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 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.
- 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, Uslu, and Mehler 2014); in addition, also the metadata was extracted.
- GenPaRCor 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
- abrami.mp unr

Workflow

- Start
- OCR
- Convert every page of every document into a picture
- Divide each page into groups of good quality and bad quality using separate folders
- Scale each box to a larger size
- Extract color of every box from RGB to Gray
- Edute and dilate every box
- Filter every box to remove noises
- OCR every box and every document with Tesseract
- Spellcheck every box for reasons of quality checking
- Use spaCy3 via TextImage for pre-processing of every document

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 Spellchecker using SymSpell (manufactor 2018).
- Every token which is a combination of numbers and letters, will be checked.
  - In some casesSymSpell 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 Gen-PaRCoR (Table 2).

References


Future Work

- Automatic adding new protocols and its NLP preprocessing.
- 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 GenPaRCor with other parliamentary documents (e.g. include the protocols of the GDR People’s Chamber).