Modeling Linguistic Research Data for a Repository for Historical Corpora

Carolin Odebrecht
Humboldt-Universität zu Berlin
LAUDATIO-repository.org
Motivation

• to enable the search for, the access of and the reuse of historical corpora we need:
  – a platform where different corpora are stored
    LAUDATIO-Repository
  – a generic data model for corpora and their metadata
    meta-model
  – structured metadata for different kinds of corpora
    metadata format in TEI XML
Application

- LAUDATIO-Repository (Odebrecht et al. 2014)
  - Long-term Access and Usage of Deeply Annotated Information
  - open access repository for historical corpora
  - searching for different types of corpora
  - uniform and structured display of their metadata
  - cooperation between corpus linguistics, historical linguistics and the Computer und Medienservice Humboldt-Universität zu Berlin
Application

- faceted and full-text search for corpora
Application

- display of the corpus metadata
Application

• LAUDATIO-Repository (Odebrecht et al. 2014)
  → Long-term Access and Usage of Deeply Annotated Information
  - open access repository for historical corpora
  - searching for different types of corpora
  - uniform and structured display of their metadata
  - cooperation between corpus linguistics, historical linguistics and the Computer und Medienservice Humboldt-Universität zu Berlin

• the meta-model should enable
  - retrieval of
  - structured search for
  - and a holistic and extensive documentation of the variety of corpora
Meta-model

• issues to consider:

– What do have historical linguistic corpora in common and what don’t they?
– Searching for the ‘right’ corpus: What do we need to know about a corpus?
– How can we generalize different corpus architectures for the application?
Meta-model

What do have historical linguistic corpora in common and what not? - baseline

• linguistic corpora differ with respect to
  – purpose and research questions
  – editing and analyzing formats
  – types of annotation, e.g. token or span annotation
  – content of annotation, e.g. lemmatization or part of speech (e.g. Schiller et al 1999)

➢ different formats and content standards

• linguistic corpora have in common
  – set of textual (‘primary’) documents
  – set of annotations

➢ goal: unified and structured documentation and access of different kinds linguistic corpora
Searching for the ‘right’ corpus: What do we need (or want) to know about a corpus?

• metadata for the ‘primary’ source
  – date, publication place, author (bibliographic features)
  – register, language period, linguistic phenomena (linguistic features)
  → criteria to consider when searching for corpus data

• metadata for the corpus
  – annotation guidelines
  – structure of the annotation/ type of annotation
  – corpus project, annotators
  – annotation tools and checking methods
  – revision history, versions of a corpus
  → criteria to consider when searching for and selecting a corpus

(Odebrecht & Krause 2013)
Meta-model

How can we generalize different corpus architectures for the application? – Analytic class diagram

• a corpus contains documents – no matter how the texts look like
• each document contains annotation keys (with their values) – no matter what is annotated and how it is annotated
Customization of TEI with ODD

- technical basis for the uniform display and search of every class and its attributes in the repository
  - attributes of each class are mapped into corresponding TEI elements and attributes
    - full subset of TEI p5
    - using already existing TEI modules and element groups (e.g. TEI epiDoc)
  - specification, customization and documentation via ODD (Burnard & Rahtz, 2004)

- Searching for the ‘right’ corpus: What do we need to know about a corpus?
  - Who is the author of what?
Customization of TEI with ODD

- each concept (‘Corpus’, ‘Document’, ‘AnnotationKey’) has an attribute ‘author’ (the one who published, wrote or created)
  a. there is always someone who published/created the whole corpus
  b. there is always an author of a document/ the single text
  c. there is always someone who annotated the text

- Diagram:
  - Corpus
    - contains 0 - ∞ sub-corpora
    - is part of 0 - ∞ super-corpora
  - Document
    - contains 1 - ∞ documents
  - AnnotationKey
    - contains 1 - ∞ annotation values
  - a. Research assistant of the corpus project
  - b. someone who is already dead and was a writer in the 18th century
  - c. a student who annotated the corpus in class
Customization of TEI with ODD

- meta-model is mapped to three teiHeader structures
- each header contains the basic teiHeader structure
- linking correspondences between the headers
Some open issues

• testing on non-linguistic corpora
• testing on more relational annotation schemes
• describing disconnected sources such as annotation between documents in the same or in different corpora in the metadata
• interoperability: testing the conversion to other metadata standards, e.g. CMDI, Dublin Core
Accessing a heterogeneous field of linguistic corpora with the help of the open-access repository LAUDATIO

1. How can we access and use heterogeneous research data with the help of an open-access repository?
   - Access, navigation, and search within the repository.
   - Access of data and use of data.

2. Research data and open access: examples from the Digital Humanities.
   - Open access data from various Digital Humanities projects.
   - Access and use of data.

LAUDATIO-Repository

Poster Session 2
Thank you!

And many thanks to

Malte Belz, Karin Donhauser, Hagen Hirschmann, Maxi Kindling, Thomas Krause, Anke Lüdeling, Laura Perlitz, Laurent Romary, Peter Schirmbacher, Vivian Voigt, Amir Zeldes, Dennis Zielke, Florian Zipser
References


- LAUDATIO teiHeader [http://www.laudatio-repository.org/repository/documentation](http://www.laudatio-repository.org/repository/documentation)
- Dublin Core [http://dublincore.org/](http://dublincore.org/)
- CMDI [http://www.clarin.eu/content/component-metadata](http://www.clarin.eu/content/component-metadata)
Appendix
TEI ODD

• meta-model for TEI

• provide a full structure
  – fixed, semi-fixed or open

• generates schemes
  – RELAX NG, W3C, DTD etc.

• generates a HTML documentation
  – code, free text, pictures
Customization of TEI with ODD

- the corpus header, the document header and the preparation header contain information about the author with

```xml
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader type="DocumentHeader" style="Brief">
<fileDesc xml:id="AD_JE2_1677_11_07">
<titleStmt>
<title type="Document"AD_JE2_1677_08_14</title>
<author>
<persName>
<forename>Anna Dorothea</forename>
<surname>von Sachsen-Weimar</surname>
</persName>
</author>
</titleStmt>
</fileDesc>
</teiHeader>
</TEI>
```
Corpus references

- RIDGES Herbology  
  http://hdl.handle.net/11022/0000-0000-2106-4
- Märchenkorpus  
  http://hdl.handle.net/11022/0000-0000-1F5B-9
- KAJUK  
  http://hdl.handle.net/11022/0000-0000-2102-8
- DDB  
  http://hdl.handle.net/11022/0000-0000-20A6-0
- Fürstinnenkorrespondenzkorpus  
  http://hdl.handle.net/11022/0000-0000-20A6-0
Märchenkorpus

• Treetagger Output (Schmid 1994)
  – tokenized
  – automatic pos-tagging and lemmatization
  – no checking
  – guidelines for transcription
  – open access via LAUDATIO-Repository and ANNIS
Fürstinnenkorrespondenzkorpus

- EXMARaLDA (Schmidt & Wörner 2009)
  - tokenized
  - manually annotated and corrected
  - extensive annotation guidelines
  - open access via LAUDATIO-Repository and ANNIS
Deutsche Diachrone Baumbank

- partitur, TIGER
  - tokenized
  - extensive annotation guidelines
  - automatic and semi-automatic annotation and correction
  - open access via LAUDATIO-Repository and ANNIS
Kasseler Junktionskorpus

- idiosyncratic format
  - not uniformly tokenized
  - extensive annotation guidelines
  - manual annotation and correction
  - open access via LAUDATIO-Repository and ANNIS

```
<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>
```

```
<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>
```
RIDGES Herbology

- **EXCEL (EXMARaLDA, Schmidt & Wörner 2009)**
  - manual/(semi-)automatic annotation (e.g. Schmid 1994) and correction (e.g. Dickenson & Meurers 2003)
  - multiple segmentations, step by step normalization
  - open access via LAUDATIO-Repository and ANNIS

<table>
<thead>
<tr>
<th>dipl</th>
<th>clean</th>
<th>norm</th>
<th>pos</th>
</tr>
</thead>
<tbody>
<tr>
<td>das</td>
<td>das</td>
<td>das</td>
<td>PRELS</td>
</tr>
<tr>
<td>ich</td>
<td>ich</td>
<td>ich</td>
<td>PPER</td>
</tr>
<tr>
<td>nur</td>
<td>nur</td>
<td>nur</td>
<td>ADV</td>
</tr>
<tr>
<td>verwändt</td>
<td>verwändt</td>
<td>verwende</td>
<td>VVFIN</td>
</tr>
<tr>
<td>viel</td>
<td>viel</td>
<td>viel</td>
<td>ADV</td>
</tr>
<tr>
<td>minder</td>
<td>minder</td>
<td>minder</td>
<td>ADJD</td>
</tr>
<tr>
<td>gehoffet</td>
<td>gehoffet</td>
<td>gehofft</td>
<td>VVPP</td>
</tr>
<tr>
<td>/</td>
<td>/</td>
<td>/</td>
<td>$(</td>
</tr>
<tr>
<td>mann</td>
<td>mann</td>
<td>man</td>
<td>PIS</td>
</tr>
<tr>
<td>möchte</td>
<td>möchte</td>
<td>möchte</td>
<td>VMFIN</td>
</tr>
<tr>
<td>einigerley</td>
<td>einigerley</td>
<td>einigerlei</td>
<td>ADJA</td>
</tr>
<tr>
<td>metall</td>
<td>metall</td>
<td>Metall</td>
<td>NN</td>
</tr>
<tr>
<td>dardurch</td>
<td>dardurch</td>
<td>dadurch</td>
<td>PAV</td>
</tr>
</tbody>
</table>