German Parliamentary Corpus (GerParCor) Reloaded

Giuseppe Abrami, Mevlüt Bagci, Alexander Mehler

Goethe University Frankfurt | Text Technology Lab

LREC-COLING 2024

20-25 May, 2024
Turin, Italy

Introduction

- In 2022, the largest German-language corpus of parliamentary protocols from three different centuries along for national and federal levels from Germany, Austria, Switzerland, and the Principality of Liechtenstein has been published at this time – GerParCor (Abrami et al. 2022).
- Since no further NLP-processed German-language parliament corpora have been published in the meantime, we have significantly expanded GerParCor.
- GerParCor's update includes a 26,3% increase for sentences and a 28,35% increase for tokens.

Updates

- Existing parliaments have been continued since the last update; those not already included have also been incorporated.
- Regional parliaments from Austria have been added.
- Historical parliaments have been added so that a period since 1797 is now covered.
- Provision of the corpus not only in XMI but also using other formats (CoNLL, plain text, etc).
- Conversion of the NLP pre-processing engine from TextImager (Hemati, Uslu, and Mehler 2016) to Docker Unified UIMA Interface (DUUI– Leonhardt et al. 2023); addition of sentiment annotations.
- Implementation of an API and establishment of a database for using the corpus.
- Setting up a DUUI reader for direct access to the corpus data for NLP processing.

GerParCor API

```
// Initialization of the API.
GerParCorAPI pAPI = new GerParCorAPI();
// Initialize the factory to allow access to the corpus.
Factory pFactory = pApi.getFactory();
// Possible request to query all countries deposited in the
    corpus....
pFactory.listCountries().stream().forEach(sCountry->{
    // ... with subsequent filtering.
    QueryBuilder pQuery = new QueryBuilder();
    pQuery.withCountry(sCountry).withDevision("National")
    .withStartDate(pDate);

    Set<Protocol> pResult = pQuery.build();
    pResult.stream().forEach(p->{
        // Download a protocol
        File pFile = p.download(Format.XMI);
        // ...
    })
});

// In contrast, entire batch processes can be run to download
    the requested protocols in different contexts.
// Download all protocols from Germany as well as from
    Austria on national level.
QueryBuilder pQuery = new QueryBuilder();
pQuery.withCountry("Germany").withDevision("Regional");

// After the QueryBuilder is built, the desired protocols are
    written to the location of choice in the desired format.
pFactory.download(pQuery, Format.TXT, "/opt/corpus/");

// Alternatively, if protocols have already been downloaded
    to this location, avoiding overwriting will only add new
    protocols.
pFactory.download(pQuery, Format.XMI, "/opt/corpus/", false);
```

Figure 1: Example of GerParCor Java API usage.

Statistics - existing parliaments

Parliament	Periods	Sessions	Token	Sentences
Germany - National Level				
Bundestag	1949-07-09–-2023-10-19	3 784	253 011 771	16 145 907
Bundesrat	1949-07-09–-2023-07-07	1 034	32 770 581	2 542 619
Germany - Federal level				
Baden Württemberg	1952-03-25–2023-07-19	1 411	87 432 504	6 686 017
Bayern	1946-12-16–2023-07-20	2 435	121 107 176	9 498 137
Berlin	1947-10-30–2023-06-29	615	49 261 998	4 105 261
Brandenburg	1990-10-26–2023-02-23	472	35 023 783	2 672 329
Bremen	1933-02-01–2023-03-22	1 086	63 556 596	4 444 465
Hamburg	1997-10-08–2023-05-24	618	33 194 830	2 384 464
Hessen	1947-02-04–2023-07-19	1 906	116 590 626	9 378 830
Mecklenburg Vorpommern	1990-10-26–2023-03-21	806	57 832 474	3 999 065
Niedersachsen	1982-06-22–2023-06-23	1 149	84 630 836	6 653 024
Nordrhein Westfalen	1922-07-29–2023-06-16	2 104	119 825 597	9 129 715
Rheinland Pfalz	1909-02-03–2023-05-10	1 598	77 403 705	5 794 897
Saarland	1959-06-19–2023-06-21	880	49 083 469	3 269 933
Sachsen	1990-10-27–2023-07-06	728	55 445 597	4 220 679
Sachsen-Anhalt	1990-10-28–2023-06-30	650	48 472 896	3 838 504
Schleswig Holstein	1946-02-26–2022-04-28	1 840	95 883 840	7 467 113
Thüringen	1990-10-25–2023-09-30	862	54 918 174	3 322 782
Austria - National Level				
Nationalrat	1918-10-21–2023-09-29	3 749	237 633 328	16 897 525
Bundesrat	1920-12-01–2023-07-07	1 154	54 153 318	3 390 363
Austria - Federal Level				
Kärnten	1994-04-19–2023-04-13	389	28 403 753	1 805 111
Niederöstereich	1945-12-12–2023-07-06	764	32 005 485	2 198 781
Oberöstereich	1945-12-13–2023-05-11	569	24 683 177	1 367 965
Salzburg	1994-05-02–2023-02-01	216	9 387 717	574 290
Steiermark	1848-06-13–1968-06-17	1 797	28 021 868	1 656 910
Tirol	1865-11-23–2023-09-14	2 625	59 331 644	3 409 634
Voralberg	1822-04-29–2021-06-09	1 471	42 168 655	2 345 240
Wien	1998-01-23–2023-06-21	204	414 997	31 657
Switzerland - National Level				
Nationalrat	1999-06-12–2023-09-29	1 309	28 288 768	1 668 562
Liechtenstein - National Level				
Landtag	1997-03-13–2023-04-05	556	33 853 338	2 744 172

Table 1: Session periods of the individual regional and national parliaments in GerParCor which exist at present.

Statistics - historical parliaments

Parliament	Periods	Sessions	Token	Sentences
Germany - National Level				
Reichstag (North German Union / Zollparlamente)	1867-02-25–1895-05-24	1 970	76 593 232	4 430 065
Reichstag (German Empire)	1895-03-12–-1918-10-26	2 183	60 102 498	3 096 673
Weimar Republic	1919-02-06–-1932-09-12	1 331	44 408 757	2 888 948
Third Reich	1933-21-03–-1942-04-24	9	186 955	11 998
Germany - Federal level				
Alter Landtag Württemberg	1797-04-24–1799-01-30	19 (🗃)	1 173 546	68 186
Landtag Württemberg	1820-01-18–1933-10-16	381	159 894 169	9 172 003
Landtag Württemberg-Baden	1946-12-10–1952-05-30	11 (🗃)	7 293 642	517 172
Landtag Württemberg-Hohenzollern	1947-06-03–1952-05-30	5 (🗃)	2 613 376	179 772
Ständeversammlung Württemberg	1815-03-15–1891-09-25	49	2 232 113	147 158
VGL Baden-Württemberg	1952-03-25–1953-11-11	3 (🗃)	2 516 838	189 249
VGL Württemberg	1849-12-01–1920-05-21	6 (🗃)	3 765 383	228 212
VGL Württemberg-Baden	1946-01-16–1946-06-19	1 (🗃)	304 177	16 596
VGL Württemberg-Hohenzollern	1946-11-22–1947-05-09	1 (🗃)	299 201	18 784

Table 2: Overview of historical parliamentary protocols in GerParCor, for parliaments which no longer exist. VGL is a acronym for "Verfassungsgebende Landesversammlung", which meens Constitutional State Assembly. If there is 🗃 in the column with the number of sessions, this means that the sessions are only available in collections and not individually.

https://gerparcor.texttechnologylab.org

https://github.com/texttechnologylab/GerParCor

https://github.com/texttechnologylab/GerParCorAPI

References

Abrami, Giuseppe, Mevlüt Bagci, Leon Hammerla, and Alexander Mehler (2022). "German Parliamentary Corpus (GerParCor)." In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 1900–1906.

Hemati, Wahed, Tolga Uslu, and Alexander Mehler (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. Osaka, Japan.

Leonhardt, Alexander, Giuseppe Abrami, Daniel Baumartz, and Alexander Mehler (2023). "Unlocking the Heterogeneous Landscape of Big Data NLP with DUUI." In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 385–399.