Towards ParlaMint-DE: Improving the Interoperability of the GermaParl Corpus of Plenary Protocols of the German *Bundestag*

Christoph Leonhardt and Andreas Blätte
University of Duisburg-Essen
{christoph.leonhardt, andreas.blaette}@uni-due.de

Abstract

With the number of machine-readable corpora of plenary protocols continuously increasing, concerns about the potentials of harmonisation and shared encoding standards gain prominence. Interoperability of corpora can contribute to innovative research, in particular when comparative analyses are concerned. The ParlaMint encoding schema introduced by CLARIN provides comprehensive guidelines towards this goal. This contribution shows how GermaParl, a large corpus of plenary protocols of the German *Bundestag*, is transformed from a TEI-inspired XML format to the ParlaMint encoding schema. Based on previous work, this paper presents an adjusted preparation pipeline and discusses challenges of advancing an established resource into a new data format. The prospective ParlaMint-DE corpus will make the plenary debates in Germany from 1949 to 2025 available in a highly interoperable data format. Clear documentation and taxonomies increase the usefulness of the resource in comparative analyses, whereas additional metadata and linguistic annotation broaden its general applicability.

Keywords: ParlaMint, Parliamentary Data, Plenary Protocols, Corpus Creation, German Bundestag

1. Introduction

Many studies in the social sciences and beyond rely on comprehensive corpora of plenary protocols (e.g., Skubic and Fišer, 2022; Skubic and Fišer, 2024). Following a trend towards more accessible and reusable data, an increasing share of these resources is released in machine-readable formats (e.g., Agnoloni et al., 2022, p. 117; Erjavec et al., 2025a, p. 2072; Sebők et al., 2025, p. 33). However, despite the obvious merits of these efforts, the interoperability of parliamentary resources remains a challenge (e.g., Sebők et al., 2025, p. 19).

Against this backdrop, the CLARIN infrastructure has worked towards shared encoding guidelines for parliamentary corpora. Arguing that the described lack of a harmonised standard "present[s] a barrier to their interchange, re-use and comparison" (Erjavec et al., 2023, p. 418), Erjavec et al. introduced the ParlaMint encoding guidelines to facilitate comparative research and the development of new tools and methods (Agnoloni et al., 2022, p. 117; Erjavec et al., 2023). Since then, the ParlaMint project created 29 corpora of European national and regional parliaments (Erjavec et al., 2025a, p. 2072).

One parliament not yet represented in this shared format is the German *Bundestag*. The relevance of German parliamentary debates in the ParlaMint encoding standard motivates our work: The inclusion of an additional, large corpus of parliamentary debates would contribute to comparative parliamentary research at large. At the same time, the adoption of ParlaMint would benefit German parliamentary research. Aside from the immediate gain of interoperability, it would increase usability by adding many innovative features which go beyond those included in comparable available corpora of German parliamentary proceedings while also contributing to the findability of the resource by bringing it closer to the efforts of ParlaMint and CLARIN.

We base our work on GermaParl, a comprehensive and richly annotated corpus of parliamentary debates in the German *Bundestag* (Blätte and Leonhardt, 2025). GermaParl is currently provided in an XML format inspired by a standard of the Text Encoding Initiative (TEI)[1] as well as in the format of the Corpus Workbench (CWB) (Evert and Hardie, 2011). While these formats contribute to the usability of GermaParl in many aspects (Blätte and Blessing, 2018; Blätte et al., 2022), the emergence of the ParlaMint encoding guidelines presents an opportunity to further strengthen its FAIRness.[2]

Following up on an earlier description of the corpus (Blätte et al., 2022), this work presents the transformation of GermaParl to ParlaMint-DE.[3] We focus on the innovations of the new representation and updates of the corpus preparation workflow. Beyond the specific use case, we provide insights into challenges and lessons learned and discuss the updated workflow and toolset as resources which might benefit similar corpus curation projects in the future.

[1] https://tei-c.org.

[2] Referring to the FAIR data principles of Findability, Accessibility, Interoperability and Reusability (Wilkinson et al., 2016).

[3] At the time of writing, the full adoption of ParlaMint is not completed. The following description should not pre-empt the regular collaboration process of ParlaMint described by Erjavec et al. (2025a), but illustrate the workflow, challenges and potentials of transforming an existing resource into a shared data format.

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, pages 31–43

2. Background

The current version of GermaParl as well as the ParlaMint encoding guidelines present the starting point and the target of our efforts respectively. Discussing them briefly will inform the subsequent presentation of the corpus preparation workflow.

2.1. GermaParl

GermaParl is a comprehensive corpus of plenary protocols in the German *Bundestag*. Since the release of the first version which included all plenary protocols between 1996 and 2016 (Blätte and Blessing, 2018), it has been continuously maintained and advanced. GermaParl v2.0.0, presented in Blätte et al. (2022) and fully released in 2023 covered all protocols between 1949 and 2021. Since then, the quality of the corpus has been evaluated and improved (Leonhardt and Blätte, 2023) and new features like the addition of Named Entity Linking via `DBpedia Spotlight` (Mendes et al., 2011; Daiber et al., 2013) have been included in multiple subsequent updates. In its most recent version (Blätte and Leonhardt, 2025), GermaParl contains 290 million tokens and covers all 4559 sessions between September 1949 and March 2025.[4] The corpus includes comprehensive metadata on the level of protocols (e.g., session date, legislative period) and speeches (e.g., the name or party affiliation of a speaker) as well as linguistic mark-up in its CWB variant (Blätte et al., 2022).

Following the considerations of Blätte and Blessing (2018), GermaParl is provided in two data formats: An XML format serves as an interchange format, aimed at interoperability and especially used in more advanced workflows of users who wish to use their own toolchain. It is inspired by the TEI encoding standard for performance text (Blätte and Blessing, 2018, p. 811). Protocols are represented as nested structures of agenda items, speeches and paragraphs. To further increase usability (e.g., Blätte and Blessing, 2018, p. 813; Skubic and Fišer, 2024, p. 9), GermaParl is also provided in the format of the Corpus Workbench (Evert and Hardie, 2011) which can be analysed using a many different tools, including graphical user interfaces like `CQPweb`[5] or script-based analysis environments like `polmineR` (Blätte, 2023) which is implemented in the statistical programming language R (R Core Team, 2025). This takes into account the relevance of parliamentary corpora for social science research as well as the prevalence of R in this field (Skubic and Fišer, 2024, p. 9).[6]

GermaParl is not the only corpus of parliamentary protocols of the German *Bundestag* (see for example Abrami et al. (2024) or Richter et al. (2023)). Just like GermaParl, these resources regularly face potential trade-offs between coverage, data quality, the availability of metadata and usability. We argue that GermaParl as a high-quality resource which combines longitudinal coverage with comprehensive metadata on various levels of granularity (similarly Blätte et al., 2022) constitutes a good foundation for ParlaMint.

2.2. ParlaMint

The increased interoperability generated by a shared encoding standard provides many opportunities to lower barriers and facilitate innovative research. In this regard, the ParlaMint encoding standard markedly advanced efforts towards more interoperable data. Following the ParlaMint II project, "29 European countries and autonomous regions" have become available in the ParlaMint encoding (Erjavec et al., 2025a, p. 2072; Erjavec et al., 2025b). The temporal coverage of these corpora starts in the 1990s, with most of them beginning in the 2010s and all corpora ending between 2022 and 2024 (Erjavec et al., 2025a, p. 2080). The guidelines[7] formulate comprehensive recommendations and requirements which are sufficiently granular to allow for a truthful representation of different parliamentary proceedings in an interoperable fashion. Comprehensive metadata further contributes to the usefulness and comparative potential of the resources while the inclusion of linguistic annotation in the TEI makes the linguistic mark-up accessible for a broad audience. Furthermore, the ParlaMint project has evolved beyond corpora by making various tools and workflows easily usable for all compatible resources. This broad landscape of data, tools and workflows improves the findability of its individual resources. This yields great potentials for a German parliamentary corpus.

As summarised by Erjavec et al. (2025a, p. 2073), the current ParlaMint encoding guidelines are the result of the iterative revision of the Parla-CLARIN encoding recommendations. In ParlaMint, each corpus has the following structure:

- **Root File**: The root file of the corpus containing metadata on the corpus level including title, creators, extent, data sources and references to other files (person metadata, organisation metadata and taxonomies) (Erjavec et al., 2023, p. 435; Erjavec et al., 2025a, p. 2075).

[4] Number of tokens is based on the CWB version.

[5] `https://cwb.sourceforge.io/cqpweb.php`.

[6] Data of GermaParl is also made available interactively in `PoliCorp`, a web portal currently developed and maintained at GESIS (Smirnova et al., 2025).

[7] `https://clarin-eric.github.io/ParlaMint/`.

- **Common taxonomies**: Taxonomies shared by all corpora following the ParlaMint encoding standard (e.g., speaker types) (Erjavec et al., 2023, p. 436).
- **Local taxonomies**: Taxonomies used by a single corpus (Erjavec et al., 2025a, p. 2075).
- **List of Persons**: A list of metadata on persons, including full names, gender, affiliations to institutions (e.g., the German *Bundestag*), parties and parliamentary groups where applicable (Erjavec et al., 2023, p. 436).
- **List of Organisations**: A list of metadata on organisations (e.g., parties, parliamentary groups, cabinets), including abbreviations and full names, dates of existence as well as the political orientation of parties and parliamentary groups (Erjavec et al., 2023, p. 435; Erjavec et al., 2025a, p. 2086–2087).
- **Corpus Components**: XML files containing a single session. In the German *Bundestag*, this usually corresponds to all debates on a single day.[8] Each component contains corpus-specific and session-specific metadata as well as the speeches. These are encoded as utterances (`<u>`) which themselves contain paragraphs encoded as segments (`<seg>`) and transcriber comments (Erjavec et al., 2023, p. 438–439).

In ParlaMint, there are two versions of each corpus: While the *plain* version of ParlaMint contains the structural annotation described above, the *linguistically annotated* version of the corpus additionally includes linguistic mark-up. This comprises the segmentation of utterances into sentences and tokens, the annotation of lemmata, Part-of-Speech tags, morphological features and named entities as well as the addition of syntactic parsing (Erjavec et al., 2023, p. 439–440).[9] Generally, the encoding guidelines provide some flexibility and allow for the inclusion of additional attributes and mark-up.

In summary, the increased interoperability and resulting usefulness for comparative analyses in particular, but also the additional and more granular metadata, its comprehensive documentation within the corpus itself as well as a more accessible representation of linguistic mark-up are major arguments to move towards ParlaMint. It can be noted that GermaParl is substantively larger than the ParlaMint corpora presented in Erjavec et al., 2025a, p. 2080. However, as already indicated by the creation of other ParlaMint corpora such as ParlaMint-IL which contains over 400 million words (Goldin et al., 2025), the encoding schema itself seems well suited to accommodate large corpora. We follow up on this in the following presentation of an updated corpus preparation workflow.

[8] In the German *Bundestag*, there are a few exceptions to this, e.g., two sessions on December 16, 1949.

[9] Following the ParlaMint guidelines, Part-of-Speech, morphological features and syntactic parsing are annotated according to Universal Dependencies.

3. A Reproducible Corpus Preparation Pipeline (Revisited)

In Blätte et al. (2022), we presented a "Reproducible Corpus Preparation" pipeline for GermaParl. This workflow advanced the process described in Blätte and Blessing (2018) and still provides the foundation of the current version of GermaParl. In general, the corpus preparation pipeline follows the sequence of data *preprocessing*, its re-structuring into XML (or "*XMLification*") and the *consolidation* of metadata, in particular for speakers. In a final step, the linguistic annotation is added.[10]

Despite the differences in the encoding schemas of the current GermaParl corpus and ParlaMint, this established workflow constitutes a good foundation for the preparation of ParlaMint corpora. This benefits from the genuinely generic approach of the established workflow which relies on frameworks, scripts and R packages which can be easily adjusted for different input and output formats. In consequence, we adopt the existing workflow and resources to create a German ParlaMint corpus. To match the encoding guidelines of ParlaMint, some minor adjustments are necessary. In line with Blätte et al. (2022), we structure these updates along the dimensions of *preprocessing*, *XMLification* and *consolidation* and focus on new challenges and lessons learned during the adoption of ParlaMint.

3.1. Preprocessing

Most of the raw input is based on the data collection described in Blätte et al. (2022, p. 10–11). Mostly unstructured XML (for the period of 1949–1996), unstructured plain text (1996–2017) and structured XML (2017–2025) constitute the majority of raw data. In addition, PDF files on which Optical Character Recognition was already performed, were used if protocols were not available in sufficient quality in other data formats.[11]

[10] The documentation of the corpus is also available online: https://polmine.github.io/GermaParl2/.

[11] The input data is retrieved from the website of the German *Bundestag*, most importantly https://www.bundestag.de/services/opendata. The precise source of each protocol is documented within the resulting XML/TEI in both the current GermaParl XML/TEI and ParlaMint.

Each of these data formats poses different challenges and trade-offs. In general, we follow the considerations described in the existing workflow (Blätte et al., 2022, p. 10–11) and only change input formats of individual protocols when issues become apparent. Aside from switching input formats selectively, we fine-tune the existing preprocessing steps which among other things include the removal non-substantive contents from the raw input and the correction of OCR errors.

3.2. XMLification

When preparing GermaParl as XML/TEI, we relied on the "Framework for Parsing Plenary Protocols" (`frappp`), an R package which facilitates the identification, encoding and enrichment of agenda items and speeches in a large collection of unstructured text (Blätte et al., 2022).[12] Using `frappp`, most information was extracted from preprocessed input documents via regular expressions.

`frappp` creates one XML/TEI file per plenary session. Since the general structure of these session-specific files in the XML/TEI format is similar to those of the ParlaMint encoding, we continue to use `frappp`. Similar to the previous XML/TEI format, each ParlaMint corpus component file begins with a TEI header containing both general and session-specific metadata. This is followed by utterances which are potentially nested within agenda items.

Compared to the current XML/TEI of GermaParl, the metadata represented in each TEI header in ParlaMint is more comprehensive and potentially multilingual to further increase interoperability. The representation of utterance elements in ParlaMint and speech elements in the earlier XML/TEI format differs in some aspects. Importantly, in ParlaMint, each utterance and segment (as well as sentence and token in the linguistically annotated version) is assigned a unique identifier. In addition, each utterance element contains three attributes in the new format: A person identifier, an identifier for the utterance itself and a reference to the parliamentary role of the speaker. At this stage, we use the speaker name we extract from the protocol itself as a temporary person identifier which we replace in a later step of the workflow with a consolidated person identifier[13] as this information can be noisy (e.g., typos or errors in the protocols), incomplete (e.g., missing given names) or ambivalent (e.g., various speakers sharing the same family name). To prepare the subsequent consolidation of person metadata, the parliamentary group affiliation for Members of Parliament (MPs) is extracted from the protocols and temporarily stored in the utterance element. Similar to the temporary person identifier, we remove this information later on as it is not part of the utterance element in ParlaMint.

The classification of transcriber comments is more fine grained in ParlaMint. This is implemented via an additional mapping of regular expressions used to identify these sequences in the protocols. In addition, in the updated workflow, some regular expressions are tuned to improve the detection of speakers and transcriber comments.

In summary, turning unstructured input into XML largely follows the same workflow as before. The code base of `frappp` was modified mainly to account for running identifiers of various elements. Otherwise, the existing framework proved to be flexible enough to facilitate the creation of other data formats such as ParlaMint.

3.3. Consolidation

3.3.1. Enrichment and Metadata Structure

To consolidate the speaker information stored in each utterance element, extracted information is matched against external data sets of known parliamentary actors. Based on this disambiguation, additional metadata is then added to the ParlaMint corpus. Like above, this process is very similar to the previous workflow and only minor adjustments are necessary from a technical perspective. This mainly concerns the representation of person metadata itself: Metadata on speakers is not stored within each utterance element, but in a separate file. We adjust the consolidation mechanism of `frappp` accordingly.

3.3.2. Metadata Collection and Sources

Much of the information included about speakers in the GermaParl XML/TEI can also be found in ParlaMint, albeit often with greater granularity: Full names are provided as "*forename*" and "*surname*" and "*affiliations*" to parties and parliamentary groups are represented by references to a list of metadata on organisations. The temporal quality of metadata is made explicit: Names as well as affiliations to parties, parliamentary groups or roles can change, indicated by attributes "*from*" and "*to*" in some elements. In addition to required metadata (gender and affiliation information in particular), we provide dates of birth where possible and several external identifiers which should facilitate easier linkage with other parliamentary data sets as well as general knowledge bases like Wikidata (Vrandečić and Krötzsch, 2014).

In general, the extent of metadata required by ParlaMint greatly exceeds the information available

[12] `frappp` is available on GitHub: `https://github.com/PolMine/frappp`.

[13] We ultimately use Wikidata IDs (Vrandečić and Krötzsch, 2014) as the central identifier for persons.

in the current version of GermaParl. The limited availability of structured and date-specific information about speakers poses particular challenges when preparing ParlaMint. To address this, we rely on four major sources for metadata:

- ***Stammdaten* file**: Provided by the German *Bundestag*,[14] the *Stammdaten* file provides demographic and biographic information on every MP in the German *Bundestag*.[15] We extract date-specific full names, time-invariant gender information, date of birth, date-specific affiliations to parliamentary groups as well as the parliament itself and a parliamentary identifier.

- **Parliaments Day-by-Day Database**: Compiled by Turner-Zwinkels et al. (2022), the Parliaments Day-by-Day Database (PDBD) comprises of date-specific information on the party affiliation of MPs in Germany, the Netherlands and Switzerland. We use the date-specific party affiliation information for German MPs which is covered between 1949 and 2017.[16]

- **Wikipedia**: Wikipedia is the source for metadata about speakers other than MPs. Particularly relevant metadata includes full names, party affiliations and affiliations to cabinets and other offices. Wikipedia is also used to extend the data of PDBD for date-specific party affiliations (2017–2025).[17]

- **Wikidata**: Gender information of speakers who never have been MPs as well as Wikidata IDs have been retrieved from Wikidata (Vrandečić and Krötzsch, 2014).

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

[15] The previous workflow (Blätte et al., 2022), but also other data curation projects rely on the *Stammdaten* as well, see Richter et al. (2023, p. 4) who translate it as "base data" or Turner-Zwinkels et al. (2022, p. 764) who translate it as "master data sheet".

[16] According to the online appendix of Turner-Zwinkels et al. (2022), the source of this information in PDBD is the *Stammdaten* file which contains time-invariant party affiliation information but date-specific information on parliamentary group affiliations.

[17] We manually extract information about speakers' affiliations to parties from Wikipedia. The main source are lists of MPs per legislative period in which changes in party affiliation are indicated in an unstructured fashion. Individual Wikipedia pages are used if the information on these overview pages is ambivalent. In instances in which no specific date can be found, we tried to provide the most granular information available (e.g., month-specific information).

3.3.3. Challenges

The integration of these various sources is challenging. At first glance, matching persons over these various data sets is comparatively easy: Each of the data sets contains some kind of identifier which could be related to another. However, different data sources (e.g., the *Stammdaten* file and Wikipedia) would potentially describe the same speaker (e.g., once as an MP and once as a governmental actor) using contradictory information (e.g., variations in speaker names, different party affiliations in overlapping time spans, partly due to different granularity of metadata). We consolidated varying representations of information and resolved contradictions by using both hard-coded interventions and heuristics. In particular, this should guarantee that no speaker has more than one affiliation to a parliamentary group or party or more than one name at any given time in accordance with ParlaMint.

3.3.4. Organisational Metadata

ParlaMint requires additional information about organisations, in particular parties and parliamentary groups, but also cabinets and other organisations. This information is generally gathered from Wikipedia. In contrast to their rather sparse annotation in the established version of GermaParl, full names in German and English, abbreviations, dates of existence, identifiers and information about the political orientation of parties and parliamentary groups are provided in the ParlaMint variant where available.[18] Like person metadata, this information is stored in a separate file.

3.4. Linguistic Annotation

The planned linguistic mark-up of ParlaMint-DE is presented in Table 1. The transition towards ParlaMint necessitates the addition of morphological features and syntactic parsing. Since neither of the tools used in the previous setup (Blätte et al., 2022, p. 12) extract morphological features, a modification of the processing pipeline is necessary.

For ParlaMint, we plan to use a combination of UDPipe (Straka, 2018) and Stanford CoreNLP (Manning et al., 2014).[19] This selection is motivated by the desire to integrate the tools in our established R-based corpus preparation workflow which should minimise transaction costs and facilitate easier maintenance as well as reproducibil-

[18] Political orientation is extracted manually from info boxes in the English Wikipedia.

[19] We conducted initial annotations with the model "german-hdt-ud-2.5-191206" for UDPipe which we plan to use alongside the workflow for Stanford CoreNLP presented in Blätte et al. (2022). Currently, we still evaluate various options to provide the required linguistic mark-up.

Layer	Annotation Tool
Sentence Segmentation	UDPipe (Straka, 2018)
Tokenization	
POS (UD)	
POS (STTS)	
Lemmata	
Morphological Features	
Syntactic Parsing	
Named Entities*	Stanford CoreNLP (Manning et al., 2014)
Named Entity Linking*	DBpedia Spotlight (Mendes et al., 2011; Daiber et al., 2013)

* Not fully implemented at time of writing.

Table 1: Prospective Linguistic Mark-Up of ParlaMint-DE

ity. This can be realised by using the R packages `udpipe` (Wijffels and Straka, 2026) and `bignlp`.[20] The selected set of tools should be sufficiently fast, given the size of the corpus and adequately accurate (Straka, 2018; Ortmann et al., 2019).

In addition to the required annotation layers, we strive to include Named Entity Linking to the German ParlaMint corpus at an early stage. Already a part of the current CWB variant of GermaParl, we argue that the inclusion of Named Entity Linking yields great potential for the analysis of parliamentary debates (Leonhardt and Blätte, 2024; similarly van Heusden et al., 2022 and Janssen and Kopp, 2024, p. 125). We plan to add this annotation layer via a workflow centred around the `dbpedia` R package presented in Leonhardt and Blätte (2024).[21]

3.5. Customisation

One particular feature of the ParlaMint encoding schema is its applicability to many different corpora (Erjavec et al., 2025a, p. 2073). To account for variation between parliaments, ParlaMint allows for the inclusion of "local taxonomies" (Erjavec et al., 2025a, p. 2075). Making use of this, we suggest multiple custom taxonomies to represent country-specific speaker types, agenda item types and the sources of metadata. This might require adjustments to the overall encoding schema and is thus currently experimental.

3.6. Finalisation

This updated corpus preparation workflow results in 4559 corpus component files. In a final step, we use additional scripts to create the corpus root file, format the metadata files for persons and organisations and add taxonomies. Most of the data of the corpus root file corresponds to the current headers in the individual XML files and is largely based on hard-coded information.

3.7. Lessons Learned

The described updates to our workflow provide insights for curation projects beyond the scope of GermaParl. Two aspects seem particular important: Flexible preparation pipelines which allow for a programmatic implementation of a new output format while not relying on singular hard-coded assignments as well as a dynamic and efficient organisation of required metadata.

Using the presented pipeline centred around `frappp`, it was possible to adopt the ParlaMint encoding standard comparatively quickly. This further emphasises the merits of the initial generic approach: The toolset is not limited to the preparation of a specific corpus but sufficiently flexible. For example, `frappp` allows for the definition of an XML template which provides scaffolding for plenary protocols in various output formats. Challenges emerged from data availability and one lesson learned is that the thorough organisation of metadata is important, especially when various sources are involved. To this end, we created multiple R packages which contain metadata on persons and organisations in a structured format. One particular advantage of R packages is that both the origin of the data as well as potential data processing steps can be comprehensively documented. Using semantic versioning, we can relate various versions of the same metadata to a particular corpus version, allowing for dynamic data management. This strategy pairs well with ParlaMint in which data is also documented and described in detail. This solution also underlines the lack of an integrated, authoritative data source for our use case, necessitating the combination of various country-specific (e.g., the *Stammdaten* file) and more generic (Wikipedia, Wikidata) resources.

4. ParlaMint-DE

In this section, we describe an initial version of ParlaMint-DE which is based on the workflow outlined above. The plain text variant of this version of the corpus is available on GitHub.[22] As will be

[20] https://github.com/PolMine/bignlp.
[21] https://github.com/PolMine/dbpedia.
[22] https://github.com/PolMine/ParlaMint-DE_beta.

discussed in more detail in subsection 4.3, this current state does not yet fully align with the ParlaMint encoding guidelines. We consider this to be a "beta" version of the corpus which will be further refined. In addition, since we currently still evaluate the best workflow to add linguistic annotation we focus on the plain text variant of the corpus in the following.

The protocol data used for the current version of ParlaMint-DE is the same as GermaParl v2.3.0-rc1 (Blätte and Leonhardt, 2025) and covers all plenary protocols of the German *Bundestag* between September 1949 and March 2025.[23] To move towards the ParlaMint encoding standard, the workflow underlying GermaParl v2.3.0-rc1 has been modified and metadata has been added according to the steps discussed above.

Covering the first twenty legislative periods of the German *Bundestag*, the resulting initial version of ParlaMint-DE comprises of about 260 million tokens in 1.03 million utterances in total. Figure 1 provides an overview of the number of tokens in millions by year.[24]

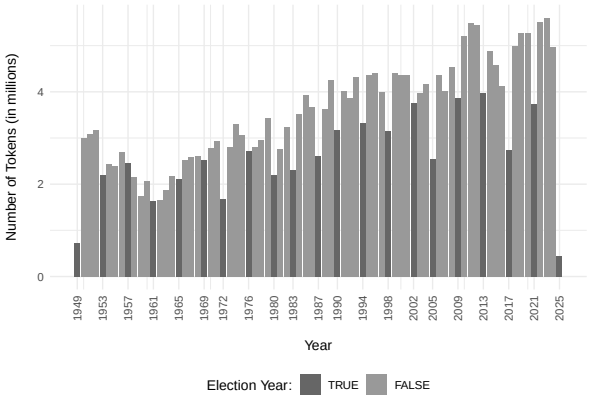

Figure 1: Number of Tokens by Year

Unsurprisingly, this is very similar to Figure 2 in Blätte et al. (2022, p. 10). As before, it is observable that election years usually contain fewer tokens than other years in the same legislative period. The trend towards more tokens per year already seen in Blätte et al. (2022, p. 10) persists.[25]

[23] GermaParl v2.3.0-rc1 is available on Zenodo as a "release candidate" due to the novel nature of the included Named Entity Linking annotations. Its general data structure is equivalent to other releases of GermaParl2. See for example GermaParl v2.1.0 which is openly available (Blätte and Leonhardt, 2024).

[24] We use `quanteda` (Benoit et al., 2018) to extract counts of tokens (including punctuation) from all segment elements in the plain text version of the corpus.

[25] Interestingly, the Open Discourse corpus exhibits the same trend (Richter et al., 2023, p. 6), underlining that this is not merely an artefact of our processing pipeline.

4.1. Metadata

Table 2 provides an overview over the metadata for persons in the German ParlaMint corpus. All in all, we extracted metadata for 4660 persons who actually take the plenary floor.

Concerning organisations, we collected metadata on 42 parties (including explicit references to unknown and missing party affiliations). For 25 out of 42 parties, we were able to extract the party's political orientation based on their English Wikipedia pages. The number of parties further underlines the extensive temporal coverage of the corpus. Information on 21 parliamentary groups (including an explicit reference for speakers without an affiliation to any parliamentary group) was gathered in a similar fashion.

4.2. Linguistic Annotation

Linguistic mark-up which will be available in the final version of ParlaMint-DE is presented in Table 1. Aside from the necessary annotations, we plan to add language-specific Part-of-Speech tags and Named Entity Linking.[26]

4.3. Data Availability

Currently, an initial version of the plain text variant of ParlaMint-DE has been prepared and is made available on GitHub.[27] Aside from all protocols between 1949 and 2025, metadata on speakers and organisations for the entire corpus as well as a draft of the root file and taxonomy files are included in this repository. We provide it as a "beta" version of an emerging ParlaMint-DE corpus to provide insights into the current state of the curation project. We identify three major aspects necessary to fully integrate the data into the ParlaMint encoding schema: Validation, the need to add further metadata and annotation layers as well as customisation.

Complete validation is a major next step towards a fully compatible ParlaMint-DE version. While we paid close attention to the ParlaMint encoding guidelines, we did not yet set up the validation pipeline in accordance with the process used within the ParlaMint project itself (Erjavec et al., 2025a, p. 2075–2076). However, given the comprehensive nature of the schema and potential country-specific characteristics, some complications cannot be ruled out before this validation has been completed. In this regard, changes to the current

[26] At the time of writing, we did not yet finalise the annotation of Named Entities and Named Entity Linking and the precise workflow used for linguistic annotation still needs to be consolidated.

[27] https://github.com/PolMine/ParlaMint-DE_beta.

Metadata	Values	Source
Speaker Name	surname, forename(s)	*Stammdaten* (MPs), Wikipedia (other speakers)
Sex	female (F) / male (M) / unknown (U)	*Stammdaten* (MPs), Wikidata (other speakers)
Date of Birth	Date in YYYY-MM-DD	*Stammdaten* (MPs)
Affiliation/Party	Party, including full labels in German and English, political orientation	PDBD (MPs), Wikipedia (MPs after 2017, other speakers, Metadata on parties themselves)
Affiliation/Parliamentary Group	Parliamentary Group, including full labels in German and English, political orientation	*Stammdaten* (MPs), Wikipedia (Metadata on parliamentary groups themselves)
Affiliation/Role	References to Organisations (e.g., cabinets, other offices)	Protocol Data, Wikipedia for duration data
Identifiers	Wikidata ID, Wikipedia URI, Parliamentary ID	Wikidata, Wikipedia, *Stammdaten*

Table 2: Metadata of ParlaMint-DE

structure of the data available on GitHub are to be expected to ensure full interoperability.

Furthermore, we are aware that some additional metadata and annotation layers still need to be added, in particular with regards to the additions of ParlaMint II (Erjavec et al., 2025a, p. 2085–2090). By making the current version available on GitHub, we aim to make these additions in a more iterative and transparent fashion in the future. At the same time, this also means that the current state of the corpus does not yet include all features available in other ParlaMint related resources.

Finally, as suggested at above, we introduced some customisations to the default ParlaMint encoding schema. In how far these adjustments should be part of the final version of ParlaMint-DE is still to be evaluated.

The release of the current version marks the start of comprehensive and strict technical validation. Going forward, this will enable us to identify remaining issues more quickly. We also aim to increase the transparency of the preparation process. Moving this process to GitHub allows for the public documentation of future adjustments and has the potential to facilitate discussions of encoding decisions with potential users and other stakeholders.

We strive to release complete versions of both the plain and the linguistically annotated variants of ParlaMint-DE in the summer of 2026. In keeping with the established dissemination strategy of GermaParl, we plan to make the corpora available on Zenodo using an open licence (CC BY). Upon release, we would encourage other repositories and platforms to make use of the resource as well.

5. Applications

As initially discussed, the potentials of ParlaMint are underlined by the broad array of workflows and tools which were developed in the context of or are compatible with ParlaMint. This section highlights but a few of these resources.

The `ParlaCAP`[28] and `ParlaSent` (Mochtak et al., 2025) classification models make topic classification and sentiment analysis, two very common tasks in substantive analyses, more accessible for parliamentary research. GermaParl as ParlaMint-DE further simplifies the adoption and comparative application of these resources for German parliamentary debates. Combined with the comprehensive metadata available in ParlaMint-DE, approaches like the sentiment analysis shown by Mochtak et al. (2025) become feasible for scholars of various technical backgrounds.

Similarly, the `dbpedia` R package was designed to lower barriers for Named Entity Linking in social science research (Leonhardt and Blätte, 2024). This can contribute to more fine-grained comparative analyses over multiple ParlaMint corpora (similarly Janssen and Kopp, 2024, p. 125; van Heusden et al., 2022). `dbpedia` makes the adoption of this approach easier by providing a wrapper and workflows for `DBpedia Spotlight` (Mendes et al., 2011; Daiber et al., 2013) for the R programming language with specific support for ParlaMint corpora (Leonhardt and Blätte, 2024).

Finally, ParlaMint corpora can be analysed and visualised with numerous tools. Platforms like `NoSketch Engine`[29] and `KonText` (Machálek,

[28] See the tutorial on GitHub: https://github.com/clarinsi/ParlaCAP-Analysis-Tutorials.

[29] `NoSketch Engine` is an open-source version of

2020) are utilized for ParlaMint corpora and particularly provide access to linguistic analyses (Janssen and Kopp, 2024, p. 121).[30] To make substantive and comparative analyses even more accessible, several additional tools like `TEITOK` (Janssen and Kopp, 2024) or the `ParlaMint NGram Viewer` (de Jong et al., 2024) are available for ParlaMint corpora (see also Erjavec et al., 2025a, p. 2079).

6. Discussion and Conclusion

This contribution presented the transition of GermaParl, a large corpus of plenary protocols of the German *Bundestag*, towards ParlaMint-DE, following a widely adopted TEI encoding standard for parliamentary proceedings. We focused on the potentials of the new standard, the challenges of the transition and possible applications. At the time of writing, ParlaMint-DE is still in development, but it should be released in the near future. A beta version of the corpus is made available on GitHub.

Once finished, the corpus will include all debates in the German *Bundestag* from September 1949 until March 2025 along with comprehensive metadata and linguistic mark-up in an interoperable format and provide access to the resources and ever growing toolset associated with ParlaMint. This immediately addresses some of the gaps presented in current-day discussions on legislative research (e.g., Sebők et al., 2025; Baden et al., 2022). By harmonising our resources, we create comparability and open up avenues for new research. ParlaMint-DE would further add to the coverage of the overall collection of ParlaMint corpora. Furthermore, shared encoding standards also contribute to the integration of tools and workflows and increase the accessibility and findability of resources, in particular for languages other than English.

We further showed how the workflow first presented in Blätte et al. (2022) could be adjusted for ParlaMint. While applied to proceedings of the German *Bundestag* in this paper, its relevance goes beyond this specific use case. By describing both the data we work with and the specific requirements of the targeted encoding guidelines, we were able to identify some generally relevant learnings: Due to its genuine flexibility, the generic approach of our corpus preparation pipeline centred around the `frappp` R package made the adoption of ParlaMint generally possible with comparatively little technical effort. The collection and representation of metadata remained more challenging. The ParlaMint encoding guidelines require comprehensive metadata while also encouraging thorough documentation. Especially when multiple sources of metadata are concerned, this can entail complex data structures. While our solution to create R data packages is specific for our R-based workflow, the need to document the provenance of data included in large corpora is important and potentially deserves more attention by data providers and curators. Ultimately, the applicability of our workflow depends on the structure and quality of both parliamentary proceedings and metadata in each use case. If both the debates themselves and metadata are not available in sufficiently structured data formats, the proposed workflow might be suitable. Lastly, aside from the file size of the resulting corpus which requires adequate infrastructure, the transition of GermaParl underlines that the volume of data is not a limiting factor of future corpus curation projects. From this perspective, parliamentary proceedings which are available in unstructured formats in large volumes might potentially be considered a next use case. The parliaments of the German regional states might be suitable candidates in this regard.

The finalisation of ParlaMint-DE should only mark a starting point for new research. Once completed, we envision a machine-translated version in English like other ParlaMint corpora (Kuzman Pungeršek et al., 2025), a closer integration of tools and workflows afforded by ParlaMint and, ultimately, more substantive analyses in the field of parliamentary research and beyond.

7. Acknowledgements

This work has been made possible by funding from the German National Research Data Infrastructure (Nationale Forschungsdateninfrastruktur/NFDI). We gratefully acknowledge funding from KonsortSWD – NFDI4Society which is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft/DFG) as part of NFDI under project number 442494171 as well as from the Text+ consortium which is funded by the German Research Foundation as part of NFDI under project number 460033370.

Furthermore, the development of GermaParl towards ParlaMint has benefited greatly from a Visiting Fellowship at the Institute of Contemporary History in Ljubljana, Slovenia. This support is appreciated.

Finally, we want to thank the anonymous reviewers for their insightful comments and suggestions.

the commercial `Sketch Engine` corpus management software by Lexical Computing (see https://www.sketchengine.eu/nosketch-engine/). See also Machálek (2020, p. 7003).

[30] The CLARIN.SI research infrastructure provides access to ParlaMint corpora in both `NoSketch Engine` (https://www.clarin.si/ske/) and `KonText` (https://www.clarin.si/kontext/).

8. Bibliographical References

Giuseppe Abrami, Mevlüt Bagci, and Alexander Mehler. 2024. German Parliamentary Corpus (GerParCor) Reloaded. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7707–7716, Turin, Italy. ELRA and ICCL.

Tommaso Agnoloni, Roberto Bartolini, Francesca Frontini, Carlo Marchetti, Simonetta Montemagni, Valeria Quochi, Manuela Ruisi, and Giulia Venturi. 2022. Making Italian Parliamentary Records Machine-Actionable: The Construction of the ParlaMint-IT Corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 117–124, Marseille, France. European Language Resources Association.

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G. van der Velden. 2022. Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, 16(1):1–18.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

Andreas Blätte. 2023. polmineR. Verbs and Nouns for Corpus Analysis. R package version 0.8.9.

Andreas Blätte and André Blessing. 2018. The GermaParl Corpus of Parliamentary Protocols. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 810–816, Miyazaki, Japan. European Language Resources Association.

Andreas Blätte, Julia Rakers, and Christoph Leonhardt. 2022. How GermaParl Evolves: Improving Data Quality by Reproducible Corpus Preparation and User Involvement. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 7–15, Marseille, France. European Language Resources Association.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124, Graz, Austria. Association for Computing Machinery.

Asher de Jong, Taja Kuzman, Maik Larooij, and Maarten Marx. 2024. ParlaMint Ngram Viewer: Multilingual Comparative Diachronic Search Across 26 Parliaments. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 110–115, Turin, Italy. ELRA and ICCL.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Darģis, Sascha Diwersy, Maria Gavriilidou, Ruben van Heusden, Mikel Iruskieta, Neeme Kahusk, Anna Kryvenko, Noémi Ligeti-Nagy, Carmen Magariños, Martin Mölder, Costanza Navarretta, Kiril Simov, Lars Magne Tungland, Jouni Tuominen, John Vidler, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, and Darja Fišer. 2025a. ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*, 59:2071–2102.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57:415–448.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

Maarten Janssen and Matyáš Kopp. 2024. ParlaMint in TEITOK. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 121–126, Turin, Italy. ELRA and ICCL.

Christoph Leonhardt and Andreas Blätte. 2023. Evaluating the Quality of the GermaParl Corpus of Plenary Protocols (v2.0.0). In *Proceedings of the 3rd Workshop on Computational Linguistics*

for the Political and Social Sciences, pages 88–100, Ingolstadt, Germany. Association for Computational Linguistics.

Christoph Leonhardt and Andreas Blätte. 2024. The dbpedia R Package: An Integrated Workflow for Entity Linking (for ParlaMint Corpora). In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 133–144, Turin, Italy. ELRA and ICCL.

Tomáš Machálek. 2020. KonText: Advanced and Flexible Corpus Query Interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France. European Language Resources Association.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*, Graz, Austria. Association for Computing Machinery.

Michal Mochtak, Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2025. Parlasent: mapping sentiment in political discourse with large language models. *Political Research Exchange*, 7(1):2508377.

Katrin Ortmann, Adam Roussel, and Stefanie Dipper. 2019. Evaluating Off-the-Shelf NLP Tools for German. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 212–222, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

R Core Team. 2025. R. A Language and Environment for Statistical Computing.

Florian Richter, Philipp Koch, Oliver Franke, Jakob Kraus, Lukas Warode, Fabrizio Kuruc, Stella Heine, and Konstantin Schöps. 2023. Open Discourse: Towards the first fully Comprehensive and Annotated Corpus of Parliamentary Protocols of the German Bundestag. SocArXiv.

Miklós Sebők, Sven-Oliver Proksch, Christian Rauh, Péter Visnovitz, Gergő Balázs, and Jan Schwalbach. 2025. Comparative European legislative research in the age of large-scale computational text analysis: A review article. *International Political Science Review*, 46(1):18–39.

Jure Skubic and Darja Fišer. 2022. Parliamentary Discourse Research in Sociology: Literature Review. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 81–91, Marseille, France. European Language Resources Association.

Jure Skubic and Darja Fišer. 2024. Parliamentary Discourse Research in Political Science: Literature Review. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 1–11, Turin, Italy. ELRA and ICCL.

Nina Smirnova, Muhammad Ahsan Shahid, and Philipp Mayr. 2025. Open Political Corpora: Structuring, Searching, and Analyzing Political Text Collections with PoliCorp. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 983–992, Suzhou, China. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Tomas Turner-Zwinkels, Oliver Huwyler, Elena Frech, Philip Manow, Stefanie Bailer, Niels D. Goet, and Simon Hug. 2022. Parliaments Day-by-Day: A New Open Source Database to Answer the Question of Who Was in What Parliament, Party, and Party-group, and When. *Legislative Studies Quarterly*, 47(3):761–784.

Ruben van Heusden, Maarten Marx, and Jaap Kamps. 2022. Entity Linking in the ParlaMint Corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 47–55, Marseille, France. European Language Resources Association.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Jan Wijffels and Milan Straka. 2026. udpipe: Tokenization, Parts of Speech Tagging, Lemmatiza-

tion and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. R package version 0.8.16.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzales-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

9. Language Resource References

Blätte, Andreas and Leonhardt, Christoph. 2024. *GermaParl Corpus of Plenary Protocols (v2.1.0)*. Zenodo. PID https://doi.org/10.5281/zenodo.12794676.

Blätte, Andreas and Leonhardt, Christoph. 2025. *GermaParl Corpus of Plenary Protocols (v2.3.0-rc1)*. Zenodo. PID https://doi.org/10.5281/zenodo.15495748.

Erjavec, Tomaž and Kopp, Matyáš and Kuzman Pungeršek, Taja and Ljubešić, Nikola and Ogrodniczuk, Maciej and Osenova, Petya and Agirrezabal, Manex and Agnoloni, Tommaso and Aires, José and Albini, Monica and Alkorta, Jon and Antiba-Cartazo, Iván and Arrieta, Ekain and Barcala, Mario and Bardanca, Daniel and Barkarson, Starkaður and Bartolini, Roberto and Battistoni, Roberto and Bel, Núria and Bonet Ramos, Maria del Mar and Calzada Pérez, María and Cardoso, Aida and Çöltekin, Çağrı and Coole, Matthew and Dargis, Roberts and de Libano, Ruben and Depoorter, Griet and Diwersy, Sascha and Dodé, Réka and Fernandez, Kike and Fernández Rei, Elisa and Frontini, Francesca and Garcia, Marcos and García Díaz, Noelia and García Louzao, Pedro and Gavriilidou, Maria and Gkoumas, Dimitris and Grigorov, Ilko and Grigorova, Vladislava and Haltrup Hansen, Dorte and Iruskieta, Mikel and Jarlbrink, Johan and Jelencsik-Mátyus, Kinga and Jongejan, Bart and Kahusk, Neeme and Kirnbauer, Martin and Kryvenko, Anna and Ligeti-Nagy, Noémi and Luxardo, Giancarlo and Magariños, Carmen and Magnusson, Måns and Marchetti, Carlo and Marx, Maarten and Meden, Katja and Mendes, Amália and Mochtak, Michal and Mölder, Martin and Montemagni, Simonetta and Navarretta, Costanza and Nitoń, Bartłomiej and Norén, Fredrik Mohammadi and Nwadukwe, Amanda and Ojsteršek, Mihael and Pančur, Andrej and Papavassiliou, Vassilis and Pereira, Rui and Pérez Lago, María and Piperidis, Stelios and Pirker, Hannes and Pisani, Marilina and Pol, Henk van der and Prokopidis, Prokopis and Quochi, Valeria and Rayson, Paul and Regueira, Xosé Luís and Rii, Andriana and Rudolf, Michał and Ruisi, Manuela and Rupnik, Peter and Schopper, Daniel and Simov, Kiril and Sinikallio, Laura and Skubic, Jure and Tungland, Lars Magne and Tuominen, Jouni and van Heusden, Ruben and Varga, Zsófia and Vázquez Abuín, Marta and Venturi, Giulia and Vidal Miguéns, Adrián and Vider, Kadri and Vivel Couso, Ainhoa and Vladu, Adina Ioana and Wissik, Tanja and Yrjänäinen, Väinö and Zevallos, Rodolfo and Fišer, Darja. 2025b. *Multilingual comparable corpora of parliamentary debates ParlaMint 5.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/2004. ISSN: 2820-4042.

Goldin, Gili and Howell, Nick and Ordan, Noam and Rabinovich, Ella and Wintner, Shuly. 2025. *Comparible corpus of parliamentary debates ParlaMint-IL 1.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/2032. ISSN: 2820-4042.

Kuzman Pungeršek, Taja and Ljubešić, Nikola and Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej and Osenova, Petya and Rayson, Paul and Vidler, John and Agerri, Rodrigo and Agirrezabal, Manex and Agnoloni, Tommaso and Aires, José and Albini, Monica and Alkorta, Jon and Antiba-Cartazo, Iván and Arrieta, Ekain and Barcala, Mario and Bardanca, Daniel and Barkarson, Starkaður and Bartolini, Roberto and Battistoni, Roberto and Bel, Núria and Bonet Ramos, Maria del Mar and Calzada Pérez, María and Cardoso, Aida and Çöltekin, Çağrı and Coole, Matthew and Dargis, Roberts and de Does, Jesse and de Libano, Ruben and Depoorter, Griet and Depuydt, Katrien and Diwersy, Sascha and Dodé, Réka and Fernandez, Kike and Fernández Rei, Elisa and Fron-

tini, Francesca and Garcia, Marcos and García Díaz, Noelia and García Louzao, Pedro and Gavriilidou, Maria and Gkoumas, Dimitris and Grigorov, Ilko and Grigorova, Vladislava and Haltrup Hansen, Dorte and Iruskieta, Mikel and Jarlbrink, Johan and Jelencsik-Mátyus, Kinga and Jongejan, Bart and Kahusk, Neeme and Kirnbauer, Martin and Kryvenko, Anna and Ligeti-Nagy, Noémi and Luxardo, Giancarlo and Magariños, Carmen and Magnusson, Måns and Marchetti, Carlo and Marx, Maarten and Meden, Katja and Mendes, Amália and Mochtak, Michal and Mölder, Martin and Montemagni, Simonetta and Navarretta, Costanza and Nitoń, Bartłomiej and Norén, Fredrik Mohammadi and Nwadukwe, Amanda and Ojsteršek, Mihael and Pančur, Andrej and Papavassiliou, Vassilis and Pereira, Rui and Pérez Lago, María and Piperidis, Stelios and Pirker, Hannes and Pisani, Marilina and Pol, Henk van der and Prokopidis, Prokopis and Quochi, Valeria and Regueira, Xosé Luís and Rii, Andriana and Rudolf, Michał and Ruisi, Manuela and Rupnik, Peter and Schopper, Daniel and Simov, Kiril and Sinikallio, Laura and Skubic, Jure and Tamper, Minna and Tungland, Lars Magne and Tuominen, Jouni and van Heusden, Ruben and Varga, Zsófia and Vázquez Abuín, Marta and Venturi, Giulia and Vidal Miguéns, Adrián and Vider, Kadri and Vivel Couso, Ainhoa and Vladu, Adina Ioana and Wissik, Tanja and Yrjänäinen, Väinö and Zevallos, Rodolfo and Fišer, Darja. 2025. *Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 5.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/2006. ISSN: 2820-4042.