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 Oedrebrt*, Dennis Zielke*
Humboldt-Universität zu Berlin*, Inria France*

www.laudatio-repository.org

Digital Humanities, Lausanne – Switzerland ‘14

1. LAUDATIO-Repository:
- displayed 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
- customized web interface
- faceted search with ElasticSearch
- automatic registration of PID’s for every version of a corpus
- access to corpora via CC-Licenses
- connected to the search and visualization system ANNIS for corpus analysis
- metadata schema with TEI ODD
- conversion between different formats with SaltNPeeper

2. Research data
Examples for corpus formats for editing and accessing data
- Historical German corpora already included in the LAUDATIO-Repository
  - Deutsche Diachrone Baumbank
  - Fürstenkorrespondenzkorpus
  - Manchester Corpus
  - Kasseler Junktionskorpus
  - Märchenkorpus
  - Referenzkorpus Althochdeutsch
  - Register in German Science
- many different formats and content standards are used

3. Metadata model for corpora
- structured metadata for corpora, documents and annotations metadata for textual corpora
- used for searching and displaying corpora
- describes the corpus, the textual primary sources and the annotation
- specification with TEI XML + ODD
- ODD customization and documentation and validation with RELAX NG XML syntax schema
- rich and structured metadata for heterogeneous data

4. Specific corpus architecture – exemplified with KAJUK
- historical corpus of German ‘Nähesperche’ 17th – 19th e.g. private letters
- syntactically annotated, e.g. cohesion and clauses
- token and span annotation, pointing annotation

KAJUK Format
- 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

- 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, semi-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
- description of the annotation guidelines (e.g. what is meant by sub)
- 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’)

6. References
[7] Forstinnenkorrespondenzkorpus
[14] Handle-PIDs, Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
[15] RIDGES
[17] JSON (http://json.org/)
[18] RELAX NG (http://relaxng.org/)
[19] ElasticSearch (http://www.elasticsearch.org/)
[20] Creative Commons Licenses (https://creativecommons.org/licenses/)

KALKUK represented in reANNIS

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.