Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1900–1906
Marseille, 20-25 June 2022

German Parliamentary Corpus (GERPARCOR)

Giuseppe Abrami, Mevlüt Bagci, Leon Hammerla, Alexander Mehler
Goethe University Frankfurt
Robert-Mayer-Straße 10, 60325 Frankfurt am Main
{abrami, bagci, mehler}@em.uni-frankfurt.de, hammerla@stud.uni-frankfurt.de

Abstract

Parliamentary debates represent a large and partly unexploited treasure trove of publicly accessible texts. In the German-speaking area, there is a certain deficit of uniformly accessible and annotated corpora covering all German-speaking parliaments at the national and federal level. To address this gap, we introduce the *German Parliamentary Corpus* (GERPARCOR). GERPARCOR is a genre-specific corpus of (predominantly historical) German-language parliamentary protocols from three centuries and four countries, including state and federal level data. In addition, GERPARCOR contains conversions of scanned protocols and, in particular, of protocols in Fraktur converted via an OCR process based on TESSERACT. All protocols were preprocessed by means of the NLP pipeline of spaCy3 and automatically annotated with metadata regarding their session date. GERPARCOR is made available in the XMI format of the UIMA project. In this way, GERPARCOR can be used as a large corpus of historical texts in the field of political communication for various tasks in NLP.

Keywords: Parliament, German, Corpus, UIMA

1. Introduction

The creation of language resources that are fully annotated in an optimal way is a major issue which consumes a lot of time and effort. Nevertheless, in the current era, with increasing digitization and open access strategies, new treasures of corpora can be unearthed. This includes parliamentary documents, which are available in various types:

- **Plenary protocols**: Plenary protocols are stenographic documentations of the plenary session, including speeches, comments and other contributions such as applause. In the plenary protocols there are references to printed matters which are being debated.
- **Printed matter**: All processes which are dealt with in a parliament are referred to as printed matter. These can be draft bills, proposals, reports or questions.
  - **Minor Questions**: Members of a parliament may ask their government Minor Questions, which the government must answer and publish in a timely manner.
  - **Major Question**: In addition, Members of Parliament can use a Major Question to request information and clarification from the government on political issues and facts. At least in the German Bundestag, the government's answer can be discussed publicly in the plenary session.
- **Committee protocols**: Most parliaments discuss issues beforehand in individual committees, which then (among other things) prepare proposals for the plenum. These meetings are usually open to the public and are also minuted.

Currently, the latter documents are not yet fully available, which has several reasons: many of them are not accessible via a direct path (API), only as scanned images, or not available at all because they have not been digitized. Since not all of the above-mentioned types of documents are equally available from all German-speaking parliaments, GERPARCOR includes only the plenary protocols on a national and federal level in order to create as broad a German parliamentary corpus as possible. For the distributed processing this corpus, we used TEXTIMAGER (Hemati et al., 2016) which utilized spaCy3[1] for NLP-related preprocessing. Using spaCy3, we executed the following preprocessing pipeline to enrich GERPARCOR with linguistic annotations: tokenization, sentence recognition, PoS tagging, lemmatization, named entity recognition, morphology recognition and dependency parsing.
We make all of the annotated documents available using UIMA (Ferrucci et al., 2009) and the XMI format. In addition, for each document, we extract metadata from the documents and add it to the XMI files based on UIMA – this includes the session date, location, and title, if available. In this way, GERPARCOR enables a time-related analysis of parliamentary text data. The final corpus, GERPARCOR, is available via GitHub (`https://github.com/texttechnologylab/GerParCor`).

2. Related Work

Several German-language parliamentary corpora already exist, although some are not primarily based on plenary sessions. Barbaresi (2018) collects speeches by the German President, the President of the Bundestag, the German Chancellor and the Foreign Minister from the years 1982–2020. Another collection of tokenized parliamentary debates of the German Bundestag between 1998 and 2015 is presented by Truan (2019). The *GermaParl* corpus makes available plenary debates between February 1996 and December

[1] `https://spacy.io/`

2016 (Blätte and Blessing, 2018). For the National Council in Austria, Wissik and Pirker (2018) created a parliamentary corpus for the years 1996–2016. For Austria, there is also a corpus of plenary debates from 2013–2015 Sippl et al. (2016), which was processed using Stanford Tagger. *ParlSpeech V2* (Rauh and Schwalbach, 2020) contains the parliamentary protocols of the national chambers of Austria, Germany, Denmark and other countries for several periods between 21 and 32 years. The *DeuParl* corpus of (Kirschner et al., 2021) contains the plenary minutes of the Reichstag and the Bundestag, in total from 1867 to June 2021.

Since there is no complete corpus of protocols of the national parliaments for Austria, Switzerland, Liechtenstein, or Germany, which would also be constantly updated to include the ever new protocols, we generated GERPARCOR to fill this gap. To round off this task, GERPARCOR also contains the minutes of the German federal parliaments. In this way, a very large corpus of genre-specific (predominantly historical) German-language texts from three centuries from different countries and different political levels is created (in future work we plan to include the minutes of the GDR People's Chamber).

3. Corpus Building

We downloaded all the parliamentary speeches available online to collect the texts of GERPARCOR. We used the APIs of the individual parliaments for this purpose, although this could not be done in a uniform manner. In some cases, parliaments do not even have an API, but only a website that offers their minutes as downloads, separated by session, often mixing minutes and other material, as described in Section 1. Only a few parliaments, such as the Bundestag, offer complete archives for past periods for download. In addition, the available plenary minutes can often only be downloaded individually, with interfaces differing between parliaments. Although there is a joint project of the German state parliaments[2], only a few of the protocols are available there. As a consequence, we developed a *separate* download function for each state parliament. The software is available via GitHub.

Some protocols were not available online, but could be made available thanks to the support of the Stenographic Services of the Saarland Parliament, the Bremen Parliament, as well as the Parliament of Rheinland-Pfalz. However, the plenary minutes of the Niedersachen State Parliament of the 1st to 9th legislative periods were not available in digital form and could not be digitized. An overview of the automatically recorded protocols of the respective parliaments can be found in Table 1. The distribution of the corresponding parliamentary sessions is shown in Figure 1 to Figure 3.

[2] https://www.parlamentsspiegel.de

Depending on the dissemination method, the individual protocols were downloaded individually or as a package and preprocessed using spaCy3 (Honnibal et al., 2020) via TEXTIMAGER (Hemati et al., 2016). We used TEXTIMAGER because the amount of data required distributed processing, as enabled by TEXTIMAGER. We additionally extracted metadata from the protocols and annotated this data as instances of the class DOCUMENTANNOTATION. Besides a possible subtitle that contains the legislative period, this DOCUMENTANNOTATION also contains the date of the protocol. A sample XMI annotation is shown in Figure 6.

Parliament	Period
Germany	
Reichstag (North German Union / Zollparlamente)	1867-02-25–1895-05-24
Reichstag (German Empire)	1895-03-12–1918-10-26
Weimar Republic	1919-02-06–1932-09-12
Third Reich	1933-21-03–1942-04-24
Bundestag	1949-07-09–2021-07-09
Bundesrat	1949-07-09–2021-08-10
German Regional Parliaments	
Berlin	1989-04-02–2021-09-16
Bremen	1995-04-07–2021-09-16
Hamburg	1997-10-08–2021-03-11
Baden-Württemberg	1984-06-05–2021-09-29
Bayern	1946-12-16–2021-10-14
Brandenburg	1990-10-16–2021-08-27
Hessen	1947-02-04–2021-09-29
Mecklenburg-Vorpommern	1990-10-26–2021-06-11
Niedersachsen	1982-06-22–2021-09-15
Nordrhein-Westfalen	1947-05-21–2021-10-08
Rheinland-Pfalz	1947-07-24–2021-09-22
Saarland	1959-07-23–2021-09-15
Sachsen	1990-10-27–2021-11-18
Sachsen-Anhalt	1990-10-28–2021-09-17
Schleswig-Holstein	1946-02-26–2021-02-11
Thüringen	1990-10-25–2021-11-19
Liechtenstein	
Landtag Fürstentums Liechtenstein	1997-03-13–2021-11-06
Austria	
Nationalrat (AT)	1918-10-21–2021-05-17
Switzerland	
Nationalrat (CH)	1999-06-12–2021-12-09

Table 1: Parliamentary protocols of regional and national parliaments included in GERPARCOR.

4. OCR

Some parliamentary minutes were only available as scanned copies, so they had to be pre-processed with OCR (see Tab. 2). Moreover, some of these scans are

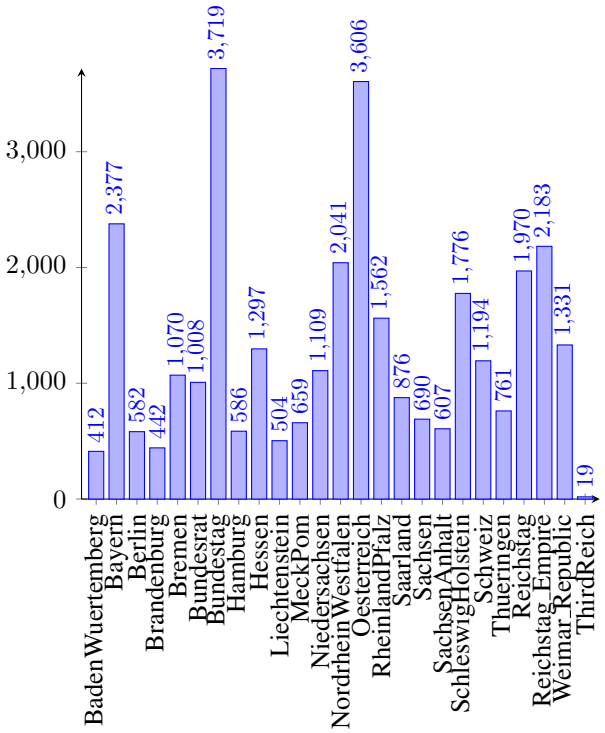

Figure 1: Number of sessions.

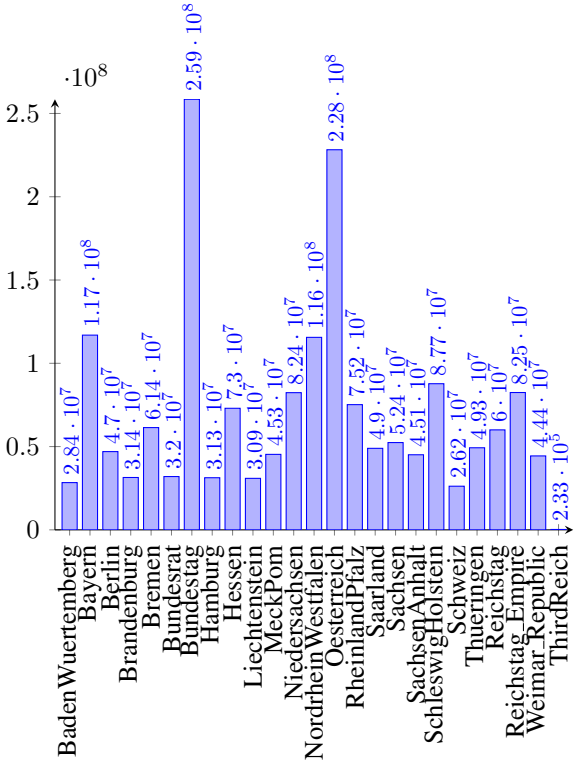

Figure 2: Number of tokens in the parliaments protocols.

only available in Fraktur. To convert these scans into text, *Optical Character Recognition* (OCR) was performed using TESSERACT (Kay, 2007) from Google.

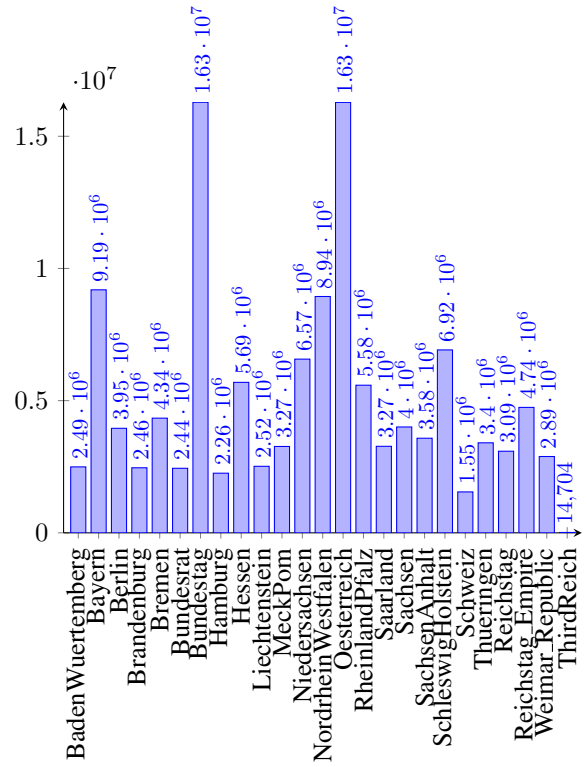

Figure 3: Number of sentences in the parliaments protocols.

TESSERACT provides various language models for text recognition, including German Fraktur.[3] To perform OCR, the individual PDF documents must be converted into images page by page as shown in the workflow in Figure 4 and described by the following procedure:

1. Divide all downloaded PDF documents into readable PDF documents (📄) and into scanned documents (📷).
2. Convert every page of every scanned document (📷) (python library: pdf2image (Belval, 2017)).
   (a) Divide the documents into good and poor quality scans; if there are only good quality scans, proceed to point 3.
   (b-e) For each image of the group of bad scans: rescale, convert the color from RGB to gray, erode, dilate and remove/reduce noise with a filter (Python library: OpenCV (Heinisuo et al., 2016)).
3. Text-extraction:
   (a) Extract the text of every readable PDF document with a PDF extractor (python library: textract[4]).
   (b) Extract the text of every scanned document from converted images using TESSERACT.

[3] https://tesseract-ocr.github.io/tessdoc/Data-Files-in-different-versions.html

[4] https://textract.readthedocs.io/en/stable/

4. NLP-Processing of every text extraction using spaCy 3 via TEXTIMAGER.
5. Check OCR output quality using a spellchecker (*SymSpell* (Garbe, 2014)).

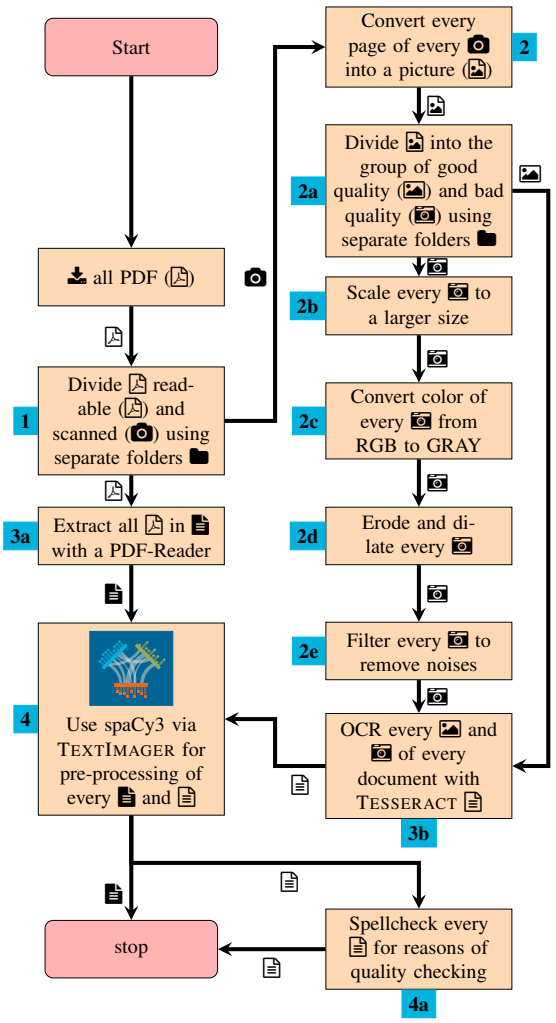

Figure 4: Workflow of GERPARCOR's OCR process including NLP preprocessing.

GeeksforGeeks have a good example to extract the text of an PDF document, which we used as a basis for our code [5]. To this end, we removed each converted image and used multithreading to speed up the extraction. By default, TESSERACT uses four cores to extract text from images[6] Thus, there are two alternatives to prevent overthreading:

1. Change the number of cores for text extraction from four to one and start the application with multithreading.
2. Divide the number of existing threads by four, round the result and start the application with x threads ($x = result$).

[5] https://www.geeksforgeeks.org/python-reading-contents-of-pdf-using-ocr-optical-character-recognition

[6] https://tesseract-ocr.github.io/tessdoc/FAQ.html

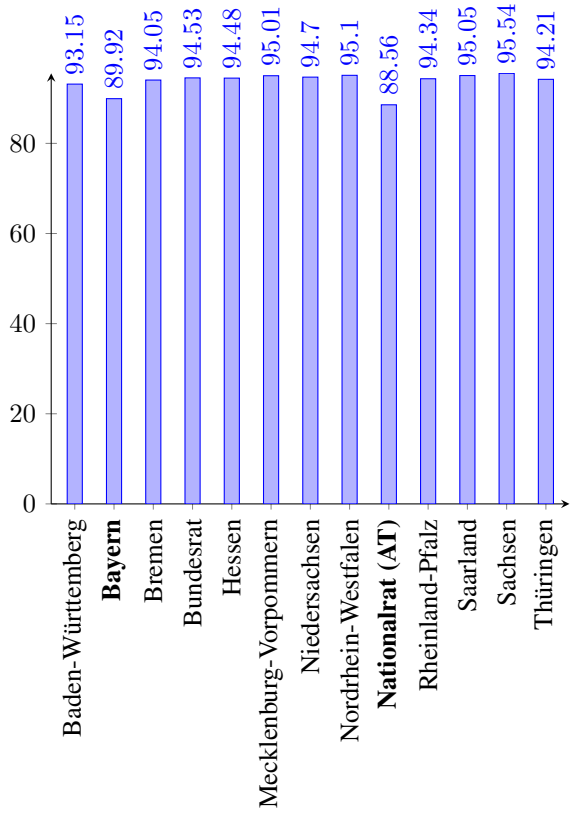

Figure 5: Testing OCR quality based on TESSERACT. Bold face refers to Fraktur. y-axis shows the percentage of correct tokens to the total number of tokens (exclude the unknown tokens).

PDFs that contain Fraktur are a challenge for OCR. For this reason, we rescaled, eroded, and dilated them and tried to reduce noise with a filter to improve extraction, as recommended by TESSERACT.[7] Figure 5 shows the results of testing the OCR output quality. A spell checker was used for this test. Bold face columns concern extractions in Fraktur. Most of the quality outputs are close to equal at around 94%. For spell checking, we used *SymSpell*. For this we used the Python library *sysmspellpy* (mammothb, 2018). We checked every token which consists of letters or is a combination of numbers and letters. Otherwise it was skipped, because *SymSpell* processes only words or word-like tokens. *SymSpell* has three possible outputs in our case:

1. The input and the output are equal to each other (which increases the number of correct words).
2. The input and the output are **unequal** to each other (which increases number of wrong words).
3. The output is empty; in this case *SymSpell* cannot correct the input (which increases the number of unknown words).

Moreover, it should be noted that *good quality* says nothing about the number of unknown words and that *unknown good quality* contains all words that are not

[7] https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html

Parliament	Period	good quality	unknown good quality	unknown words %	right words %	wrong words %
Baden Württemberg	1985-06-05–1996-02-08	93.15%	87.52%	6.05%	87.52%	6.43%
Bayern	1946-12-16–1950-11-20	89.92%	86.60%	3.70%	86.60%	9.70%
Bremen	1967-11-08–1995-09-05	94.05%	88.73%	5.66%	88.73%	5.62%
Bundesrat	1949-09-07–1996-12-21	94.53%	86.60%	8.39%	86.60%	5.02%
Hessen	1946-12-19–1998-12-16	94.48%	88.86%	5.95%	88.86%	5.19%
Mecklenburg-Vorpommern	1990-10-26–2002-06-27	95.01%	88.44%	6.92%	88.44%	4.64%
Niedersachsen	1982-06-22–1998-02-19	94.70%	88.56%	6.47%	88.56%	4.96%
Nordrhein Westfalen	1947-05-19–2005-04-21	95.10%	89.18%	6.23%	89.18%	4.59%
Nationalrat (AT)	1918-10-21–1930-07-16	88.56%	85.15%	3.84%	85.15%	11.01%
RheinlandPfalz	1947-06-04–2006-02-17	94.34%	88.30%	6.41%	88.30%	5.30%
Saarland	1994-09-11–1999-08-25	95.05%	89.44%	5.91%	89.44%	4.65%
Sachsen	1990-10-27–2004-06-25	95.54%	89.17%	6.67%	89.17%	4.16%
Thüringen	1990-10-25–1994-08-09	94.21%	87.61%	7.01%	87.61%	5.38%

Table 2: Testing OCR quality based on TESSERACT. Bold face refers to Fraktur.

skipped. However, Table 2 illustrates that the number of unknown words is significantly lower than the number of correct words. The percentages of the numbers of correct, wrong and unknown words are based on all words, which are not skipped. For this reason, *unknown good quality* is equal in percentage to the percentage of correct words. The National Council has the worst quality score (88.30% – unknown good quality) and Sachsen/Saxony the best one (95.54% – good quality). Our test shows that OCR is sufficiently good to support NLP based on GERPARCOR.

With the preprocessed version of GERPARCOR it is possible to create different subcorpora to support different research endeavors:

- one can use GERPARCOR as a whole,
- without OCR-based documents,
- only with OCR-based documents,
- or only with those documents based on Fraktur.

In particular, we expect time-related approaches (concerning studies of language change); but also analyses of political language should become possible with these data on a scale that encompasses parliamentary texts from several parliaments and, at the same time, several countries.

5. Future Work

Once the basic corpus has been created, it must be ensured that new releases of parliamentary minutes are continually added to the corpus. This requires automated retrieval of the protocols and their processing. In addition, a web-based search portal is needed to search and extract the minutes in different subsets and different formats. To enable this, the *UIMADatabaseInterface* (Abrami and Mehler, 2018) can be used, which enables storage and retrieval of UIMA documents with a number of data and document-based database systems. Moreover, for improving the quality of OCR recognition, it is planned to train a model capable of reconstructing unknown words, which should be possible given words and their contexts. Finally, GERPARCOR should be extended to include other parliamentary documents as listed in section 1.

6. Conclusion

We presented, GERPARCOR, the currently largest German-language corpus for parliamentary protocols. It includes the protocols of parliaments in Austria, Germany, Liechtenstein and Switzerland. For this purpose, the online available minutes of federal parliaments (for Germany also for state parliaments) were automatically extracted, OCRed and preprocessed with spaCy3. Since some protocols were only available as scans, some in Fraktur, they were converted with the help of TESSERACT. The complete corpus with its annotations and all the programs created for it are available via GitHub.

7. Bibliographical References

Abrami, G. and Mehler, A. (2018). A UIMA database interface for managing NLP-related text annotations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Barbaresi, A. (2018). A corpus of german political speeches from the 21st century. In *LREC*.

Belval, E. (2017). symspellpy. https://github.com/Belval/pdf2image. Accessed: 2022-01-17.

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)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Ferrucci, D., Lally, A., Verspoor, K., and Nyberg, E. (2009). Unstructured Information Management Architecture (UIMA) Version 1.0. OASIS Standard, Mar.

Garbe, W. (2014). Symspell. https://github.com/wolfgarbe/SymSpell. Accessed: 2022-01-17.

Heinisuo, O.-P., Smorkalov, A., and Serebryakov, G. (2016). Opencv on wheels. https://github.com/opencv/opencv-python. Accessed: 2022-01-17.

Hemati, W., Uslu, T., and Mehler, A. (2016). Textimager: a distributed uima-based system for nlp. In *Proceedings of the COLING 2016 System Demonstrations*. Federated Conference on Computer Science and Information Systems.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Kay, A. (2007). Tesseract: An open-source optical character recognition engine. *Linux J.*, 2007(159):2, jul.

Kirschner, C., Walter, T., Eger, S., Glavas, G., Lauscher, A., and Ponzetto, S. P. (2021). Deuparl.

mammothb. (2018). symspellpy. https://github.com/mammothb/symspellpy. Accessed: 2022-01-17.

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.

Sippl, C., Burghardt, M., Wolff, C., and Mielke, B. (2016). Korpusbasierte analyse österreichischer parlamentsreden. In Erich Schweighofer, editor, *Netzwerke : Tagungsband des 19. Int. Rechtsinformatik Symposions IRIS 2016 : 25.- 7. Feb. 2016, Univ. Salzburg*, volume 320 of *Books_372ocg.at*, pages 139–148.

Truan, N. (2019). Débats parlementaires sur l'europe au bundestag (1998-2015). ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Wissik, T. and Pirker, H. (2018). Parlat beta corpus of austrian parliamentary records. In Darja Fišer, et al., editors, *LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora In Proceedings of the Eleventh International Conference on Language Resources and Evaluation LREC2018*, Miyazaki. European Language Resources Association.

```
<annotation2:DocumentAnnotation xmi:id="23" sofa="1" dateDay="11" subtitle=   1
    "17.Wahlperiode__1.Sitzung" dateMonth="5" dateYear="2021" timestamp="
    1620691200000"/>
<type4:DocumentMetaData xmi:id="33" sofa="1" begin="0" end="48634" language   2
    ="de" documentTitle="Landtag von Baden-Württemberg-Plenarprotokoll vom
    11.05.2021" documentId="Plenarprotokoll_17_1_11.05.2021_S._1-13.xmi.gz"
    documentUri="file:/resources/corpora/parlamentary_germany/
    BadenWuertemberg/xmi/17/Plenarprotokoll_17_1_11.05.2021_S._1-13.xmi.gz"
    documentBaseUri="file:/resources/corpora/parlamentary_germany/"
    isLastSegment="false"/>

<type6:Sentence xmi:id="757" sofa="1" begin="2733" end="2841"/>               4

<type6:Lemma xmi:id="284068" sofa="1" begin="2733" end="2748" value="         6
    Alterspräsident"/>
<type6:Lemma xmi:id="284080" sofa="1" begin="2749" end="2757" value="         7
    Winfried"/>
<type6:Lemma xmi:id="284092" sofa="1" begin="2758" end="2769" value="         8
    Kretschmann"/>
<type6:Lemma xmi:id="284104" sofa="1" begin="2769" end="2770" value=":"/>     9
<type6:Lemma xmi:id="284116" sofa="1" begin="2771" end="2776" value="Meine"   10
    />
<type6:Lemma xmi:id="284128" sofa="1" begin="2777" end="2781" value="sehr"/   11
    >
<type6:Lemma xmi:id="284140" sofa="1" begin="2782" end="2791" value="         12
    verehren"/>
<type6:Lemma xmi:id="284152" sofa="1" begin="2792" end="2797" value="Dame"/   13
    >
<type6:Lemma xmi:id="284164" sofa="1" begin="2798" end="2801" value="und"/>   14
<type6:Lemma xmi:id="284176" sofa="1" begin="2802" end="2808" value="Herr"/   15
    >

<type6:Token xmi:id="19853" sofa="1" begin="2733" end="2748" lemma="284068"   17
     pos="178914" order="0"/>
<type6:Token xmi:id="19873" sofa="1" begin="2749" end="2757" lemma="284080"   18
     pos="178927" morph="389238" order="0"/>
<type6:Token xmi:id="19893" sofa="1" begin="2758" end="2769" lemma="284092"   19
     pos="178940" morph="389268" order="0"/>

<morph:MorphologicalFeatures xmi:id="389238" sofa="1" begin="2749" end="      21
    2757" gender="Masc" number="Sing" case="Nom" value="Case=Nom|Gender=Masc
    |Number=Sing"/>

<dependency:Dependency xmi:id="606959" sofa="1" begin="2733" end="2748"       23
    Governor="19893" Dependent="19853" DependencyType="PNC" flavor="basic"/>
<dependency:Dependency xmi:id="606974" sofa="1" begin="2749" end="2757"       24
    Governor="19893" Dependent="19873" DependencyType="PNC" flavor="basic"/>
```

Figure 6: Excerpt from an annotated XMI document: Line 1 and 2 shows meta information from the minutes of the Baden-Würtemberg state parliament on 2021-11-05. This contains the title (2) as well as the date and a subtitle (1). For this protocol, the sentence "Alterspräsident Winfried Kretschmann: Meine sehr verehrten Damen und Herren, liebe Kolleginnen und Kollegen!" (English: "Senior President Winfried Kretschmann: Ladies and gentlemen, dear colleagues!") is shown here in XMI. Line 4 shows the sentence annotation and lines 6 - 15 an exerpt of the lemma annotations; and lines 17 - 19 an excerpt from the token annotations. Within the serialization of the CAS document (XMI) references can be recognized, which are specified via the ID's of the respective attributes. In line 21 the morphological annotation is given for the token in line 17. Lines 23 and 24 show an excerpt of the dependency annotations for the sentence.