Proceedings of ParlaCLARIN III @ LREC2022 pages 7–15
Marseille, 20 June 2022

How GermaParl Evolves: Improving Data Quality by Reproducible Corpus Preparation and User Involvement

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

Abstract

The development and curation of large-scale corpora of plenary debates requires not only care and attention to detail when the data is created but also effective means of sustainable quality control. This paper makes two contributions: Firstly, it presents an updated version of the GermaParl corpus of parliamentary debates in the German *Bundestag*. Secondly, it shows how the corpus preparation pipeline is designed to serve the quality of the resource by facilitating effective community involvement. Centered around a workflow which combines reproducibility, transparency and version control, the pipeline allows for continuous improvements to the corpus.

Keywords: corpus creation, reproducibility, FAIR, community involvement, parliamentary debates, German Bundestag

1. Introduction[1]

Parliaments are at the heart of democracy and institutions with rich traditions. Nonetheless, the datafication of parliamentary resources is a relatively recent trend for research on parliaments and representative democracy. Plenary protocols prepared as corpora serve many research objectives – such as assessing party positions between elections, the (substantial) representativeness of parliamentarians, and much more (Fernandes et al., 2021).

One reason for the increasing use of parliamentary debates as research data – apart from their substantial meaning for democracy – is that tools and techniques to process large amounts of plenary data have become widely accessible and affordable. If data quality does not matter too much, the technically savvy will soon attain large-scale data, yet with a hacky prototyping approach. But the challenges to develop corpora of plenary data as a sustainable, multifunctional research resource are not to be underestimated.

Making data “FAIR” (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al., 2016) has emerged as a new gold standard. In the case of parliamentary data, being FAIR means that corpora need to be findable in common repositories, data and workflows should be publicly available for download - in our case as open access resources -, should function across different technical environments, and should be applicable to different research objectives.

For good research data, just being “FAIR” is not enough. Data quality is as important as it has always been. We posit that in the case of large-scale corpora, data quality is hard to reach without explicit community involvement and a reproducible data preparation workflow. Moreover, without usability, there will not be users for the data. To be widely usable (and used), community efforts are needed to spot errors and flaws in a large quantity of data and to improve data quality. Furthermore, community efforts can improve the usability of tools. Even the best data and tools will not be used widely if their use is not intuitive. Scholars focused on a particular research question favor data and tools which are intuitive, easy to use, reliable, and with good performance.

Against this background, we suggest a process for evolving data quality with a reproducible data preparation workflow and community feedback as central building blocks. Issue tracking, transparent workflows, reproducibility, and versioning play important roles in this process. The use case for presenting this model is the GermaParl corpus of parliamentary debates in the German *Bundestag* that covers Germany’s entire postwar history (1949 to 2021) in terms of parliamentary proceedings. We introduce this resource in the context of other corpora of debates in the *Bundestag*, and then discuss how reproducible data preparation and user involvement contribute to an evolving data quality.

In this contribution, a broad understanding of reproducibility is applied. We assume that data is created in a reproducible manner when a specific workflow reliably results in the same output given the same raw data. This definition is less granular than those of oth-

[1] The development of this version of the GermaParl corpus was supported by the team of the SOLDISK project at the University of Hildesheim (https://www.uni-hildesheim.de/soldisk/en/soldisk/). GermaParl has benefitted significantly from SOLDISK’s comprehensive manual quality control of the data. Our special thanks goes to Hannes Schammann, Max Kisselew, Franziska Ziegler, Carina Böker, Jennifer Elsner and Carolin McCrea. Funding from KonsortSWD has advanced the data preparation tool set (namely the bignlp R package) to facilitate annotation layers relevant for data linkage. Funding from the Text+ consortium warrants updates of the corpus, quality control and keeping data formats up with current and future developments.The authors also want to thank Isabelle Borucki and the three anonymous reviewers for their valuable feedback on the first draft of this paper.

ers which differentiate, for example, between replicability, reproducibility and repeatability (Pawlik et al., 2019, p. 107). In practical terms, this means that the GermaParl corpus can be rebuilt from scratch in an automated and transparent fashion.

2. Existing Resources

The making and evolution of new the version of the GermaParl corpus of parliamentary debates shall illustrate how a reproducible data preparation pipeline is the essential counterpart to community involvement to improve data quality and usability. In line with the relevance of this conceptual and technical approach and the increasing popularity of plenary debates in research, multiple corpora of the German *Bundestag* exist nowadays. Before turning to GermaParl, the benefits and limitations of the siblings of GermaParl shall be considered.[2]

Three of those are summarized in table 1: The DeuParl corpus (Kirschner et al., 2021) covers the most extensive period of all corpora and covers the period from 1867 to 2020. However, the corpus lacks structural annotations that might be needed for more fine-grained analyses. In contrast to that, the ParlSpeech corpus (Rauh and Schwalbach, 2020) encompasses extensive metadata – for example the name of agenda items and speakers as well as their party affiliations, including Party Facts IDs (Döring and Regel, 2019) – and covers nine parliamentary chambers including the German *Bundestag* (from 1991 until 2018). While the ParlSpeech corpus is a significant resource for comparative parliamentary research, the authors' intention to continuously update the corpus and improve its quality is not clear and thus future availability and sustained improvements are not guaranteed. Another non-profit resource for parliamentary data is the Open Discourse Project (Richter et al., 2020). It includes plenary protocols and metadata from the so-called *Stammdaten* of the German *Bundestag* (Deutscher Bundestag, 2021) and Wikipedia. The authors offer a graphical user interface, the code how to build the corpus, the data itself, and a GitHub page to encourage user pull requests. Furthermore, the authors aim to continuously update the corpus. However, the Open Discourse Project has one important drawback for scientific use: While it is thoroughly documented from a technical perspective and comprehensively presented on their website, the data paper is not available as of the time of writing. As a result, substantial documentation about design decisions is still missing.

Apart from scientific projects, commercial newspapers leaping into data journalism provide corpora. The German weekly newspaper *Die Zeit* covers 70 years of German parliamentary activity in its corpus (Biermann et al., 2019). However, this is not an open research resource and only accessible through a graphical user interface with limited functionality. It is a great information tool for interested newspaper readers but not for researchers with specific questions in mind. Apart from *Die Zeit*, the *Süddeutsche Zeitung* offers different corpora covering German parliamentary debates. Under #sprachemachtpolitik, the newspaper offers different analyses about discursive changes and topics in the *Bundestag* (Schories, without year), covering 70 years of parliamentary activity. However, this larger corpus is not publicly accessible. In addition, the *Süddeutsche Zeitung* compiled an earlier corpus of the German *Bundestag* and published their code on GitHub (Brunner and Schories, 2018). Of all corpora mentioned, the smaller *Süddeutsche* corpus is the smallest one covering six months of plenary activity to assess changes of parliamentary habits after the advent of the right-wing populist AfD in Germany's national parliament in 2017/2018. The analysis was updated in 2020 to cover 2019 as well. Despite this transparency, the limited coverage is a limitation of this corpus for many research questions.

Besides these general-purpose corpora for the German *Bundestag*, there are specialized corpora covering German politics. A corpus prepared by Barbaresi (2018) includes political speeches of the four highest ranked political functionaries in Germany. The MigParl corpus (Blätte and Leonhardt, 2020) focuses on migration and integration related speeches in the German *Länder*. Corpora like these may be a suitable option for research projects closely related to the authors' initial projects. Nonetheless, many projects will require a general-purpose, multi-functional resource.

A final flavor of *Bundestag* debates to be addressed are XML documents of parliamentary protocols directly issued by the German *Bundestag*. Of course, it would be a great relief for researchers if standardized XML was prepared right at the origin. The XML offered by the German *Bundestag* is a disappointment in this respect. Documents for older legislative periods are just plain text wrapped into a very slim header. New documents need considerable transformation and consolidation to serve as a research resource.

Preparing corpora or parliamentary debates for scientific purposes costs time and demands technical knowledge. Distinguishing it from the resources that have been introduced, the GermaParl is comprehensively ambitious to serve as a sustainable research resource. Firstly, by providing extensive coverage and metadata, we aim to provide a resource that is suitable for many different research projects on parliamentary debates in Germany. Secondly, we aim to actively engage the scientific community to enhance data quality. Thirdly, we offer our data as an open access resource for research.

As a follow-up to the previous release of GermaParl covering twenty years of parliamentary debates in Germany (1996-2016), the first comprehensive version of GermaParl is published in 2022. The release of the cor-

[2] While the following overview is our own, the OPTED project, for example, currently works on a systematic inventory of available parliamentary corpora (Sebők et al., 2021).

	DeuParl (Kirschner et al., 2021)	ParlSpeech (Rauh and Schwalbach, 2020)	Open Discourse (Richter et al., 2020)
Size	5,446 protocols from the Reichstag; 4,260 from the Bundestag	more than 6.3 million speeches	more than 4,000 protocols; 907,644 speeches
Scope	German Reichstag and German Bundestag	9 parliamentary chambers: Austrian Nationalrat, the Czech Poslanecká sněmovna Parlamentu, the Danish Folketing, the Dutch Tweede Kammer, the German Bundestag, the New Zealand House of Representatives, the Spanish Congreso de los Disputados, the Swedish Riksdag, the UK House of Commons	German Bundestag
Time periods	1867 - 2020	Differs per parliament: 1987-2019	1949 – 2021 (19 legislatures)
Meta data	year/date	Date, speech number, speaker, party, Party Facts ID, speaker's position as chair, speech length, name of the agenda item	among others: id; session; electoral term; first name; last name; politician id; speech_content; faction id; document url; position short; position_long; date; search_speech_content; multiple variables on politicians, electoral terms and factions
Raw text available	yes	yes	yes
Publicly available	Via university's repository	Via Harvard Dataverse	Via Harvard Dataverse
Context of origin	Data is part of a research paper	Along authors' research goals	Non-profit project

Table 1: Other Resources concerned with German Parliamentary Data

pus follows a two-stage scheme that is laid out in detail in section 5 of this paper.

3. The GermaParl Corpus 1949 - 2021

Covering the years from 1996 to 2016, the initial release of GermaParl corpus is an established resource for the analysis of parliamentary debates in the German *Bundestag*. It is available in two editions. The first is an interoperable XML format inspired by the standards of the Text Encoding Initiative (TEI).[3] In addition, the data has been imported into the IMS Open Corpus Workbench (CWB) (Evert and Hardie, 2011) which facilitates the management of large corpora and provides a powerful query language (the Corpus Query Processor / CQP) to make use of additional linguistic annotation layers which come with this version of the corpus.[4] The corpus has been introduced by Blätte and Blessing (2018) and has been used, inter alia, to investigate discourses on economic inequality and taxation (Smith Ochoa, 2020; Hilmar and Sachweh, 2022) and the politics of parliamentary speech-making (Müller et al., 2021).

Users of the R programming language will just need the following snippet to install GermaParl locally.

```
install.packages("polmineR")
install.packages("cwbtools")
doi <- "10.5281/zenodo.3742113"
cwbtools::corpus_install(doi = doi)
```

After the installation, users are ready to load polmineR as a toolset for corpus analysis and run some initial queries.

```
library(polmineR)
kwic(
  "GERMAPARL",
  query = "Integration"
)
```

The Comprehensive R Archive Network (CRAN) takes extraordinary care that all published R packages are interoperable. So this code is proven to work on Windows, macOS and several flavors of Linux.

In the following section, we present the corpus which extends the coverage of the previous one, describe the workflow used to create it and discuss how this addresses the need for reproducible workflows to facilitate community involvement. This workflow might provide some inspiration for the creation of other corpora with comparable goals.

3.1. Data Report

The 2022 release of GermaParl covers all debates of the first 19 legislative periods of the German *Bundestag*. In its current state, the corpus comprises about 271 million tokens in total. Figure 1 shows the size of the corpus per legislative period.[5] In both the TEI and the CWB version, the corpus is enriched with a number of metadata which makes it possible to create meaningful subcorpora. In the terminology of the CWB, these are called structural attributes. These are presented in the subsequent section. The CWB version also contains additional linguistic annotation layers which are mostly added at the level of individual tokens. These are called positional attributes in the CWB terminology and are presented thereafter.

3.2. Structural Attributes

To create useful subcorpora, a number of structural attributes is available. An overview is provided in table 2. Firstly, there is document level metadata such as the legislative period and session number, the date and, derived from that, the year. Figure 1 already showed the corpus size by legislative period. Figure 2 adds granularity by providing the same information by year and

[3] See https://github.com/PolMine/GermaParlTEI.

[4] The CWB corpus can be downloaded from Zenodo: https://zenodo.org/record/3742113.

[5] All reported numbers and visualizations in this contribution are based on the CWB version of the corpus.

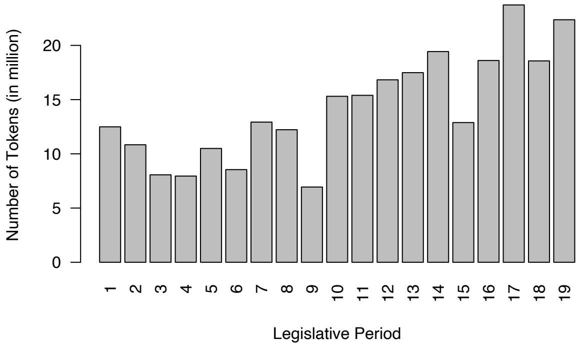

Figure 1: Number of Tokens by Legislative Period

reveals that election years usually contain less tokens than regular years. In addition, it also shows that the first and the last legislative period covered by the corpus include a smaller number of tokens because legislative periods do not align with calendar years. Finally, the long-term trend towards more words in parliament indicates a general increase in the number of delivered speeches per year.

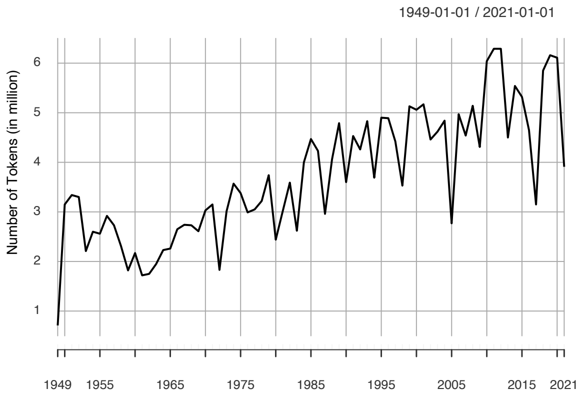

Figure 2: Number of Tokens by Year

Secondly, there is information on speaker level. This means that subsets of the corpus can be created for individual speakers, parliamentary groups and other attributes. Some of these attributes are part of the initial plenary protocols while others were added using additional external resources. Prominent examples for these attributes are full speaker names as well as party affiliations. The assignment of these attributes is discussed in detail later, but to provide a glimpse into the data, there are 4303 unique speaker names and 22 unique parliamentary groups included in the current version of the corpus. Thirdly, some linguistic features (named entities with types for persons, organizations, locations and miscellaneous named entities as well as paragraphs and sentences) are encoded as structural attributes. Finally, there are some technical structural attributes like the URL or the type of the source material.

3.3. Positional Attributes

The corpus contains linguistic features beyond the word form of a token. Two different Part-of-Speech tag sets (the Universal Dependencies tag set and the Stuttgart-Tübingen tag set) and lemmata have been added to the corpus at the token level. Table 3 represents a token stream extracted from the corpus to illustrate the available positional attributes.

4. A Reproducible Corpus Preparation Pipeline

While the corpus has been prepared with great care, remaining flaws in the data cannot be ruled out. Unknown variations or simply typos in the original data can cause speakers to be missed, for example. Despite systematic checks, some of these flaws will be encountered only after release by researchers who actually work with the corpus. The data is simply too big to be aware of all potential shortcomings.

We see it as a precondition for a culture of suggesting improvements, reporting bugs and an approximation to fundamental Open Science principles, that the corpus is prepared in a transparent and reproducible fashion.[6] This is the technical basis for a feedback loop for quality control (see also Blätte and Blessing (2018, p. 813)). Reproducibility facilitates that community involvement and feedback can improve resources and tools.

The following workflow is based on the preparation pipeline initially presented by Blätte and Blessing (2018). The initial steps are thus similar to the workflow presented there. Due to changes in data coverage and availability, some stages differ. In general, the corpus preparation workflow still comprises the three steps described by Blätte and Blessing (2018, p. 812):

- Preprocessing
- XMLification
- Consolidating

4.1. Preprocessing

The corpus preparation starts with the download of the raw data from the website of the German *Bundestag*.[7] The first 13 legislative periods as well as the 18th legislative period are downloaded as XML files. The existing GermaParl data is incorporated into the new version of the corpus. The existing corpus can be retrieved from GitHub and covers the years between 1996 and 2016 (about the second half of the 13th legislative period until about the first half of the 18th legislative period).[8] For reasons explained below, protocols of the 18th legislative period are extracted from PDF files.[9]

[6] See https://openscience.org/what-exactly-is-open-science/.

[7] See https://www.bundestag.de/services/opendata.

[8] See https://github.com/PolMine/GermaParlTEI.

[9] The XML files for the 18th legislative period are used to retrieve the metadata of the documents while the PDF files are used to retrieve the text of the protocols.

Structural Attribute	Level	In initial protocols	Description
lp	document level	yes	Legislative period
protocol_no	document level	yes	Session number
date	document level	yes	Date of the protocol
year	document level	yes	Year derived from date
speaker	text level	partially	Full name of the speaker, including regional specification when necessary
parliamentary_group	speaker level	yes	Parliamentary group of a speaker, corrected errors when necessary
party	speaker level	no	Party affiliation of a speaker, retrieved from Wikipedia
role	speaker level	yes	Parliamentary role of a speaker, derived from speaker call
stage_type	text level	yes	Type of stage comment, if segment is not speech but some form of comment or interjection
ner_type	text level	no	Type of named entity, if a sequence is a named entity
p	text level	partially	paragraph
s	text level	yes	sentence

Table 2: Structural Attributes in the GermaParl Corpus

cpos	word	upos	xpos	lemma
0	Meine	PRON	PPOSAT	mein
1	Damen	NOUN	NN	Dame
2	und	CCONJ	KON	und
3	Herren	NOUN	NN	Herr
4	!	PUNCT	$.	!
5	Abgeordnete	NOUN	NN	Abgeordnete
6	des	DET	ART	die
7	Deutschen	PROPN	ADJA	deutsch
8	Bundestags	PROPN	NN	Bundestag
9	!	PUNCT	$.	!

Speech by Paul Löbe on 1949-09-07

Table 3: Beginning of GermaParl as a Token Stream

During preprocessing, the protocols of the first 13 legislative periods are extracted from the downloaded XML files. Aside from some document-level metadata, these files only contain a single text node in which the entire text of the protocol is found. Compared to the PDF versions of the document, this has the advantage that the initial two column layout is already resolved. However, header lines as well as the table of contents and appendices are still part of the text and have to be removed. This is also a reason why we still use the PDF files for the 18th legislative period because they are sufficiently formatted to be extracted via the trickypdf R package, removing margin columns as well as header and footer lines.[10]

4.2. XMLification

After extracting the raw text from the XML and PDF files and removing header lines, table of contents and appendices where necessary, the data for legislative periods 1 to 13 and 18 is processed in the same workflow. Using the Framework for Parsing Plenary Protocols (frappp) which provides a generic workflow to parse unstructured protocols into structured XML (implemented as an R package), the raw text is XMLified. This follows the process described in Blätte and Blessing (2018): Based on the notion that regular expressions can be used to identify metadata as well as speaker calls, interjections or agenda items, an iterative process is used to formulate a battery of specific regular expressions for different speaker types and other structural elements such as interjections. The result is an XML format which resembles the standards of the Text Encoding Initiative (TEI) for performance text.[11] It is envisioned to extend the output format to also include a format compatible with the ParlaMint project (Erjavec et al., 2022). This would further increase the interoperability of the data.

4.3. The 19th Legislative Period as a Special Case

The 19th legislative period is a special case: Compared to earlier legislative periods, the format of the XML files issued by the German *Bundestag* changes completely, from an essentially unstructured plain text format with XML headers to a comprehensively annotated, structured XML format. Thus, the preprocessing for this legislative period follows a separate pro-

[10] See `https://github.com/PolMine/trickypdf`.

[11] See `https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DR.html`.

cess. Providing an already comprehensively annotated XML, the central task is to convert this XML to the same TEI-inspired output format used for the other legislative periods. The most significant hurdle is the decision in the original XML specification to include presidential speakers of the *Bundestag* as child nodes of the current speaker. Resolving this robustly can be challenging. In addition, specific conditions apply for specific types of debates, in particular question time, in which not all utterances are actually speech nodes in their own right. In consequence, we flatten this structure by assigning speech nodes to each individual utterance.

4.4. Consolidating

When consolidating the data, apart from the volume and the resulting variations in the data, there is one difference: While the previous version of GermaParl used Wikipedia for all parliamentary actors, in this new iteration, members of parliament are consolidated using the so-called "*Stammdaten*" of the German *Bundestag* (Deutscher Bundestag, 2021) to provide canonical names. The comprehensive *Stammdaten* file contains information for each member of parliament in the history of the *Bundestag*, including biographical as well as political data. In contrast, the plenary protocols issued by the *Bundestag* itself contain the speaker name in various formats. From the first to the 11th legislative period, the protocols contain only family names (sometimes with a speaker's constituency for disambiguation) when speakers are called. At the beginning of the 12th legislative period, this changed and full names are used thereafter in the speaker call. To harmonize the names of speakers in the corpus, the names found in the initial protocols were matched against the data in the *Stammdaten* file, using an approach to match the name, parliamentary group and legislative period of a speaker found in the speaker call with the *Stammdaten*. Like in the previous version of GermaParl, it was necessary to account for alternative names in some cases as well as to deploy fuzzy matching and make manual interventions in case of typos, missing information in either the protocols or the *Stammdaten* and other errors or divergences. To keep the results of these interventions reproducible, they are done programmatically during the corpus preparation. Similarly, the party affiliation is not part of the initial protocols. We use the data provided on the Wikipedia pages for each legislative period to add this information to the corpus. This has been done for the previous version of GermaParl as well (Blätte and Blessing, 2018, p. 813). This enables us to add a party affiliation specific to the legislative period to a speaker. In contrast, the *Stammdaten* file only reports a single party assignment for each member of parliament for the entire time, not documenting switches between parties. This still does not equate to date-specific party assignments, though: It must be noted that the information extracted from these tables on Wikipedia does not account for changes during legislative periods in a structured fashion. In addition, they are not entirely homogeneous when it comes to the point in time (beginning or end of the legislative period) which is used to determine the current party affiliation of a speaker. A remaining challenge is the annotation of agenda items. While these will be of great interest for a number of analyses, their identification is challenging as they are called in a great variety of forms. Using sentence similarities to find agenda item calls which are similar to those found in the 19th legislative period, a first implementation of an agenda item annotation is included in the TEI version of the corpus. The CWB version does not contain agenda items yet.

4.5. Linguistic Annotation and Import into the Corpus Workbench

As a result, we end up with 4340 structurally annotated plenary protocols in the TEI format described earlier. For the linguistic annotation which is part of the CWB corpus we first use Stanford CoreNLP (version 4.2) (Manning et al., 2014) to segment the textual data into tokens, sentences and paragraphs, add Part-of-Speech Tags (in the Universal Dependencies tag set) and perform named entity recognition. To use Stanford CoreNLP from within R, a wrapper called bignlp was developed that exposes the Java implementation of Stanford CoreNLP in a way that allows the processing of large amounts of text in parallel.[12] This both should speed up the process and increase robustness, at least vis-a-vis problems concerned with limited memory. The intermediate result is a vertical XML format which contains segmented tokens as well as named entities and part-of-speech annotation. This vertical XML format can then be imported into the Corpus Workbench. Finally, based on the CWB corpus, we use the TreeTagger (Schmid, 1995) to add Part-of-Speech tags in the Stuttgart-Tübingen tag set as well as lemmata to the corpus. More recent developments like the RNNTagger (Schmid, 2019) may be used in the future.

4.6. Reproducibility

To ensure that feedback can be incorporated into the data preparation and maintenance workflow, the process needs to be designed in a reproducible fashion and should be centered around open source tools. To this end, the entire workflow is set up in R (R Core Team, 2021) (as explained earlier, also accessing resources implemented in other programming languages via wrappers) and can theoretically be executed in a single R script. While this might not be advisable for each phase of corpus creation - especially when a large amount of quality control including iterative and manual optimization is involved at the beginning of the process - this facilitates a reproducible workflow in later stages. For example, it is possible to adjust a regular expression to improve the matching of specific speak-

[12] See `https://github.com/PolMine/bignlp`.

ers and re-run the script to create an updated version of the corpus. Combined with dissemination methods like GitHub and Zenodo (providing digital object identifiers), the process is transparent and both workflow and output are subject to version control. With most steps being realized via documented R packages which are under version control, this workflow was developed with a long-term perspective in mind.

5. The Role of Community Involvement

As argued above, aside from a workflow that allows for reproducibility, involving an active community is a crucial precondition for high quality data on a large scale. This involvement includes both the creation of the data as well as later stages. During the development, the community takes part in a two-stage release process. Firstly, researchers are offered access to the corpus during a beta phase starting in May 2022. Researchers can get access after expressing their interest.[13] During this stage, we encourage feedback from beta users to improve data quality and workflows. Apart from established feedback mechanisms such as GitHub issues, a community workshop provides an opportunity to gain more detailed insights about the user experience when working with the corpus. These insights go beyond dealing with outright flaws and errors in the data. Participants discuss aspects like the (non)intuitive conventions and workflows as well as potential difficulties when using the corpus and tools for analysis. Secondly, after this initial stage of testing and improving the corpus, a general release is planned in October 2022. Subsequently, GermaParl is available as an open research resource with a proportionately open license. The corpus will be available from GitHub (in the specific XML format described above) and Zenodo (as a CWB corpus). More information and documentation will be provided on the GermaParl website. Workflows used when the corpus was built will be documented on GitHub to increase transparency.

This two-stage release process aims at improving both the quality of the data and its usability before the general release of the data. After the initial open release, feedback mechanisms such as issues via GitHub are available to report remaining flaws, improve the documentation of the data or to suggest additional features which should be considered and incorporated on a regular basis in subsequent releases. Closely related to community involvement is community outreach: While GermaParl is an established resource, its active community should be engaged, maintained and grown. Amongst others, we use GermaParl and related R-packages in university courses. Furthermore, we present GermaParl-based research at national and international political science conferences. Talks and forums of the National Research Data Infrastructure Germany (NFDI) are an important dissemination mechanism. These events are only the most visible among a number of different exchange formats for the PolMine Project to reach out.

6. Discussion and Outlook

Reproducibility is the core idea of the workflow behind GermaParl. Being based around a set of generic tools, especially the Framework for Parsing Plenary Protocols (frappp), it should facilitate an iterative process of data creation and quality control. Given the size of the corpus and the number of protocols, even the most thorough checks during the creation of the corpus cannot guarantee the identification of all possible flaws. The names of speakers might contain typos which prevent regular expressions to match them, for example. These are scenarios which benefit from an active community in which researchers and other interested persons use the data and report errors when they encounter them. However, reporting errors is not enough when these errors cannot be fixed. And here, a reproducible workflow is a central requirement.

We conceive GermaParl as a comprehensively annotated and thoroughly checked high-quality research resource. Going beyond other existing resources for parliamentary debates in Germany, the focus is on reproducibility and community involvement, transparency and long-term perspectives as well as multifunctionality - the usability in different research projects. Unlike other resources on the *Bundestag*, GermaParl is available in two editions: 1) a TEI-inspired XML edition which makes it interoperable, 2) a linguistically annotated Corpus Workbench (CWB) corpus. This opens up the potentials of the CWB as a powerful corpus management tool and query engine. Moreover, it makes the analysis of large amounts of textual data accessible when analyzed with the polmineR (Blätte, 2020) analysis environment shown earlier.

The development does not stop here. For instance, Wikidata-IDs for persons will be added to the corpus to facilitate the linkage of parliamentary data to other resources such as, for example, roll call vote data. This would allow even more comprehensive analyses, for example concerning the relationship between parliamentary speech and other public arenas or how specific characteristics on the individual level contribute to parliamentary discourse.

7. References

Barbaresi, A. (2018). A corpus of German political speeches from the 21st century. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 792–797, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Biermann, K., Blickle, P., Drongowski, R., Ehmann, A., Erdmann, E., Gortana, F., Lindhoff, A., Möller, C., Rauscher, C., Scheying, S., Schlieben, M., Stahnke, J., Tröger, J., and Venohr, S.

[13] See `https://zenodo.org/record/6539967` for further information.

(2019). Darüber spricht der Bundestag. Zeit Online. https://www.zeit.de/politik/deutschland/2019-09/bundestag-jubilaeum-70-jahre-parlament-reden-woerter-sprache-wandel. Accessed: 2022-05-20.

Blätte, A. and Blessing, A. (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, May. European Language Resources Association (ELRA).

Blätte, A. and Leonhardt, C. (2020). MigParl. A Corpus of Speeches on Migration and Integration in Germany's Regional Parliaments. https://doi.org/10.5281/zenodo.3872263.

Blätte, A. (2020). polmineR: Verbs and Nouns for Corpus Analysis. https://doi.org/10.5281/zenodo.4042093.

Brunner, K. and Schories, M. (2018). Das steckt in den Bundestagsprotokollen. Süddeutsche Zeitung Online. https://www.sueddeutsche.de/politik/bundestag-analyse-plenarprotokolle-1.3944784. Accessed: 2022-05-20.

Deutscher Bundestag. (2021). Stammdaten aller Abgeordneten seit 1949 im XML-Format. https://www.bundestag.de/resource/blob/472878/d5743e6ffabe14af60d0c9ddd9a3a516/MdB-Stammdaten-data.zip.

Döring, H. and Regel, S. (2019). Party Facts: A database of political parties worldwide. *Party Politics*, 25(2):97–109.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darģis, R., Ring, O., van Heusden, R., Marx, M., and Fišer, D. (2022). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*.

Evert, S. and Hardie, A. (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.

Fernandes, J. M., Debus, M., and Bäck, H. (2021). Unpacking the politics of legislative debates. *European Journal of Political Research*, 60(4):1032–1045.

Hilmar, T. and Sachweh, P. (2022). "Poison to the Economy": (Un-)Taxing the Wealthy in the German Federal Parliament from 1996 to 2016. *Social Justice Research*.

Kirschner, C., Walter, T., Eger, S., Glavas, G., Lauscher, A., and Ponzetto, S. P. (2021). DeuParl. https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2889.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (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, June. Association for Computational Linguistics.

Müller, J., Stecker, C., and Blätte, A. (2021). Germany: Strong Party Groups and Debates among Policy Specialists. In Hanna Bäck, et al., editors, *The Politics of Legislative Debates*, pages 376–398. Oxford University Press, Oxford.

Pawlik, M., Hütter, T., Kocher, D., Mann, W., and Augsten, N. (2019). A Link is not Enough – Reproducibility of Data. *Datenbank Spektrum*, 19:107–115.

R Core Team. (2021). R: A Language and Environment for Statistical Computing. https://www.R-project.org/.

Rauh, C. and Schwalbach, J. (2020). The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. Harvard Dataverse, V1. https://doi.org/10.7910/DVN/L4OAKN.

Richter, F., Koch, P., Franke, O., Kraus, J., Kuruc, F., Thiem, A., Högerl, J., Heine, S., and Schöps, K. (2020). Open Discourse. Harvard Dataverse, V3. https://doi.org/10.7910/DVN/FIKIBO.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging With an Application To German. Revised version of a paper originally presented at the EACL SIGDAT workshop in Dublin in 1995. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf.

Schmid, H. (2019). Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2019, New York, NY, USA. Association for Computing Machinery. https://www.cis.uni-muenchen.de/~schmid/papers/Datech2019.pdf.

Schories, M. (without year). So haben wir den Bundestag ausgerechnet. Süddeutsche Zeitung Online. https://www.sueddeutsche.de/projekte/artikel/politik/so-haben-wir-den-bundestag-ausgerechnet-e893391/. Accessed: 2022-05-20.

Sebők, M., Proksch, S.-O., and Rauh, C. (2021). OPTED. Review of available parliamentary corpora. Deliverable D5.1. https://opted.eu/fileadmin/user_upload/k_opted/OPTED_Deliverable_D5.1.pdf.

Smith Ochoa, C. (2020). Trivializing inequality by narrating facts: a discourse analysis of contending storylines in Germany. *Critical Policy Studies*, 14(3):319–338.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.