How GermaParl Evolves
Improving Data Quality by Reproducible Corpus Preparation and User Involvement
Andreas Blätte, Julia Rakers and Christoph Leonhardt
ParlaCLARIN III Workshop, Marseille, 2022-06-20
Motivation

The Need for a Reproducible Workflow for Corpus Creation

- large parliamentary text corpora are widely used
- being FAIR (Wilkinson et al. 2016) is increasingly intended
- data quality is crucial
- but not all flaws can be ruled out in large data collections

Requirement for large high-quality data: A process for evolving data quality with a reproducible data preparation workflow and community feedback as central building blocks
Outline

1. The GermaParl Corpus of Parliamentary Debates
2. A Reproducible Corpus Preparation Pipeline
3. Data Quality by Reproducibility and User Involvement
The GermaParl Corpus of Parliamentary Debates

GermaParl: An Overview

- corpus of parliamentary debates in the German Bundestag
- released version: 1996 – 2016 (Blätte and Blessing 2018)
- **GermaParl v2: 1949 – 2021**
- Project Contexts
  - PolMine project (see also R package polmineR, Blätte 2020)
  - KonsortSWD and Text+
- not the only corpus of parliamentary debates in the German Bundestag
  - ParlSpeech (Rauh and Schwalbach 2020), DeuParl (Kirschner et al. 2021) or Open Discourse (Richter et al. 2020) (among others) as meaningful contributions
- but: GermaParl as a comprehensive, universally applicable and evolving resource
GermaParl v2: Comprehensive in Volume

- 19 legislative periods
- 72 years (1949 - 2021)
- 271 million tokens
GermaParl v2: Comprehensively Annotated

- Structurally and linguistically annotated

- Structural Annotation (“structural attributes”):
  - mostly metadata
  - nested structures like named entities

<table>
<thead>
<tr>
<th>Structural Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>lp</td>
<td>legislative period</td>
</tr>
<tr>
<td>protocol_no</td>
<td>session number</td>
</tr>
<tr>
<td>date</td>
<td>date</td>
</tr>
<tr>
<td>year</td>
<td>year</td>
</tr>
<tr>
<td>speaker</td>
<td>speaker name</td>
</tr>
<tr>
<td>parliamentary group</td>
<td>parliamentary group of speaker</td>
</tr>
<tr>
<td>party</td>
<td>party of speaker</td>
</tr>
<tr>
<td>role</td>
<td>role of speaker</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Source: GermaParl, see Blätte et al. 2022, Table 2
Linguistic Annotation ("positional attributes"):
- Word
- POS-Tags
  - Universal Dependencies tag set
  - language specific STTS
- Lemmata

<table>
<thead>
<tr>
<th>cpos</th>
<th>word</th>
<th>upos</th>
<th>xpos</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Meine</td>
<td>PRON</td>
<td>PPOSAT</td>
<td>Mein</td>
</tr>
<tr>
<td>1</td>
<td>Damen</td>
<td>NOUN</td>
<td>NN</td>
<td>Dame</td>
</tr>
<tr>
<td>2</td>
<td>und</td>
<td>CCONJ</td>
<td>KON</td>
<td>und</td>
</tr>
<tr>
<td>3</td>
<td>Herren</td>
<td>NOUN</td>
<td>NN</td>
<td>Herr</td>
</tr>
<tr>
<td>4</td>
<td>!</td>
<td>PUNCT</td>
<td>$.$</td>
<td>!</td>
</tr>
</tbody>
</table>

Source: GermaParl, see Blätte et al. 2022, Table 3
A Reproducible Corpus Preparation Pipeline

- Size of data as a **challenge**
- **Reproducibility** an important aspect
  - the technical basis for a **feedback loop for quality control** during creation as well as after the initial release of a corpus

  ➢ Evolving Data Quality
A Reproducible Corpus Preparation Pipeline

General Workflow

- **Goal:** An iterative workflow for corpus creation
- Successfully used in other corpus preparation processes, especially GermaParl (Blätte and Blessing 2018)
- Flexible, portable, local

- **Output formats:**
  - TEI-XML (sustainable, interoperable)
    - perspectively: ParlaMint format (Erjavec et al. 2022)
  - CWB Corpus (powerful corpus management and query tool, Evert and Hardie 2011)
A Reproducible Corpus Preparation Pipeline

General Workflow

- Raw Data
- Plain Text
- TEI/XML
- Consolidating
- CWB Corpus

Preprocessing
XMLification
Linguistic Annotation / Import into CWB

External Data
A Reproducible Corpus Preparation Pipeline

Preprocessing

- **Input / Raw Data:**
  - new parliamentary protocols retrieved from https://www.bundestag.de/services/opendata
  - (mostly unstructured) XML
  - PDF

- **Notes:**
  - existing GermaParl TEI added in consolidation step
  - Special Case 19th Legislative Period

- **Output:** Clean Plain Text

- **Steps:**
  - Removal of header and footer lines, removal of table of contents and appendices...
A Reproducible Corpus Preparation Pipeline

**XMLification**

- Process to **reconstruct debate structure from unstructured text**
- Essential idea: a battery of regular expressions for speakers, interjections, etc.
- See Blätte and Blessing 2018

- Using the **Framework for Parsing Plenary Protocols** (frappp)
  - generic tool set, integrating standardized steps
  - further information and code examples: https://polmine.github.io/frappp_slides
A Reproducible Corpus Preparation Pipeline

**Consolidating**

- Initial correction of missing speakers and other flaws in the data
- Enrichment of speaker attributes not included in the protocols via external data sources (Wikipedia and the *Stammdaten* file of the German Bundestag (Deutscher Bundestag 2021))
A Reproducible Corpus Preparation Pipeline

Consolidating

- Wikipedia for speakers' **party affiliations** and **full speaker names of speakers which are not members of parliament**

- **Stammdaten** file of the German Bundestag for **full speaker names of members of parliament**

```xml
<sp who="Adenauer" parliamentary_group="CDU/CSU" role="mp" position="NA" party="CDU" name="Konrad Adenauer">
  <speaker>Dr. Adenauer (CDU/CSU) :</speaker>
  ...
</sp>

(04/002.xml)
```
A Reproducible Corpus Preparation Pipeline

TEI-XML Output

- Available on GitHub for released data

(https://github.com/PolMine/GermaParlTEI/blob/master/13/BT_13_087.xml)

20 June 2022
Linguistic Annotation and Import into the Corpus Workbench

- **Linguistic Annotation**
  - Segmentation into Tokens and Sentences (Stanford CoreNLP, Manning et al. 2014)
  - POS-Tagging (Universal Dependencies) (Stanford CoreNLP)
  - Named Entity Recognition (Stanford CoreNLP)
  - POS-Tagging (STTS) (TreeTagger, Schmid 1995)
  - Lemmata (TreeTagger)

- Used from within R (e.g., bignlp R package)
Linguistic Annotation and Import into the Corpus Workbench

- **Import into the Corpus Workbench (CWB)**
  - CWB as a fast corpus management and querying tool ([https://cwb.sourceforge.io/](https://cwb.sourceforge.io/))
  - Import via the cwbttools R package (on CRAN)

- **Dissemination** of linguistically annotated, CWB-indexed corpora
  - GermaParl stored on Zenodo
  - Beta Version of GermaParl v2 (1949-2021) currently as restricted access on Zenodo
Data Quality by Reproducibility and User Involvement

- **User Feedback** important to detect and remedy remaining flaws and bugs in the data
- **User Involvement** via
  - GitHub Issues
  - User Workshops
- **Reproducibility**
  - Feedback most useful when it can be incorporated into the data efficiently
  - Entire process fully reproducible
  - Transparency through Versioning/DOIs via Zenodo (for data) and GitHub
Release Plan of GermaParl v2.0

- Restricted Beta Release for interested persons (request access via Zenodo) in May 2022
- Issue-Only Repository on GitHub accompanies the release for feedback during beta phase
- User Workshop collects feedback on bugs, but also usability, convenience and further feature requests
- Full, public release in October 2022

https://zenodo.org/record/6539967
Conclusion

- **Contribution**
  - GermaParl as a large, comprehensively annotated resource
  - Thoroughly checked but flaws will remain, given its size
  - Reproducible, replicable data preparation as a precondition for high-quality data, especially in large datasets and corpora
  - Workflow facilitates this for GermaParl, allowing efficient user feedback

- **Next Steps**
  - Increasing interoperability by providing XML as ParlaMint corpus (Erjavec et al. 2022)
  - GermaParl as a basis for linkage to other types of data (“Linking Textual Data” within KonsortSWD)
References


20 June 2022
References


References


Thank you for your attention!

Christoph Leonhardt | christoph.leonhardt@uni-due.de

ParlaCLARIN III Workshop, Marseille, 2022-06-20