German Parliamentary Corpus (GerParCor)

Giuseppe Abrami, Mevlüt Bagci, Leon Hammerla, Alexander Mehler

Goethe University Frankfurt | Text Technology Lab

LREC 2022 Marseille

21 - 23 June 2022
Marseille, France

Motivation

- 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.
- In order to close this gap we create GerParCor: the largest genre-specific corpus of (predominantly historical) German-language parliamentary protocols from three centuries and four countries, including state and federal level data.
- GerParCor contains conversions of scanned protocols and, in particular, of protocols in Fraktur converted via an OCR process. based on Tesseract.
- All protocols are preprocessed with spaCy3 (Honnibal et al. 2020) via TextImager (Hemati, Uslu, and Mehler 2016); in addition, also the metadata was extracted.
- GerParCor is publicly available as annotated XMI documents and is updated periodically with new parliamentary protocols.

Resources

www.gerparcor.texttechnologylab.org

https://github.com/texttechnologylab/GerParCor

Statistics

Parliament	Period	Sessions	Sentences	Tokens
Germany				
Reichstag (North German Union / Zollparlamente)	1867-02-25–1895-05-24	1 970	3 086 888	60 023 446
Reichstag (German Empire)	1895-03-12–1918-10-26	2 183	4 744 901	82 456 344
Weimar Republic	1919-02-06–1932-09-12	1 331	2 887 216	44 389 348
Third Reich	1933-21-03–1942-04-24	19	14 704	233 421
Bundestag	1949-07-09–2021-07-09	3 719	16 286 016	258 521 349
Bundesrat	1949-07-09–2021-08-10	1 008	2 441 772	31 999 748
German Regional Parliaments				
Berlin	1989-04-02–2021-09-16	582	3 954 101	46 981 044
Bremen	1995-04-07–2021-09-16	1 070	4 338 171	61 396 356
Hamburg	1997-10-08–2021-03-11	586	2 256 178	31 294 553
Baden-Württemberg	1984-06-05–2021-09-29	412	2 494 970	28 365 464
Bayern	1946-12-16–2021-10-14	2 377	9 191 955	116 914 415
Brandenburg	1990-10-16–2021-08-27	442	2 460 840	31 439 980
Hessen	1947-02-04–2021-09-29	1 297	5 692 122	72 994 750
Mecklenburg-Vorpommern	1990-10-26–2021-06-11	659	3 267 241	45 320 645
Niedersachsen	1982-06-22–2021-09-15	1 109	6 570 416	82 367 685
Nordrhein-Westfalen	1947-05-21–2021-10-08	2 041	8 939 350	115 581 074
Rheinland-Pfalz	1947-07-24–2021-09-22	1 562	5 584 254	75 178 248
Saarland	1959-07-23–2021-09-15	876	3 273 321	48 950 664
Sachsen	1990-10-27–2021-11-18	690	4 004 190	52 404 321
Sachsen-Anhalt	1990-10-28–2021-09-17	607	3 578 857	45 083 355
Schleswig-Holstein	1946-02-26–2021-02-11	1 776	6 918 739	87 739 660
Thüringen	1990-10-25–2021-11-19	761	3 404 991	49 281 475
Liechtenstein				
Landtag Fürstentums Liechtenstein	1997-03-13–2021-11-06	504	2 516 530	30 927 117
Austria				
Nationalrat (AT)	1918-10-21–2021-05-17	3 606	16 282 052	228 214 975
Switzerland				
Nationalrat (CH)	1999-06-12–2021-12-09	1 194	1 548 029	26 203 941

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

Workflow

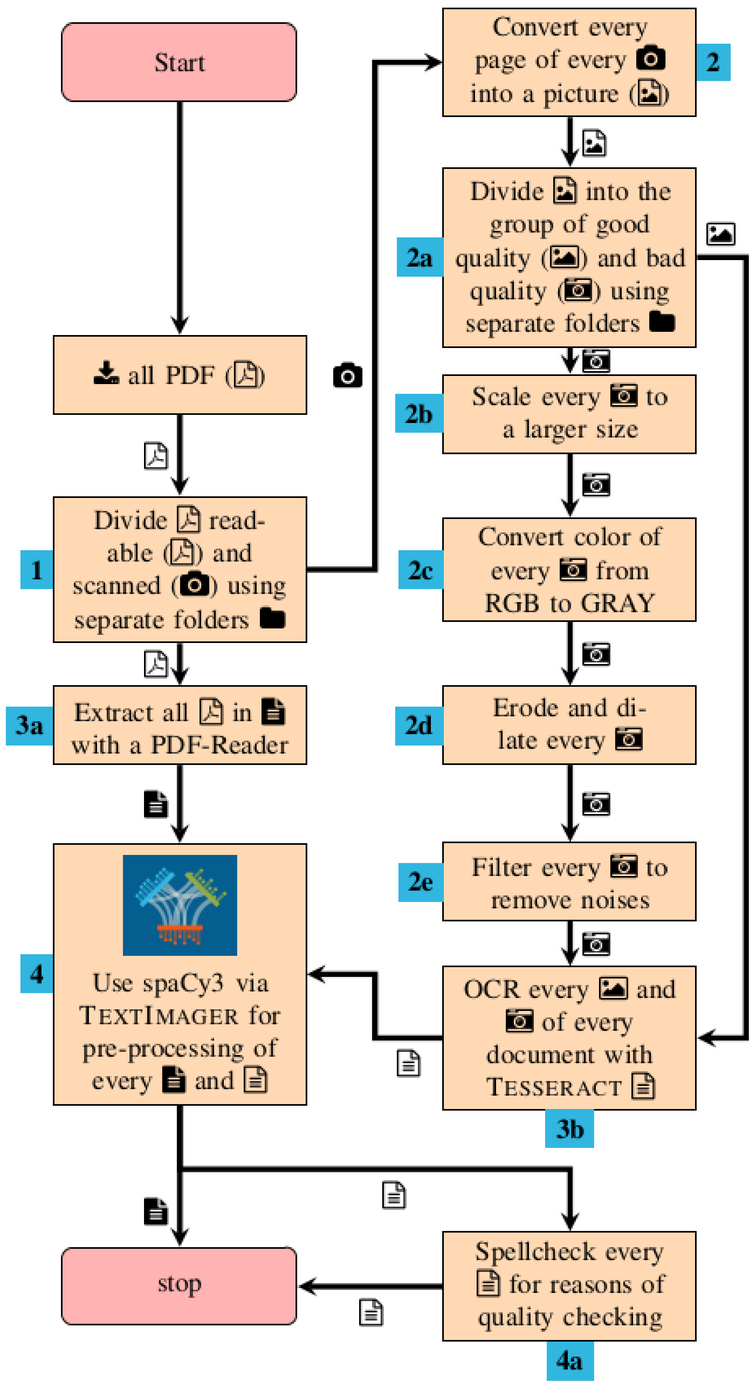

Figure 1: Workflow of GerParCor's OCR process including NLP pre-processing.

OCR

- Some parliamentary minutes were only available as scanned copies (Table 2).
- OCR (Optical Character Recognition) is a process to convert scans into text using Tesseract (Kay 2007), which also provides a language model for *German Fraktur*.
- The OCR quality is are controlled with a Spellchecker using SymSpell (mammothb 2018).
- Every token which is a combination of numbers and letters, will be checked.
  - In some cases SymSpell can not correct the words (unknown words).
  - good quality contains the right and wrong recognize words.
  - unknown good quality contains all words, which are not skipped.
- The results are suited to support NLP approaches based on GerParCor (Table 2).

References

Abrami, Giuseppe and Alexander Mehler (May 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: European Language Resources Association (ELRA).

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.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd (2020). "spaCy: Industrial-strength Natural Language Processing in Python." In.

Kay, Anthony (2007). "Tesseract: An Open-Source Optical Character Recognition Engine." In: *Linux J.* 2007.159, p. 2.

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

OCR Results

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

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

Future Work

- Automatic adding new protocols and its NLP pre-processing.
- Implementation of a web-based search portal for searching and extraction of the protocols in different subsets and formats.
- This process can be enabled with the *UIMADatabaseInterface* (Abrami and Mehler 2018)
- Improve OCR recognition with trained model, which predict the unknown words
- Extending GerParCor with other parliamentary documents (e.g. include the protocols of the GDR People's Chamber).