A Raw XML Formats

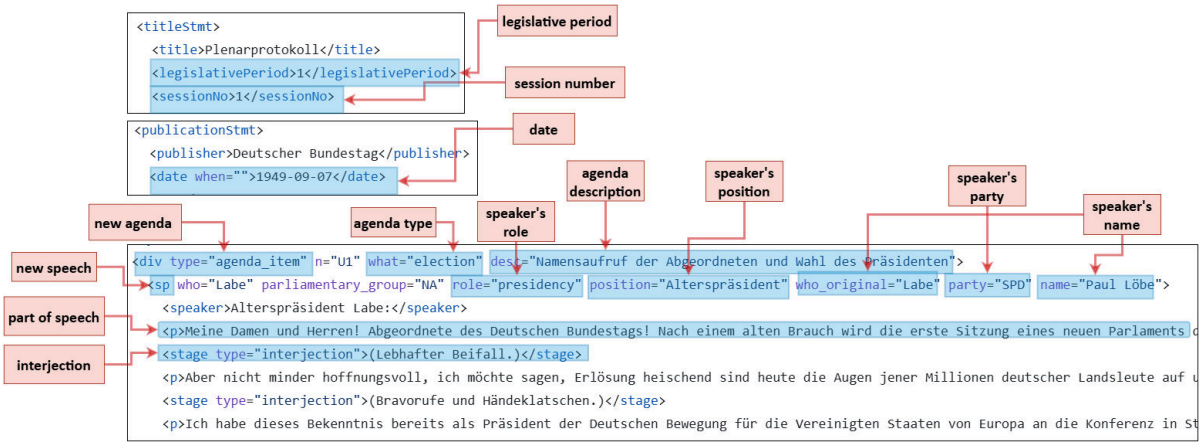

Figure 11: Example of the raw GermaParl data format

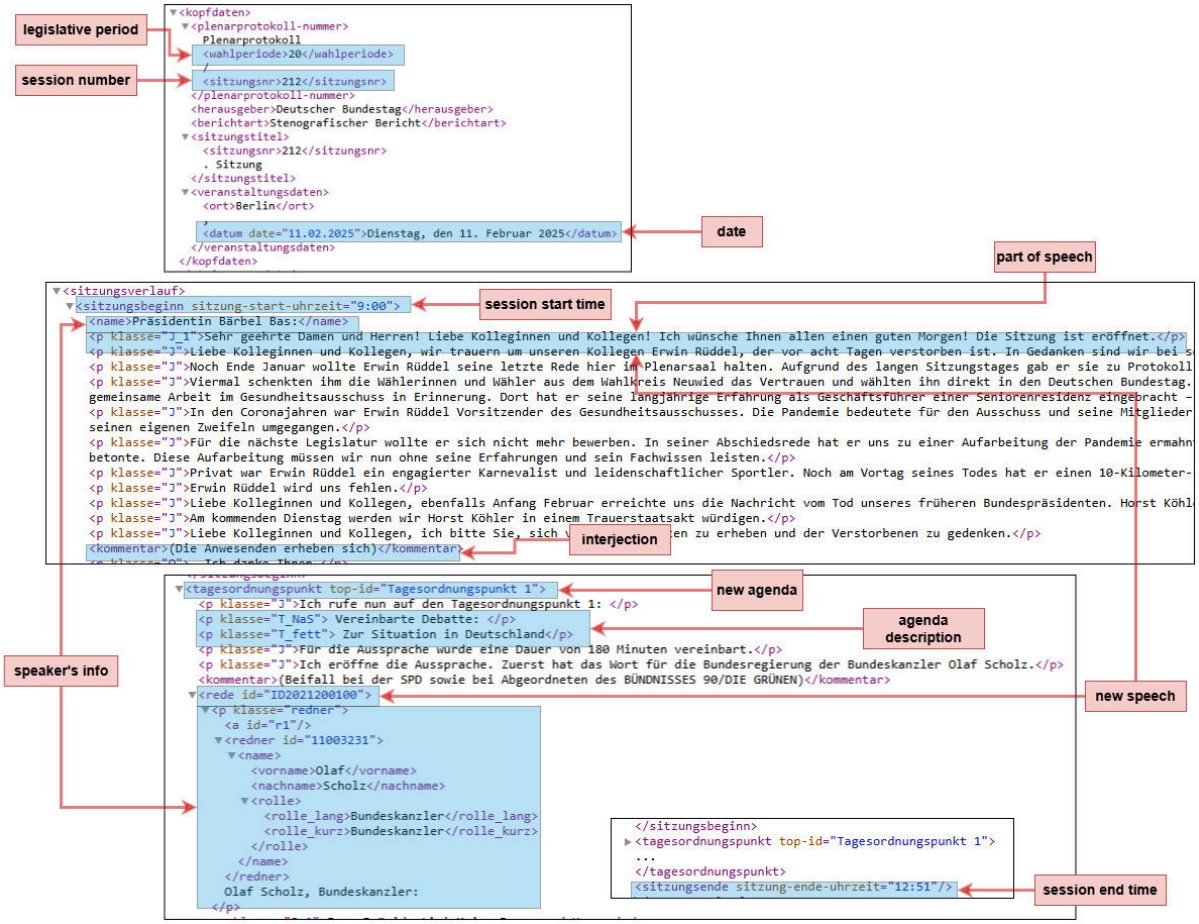

Figure 12: Example of the raw Bundestag Open Data data format

B Translation of the German text in Figure 3

- **CALL TO ORDER:** Mr. Baumann, I call you to order for the interjection "You agitator!".
- **INTERJECTION:** (Shout from the AfD: So, people can say "Nazi" here!) SPEECH: - One moment. Watch out! The President's calls to order will not be commented on.

- **INTERJECTION:** (Britta Haßelmann [Alliance 90/The Greens], addressing the AfD: Mr. Hilse, that goes for you too! - Counter-cry from Tino Chrupalla [AfD]: But also for you! - More shouts from the AfD)
- **CALL TO ORDER:** I call you to order for that, Mr. Hilse.
- **INTERJECTION:** (Stephan Brandner [AfD]: Unbelievable what's going on here! Unbelievable! - Shout-out: Yes, indeed!)