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 parlaments at the national and federal level.
- In order to close this gap we create GenParCor: the largest genre-specific corpus of (predominantly historical) German-language parliamentary protocols from three centuries and four countries, including state and federal level data.
- GenParCor 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 TextImage (Hernani, Ushl, and Mehler 2018); in addition, the metadata was extracted.
- GenParCor is publicly available as annotated XML documents and is updated periodically with new parliamentary protocols.

Resources

[link to website]
[link to GitHub]

Workflow

OCR

- Some parliamentary minutes were only available as scanned copies (Table 1).
- 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 controlled with a Spellchecker using SymSpell (mammoth 2013).
- Every token which is a combination of numbers and letters, will be checked.
- In some cases SymSpell can not recognize 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 GenParCor (Table 2).

References