Accessing a heterogeneous field of linguistic corpora with the help of the open access repository **LAUDATIO**

Thomas Krause*, Anke Lüdeling*, Laurent Romary+, Peter Schirmbacher*, Carolin Odebrecht*, Dennis Zielke*
Humboldt-Universität zu Berlin*, Inria France+
www.laudatio-repository.org

Digital Humanities, Lausanne – Switzerland ‘14

How can we access and (re-)use heterogeneous research data with the help of an open access repository?

1. LAUDATIO-Repository[21]

- **display structured metadata, search for and download corpora without registration**
- **upload new annotations to existing corpora and upload new corpora**
- based on the Fedora Repository Software[4]
- customized web interface
- faceted search with Elastic Search[17]
- automatic registration of PIDs for every version of a corpus[18]
- open access to corpora via CC-Licenses[16]
- connected to the search and visualization system ANNIS[8] for corpus analysis
- metadata schema with TEI ODD[3]
- conversion between different formats with SaltNPepper[7]

2. Research data

Examples for corpus formats for editing and accessing data

Examples for corpus registers
Private letters, chronicles, science, fairy tales, prose, newspapers, lyrics, songs

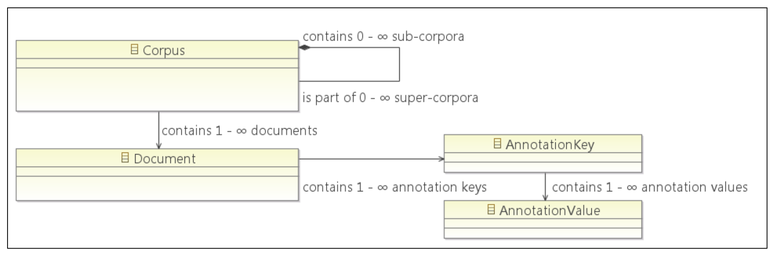
Historical German corpora already included in the LAUDATIO-Repository
- Deutsche Diachrone Baumbank[9]
- Fürstinnenkorrespondenzkorpus[10]
- German Manchester Corpus[11]
- Kasseler Junktionskorpus[12]
- Märchenkorpus[13]
- Referenzkorpus Althochdeutsch[14]
- Register in German Science[15]

➢ **plenty of different formats and content standards are used**

3. Metadata model for corpora

- structured metadata for corpora, documents and annotations metadata for textual corpora[5,6]
  - used for searching and displaying corpora
  - describes the corpus, the textual primary sources and the annotation
- specification with TEI XML + ODD[3,4]
  - ODD customization and documentation and validation with RELAX NG XML syntax schema[20]

➢ **rich and structured metadata for heterogeneous data**



4. Specific corpus architecture – exemplified with KAJUK

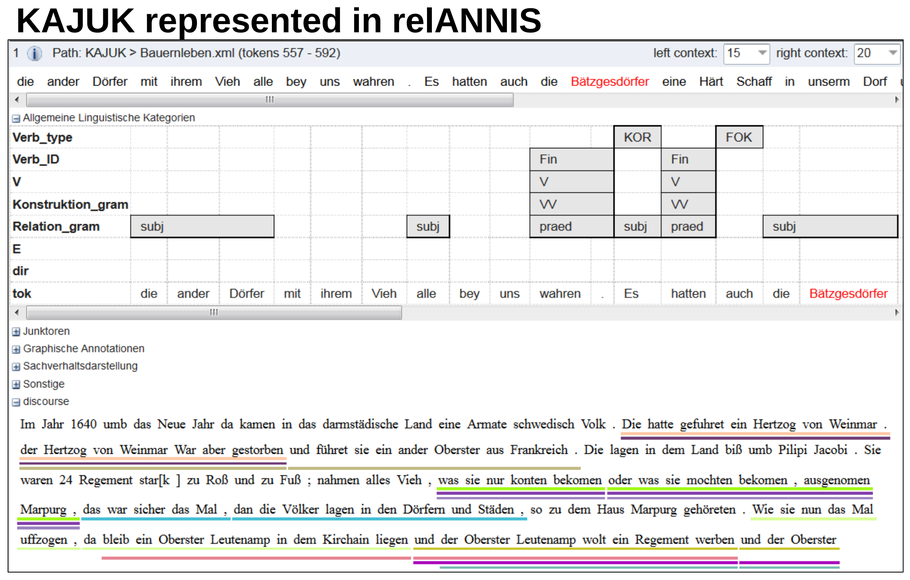
- historical corpus of German ‘Nähesprache’[1] of 17th – 19th e.g. private letters
- syntactically annotated , e.g. cohesion and clauses
- token and span annotation, pointing annotation

KAJUK Format

```
<lb n="28,3009">
<KOR type="subj">Es</KOR>
<praed><V ID="Fin"><VV>hatten</VV></V></praed>
<FOK>auch</FOK>
<subj>die Bätzgesdörfer</subj>
eine Härt Schaff in unserm Dorf</lb>
<line n="12"/>
<lb n="28,29,3009">
<J IR="kop"><KON>und</KON></J>
<subj type="E" dir="V">die Bätzgesdörfer</subj>
<praed><V ID="Fin"><VV>erhilten</VV></V></praed>
sie mit Gottes Hulf
<FOK>auch</FOK>
in dero Zeit bey uns,</lb>
<lb n="29,30,31">
<J IR="kaus" norm="denn"><KON>dan</KON></J>
<subj real="Pron">sie</subj>
<praed><V ID="Fin"><VV>gaben</VV></V></praed>
dem Hauptman darvon,</lb><!--hier line-->
<line n="13"/>
```

- annotated in an idiosyncratic SGML style
- without tokenization
- the very same category such as ‘subj’ can be either annotated as an element or a value of an attribute
- mixture of unary and binary tags
- comments are annotation
- attribute values of ‘lb’ elements refer to each other

KAJUK represented in relANNIS



- tokenization of the corpus with a mapping of categories in token and span annotations
- renaming and structured grouping of attributes, element names and their values
- attribute values of ‘lb’ elements are realized as pointing relations

➢ **documentation of both: the formats and their content in LAUDATIO**
➢ **LAUDATIO is open for every format and any content**
➢ **applicable for structured and unstructured formats**

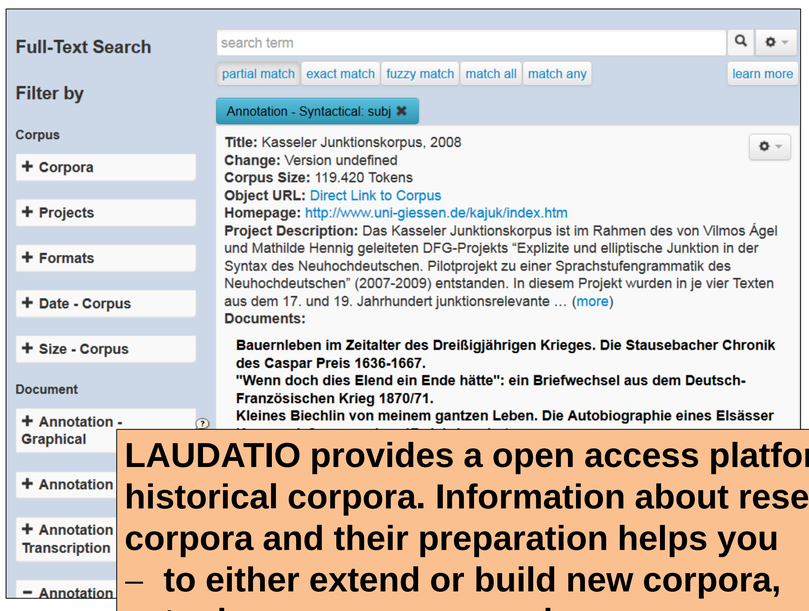
5. One metadata format for heterogeneous corpora

- import of the TEI XML metadata in the LAUDATIO-Repository
- configuration of the interface with the help of a JSON configuration file[19]
- description of the annotation guidelines (e.g. what is meant by ‘subj’)
- description of the type and preparation of the annotation in each format
  - manual or automatic annotation, which annotation tool was used
  - details (e.g. ‘subj’ is either element or value of attributes ‘EB’ or ’type’)

```
<namespace name="subj" rend="Syntactical"
xml:id="subj" corresp="Wortgruppenkategorie">
        <tagUsage gi="subj">Subjekt.</tagUsage>
</namespace>
```

➢ **structured metadata for all parts of a corpus**

```
<appInfo n="1" style="Manual">
      <application ident="XML" style="XML" version="1.0"
     type="XMLAnnotation" subtype="NA">
            <label>XML Editor</label>
            <p>XML Element 'subj'. Wert des XML Attribut
                 'type' und 'EB'.</p>
      </application>
</appInfo>
```



➢ **Open Access and (Re-)use via structured access over all types of corpora**

LAUDATIO provides a open access platform for sharing historical corpora. Information about research on the existing corpora and their preparation helps you
- **to either extend or build new corpora,**
- **to do your own research,**
- **to evaluate and re-use already existing, research and replicate results.**

6. References [1] Ágel, V., Hennig, M. (2007) DFG-Projekt “Explizite und elliptische Junktion in der Syntax des Neuhochdeutschen” Forschungsnotiz. In *Zeitschrift für Germanistische Linguistik 35*, 185-189. [2] Burnard, L., Bauman, S. (Ed.) (2008) *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Oxford.* [3] Burnard, L., Rahtz, S. (2004) *RelaxNG with Son of ODD. Extreme Markup Languages.* [4] Lagoze, C., Payette, S., Shin, E., Wilper, C. (2006). Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries, 6(2)*. 124-138. [5] Odebrecht, C. (2014) Modeling Linguistic Research Data for a Repository for Historical Corpora. *Digital Humanities 2014 Conference.* 8.7.-12.7.2014, Lausanne. [6] Odebrecht, C., Krause, Th. (2013) Metadata in an Infrastructure for Historical Corpora. *SFB 732 Incremental Specification in Context. Kolloquium.* 20.6.2013, Stuttgart. [7] Zipser, F., Romary, L. (2010) A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*. La Valette, Malta. [8] Zeldes, A., Ritz, J., Lüdeling, A., Chiarcos, C. (2009) ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009.* Liverpool. [9] DDB http://hdl.handle.net/11022/0000-0000-2107-3 [10] Fürstinnenkorrespondenzkorpus http://hdl.handle.net/11022/0000-0000-20A6-0 [11] GerManc http://hdl.handle.net/11022/0000-0000-24E3-7 [12] KAJUK http://hdl.handle.net/11022/0000-0000-2102-8 [13] Märchenkorpus http://hdl.handle.net/11022/0000-0000-1F5B-9 [14] DDD- AHD http://hdl.handle.net/11022/0000-0000-1F9C-F [15]RIDGES http://hdl.handle.net/11022/0000-0000-2106-4 [16] Creative Commons-Licenses https://creativecommons.org/licenses/ [17] ElasticSearch http://www.elasticsearch.org/ [18] Handle-PIDs, Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen http://epic.gwdg.de/wiki/index.php?title=EPIC:API:v2:contribution [19] JSON http://json.org/ [20] RELAX NG http://relaxng.org/ [21] LAUDATIO stands for **L**ong-term **A**ccess and **U**sage of **D**eeply **A**nnotated Informa**tio**n; Technical documentation LAUDATIO-Repository http://rtd.cms.hu-berlin.de/docs/laudatio-repository-documentation/