Evaluating the Quality of the GermaParl Corpus of Plenary Protocols (v2.0.0)

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

Abstract

Parliamentary debates play a key role for the democratic process and for law-making. Scholarly interest in this material benefits greatly from the emergence of new datasets and corpora of parliamentary protocols. Here we combine the presentation of a second, extended version of GermaParl with an evaluation of the data quality of this corpus of plenary protocols in the German Bundestag. For this purpose, about 1 per cent of all protocols have been annotated manually to create a gold standard against which the structurally annotated corpus is compared. Results indicate that GermaParl can be considered a trustworthy resource for a broad set of research questions.

1 Introduction

The increasing availability of large collections of text enables researchers to address new substantive research questions and paves the way for a multitude of new methodological approaches (Hurtado Bodell et al., 2022, p. 1). Whether qualitative discourse analysis or computationally elaborate text-as-data approaches, corpora are the foundation for many new research avenues. In particular, research on legislative debates (Fernandes et al., 2021) benefits from the emergence of corpora of parliamentary proceedings (Sebők et al., 2021).

It has become common sense that the pure existence of new (parliamentary) data is not enough. Availability and reusability matters, and the FAIR principles (Wilkinson et al., 2016) are becoming a cornerstone of data-driven research. But as research moves beyond experimental explorations of new methods, concerns about data quality receive increasing attention: Sound data quality is a precondition for trustworthy research and valid findings. While this is not at all unique for new, large datasets, the volume of the data, their often complex structure and intricate processing pipelines make quality control particularly important for big data.

Relevant key concerns for data quality depend on the type of data. Analyzing Twitter tweets requires awareness for and scrutiny of the technical sampling issues faced. For large corpora of newspaper articles, the presence of (near) duplicates can heavily distort results and needs to be controlled. For parliamentary data, turning raw material into semi-structured data formats (such as XML) in an automated process without a realistic possibility to hand-check output manually throughout entails many potential sources of errors. This is increasingly debated and there is an emerging concern with the quality of resources used for conducting large-scale data-driven research.

The emerging literature on data quality in big data settings emphasizes the need for rigorous quality control. The "Total Error Framework" (RatSWD, 2023, pp. 9–10) and the "Framework for Total Corpus Quality" (Hurtado Bodell et al., 2022) are important contributions to the evolving practice of evaluating data quality. They build on the "Total Survey Error Framework" established in survey methodology (Hurtado Bodell et al., 2022; RatSWD, 2023). These frameworks have an integrative view for the quality of the data. The "Framework for Total Corpus Quality" includes a concern for the transparency of its preparation and its usability to assess the way they facilitate fruitful research. As Hurtado Bodell et al. (2022, p. 12) put it:

"We suggest that it is now time to turn to a systematic analysis of the role of data quality in scientific inference from textual data. It is time to open the door into the messy data kitchen".

We here apply these considerations to a corpus we prepared and released earlier this year.[1]

[1] An evaluation of a resource conducted by its authors is not independent and can be perceived to have limited value due to the obviously lacking critical distance. However, we think

Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences, pages 88–100

GermaParl v2.0.0, released in May 2023, is a comprehensive update of GermaParl, as an established corpus of parliamentary protocols described in Blätte and Blessing (2018). As data quality has always been a key concern of the curation project, previous presentations of GermaParl had a focus on the data preparation workflow, which is designed to facilitate continuous improvements of the data (Blätte et al., 2022). Going beyond our earlier work with its procedural focus, this contribution addresses the question which level of substantive data quality has been achieved.

Our analysis is also inspired by the work of Paul Ramisch who recently addressed the issue of data quality from the perspective of historical source criticism (2023). In his work, he evaluated the quality of another corpus of debates of the *Bundestag* using a gold standard approach. While his approach – by explicitly taking into account the representation of the contents of speeches – is more comprehensive than the one we will employ, it inspired us to evaluate the quality of our corpus by comparing the processed data with a sample of the raw data. We thus feel intellectually indebted to the work presented by Ramisch (2023).

We proceed as follows: After a brief overview of the GermaParl corpus and its preparation process, the framework used to estimate data quality is introduced. Based on an explanation how a benchmark dataset has been prepared, the actual assessment of the data quality of the corpus is presented. The contribution concludes with a discussion of the results and an outlook.

2 The GermaParl corpus of parliamentary debates

2.1 Data Formats and Preparation

The GermaParl corpus includes all proceedings of the German *Bundestag* from 1949 to 2021 and is published in two different formats:

- **TEI/XML:** A structurally annotated TEI/XML format. Text is segmented into individual utterances. This version is available on GitHub.[2]
- **CWB:** An indexed version of the corpus, imported into the Corpus Workbench (CWB) (Evert and Hardie, 2011). It is structurally and linguistically annotated and available via Zenodo.[3]

An outline of preparation procedures is important to convey where potential errors in the data might be introduced. In a nutshell, the data preparation process starts with downloading the raw data from the website and online archives of the *Bundestag*. It is processed in a pipeline that includes cleaning, preprocessing, the structural annotation of the text as well as the enrichment of the data with additional information. For the CWB version of the corpus, the text is linguistically annotated. Finally, the data is imported into the CWB.

Three aspects of the data preparation are particularly important:

Preprocessing: The raw data is retrieved from the websites of the German *Bundestag* using different file formats (TXT, PDF and XML). All file formats already include digitized text one way or the other. Concerning PDF, we did not have to perform any form of Optical Character Recognition, as the *Bundestag* has already done that. When the raw protocols were available in more than one file format, data quality was the key consideration to opt for a file format. Each file format required some adjustments to the processing pipeline.

Speaker Annotation: GermaParl is structurally annotated, making it possible to variably create corpus subsets. Most importantly, it is possible to zoom in on individual speeches. The beginning of speeches is detected by matching specifically marked up lines in the protocols using a set of regular expressions. This may result in false positives and negatives. To omit false positives, a list of manually identified mismatches is used rather than refining the regular expressions until they cover all specific cases, making the expression incomprehensibly and error-prone.

Enrichment: To add information to identified speakers which is not part of the initial protocols – such as a speaker's party affiliation or the speaker's full name in some legislative periods – external data sources are used. Predominantly, additional information can be added using deterministic matching

that we offer insights into the specific challenges of curating a corpus of GermaParl's characteristics. By making the evaluation exercise fully transparent, we generate opportunities for third-party checks and a safeguard against manipulation. That being said, we would welcome future independent work comparing different corpora and using different approaches like the one suggested by Ramisch (2023).

[2] https://github.com/PolMine/GermaParlTEI.

[3] https://zenodo.org/record/7949074.

of shared attributes between the protocol and external data. But plenary protocols include errors and inconsistencies, so fuzzy matching is used to consolidate the name of a speaker. Most external information is retrieved from Wikipedia or the *Stammdaten* file of the German *Bundestag*.[4] If a speaker could not be identified on Wikipedia, alternative resources such as the Munzinger encyclopedia[5] are used selectively. To increase the usability of GermaParl, metadata at the speaker level has been harmonized. Most importantly, variations of parties and parliamentary groups are consolidated.

As elaborated on in Blätte et al. (2022), the workflow includes manual steps, yet it is fully automated and reproducible by design (see Blätte and Leonhardt (2023) for a full description). This is the prerequisite for an efficient and sustainable evolution of the resource, including successive improvements of data quality.

2.2 Data Report

GermaParl v2.0.0 comprises 273 million tokens, covering 72 years of parliamentary debates in 4341 individual protocols.[6] It provides a number of different annotation levels which are comprehensively documented in the online documentation of the resource (Blätte and Leonhardt, 2023).

The structural annotation of GermaParl covers metadata at the protocol and the speaker level. One important purpose of these attributes is to create subcorpora for synchronic and diachronic analyses according to relevant criteria. Table 1 provides an overview of the key structural attributes of GermaParl.[7]

The corpus is linguistically annotated. Aside from tokenization and sentence segmentation, Part-of-Speech tags (Universal Dependencies Tag Set provided by Stanford CoreNLP (Manning et al., 2014) and the Stuttgart-Tübingen Tag Set provided by TreeTagger (Schmid, 1994)) and lemmata (provided by TreeTagger (Schmid, 1994)) are added at the token level. While named entities, added by Stanford CoreNLP (Manning et al., 2014), are part of the linguistic annotation, they are implemented as structural attributes, reflecting that this annotation layer can span more than one token. The same applies to the annotation of sentences.

2.3 Getting Started with GermaParl

The XML version of GermaParl serves as a persistent interchange data format. It is relevant for technically oriented users that are used to process XML and that have own pipelines and infrastructures for handling large corpora. Yet given the size and the structure of the data, many users from the social sciences and the humanities will find the XML variant of GermaParl overwhelming. The CWB version provides this group of users with a linguistically annotated resource in a data format suitable for efficient data analysis.

The CWB version of the corpus can be analyzed with different compatible tools such as the Corpus Workbench itself (Evert and Hardie, 2011) or the Graphical User Interface CQPweb.[8] To access the CWB using the statistical programming language R, we offer the `polmineR` R package which is created and maintained by one of the authors of this contribution (Blätte, 2023). `polmineR` provides fast and reliable access to the functionality of the Corpus Workbench, including the powerful CQP query language. Analyzing large corpora and making use of the rich structural and linguistic annotation layers thus becomes accessible for scholars comfortable with the R programming language. `polmineR` is interoperable and tested to run out of the box and fast on (local) Windows, macOS and Linux machines, even for large corpora such as GermaParl. To download and install the corpus from Zenodo, the R package `cwbtools` (Blätte, 2022) provides convenient auxiliary functionality.

On a system with a working installation of R, the following lines of code suffice to install and run GermaParl.[9]

```
# install cwbtools and polmineR
install.packages("cwbtools") # >= v0.3.8
install.packages("polmineR") # v0.8.8
```

[4] The Stammdaten file can be retrieved from the open data website of the German *Bundestag* (https://www.bundestag.de/services/opendata). It contains comprehensive information on all members of parliament.

[5] https://www.munzinger.de/

[6] GermaParl is an evolving resource; future updates will extend its temporal coverage, and fix errors in the data either found by ourselves or reported by users.

[7] This overview describes the CWB version of the corpus. While the structural attributes are essentially identical in the TEI/XML version of the corpus, linguistic annotation was performed only for the CWB version.

[8] https://cwb.sourceforge.io/cqpweb.php.

[9] This will install the v2.0.0 release version of the corpus. For future updates, the Zenodo landing page (https://doi.org/10.5281/zenodo.3735140) will resolve to the latest version.

Structural Attribute	Description
protocol_lp	Legislative period
protocol_no	Session number
protocol_date	Date of the protocol
protocol_year	Year derived from date
speaker_name	Full name of the speaker
speaker_parlgroup	Parliamentary group of a speaker, corrected errors when necessary
speaker_party	Party affiliation of a speaker, retrieved from Wikipedia or other external resources
speaker_role	Parliamentary role of a speaker, derived from speaker call
p / p_type	paragraph / type of paragraph (speech or stage)
ne / ne_type	named entity / type of named entity

Table 1: Structural Attributes in the GermaParl Corpus

```
# install GermaParl2
cwbtools::corpus_install(
  doi = "10.5281/zenodo.7949074"
)

# test GermaParl2 installation
polmineR::corpus("GERMAPARL2") |>
  size()
```

3 Measurement of Data Quality - Method and Design

3.1 Data Quality as truthful textual representation

GermaParl v2 covers 72 years of parliamentary history, significantly extending the time covered by the v1 release of GermaParl which was limited to 1996 to 2016. The question of data quality needs to be addressed anyway, but given the additional error sources that enter the game for data that is not born-digital (scanning quality, OCR errors), historical data make data quality issues more pressing. If systematic errors remain unknown, the potential of data covering several decades of parliamentary history to uncover long-term trends is significantly impeded.

In this section, we discuss our understanding of corpus quality and how it can be measured. The approach borrows heavily from the "Framework for Total Corpus Quality" presented by Hurtado Bodell et al. (2022). The framework is proposed as "a conceptual framework for assessing the quality of textual data that enables researchers to systematically diagnose a corpus' scientific value along three quality dimensions: total corpus error, corpus comparability, and corpus reproducibility" (Hurtado Bodell et al., 2022, p. 1). As such, it is part of a family of established approaches, most importantly the "Total Survey Error Framework" and related efforts to extend this framework to the realm of big data and unstructured data (Hurtado Bodell et al., 2022; RatSWD, 2023).

In this first take to assess the quality of GermaParl, we focus on the dimension of "total corpus error" (Hurtado Bodell et al., 2022, p. 1). It has three aspects: "source errors, textual representation errors (TREs), and research inference errors (RIEs)" (Hurtado Bodell et al., 2022, p. 4). Within this triad, we will mainly focus on the aspect of "textual representation errors". Since we work with already digitized data, systematically checking the "source errors" (Hurtado Bodell et al., 2022, p. 4) is out of the scope of this contribution.[10] As a multi-purpose corpus which has been created to broadly serve research, "research inference errors" can not be estimated meaningfully either.

Thus, we employ a simplified version of this framework, asking how well the corpus represents the original data in the form published by the *Bundestag* and how truthfully additional information has been added to this data (Hurtado Bodell et al., 2022, pp. 4–5). To do this, we compare the processed TEI/XML version of GermaParl v2.0.0 with the initial raw protocols in form of the PDF files

[10] This does not address whether the transcripts represent everything that happens in parliament truthfully. This question is beyond the scope of this contribution. It has been analyzed and discussed for the German *Bundestag* in-depth in dedicated studies (Burkhardt, 2003, chapter 9). Also errors in the data provided by the German *Bundestag* (Ramisch, 2023, chapter 2) are not evaluated systematically.

which can be retrieved from the "Dokumentations- und Informationssystem für Parlamentsmaterialien" (DIP) of the German *Bundestag*.[11]

When focusing on the "Textual Representation Errors", we are concerned with the question of "How different [...] the processed machine-readable and observed corpus [are]" (Hurtado Bodell et al., 2022, p. 4). Hurtado Bodell et al. (2022) discuss this along four categories that lend structure to our evaluation. For each category, the error itself is described first, followed by potential causes of these errors in GermaParl.

source-to-(digital)-text errors Following Hurtado Bodell et al. (2022, pp. 4–5), transforming the source data into a machine-readable format is a first category of errors. Potential errors comprise flaws introduced by the digitization itself – scan artefacts, for example – or the inclusion of unwanted parts of the source material. We largely omit this aspect from our analysis because of our reliance on digitized text provided by the German *Bundestag*. So digitization errors like random additional or missing characters which might be caused by scan artefacts (Hurtado Bodell et al., 2022, p. 5) are mostly out of our control and are not systematically identified as long as they do not result in a missing speaker call.

text-to-documents errors Hurtado Bodell et al. (2022, p. 5) describe the identification of "cohesive units of text" as the source of "text-to-documents" errors. For the curation of qualitative corpora, the correct segmentation of text to meaningful documents such as speeches is crucial. The relevance of these errors is particularly evident for the assignment of speakers to segments of text in parliamentary corpora. If the beginning of a separate speech is missed, additional chunks of text are incorrectly assigned to the wrong speaker. The same is true for the creation of "faux documents" (Hurtado Bodell et al., 2022, p. 5) if separate speeches are detected where they should not.

These errors concern a step of the corpus preparation pipeline of GermaParl that is truly crucial: The identification of speeches. The sequence of text preprocessing, applying regular expressions, and the handling special cases as well as false positives is essential for the correct assignment of text to speakers, and potentially error-prone.

[11] https://dip.bundestag.de/.

documents-to-corpus errors According to Hurtado Bodell et al. (2022, p. 5) the "accuracy of metadata in a corpus" gives rise to "documents-to-corpus errors".

The capabilities to enrich identified speeches with additional metadata are important for the data quality of GermaParl, as these additional annotations provide plentiful possibilities for analysis. As described in section 2, the enrichment is realized by matching attributes found in the protocols and external data; "documents-to-corpus errors" thus would materialize in mismatches, such as wrongly assigned party affiliations.

processing errors Processing errors arise when transforming the machine-readable corpus from one format to another (Hurtado Bodell et al., 2022, p. 6). For GermaParl, this might be the case when importing the processed XML files into the Corpus Workbench. It must be noted that the TEI/XML version and the CWB version of the corpus differ by design, with the latter including an additional consolidation step to increase usability while the TEI/XML contains some more variations within party and parliamentary group names.

3.2 Research Inference Errors

GermaParl is designed as a multi-purpose resource and is, as such, not concerned with a single research question in mind. As a consequence, other errors identified by Hurtado Bodell et al. (2022, p. 6) are not entirely applicable for our curation project. While "coverage errors" – how far the data represents its stated population – and "text curation errors" – issues caused by the modification and preprocessing of text – might be relevant for corpora like GermaParl as well, this is not systematically addressed in the upcoming evaluation.

3.3 Corpus Comparability and Corpus Reproducibility

Aside from estimating the Total Corpus Error as discussed above, Hurtado Bodell et al. (2022) suggest two more dimensions of corpus quality: Corpus comparability and corpus reproducibility.

Corpus comparability is concerned with how findings based on one resource compare to findings based on another or how findings based on different sections of the same resource are comparable (Hurtado Bodell et al., 2022, p. 6). This is particularly relevant in terms of errors in the data. For diachronic analyses, missing a lot more ob-

servations in one period that in another should be avoided (Hurtado Bodell et al., 2022, p. 7). Concerning corpus comparability, as shown in the previous sections, the data sources – while all provided by the German *Bundestag* itself – are not completely homogeneous. While not in our control, it seems obvious that the "within-corpus comparability" (Hurtado Bodell et al., 2022, p. 7) might be limited by different processes to retrieve the data as text. The data quality of the raw data at different points in time is also discussed in more detail by Ramisch (2023, chapter 2). These potential challenges require thorough empirical evaluation in the upcoming sections.[12]

Regarding corpus reproducibility described by Hurtado Bodell et al. (2022, p. 7) as the goal that "two different researchers should be able to create the same corpus from the same observed material", we already presented our approach to reproducibility (Blätte et al., 2022). We are strongly opinionated in this respect: Reproducibility of the data preparation process contributes to the quality of the data not only in the sense that reproducibility is desirable in its own right. Much more than that, it is a way to ensure that a resource can evolve, incrementally increasing data quality. If the preparation workflow is not reproducible, the maintaining a resource is excessively costly.

4 Applying the Total Corpus Error framework

In the previous section, we described what potential errors might be expected. Our focus on the textual representation error informs the need to develop an understanding on what a truthful representation of the debates in the German *Bundestag* would look like. In other words, we need to create a "ground truth" that contains information about which speeches actually occur in the debates, when these debates actually occurred and what additional information should be added. A compiled representation of the true debates allows us to compare these expected speeches with the speeches in the processed corpus. In contrast to the approach by Ramisch (2023, chapter 3) who is also interested in the extent of speeches, we focus on the metadata of each speech by annotating and enriching each line indicating the beginning of an individual speech. Implicitly, these errors correspond to the false assignment of tokens to speakers where the beginning of a new speech is missed. Instead of assigning tokens to the expected but missed speaker, in most cases they will be assigned to the previous speaker instead (see Ramisch (2023, chapter 3.5.2) as well).

The precise steps are discussed in more detail in the following sections.

4.1 Sampling and Ground Truth

When creating this "ground truth", it would be unfeasible to collect the necessary information for each protocol in a larger corpus. Indeed, it is enough to evaluate a representative sample of documents. Hurtado Bodell et al. (2022, p. 8) assessed a stratified random sample of newspaper pages. We also annotate a representative sample of parliamentary protocols. To account for the changing appearance of the protocols, changes in parliamentary procedures or the changing composition of parties in the *Bundestag*, each legislative period should be included in the sample with at least two sessions. Our overall target was to annotate one per cent of the entire corpus.[13]

To organize the collection of information, a codebook outlining the annotation task was created. It contained information about how document-level metadata and speeches should be identified and documented (allowing the identification of potential text-to-documents errors) and how the metadata of speeches should be enriched with additional information (the full name and the party affiliation) to facilitate the identification of documents-to-corpus errors. The coders were provided with specific instructions about which resources to use to add metadata if possible.

The annotation task was assigned to four coders: one author of this contribution and three student assistants with a background in political science. Each protocol was initially coded by a single coder. With the categories being formal rather than evaluative and the codebook quite specific, the risk of "coder bias" – an important limitation in quantitative content analysis (Riffe et al., 2005, p. 123) – was considered as neglectable. To guarantee that the corpus is compared against an accurate ground truth, obvious remaining flaws such as missing

[12] The comparability to other corpora is no aspect of the data quality of GermaParl. However, it can be noticed that the XML version is currently TEI-inspired. Future versions of GermaParl are envisaged to adhere to the encoding standards of the ParlaMint project (Erjavec et al., 2022)

[13] Similarly, Ramisch (2023, chapter 3) manually annotated two protocols per legislative period, using the XML files provided by the German *Bundestag*.

speakers, typos in the gold standard or falsely assigned additional information were consolidated. Some of these flaws were noticed when the initial gold standard annotation was initially compared against the processed data and corrected accordingly. In sum, to ensure completeness and accuracy, after the initial annotation each protocol was looked at by at least one, sometimes also a second, additional coder – i.e. with access to the previous annotation – to iteratively create a complete and accurate gold dataset.

To ensure the comparability of the added data in the ground truth and the processed data in the corpus, minor harmonization steps were performed on the ground truth such as the adoption of a party abbreviation from GermaParl as well as the adoption of variations in speaker names – the removal or addition of middle initials, for example. The goal is to identify corresponding entities, not necessarily verbatim matches.

It has to be noted that this approach potentially comes with some limitations and biases which are discussed in the respective section on limitations at the end of this contribution.

Ultimately, the coded sample comprised 51 protocols (1.17 per cent of all protocols). Table 3 (see appendix) shows the number of annotated speakers per legislative period. For each protocol, the occurring speakers and additional metadata were documented in order of occurrence along with document-level metadata.

4.2 Estimation of Corpus Quality

The final measure of corpus quality is the proportion of correct assignments over different subsets of the corpus. First, we analyze the metadata at protocol level to estimate documents-to-corpus errors.[14]

For the speaker level, this measure includes both the assignment of tokens to the correct speaker (addressing potential text-to-documents errors) as well as assigning the correct metadata to the correct speaker (addressing potential documents-to-corpus errors). We compare each speaker in the gold standard representing the initial data with the corresponding observation in the processed data. This comparison can result in five different states:

- **full match:** Same speaker matched in processed data, metadata identical.
- **partial match:** Same speaker matched in processed data, metadata (partially) different.
- **missing:** Speaker not matched in the processed data.
- **mismatch:** Different, unexpected speaker matched.
- **only in GermaParl:** Speaker occurs in the processed data but not in the gold standard, indicating false positives or overlooked speakers when creating the ground truth

In particular, we are interested in the accuracy of the representation of the data split according to different comparative dimensions. These dimensions are the general accuracy of the data, as well as the accuracy per legislative period, parliamentary role and parliamentary group.

4.2.1 Protocol Level Annotation

To assess documents-to-corpus errors at the level of the entire protocol, the question is whether each protocol is enriched with the correct metadata. Thus, the metadata of the protocols – the legislative period, the session number and the session date – was documented for all protocols which were included in the gold standard evaluation. As table 2 indicates, this error is not very prominent in our sample. One wrong date resulted from a session taking place on two separate days – only the first date is reported in the processed data.

Attribute	Matching	Documents	Correct Matches in %
Legislative Period	51	51	100.00
Session	51	51	100.00
Date	50	51	98.04

Table 2: Accuracy of Document Level Metadata in GermaParl

4.2.2 Speaker Level Annotation

Out of 10725 annotated speakers, 10398 are fully matched in the processed data. This represents 96.95 per cent of all speakers. 194 speakers (1.81

[14] For the overwhelming majority of protocols, a single protocol corresponds to a single parliamentary session. While we know that this does not apply for all protocols, we did not encounter multiple sessions in one protocol in our sample, thus making the text-to-documents-error less important at this level.

per cent) were identified, but annotated with metadata which differs from the expected values. 69 speakers (0.64 per cent) are not matched at all. 64 speakers were mismatches. This represents 0.6 per cent. 68 speakers occur only in the processed data and not in the gold standard.

While these overall values are relevant, the accuracy of the data might vary along a set of dimensions. Table 4 in the appendix shows the results of this comparison along these different dimensions. Considering variation over time, we see that the proportion of complete matches is relatively stable over different legislative periods. Noteworthy outliers are the second, the seventh and the 14th legislative period, with a comparatively high number of partial and missing matches. Regarding the parliamentary role of speakers, the accuracy to identify speakers of the federal council (i.e. members of the German *Bundesrat*) is comparatively low. For parliamentary groups, we do not see major deviations. Focusing specifically on mismatches, we identify an increased number of mismatches in the 14th legislative period and for presidential speakers.

Regarding the documents-to-corpus errors, there is relevant variation in the proportion of partial matches. For some cases, the explanation is quite simple: For some governmental and presidential speakers, parliamentary groups are reported in the processed data where they should not. This also explains the high number of partial matches in the "NA" category in the parliamentary group section. Other speakers have false assignments of parliamentary groups or parties. While this might be due to switching parties, this deserves further investigation. While mismatches do not occur very often, they can represent crucial errors in the data. For some instances, these errors are false positives in the sense that the expected speaker and the speaker detected are actually the same person with a different name, for example because of marriage. In our case, this accounts for quite a large number of mismatches: 48 mismatches are caused by a mismatch between the expected speaker "Petra Bläss" and the observed speaker "Petra Bläss-Rafajlovski", for example. For this reason, a more granular analysis of the nature of these mismatches might be relevant. For other cases, more investigation is needed. Speakers found only in GermaParl often correspond with these mismatches. In this case, instead of the expected value in the gold annotation, other speakers were added in the processed data, leaving them unmatched. Currently, errors in the gold standard cannot be ruled out, so that these instances might point to speakers which are in the protocols but were overlooked in the gold standard annotation. But in general, the number of these cases is relatively low.

4.3 Processing Error

The *processing error* is estimated by comparing the observations in both versions, with the proportion of corresponding observations as the central measure. All errors reported for the TEI/XML version will also be part of the CWB corpus.

We assume that the CWB corpus is equivalent to the TEI/XML version of the corpus. There are just cases of a minor harmonization to increase the usability of the CWB resource. The empirical analysis supports this: While most speakers (98.86 per cent) are identical in both versions of the corpus, there are differences in 122 of the speakers identified in the evaluated protocols. For the most part, this concerns the assignment of parliamentary groups (0.62 per cent of all speakers) and parties (0.51 per cent). A preliminary glance at the deviations suggests that both are indeed caused by minor variations in the names of the same entities with the most noteworthy deviation being the inclusion of the CDU as a parliamentary group in the first legislative period in the XML/TEI version of the corpus whereas it has been harmonized to CDU/CSU in the CWB corpus.

5 Discussion

Our overall result of this evaluation exercise is: The overwhelming majority of speakers is identified – representing little text-to-documents errors – and assigned to the correct metadata – suggesting few documents-to-corpus errors. That being said, the data is not yet perfect: Specific groups of speakers are identified more robustly than others.

While for some research questions, the assignment of tokens to reasonable documents will be sufficient, for others the correct assignment of metadata throughout is imperative. Thinking about a continuum between in-depth qualitative analysis of a limited set of debates and speeches and quantitative text-as-data approaches to the data: The latter strand of research will find some noise that does not systematically distort results to be anticipated and acceptable, whereas in-depth qualitative research may require a zero-tolerance take on errors – a stan-

dard only a genuine edition could meet. Our findings on data quality convey that GermaParl may be considered a resource meeting sound quality standards for a broad set of analytical approaches to parliamentary speech, though not for all.

While the percentages shown table 4 indicate that the general workflow works well, improvements are possible and will be made. Evaluating missing speakers qualitatively suggests that the quality of the raw data is a limiting factor in this regard. Typos, missing or additional punctuation marks and whitespace as well as missing line breaks limit the effectiveness of our approach. In some instances, the preparation pipeline is able to account for this. However, the occurrence of noise is difficult to anticipate. Other errors concerning partially matched speakers seem to indicate plain inaccuracies in the preparation of the data used to enrich the corpus. Our findings also confirm the prior intuition that rare speakers are more difficult to match than common ones: Speakers from the federal council occur comparatively rarely and in quite a variety of different forms, making the formulation of regular expressions matching all relevant cases challenging.

Finally, while the sample used to generate the ground truth covers a large proportion of the data, we did not encounter all errors which are known to us at the time of writing. For instance, a known data error in GermaParl v2.0.0 is the unintended inclusion of appendices in the final dataset. Depending on the legislative period and the specific document, this either assigns additional content to the last speech – most of the time a presidential speaker – or adds speeches which were only added to the minutes, suggesting these were ordinary speeches. While the first issue seems related to a text-to-documents error, the second issue can be understood as a case of a coverage error because the intended coverage – speeches held in the German *Bundestag* – is exceeded in a portion of the protocols. Errors such as these are publicly documented in the GitHub repository of the resource.[15] Future versions of the will improve data quality by addressing these known errors.

6 Outlook

We envision GermaParl as both a trustworthy and useful resource for a broad set of research questions, and as an evolving resource which allows for continuous updates and improvements. We did not compare the quality of GermaParl to similar resources, i.e. other corpora of parliamentary debates. A comparative contextualization of the reported measures would ideally be provided for by independent researchers. Yet our own evaluation of our resource leaves us with newly-won, quantitatively grounded confidence that – remaining errors notwithstanding – the quality of GermaParl achieved is a solid foundation for current research and further developments.

The qualitative inspection of errors encountered underlines the need to improve the resource continuously in a collaborative and sustainable fashion: It is impossible to anticipate all errors in a corpus as large as GermaParl: It covers 72 years of parliamentary proceedings, 19 legislative periods and includes more than 273 million tokens in 4341 protocols. Thus, user feedback and suggestions are an important aspect for the future development of the corpus, including its data quality.

Limitations

This contribution systematically compares an accurate account of the debates in the German *Bundestag* and its representation in the GermaParl corpus. The "gold standard" has been generated in an iterative process that may have introduced a bias: The identification and correction of speakers which are missing in the ground truth (but are available in the processed data) is potentially easier than the identification of errors which occur in both the ground truth and the processed data. To avoid a potentially lopsided correction of errors which would flatter the results presented, the gold standard dataset was checked iteratively in the process outlined. Our reasoning was to design a process to obtain a gold standard annotation for a technical annotation task with little interpretative leeway that might have caused intercoder disagreement. Still, random noise and annotation errors cannot be ruled out. A consequence of our process is that we do not offer a measure of the intercoder reliability between the four coders in the initial annotation, nor a measure of the difficulty of the annotation task.[16]

A further aspect we do not discuss in depth is that we encountered errors in the PDF files such as missing pages resulting in missing speakers. Relying on the PDF files to create the gold standard

[15] https://github.com/PolMine/GermaParl2.

[16] We gratefully acknowlege our reviewers' discussion of this limitation.

annotation then results in additional errors which are not necessarily caused by errors in GermaParl.

Finally, the current implementation of the algorithm used to compare the gold standard and the processed data is very sensitive for a large number of missing speakers occurring consecutively, flagging all speakers after a specific gap as mismatches even though valid speaker matches would be available later. While the chosen parameters worked well, it is conceivable that this could overestimate the number of mismatches if a number of consecutive speakers is missing in GermaParl.

Ethical Considerations

The parliamentary data we prepared is entirely in the public domain and the data preparation process is fully transparent. We are not aware of a scenario how our work might negatively affect relevant principles of research ethics. As we see it, our contribution is also technically improved access to parliamentary debates that strengthens democratic accountability.

Acknowledgments

The preparation of GermaParl and this evaluation was made possible by funding from KonsortSWD within the National Research Data Infrastructure (NFDI) (Project Number 442494171) as well as from the Text+ consortium within the NFDI (Project Number 460033370). We gratefully acknowledge this support. In addition, we would like to thank our three student assistants Silvia Mommertz, Jan Erik Lutz and Jan Borcherding who supported this work by conducting a part of the manual annotation for this quantitative evaluation with great scrutiny. Finally, the insightful comments of our reviewers are gratefully acknowledged.

References

Andreas Blätte. 2022. cwbtools: Tools to create, modify and manage CWB Corpora. R package version 0.3.8.

Andreas Blätte. 2023. polmineR: Verbs and Nouns for Corpus Analysis. R package version 0.8.8.

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

Andreas Blätte and Christoph Leonhardt. 2023. The GermaParl Corpus of Plenary Protocols (v2.0.0) - Documentation. Version 2023-05-23. Technical report.

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.

Armin Burkhardt. 2003. *Das Parlament und seine Sprache. Studien zu Theorie und Geschichte parlamentarischer Kommunikation*. Max Niemeyer Verlag, Berlin, New York.

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. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*.

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, GBR.

Jorge M. Fernandes, Marc Debus, and Hanna Bäck. 2021. Unpacking the politics of legislative debates. *European Journal of Political Research*, 60:1032–1045.

Miriam Hurtado Bodell, Måns Magnusson, and Sophie Mützel. 2022. From Documents to Data: A Framework for Total Corpus Quality. *Socius*, 8:1–15.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven 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. Association for Computational Linguistics.

Paul Ramisch. 2023. Goldstandard-Korpus-Evaluation als Methode digitaler Quellenkritik in den Geschichtswissenschaften am Beispiel des Open-Discourse-Korpus der Bundestagsprotokolle.

RatSWD (Rat für Sozial- und Wirtschaftsdaten). 2023. Erhebung und Nutzung unstrukturierter Daten in den Sozial-, Verhaltens- und Wirtschaftswissenschaften:

Herausforderungen und Empfehlungen. (RatSWD Output Series, 7. Berufungsperiode Nr. 2), Berlin.

Daniel Riffe, Stephen Lacy, and Frederick G. Fico. 2005. *Analyzing Media Messages. Using Quantitative Content Analysis in Research*, 2 edition. Lawrence Erlbaum Associates, Inc. Publishers, Mahwah, New Jersey.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Manchester, UK.

Miklós Sebők, Sven-Oliver Proksch, and Christian Rauh. 2021. OPTED. Review of available parliamentary corpora. Technical report.

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 Gonzalez-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).

A Appendix

Legislative Period	N Speakers	N Protocols
1	473	4
2	280	3
3	371	2
4	1000	3
5	590	3
6	673	3
7	984	4
8	563	3
9	300	2
10	792	4
11	303	2
12	864	3
13	1050	3
14	370	2
15	548	2
16	271	2
17	350	2
18	263	2
19	680	2

Table 3: Ground Truth - Sample

		Match Category					
	annotated speakers	full	partial	missing	mismatch	matched speakers*	only in GermaParl
Legislative Period							
1	473	461	7	5	0	97.46	0
2	280	258	18	4	0	92.14	0
3	371	360	1	10	0	97.04	0
4	1000	968	28	4	0	96.80	0
5	590	588	0	2	0	99.66	0
6	673	651	4	15	3	96.73	3
7	984	906	74	3	1	92.07	1
8	563	548	12	2	1	97.34	1
9	300	299	0	1	0	99.67	0
10	792	782	4	5	1	98.74	1
11	303	296	3	4	0	97.69	4
12	864	858	4	2	0	99.31	0
13	1050	1038	0	4	8	98.86	8
14	370	320	0	4	46	86.49	46
15	548	546	0	0	2	99.64	2
16	271	269	0	2	0	99.26	0
17	350	337	13	0	0	96.29	0
18	263	259	0	2	2	98.48	2
19	680	654	26	0	0	96.18	0
Role							
federal_council	17	11	1	5	0	64.71	0
government	1980	1855	104	17	4	93.69	5
mp	4254	4206	12	21	15	98.87	16
parl_commissioner	4	4	0	0	0	100.00	0
presidency	4470	4322	77	26	45	96.69	47
Parliamentary Group							
AfD	48	48	0	0	0	100.00	0
CDU	40	40	0	0	0	100.00	0
CDU/CSU	1482	1461	7	7	7	98.58	7
CSU	6	6	0	0	0	100.00	0
DIE LINKE	74	74	0	0	0	100.00	0
DP	21	20	1	0	0	95.24	0
DP/FVP	1	1	0	0	0	100.00	0
FDP	613	611	1	1	0	99.67	0
FU	22	20	0	2	0	90.91	0
GB/BHE	4	4	0	0	0	100.00	0
GRUENE	371	367	0	1	3	98.92	3
KPD	29	29	0	0	0	100.00	0
NA	6470	6192	181	48	49	95.70	52
PDS	63	60	0	0	3	95.24	3
PDS/Linke Liste	19	19	0	0	0	100.00	0
SPD	1429	1416	1	10	2	99.09	3
fraktionslos	33	30	3	0	0	90.91	0

* fully matched speakers in per cent

The leftmost column indicates the dimensions as they are expected in the gold annotation.
Role "parl_commissioner" refers to the role of parliamentary commissioner in GermaParl.
Parliamentary Group "NA" describes governmental speakers, presidential speakers and other non-MPs.

Table 4: Comparison of Ground Truth and Processed Data