The GermaParl Corpus of Parliamentary Protocols

Andreas Blätte, Andre Blessing
University of Duisburg-Essen, University of Stuttgart
andreas.blaette@uni-due.de, andre.blessing@ims.uni-stuttgart.de

Abstract

This paper introduces the GermaParl Corpus. We outline available data, the data preparation process for preparing corpora of parliamentary debates and the tools we used to obtain hand-coded annotations that serve as training data for classifying debates. Beyond introducing a resource that is valuable for research, we share experiences and best practices for preparing corpora of plenary protocols.

Keywords: corpus creation, annotation, R, TEI, Digital Humanities

1. Introduction

Parliamentary debates convey the arguments, interpretations and disputes that shape political decision-making. They are recorded and transcribed by parliamentary administrations with diligence and are published as plenary protocols. These documents are available for long periods of time – for several decades and more – and they cover the full range of issues relevant to a political system. Plenary protocols are an indispensable resource for anybody interested in the politics and policies of a democratic polity. If citizens' access to plenary protocols is impeded, the norm of democratic transparency is violated. A corpus of plenary protocols is not just a language resource, it is a crucial building block of the public digital archive of democracy.

The value of plenary protocols for research goes beyond disciplines genuinely interested in the substance of parliamentary activity. These documents are a great resource for studying the variation of language across political domains, and language change in time. Given the amount of data, a corpus of plenary protocols is a good basis for testing all kinds of algorithms. Social scientists, computational and corpus linguists, as well as data scientists may use plenary protocols productively in their research.

There is a legal advantage to plenary protocols that deserves to be mentioned: There are no substantial legal barriers for using, processing and re-using of these documents. Restrictive licensing conditions that inhibit working with various kinds of media (including social media) do not arise with plenary protocols. Corpora of plenary protocols that have been prepared can be shared, which is an essential basis for attaining the ideal of reproducible research. And as plenary protocols are open data, these documents are an outstanding resource for teaching purposes.

The digital availability of plenary protocols is excellent and poor at the same time. Documents can be downloaded without technical or legal restrictions as txt, html or pdf documents. A minimally annotated XML may be available "off the shelf". But these data formats do not yet correspond to the requirements for digital-era data processing. To substantively exploit the analytical potential of the data, original documents need to be converted into a semi-structured data format (XML) with a sufficiently fine-grained markup. Most importantly, speakers and their affiliation to parliamentary groups and parties need to be annotated to attain a useful resource. The GermaParl corpus as it has been prepared in the PolMine Project [1] implements this notion and is based on an XMLification of documents in a standardized workflow.

Preparing corpora of plenary protocols is certainly an obvious idea. Thus, corpora of plenary protocols are not new in computational linguistics (Koehn, 2005)(Vinokourov et al., 2003). The development of machine translation systems has benefitted substantively from parliamentary data. In Europe, several projects have worked on preparing corpora of plenary protocols. A particularly important inspiration for GermaParl has been the DutchParl corpus (Marx and Schuth, 2010); parts of the language used (i.e. "GermaParl", and "XMLification") are inspired by the Dutch sister project. Indeed, to bring the family of projects working with and on corpora of plenary protocols into a dialogue, the European CLARIN consortium has initiated workshops to exchange approaches and experiences.[2].

One day, all parliaments all over the world might be represented in a GlobalParl corpus. But that is still a long way to go. Our contribution to the broader development is a corpus of the plenary protocols of the German Bundestag that is available in appropriately fine-grained XML. It is compatible with the standards of the Text Encoding Initiative (TEI) and it is prepared in a generic, reproducible workflow.

2. The GermaParl Corpus

The GermaParl Corpus includes all plenary protocols that were published by the German Bundestag between Februrary 1996 and December 2016. GermaParl is not the only effort to create a corpus of debates in the Bundestag. But to the best of our knowledge, it is the most comprehensive one. The comprehensiveness and size of GermaParl – it comprises more than 100 million tokens – implies that the effort that can be invested in adding annotations and information beyond what can be extracted automatically had to be limited. A thematically specialized corpus such as the one prepared by Naomi Truan on the parliamentary discourse on Europe may offer significantly more detailed metadata and annotation (Truan, 2017).

GermaParl is made available in two ways:

[1] See `www.polmine.de`.

[2] CLARIN-PLUS Workshop "Working with Parliamentary Records", March 27-29 2017, Sofia, and Workshop "ParlaCLARIN" at 11th Language Resources and Evaluation Conference (LREC2018).

- The base format of the corpus is the XMLification of the raw data (i.e. the original protocols) that follows the TEI standard[3]. Releases of the TEI version of GermaParl are available at the PolMine presence at GitHub, in a repository called GermaParlTEI.[4]

- GermaParl is also disseminated as a R data package called 'GermaParl'. The package includes a linguistically annotated, indexed and consolidated version of the corpus that has been imported into the Corpus Workbench (CWB)[5] The R data package is designed to work smoothly with the analytical tools offered by the R package 'polmineR'[6].

Both variants of GermaParl are versioned. The version number of the TEI variant of the corpus is derived from the version number of the tool for corpus preparation, an R package to keep and maintain the code. The R data package has a different version number, but the documentation in the package will report which TEI version of the XMLified protocols has been used.

Plain text documents (txt files) issued by the German Bundestag were considered the best raw format for corpus preparation. Between May 2008 and March 2010, such txt files were not available. To fill the gap, pdf documents were processed for that period.

After implementing the corpus preparation workflow for txt and pdf documents, the German Bundestag has moved to offer XML versions of plenary protocols.[7] This official Bundestag XML actually is just plain text documents wrapped into an XML tag. Documents offer minimal metadata (legislative period, session number, date). However, the tricky part of corpus preparation is not extracting the very basic metadata of a protocol, but to attain a robust, consolidated annotation of speakers and agenda items. There is a benefit of switching to the "official XML" for corpus preparation, but it is minimal as long as plain text documents are available. Thus, the recent availability of minimal XML is not a challenge to our txt-based procedure.[8]

It is important to note that the GermaParl corpus is a collection of all debates and speeches actually given in the German Bundestag. Speeches that were included in the printed protocol, but not given in a parliamentary session, were not a part of corpus preparation.

The raw corpus data is converted into a machine-readable XML data format. More specifically, we work with a variant of the TEI standards. The basis for that scheme was the existing TEI standard for performance texts.[9] Parsing a plenary protocol into the TEI standardization for performance texts is conceptually smooth, as we have speakers and functionally equivalent things happening on the stage (interjections in parliament) in both genres; the scenes in a drama are roughly equivalent to the agenda items in a plenary session. Yet there are also good reasons to use the TEI standard for transcribed speech as the template for plenary protocols (Truan, 2016). There would have to be an added value of discussing the pros and cons of the choice between potential templates, or of even developing a specialized one for plenary protocols, but we certainly admit that the current choice does not need to be the end of history.

The important fact about GermaParl is that the structural annotation of the corpus offers a broad range of possibilities to partition the corpus into subcorpora. The following information is included as basic metadata, inter alia:

- the legislative period;
- the number of the plenary protocol;
- the date of the plenary session;
- the raw data format (txt or pdf);
- the URL where the document was downloaded from.

The following table provides an overview of the number of protocols, and tokens included in the corpus by legislative period (LP).

LP	from/to	protocols	tokens
13	1996-1998	163	11.676.618
14	1998-2002	253	19.349.263
15	2002-2005	187	12.785.509
16	2005-2009	233	18.412.812
17	2009-2013	252	23.418.060
18	2013-2016	203	15.371.446

Table 1: Number of protocols and tokens by legislative period.

There is a considerable variation of the number of protocols and the number of tokens per legislative period. Looking at the breakdown of the number of tokens per year (Figure 1) conveys the reason for this. While we see an average of 4.8 million tokens per year, with peaks of more than 6 million tokens (in 2011 and 2012), the downswings of plenary activity follow the electoral cycle. In election years (1998, 2002, 2005, 2009 and 2013), there is routinely less plenary activity, and speeches: It takes time after an election to reconvene the newly elected Bundestag.

Corpus preparation departs from plain text without any markup, and the crucial task is to detect calls to agenda

[3] See http://www.tei-c.org/index.xml.

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

[5] See http://cwb.sourceforge.net/.

[6] See https://CRAN.R-project.org/package=polmineR.

[7] See http://www.bundestag.de/service/opendata.

[8] Using the official Bundestag XML as the point of departure will however be a good choice for future versions of GermaParl, when we start processing protocols pre-dating 1996 (and the availability of and plain text raw data). The merit of the minimal official XML is that the potentially cumbersome task to extract plain text from OCRed pdf documents has already been performed by the Bundestag.

[9] See http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DR.html, and http://teibyexample.org/modules/TBED05v00.htm.

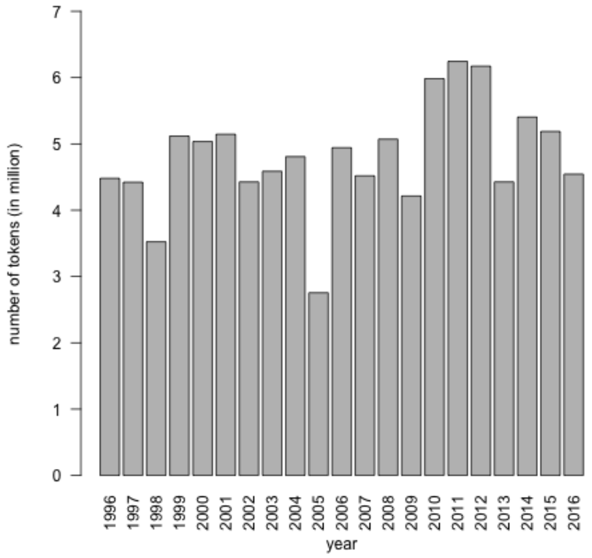


Figure 1: Number of tokens by year.

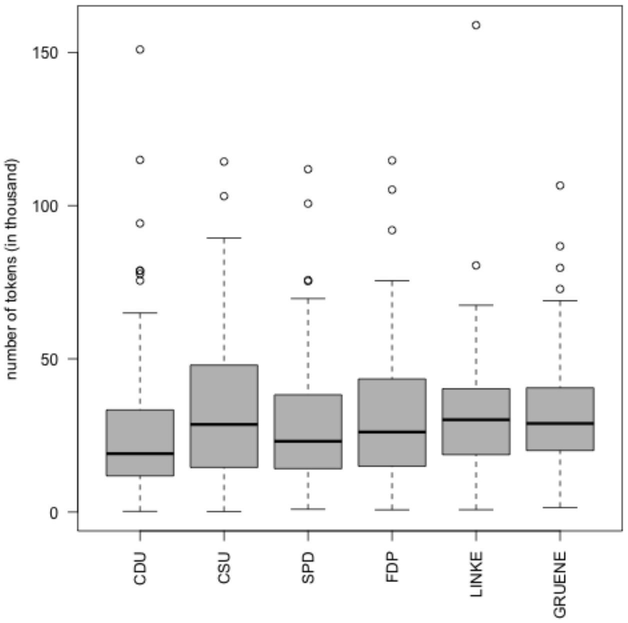


Figure 2: Boxplot – number of tokens of speakers by party (17th legislative period).

items, speakers, and to identify interjections. One important added value of the structural annotation of the corpus is the ability to distinguish between various speakers, their parliamentary groups and their parties. The following structural annotation is part of the corpus:

- the name of the speaker;
- the parliamentary group the speaker is affiliated with;
- the party affiliation of the speaker;
- whether an utterance is a speech or an interjection.

A crucial step to obtain a solid research resource is to have consolidated information at the speaker level. In the protocols, academic titles of speakers are not included in a fully consistent way, and names are sometimes reproduced with slight variations. Therefore, consolidation is necessary. Once this is achieved, it is possible to segment speeches, and to make statements about the variation between speakers of contributions to plenary sessions. The boxplot in figure 2 provides a comparison of the number of tokens contributed by individual speakers during the 17th legislative period of the German Bundestag, depending on party affiliation. The mean number of tokens spoken appears to depend on the size of the parliamentary group, i.e. the larger a parliamentary group, the smaller the average contribution of a parliamentarian. Of course the outliers raise interest. More than 150.000 words have been spoken in parliament by chancellor Angela Merkel (CDU), and the leader of the parliamentary group of *Die LINKE*, Gregor Gysi, a notoriously gifted (and fast) speaker.

Consolidating speaker information is the most time-consuming part of corpus preparation. Despite of all the efforts by parliamentary administrations, plenary protocols are not perfect. Remaining errors occur in the main body of the text, but they cause particular pain when inconsistencies occur in the names of speakers. Since it does happen, only applying regular expressions to extract speaker information is insufficient. Further consolidation is necessary. The following section outlines the workflow to obtain the consolidated corpus.

3. A Framework for Corpus Preparation

The long-term aim of the PolMine Project is to sustain corpora for all parliaments in Germany, including the 16 regional parliaments.[10] The German Bundestag is an important parliament, but it is one parliament among others. Thus, a workflow needs to be designed that is sufficiently generic to allow for a simple preparation of a corpus from any plenary protocol. Basically, the following three steps need to be carried out:

- *Preprocessing:* Prepare consolidated UTF-8 plain text documents (ensure uniformity of encodings, conversion of pdf to txt if necessary);
- *XMLification:* Turn the plain text documents into TEI format: Extraction of metadata, annotation of speakers etc.
- *Consolidating:* Check speaker names agains external data source and enriching documents with supplementary information.

[10] An early project to prepare corpora for the parliaments of the regional states of Germany was carried out in 2011/12 in cooperation with the Institut für Deutsche Sprache (IDS), see `http://www1.ids-mannheim.de/kl/projekte/korpora/archiv/pp.html`. The procedure we implemented at that time was not sufficiently object-oriented, making it difficult to customize the workflow for further parliaments, and hard to update the data.

In our case, the preprocessing step was not as trivial as it might seem. Older plain text files are offered by the German Bundestag in all kinds of encodings that are outdated. The pdf documents preserve a two-column layout that is difficult to manage. To handle these issues, respective functionality is included in the R packages 'ctk' (corpus toolkit)[11], and 'trickypdf'[12] that have evolved along with the corpus preparation tools to create GermaParl.

The core tool of the XMLification step is a set of regular expressions used to locate the beginning and the end of a debate, extract relevant metadata (such as legislative period, session number or date), to detect when speakers are called upon, and to identify the beginning and end of agenda items. The matches are used to generate the structural annotation of parliamentary speech in the XML document. However, due to the remaining haphazard variations that occur in plenary protocols – all standardization notwithstanding – we found it futile to strive for the set of universal regular expressions that would match correctly without any further document-specific interventions. Data quality is ensured in an iterative process that involves going back and forth between XMLifying all documents, and inspecting samples of the resulting data. To be able to quickly see wrong annotations, we found it useful to transform XML/TEI documents into html documents.

To handle those cases when a regular expression does not match in a desired fashion, and adjusting the regular expression might cause matches elsewhere that are not desired, we used two approaches:

- *False positives:* We run a a list of undesired matches. Before an annotation (of a speaker or an agenda item) is made, it is checked whether the match is on the mismatch list, and creating the annotation is suppressed, if it is.

- *False negatives:* There is a set of known errors (i.e. typos) in the original documents that inhibit the regular expressions to match. To ensure that the regex works, a factory of document-specific preprocessing functions is part of the code that performs (hard-coded) adjustments to make the regular expression match.

The list of mismatches and the document-specific preprocessing functions are maintained in an R package for corpus preparation that is hosted in a git repository, so that everything is under version control.

The result of the primary XMLification will still include noise. The inconsistencies that occur with names need to be handled with particular care. Accordingly, all the information that has been extracted is checked against an external data source. To be able to cope with noisy names in an automated fashion, an approximate string matching algorithm is used. To handle remaining difficulties to match names, a hand-written list of aliases is applied, and kept together with the code in the git repository, so that data quality can be improved iteratively.

There are alternative options for the external data source on parliamentary speakers. A classical source would be *Kürschners Volkshandbuch*, a book with biographical data on all parliamentarians that is published every legislative period (Holzapfel, 2018). Of course, the presentation of parliamentarians on the Website of the German Bundestag might have been considered. We opted for lists of members of parliament, cabinet members and further speakers that are available on Wikipedia.[13] This is not only because of easy digital access. The Wikipedia pages related to the German Bundestag are regularly and credibly updated by a dedicated team of volunteers. Wikipedia pages are undergoing continuous public scrutiny, ensuring permanent quality checks in a manner traditional printed material does not necessarily guarantee.

The framework for corpus preparation used for preparing GermaParl is intended to be fully generic. But of course, we need to allow for variation between parliaments, and parliaments may change layout and details of typesetting, so that alternative regular expressions may have to be used or methods to process specific details of the text are necessary. The appropriate approach to handle this is to implement things in a fully object-oriented fashion. Thus, we believe we have developed a framework for corpus preparation that might work for any parliamentary protocol. The code is included in an R package, and upon preparing a corpus, the version number of the package used is included in the TEI metadata. All interventions that may be necessary and that have been described (document-specific preprocessing functions, mismatch lists, lists with aliases) are kept as data in the R package. This way, corpus preparation is fully reproducible. This in turn is the precondition that new solutions to rectify data errors that are discovered when working with the corpus can be added, and to successively improve data quality.

4. Data Dissemination

The XML files of the corpus (according to the TEI standard) are available via a GitHub repository[14]. The data is somewhat large fo GitHub, and of course, the original purpose of git repositories has been to maintain code. But maintaining a (versioned) corpus in a git repository brings the advantage that versions can be compared, and that the effects of modifications of the corpus preparation procedure can be traced easily. What is more, GitHub repositories offer an issue tracker by default. The issue tracker of the GermaParlTEI repository is intended to organize the user feedback on data errors and on data quality.

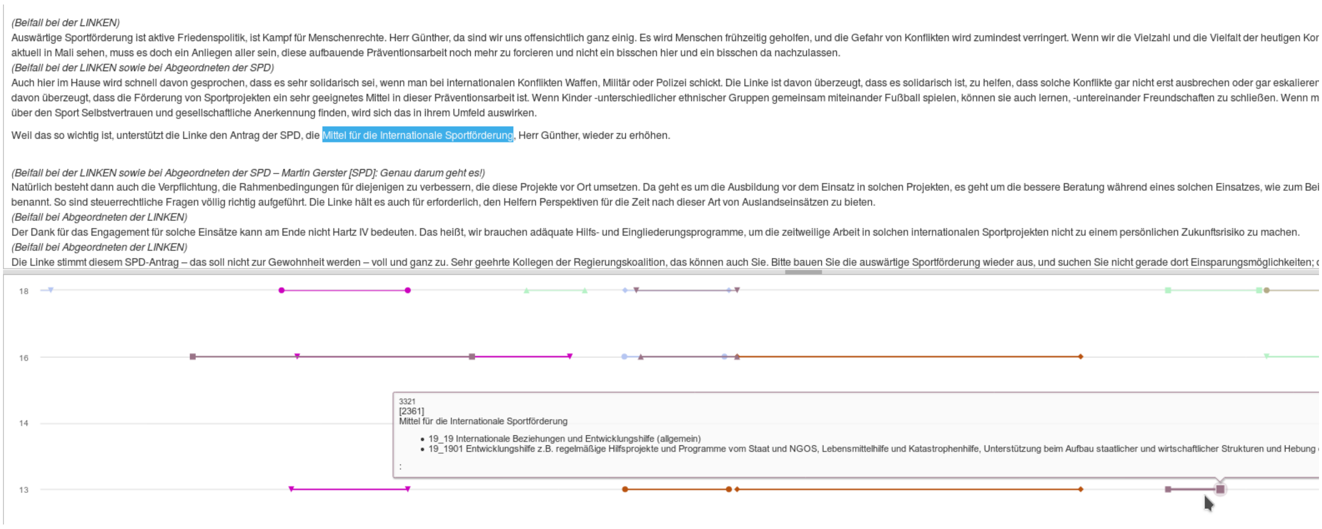
For many users in the humanities and the social sciences, XML files in general and TEI files in particular will not be overly accessible: The technical barriers to entry may be considerable for the greatest potential user group of the corpus. Therefore, the GermaParl Corpus is also disseminated as an R data package. More specifically, it is linguistically

[11] See https://www.github.com/PolMine/ctk,

[12] See https://www.github.com/PolMine/trickypdf.

[13] For instance, see the list for the 17th Bundestag at https://de.wikipedia.org/wiki/Liste_der_Mitglieder_des_Deutschen_Bundestages_(17._Wahlperiode)

[14] https://github.com/polmine/germaparltei



(Beifall bei der LINKEN)
Auswärtige Sportförderung ist aktive Friedenspolitik, ist Kampf für Menschenrechte. Herr Günther, da sind wir uns offensichtlich ganz einig. Es wird Menschen frühzeitig geholfen, und die Gefahr von Konflikten wird zumindest verringert. Wenn wir die Vielzahl und die Vielfalt der heutigen Kor
aktuell in Mali sehen, muss es doch ein Anliegen aller sein, diese aufbauende Präventionsarbeit noch mehr zu forcieren und nicht ein bisschen hier und ein bisschen da nachzulassen.
(Beifall bei der LINKEN sowie bei Abgeordneten der SPD)
Auch hier im Hause wird schnell davon gesprochen, dass es sehr solidarisch sei, wenn man bei internationalen Konflikten Waffen, Militär oder Polizei schickt. Die Linke ist davon überzeugt, dass es solidarisch ist, zu helfen, dass solche Konflikte gar nicht erst ausbrechen oder gar eskalierer
davon überzeugt, dass die Förderung von Sportprojekten ein sehr geeignetes Mittel in dieser Präventionsarbeit ist. Wenn Kinder -unterschiedlicher ethnischer Gruppen gemeinsam miteinander Fußball spielen, können sie auch lernen, -untereinander Freundschaften zu schließen. Wenn m
über den Sport Selbstvertrauen und gesellschaftliche Anerkennung finden, wird sich das in ihrem Umfeld auswirken.

Weil das so wichtig ist, unterstützt die Linke den Antrag der SPD, die Mittel für die Internationale Sportförderung, Herr Günther, wieder zu erhöhen.

(Beifall bei der LINKEN sowie bei Abgeordneten der SPD – Martin Gerster [SPD]: Genau darum geht es!)
Natürlich besteht dann auch die Verpflichtung, die Rahmenbedingungen für diejenigen zu verbessern, die diese Projekte vor Ort umsetzen. Da geht es um die Ausbildung vor dem Einsatz in solchen Projekten, es geht um die bessere Beratung während eines solchen Einsatzes, wie zum Bei
benannt. So sind steuerrechtliche Fragen völlig richtig aufgeführt. Die Linke hält es auch für erforderlich, den Helfern Perspektiven für die Zeit nach dieser Art von Auslandseinsätzen zu bieten.
(Beifall bei Abgeordneten der LINKEN)
Der Dank für das Engagement für solche Einsätze kann am Ende nicht Hartz IV bedeuten. Das heißt, wir brauchen adäquate Hilfs- und Eingliederungsprogramme, um die zeitweilige Arbeit in solchen internationalen Sportprojekten nicht zu einem persönlichen Zukunftsrisiko zu machen.
(Beifall bei Abgeordneten der LINKEN)
Die Linke stimmt diesem SPD-Antrag – das soll nicht zur Gewohnheit werden – voll und ganz zu. Sehr geehrte Kollegen der Regierungskoalition, das können auch Sie. Bitte bauen Sie die auswärtige Sportförderung wieder aus, und suchen Sie nicht gerade dort Einsparungsmöglichkeiten; c

3321
[2361]
Mittel für die Internationale Sportförderung

- 19_19 Internationale Beziehungen und Entwicklungshilfe (allgemein)
- 19_1901 Entwicklungshilfe z.B. regelmäßige Hilfsprojekte und Programme vom Staat und NGOS, Lebensmittelhilfe und Katastrophenhilfe, Unterstützung beim Aufbau staatlicher und wirtschaftlicher Strukturen und Hebung

:

Figure 3: Comparing annotations of different annotators on the same document

annotated using standard NLP tools (Stanford Core NLP)[15] and imported into the Open Corpus Workbench (CWB)[16]. The indexed and compressed corpus is then wrapped into a R data package that comes with a documentation of the data (i.e. a vignette, in the R jargon).

It can be installed easily using polmineR, an R package specifically designed to process CWB indexed corpora[17]. Once polmineR is installed, three commands are enough to get started to work with GermaParl.

```
library(polmineR)
install.corpus("GermaParl")
use("GermaParl")
```

The license chosen for both variants of GermaParl is the license PUB+BY+NC+SA[18]. Thus, the data comes with the expectation that authorship is acknowledged. The commercial use of the data is not allowed (just as academic users see restrictions to work with commercial media data). Derivatives of GermaParl may not be licensed in a more restrictive manner than the GermaParl corpus we offer.

5. Web-based Annotation

Plenary protocols cover all kinds of topics. This is a big advantage. At the same time, the thematic variety of GermaParl is a problem. Often, researchers will wish to work with a thematically defined subcorpus that fits their specific research interests. An unsupervised clustering approach (such as topic modelling) would be a relatively quick way to generate the basis for topic-specific subcorpora. To move beyond unsupervised learning, we generated training data based on a classification scheme to serve as an input for machine learning algorithms.[19]

To support the preparation of training data, we created a web-based annotation tool which offers the functionality to annotate phrases of parliamentary speeches using predefined categories. The tool contains a backend which is used to select speeches for annotation via stratified sampling, thereby ensuring a balanced sample of speeches to be annotated, i.e. that all years and parliamentary groups are represented.

Our annotation scheme and the accompanying guidelines are based on the scheme of the Comparative Agendas Project (CAP, http://www.comparativeagendas.net), a large international project to trace changing policy agendas of governments. The CAP classification scheme was originally developed for US parliamentary data. Following the guidelines of the CAP project, we extended the scheme slightly with new categories to fit the German / European context. For instance, we needed to introduce a category for debates on various aspects of European integration. The extended CAP annotation scheme contains 24 major categories and 209 subcategories.

The annotation process was realized by 5 annotators. In a preliminary step, 20 speeches were annotated by all annotators. Two reasons speak for this: (a) to train the annotators; (b) to refine the guidelines. To check intercoder reliability, we worked with a special view to simplify the comparison of the results of different annotators, ultimately helping to identify challenging cases (Figure 3).

Each of the 4 rows at the bottom depicts one annotator. The x-axis represents the word sequence of one speech. Each

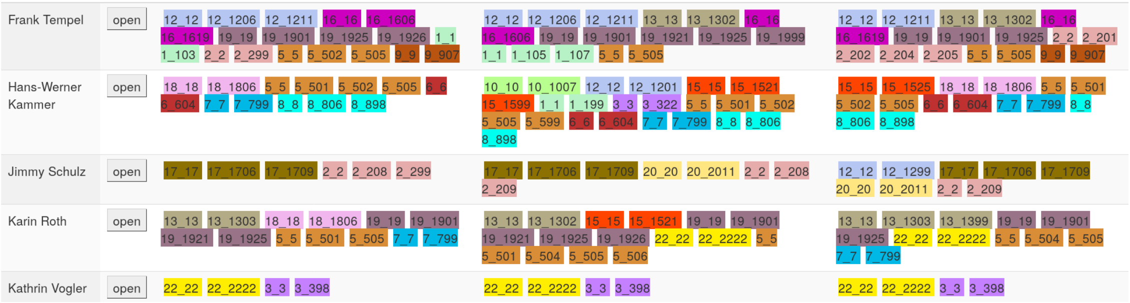
[15] See https://stanfordnlp.github.io/CoreNLP/.

[16] See http://cwb.sourceforge.net.

[17] Available at CRAN at https://CRAN.R-project.org/package=polmineR, development version at GitHub, see https://github.com/polmine/polmineR.

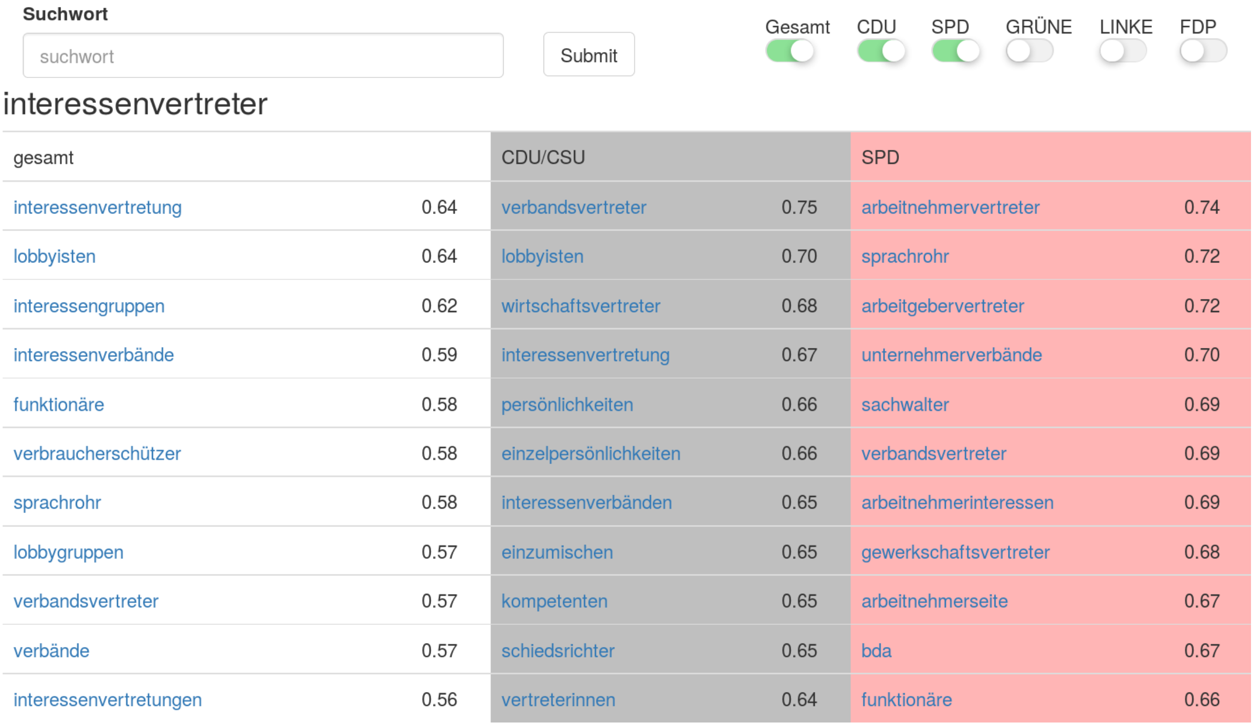
[18] See https://www.clarin.eu/content/license-categories.

[19] The context for this annotation and classification exercise was the CLARIN curation project "Plenarprotokolle als öffentliche Sprachressource der Demokratie: Klassifikation von Plenardebatten im PolMine-Plenarprotokollkorpus" (2015/16). On this occasion, we would like to thank Pawel Szczerbak, Lena Rickenberg, Vanessa Molter and Laura Dinnebier for their contribution to the project.



Frank Tempel	open	12_12 12_1206 12_1211 16_16 16_1606 16_1619 19_19 19_1901 19_1925 19_1926 1_1 1_103 2_2 2_299 5_5 5_502 5_505 9_9 9_907	12_12 12_1206 12_1211 13_13 13_1302 16_16 16_1606 19_19 19_1901 19_1921 19_1925 19_1999 1_1 1_105 1_107 5_5 5_505	12_12 12_1211 13_13 13_1302 16_16 16_1619 19_19 19_1901 19_1925 2_2 2_201 2_202 2_204 2_205 5_5 5_505 9_9 9_907
Hans-Werner Kammer	open	18_18 18_1806 5_5 5_501 5_502 5_505 6_6 6_604 7_7 7_799 8_8 8_806 8_898	10_10 10_1007 12_12 12_1201 15_15 15_1521 15_1599 1_1 1_199 3_3 3_322 5_5 5_501 5_502 5_505 5_599 6_6 6_604 7_7 7_799 8_8 8_806 8_898	15_15 15_1525 18_18 18_1806 5_5 5_501 5_502 5_505 6_6 6_604 7_7 7_799 8_8 8_806 8_898
Jimmy Schulz	open	17_17 17_1706 17_1709 2_2 2_208 2_299	17_17 17_1706 17_1709 20_20 20_2011 2_2 2_208 2_209	12_12 12_1299 17_17 17_1706 17_1709 20_20 20_2011 2_2 2_209
Karin Roth	open	13_13 13_1303 18_18 18_1806 19_19 19_1901 19_1921 19_1925 5_5 5_501 5_505 7_7 7_799	13_13 13_1302 15_15 15_1521 19_19 19_1901 19_1921 19_1925 19_1926 22_22 22_2222 5_5 5_501 5_504 5_505 5_506	13_13 13_1303 13_1399 19_19 19_1901 19_1925 22_22 22_2222 5_5 5_504 5_505 7_7 7_799
Kathrin Vogler	open	22_22 22_2222 3_3 3_398	22_22 22_2222 3_3 3_398	22_22 22_2222 3_3 3_398

Figure 4: Aggregated annotation comparison of three annotators for five speeches.



Suchwort

suchwort

Submit

Gesamt ● CDU ● SPD ● GRÜNE ○ LINKE ○ FDP ○

interessenvertreter

gesamt		CDU/CSU		SPD	
interessenvertretung	0.64	verbandsvertreter	0.75	arbeitnehmervertreter	0.74
lobbyisten	0.64	lobbyisten	0.70	sprachrohr	0.72
interessengruppen	0.62	wirtschaftsvertreter	0.68	arbeitgebervertreter	0.72
interessenverbände	0.59	interessenvertretung	0.67	unternehmerverbände	0.70
funktionäre	0.58	persönlichkeiten	0.66	sachwalter	0.69
verbraucherschützer	0.58	einzelpersönlichkeiten	0.66	verbandsvertreter	0.69
sprachrohr	0.58	interessenverbänden	0.65	arbeitnehmerinteressen	0.69
lobbygruppen	0.57	einzumischen	0.65	gewerkschaftsvertreter	0.68
verbandsvertreter	0.57	kompetenten	0.65	arbeitnehmerseite	0.67
verbände	0.57	schiedsrichter	0.65	bda	0.67
interessenvertretungen	0.56	vertreterinnen	0.64	funktionäre	0.66

Figure 5: Word embeddings trained on subsets of GermaParl (divided by parliamentary groups).

colored span in the line shows one annotation. Such colored spans can be clicked on to get additional information: The corresponding text is highlighted in the above textual view of the speech; all assigned categories are listed for the selected annotation.

In total 611 speeches were annotated. Two general issues emerged during the annotation process. First, unambiguous instructions for coders how to handle the textual boundaries of an annotation are hard to operationalize. Annotations remain a mixture of words, sentences, and paragraphs. Second, recurring topics are also hard to model. Table 2 gives an overview of the categories most frequently annotated in the corpus.

A further overview of annotated categories for each document, separated by annotators, gives additional insights about the distribution of categories in the corpus.

id	category	freq
1	Domestic Macroeconomic Issues	1404
20	Government Operations	1291
13	Social Welfare	1179
5	Labor and Employment	983
19	International Affairs and Foreign Aid	973
15	Banking Finance, and Domestic Commerce	942

Table 2: Annotation categories ordered by frequency.

6. Example Applications

The purpose of GermaParl is to serve as a sound, trustworthy basis for research in the social sciences, the humanities, computational linguistics and information science. In a paper on the parliamentary activity of parliamentarians with a migration background, GermaParl served as a data basis for measuring the salience of migration and integration issues in speeches given by parliamentarians with and without a migration background (Blätte and Wüst, 2017). The prerequisite for this research was the meticulous consolidation of speaker names in the corpus. To measure issue salience, the paper pursues a dictionary-based approach, i.e. shows how the semantic field of migration and integration can be defined in a corpus-driven manner. In another paper, Blätte uses machine learning to blindly guess the party affiliation of speakers (Blätte, 2018) based on a model trained on the annotated party membership in the corpus. In this case, classification errors are interpreted as likelihood of confusion, and political proximity of parties.

The design of the GermaParl corpus will enable researchers to implement all kinds of research questions. The corpus can be easily split by speakers, or by parliamentary groups. To explore the potential use of a recent technique in information science that requires sizeable corpora, we used such subcorpora to train different word embedding models[20]. Figure 5 shows how such embeddings can be used to investigate how terms (e.g. "Interessenvertreter" [lobbyist]) are used by different parliamentary groups. E.g.: "Verbandsvertreter" (association representatives) is the most closely related term by the Christian Democratic group, and "Arbeitnehmervertreter" (employee representatives) by the Social Democratic group. This is just one example how language use varies between parliamentary groups.

Many further uses of the corpus concern language change over time, variations or party positions etc. We are happy to offer a resource that may lower the barriers of entry to work with parliamentary protocols productively.

7. Conclusion

The current version of GermaParl is a 100-million-token corpus. The XML/TEI variant is available at a GitHub repository, a linguistically annotated and indexed version can be installed as a R data package. The data is intended to be open, versioned, reproducible, accessible and sustainable, with a focus on successively improving data quality.

However, the focus of our endeavor is not just to offer the data itself, and the annotations of parliamentary speeches. The ultimate aim of the project is to develop a generic workflow and a framework for preparing corpora of parliamentary protocols. Research on policies, politics and language change in parliamentary democracies will benefit from a public digital archive of democracy. Plenary protocols are an important part of this archive. We hope that GermaParl makes a useful contribution to a growing family of corpora of plenary protocols.

8. Bibliographical References

Blätte, A. and Wüst, A. M. (2017). Der migrationsspezifische Einfluss auf parlamentarisches Handeln. Ein Hypothesentest auf der Grundlage von Redebeiträgen der Abgeordneten des Deutschen Bundestags 1996 bis 2013. *Politische Vierteljahresschrift*, 58(2):205–233.

Blätte, A. (2018). Zum Verwechseln ähnlich? Eine Klassifikationsanalyse parlamentarischen Diskursverhaltens auf Basis des PolMine-Plenarprotokollkorpus. In Joachim Behnke, et al., editors, *Computational Social Science. Die Analyse von Big Data*. Nomos, Baden-Baden.

Holzapfel, A. (2018). *Kürschners Volkshandbuch Deutscher Bundestag. 19. Wahlperiode*. Neue Darmstädter Verl.-Anst., Rheinbreitbach.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Marx, M. and Schuth, A. (2010). Dutchparl. the parliamentary documents in dutch. In Nicoletta Calzolari, et al., editors, *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Truan, N. (2016). Corpus Annotation. Representations of the Other in the British, French and German Discourse on Europe: A Corpus-Based Contrastive Discursive Analysis. `https://repository.ortolang.fr/api/content/de-parl/3/Corpus%20Annotation.pdf`. Accessed: 2018-02-22.

Truan, N. (2017). Parliamentary Debates on Europe at the Bundestag (1998-2015). ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Vinokourov, A., Cristianini, N., and Shawe-Taylor, J. (2003). Inferring a semantic representation of text via cross-language correlation analysis. In *Advances in neural information processing systems*, pages 1497–1504.

[20] We used the word2vec implementation (Mikolov et al., 2013).