ANNIS & SaltNPepper

Infrastructure for multilayer corpus (re)search

Thomas Krause, Humboldt-Universität zu Berlin
Korpuslinguistik
krause@informatik.hu-berlin.de
Plan

1. Multilayer corpora
2. Representing data with PAULA / SaltNPepper
3. LAUDATIOD corpus repository
4. Query and visualization in ANNIS2
5. Outlook
Multilayer Corpora

• We define a multilayer corpus as a corpus:
  – with **multiple layers** of annotation
  – which can **must be created completely independently** of each other

• In theory:
  any annotations (pos+lemma => multiple layers)

• In practice:
  conflicting hierarchies, different data structures/ modalities, alternative values…
Why do we need multilayer corpora?

• Maybe you don’t!
  – Many good concordancers around
  – Some dedicated search tools, e.g. for treebanks

• It depends on the theory being tested:
  – Does it assume structures not representable in one layer? (hierarchy conflicts, discontinuity, …)
  – Do you want to preserve the work of different annotators?
  – Do you know in advance what structures are relevant?
  – Do you want to test competing theories on the same data?
  – Do you need to evaluate interactions between complex phenomena?
  – ……
Research in the SFB 632

• Project D1 “Linguistic DB” of SFB 632 on Information Structure
• Research on multilayer IS corpora
• Development of annotation schemes for IS
• Service tasks:
  – Archive data from different projects in a unified format
  – Deal with heterogeneous annotations for IS & beyond
  – Search and visualization of complex data structures
• ANNIS: ANNotation of Information Structure
  (Dipper et al. 2004; Chiarcos et al. 2008; Zeldes et al. 2009)
  http://www.sfb632.uni-potsdam.de/d1/annis/
Different Data Structures

• Different disciplines
• Variable annotation units (conflicting tokenization, subtokenization)
• Token and span annotations
• Dependency Graphs (terminal units), syntax trees / DAGs (non-terminal units), Labeled edges
• Multimodal data, Metadata, Alignment…
Different Tools/Formats

- Automatic tools with proprietary formats (taggers, parsers, machine learning)
- Diverse dedicated manual tools (transcription, annotation, treebanking, discourse analysis)
- Asynchronously produced annotations from different sources and theories referring to the same data
Annotation Formats

• Tree/Graph
  – Bracketing formats
    (e.g. Penn, Bies et al. 1995)
  – Generic Inline XML
  – Tiger XML / Negra (Synpathy, Tiger, annotate, Brants & Plaehn, 2000)
  – RST tool
    (rhetorical sentence trees, O'Donnel 2000)

• Parallel corpora
  (using PAULA XML)
  – EuroParl
    (CONLL geparste Version)
  – Stax
    (Treealigner, Lundborg et al. 2007)

• Pointing Relations
  – MMAX2
    (Müller & Strube 2006)
  – PALinkA
    (Orasan 2003)
  – Serengeti (TODO)

• Grids
  – EXMARaLDA
    (Schmidt 2004)
  – ELAN
    (Wittenburg et al. 2006)
  – Toolbox
    (Stuart et al. 2007)
  – PRAAT
    (Boersma & Weenink 2009)
Modeling Multilayer Data

• Reduction to typed and labeled nodes and edges (~DAGs)

• Independent layers (stand-off)
  – Retroactively insert / update / replace layers
  – Alternative interpretations (multiple pos-tags, parses, error analyses…)
  – PAULA XML
The dogs are faster than they
SaltNPepper (Zipser 2010)

Pepper (converter framework)

Salt (common metamodel)

Mapper

Tiger XML

EXMARaLDA

PAULA XML

⋯
Salt

• Salt is a structural model for linguistic data – dealing with several linguistic analysis
• Like LAF/GrAF (proposed ISO standard) it is also based upon a general graph structure
• Model-derivation: Every element is either a graph, a node an edge, a layer or a label → Salt is also accessible by general graph accessing mechanisms like traversing
Pepper

- Extensible via Plug-In system for Mappers (OSGi)
- Im- and Export in each supported format
- 3 kinds of modules
  - Importer: Format A → Salt
  - Manipulator: Salt → Salt
  - Exporter: Salt → Format B
LAUDATIO (1)

• “Long term Access and Usage of Deeply Annotated Information”
• Interactive corpus repository
• DFG-project application in progress
• Meant to solve the following problems
  – Stable referencing in publications
  – Reproduce searches and found matches
  – Combine corpus data from different projects
  – (re)use corpus data in your own project in different ways than intended (e.g. statistical analysis)
LAUDATIO (2)

• Cooporation between
  – Humboldt University Berlin
    • Corpus Linguistic working group (Prof. Dr. Anke Lüdeling)
    • Computer and Media Service (Prof. Dr. Peter Schirmbacher)
    • Historical Linguistics working group (Prof. Dr. Karin Donhauser)
  – INRIA Paris
    • Laurent Romary, PhD Habil.

• CLARIN-D members
ANNIS

• ANNIS: **ANNotation of Information Structure**
  (Dipper et al. 2004; Chiarcos et al. 2008; Zeldes et al. 2009)
  http://www.sfb632.uni-potsdam.de/d1/annis/

• Open-source web-based corpus search tool

• Supports diverse multilayer corpora with appropriate search & visualization
ANNIS2 – Application Logic

- Available open source under the Apache 2.0 License
Querying with AQL

• SQL queries too complex for users
• Simple query language based on nodes and relations: (cf. *NiteQL* Carletta et al. 2003, *TIGERSearch* Lezius 2002)
  
  cat="S" & node & cat="S" & pos="PRP" &
  #1 >[func="sb"] #2 &
  #3 >[func="sb"] #4 &
  #4 ->coref #2 &
  meta::language="en"
Query Builder
List of AQL Operators

<table>
<thead>
<tr>
<th>Name</th>
<th>Illustration</th>
<th>Options</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td>A B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.*</td>
<td>A x y z B</td>
<td>.n,m</td>
<td></td>
</tr>
<tr>
<td>&gt;</td>
<td>A</td>
<td>&gt;secedge</td>
<td>[func=&quot;OA&quot;]</td>
</tr>
<tr>
<td>&gt;*</td>
<td>A</td>
<td>&gt;n,m</td>
<td></td>
</tr>
<tr>
<td>-&gt;LABEL</td>
<td>LABEL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-&gt;LABEL*</td>
<td>LABEL_LABEL</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>=`</em></td>
<td>AAA BBB</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>o</em></td>
<td>AAA BBB</td>
<td><em>ol</em></td>
<td></td>
</tr>
<tr>
<td><em>.i</em></td>
<td>AAA B</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Illustration</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>l</em></td>
<td>left aligned</td>
</tr>
<tr>
<td><em>r</em></td>
<td>right aligned</td>
</tr>
<tr>
<td>&gt;@l</td>
<td>left-most child</td>
</tr>
<tr>
<td>&gt;@r</td>
<td>right-most child</td>
</tr>
<tr>
<td>$</td>
<td>Common parent node</td>
</tr>
<tr>
<td>$*</td>
<td>Common ancestor node</td>
</tr>
<tr>
<td>#x:arity=n</td>
<td>Arity</td>
</tr>
<tr>
<td>#x:length=n</td>
<td>Length</td>
</tr>
</tbody>
</table>
Negation in AQL

• New feature since the 2.1.7 release
• Key-value annotations on nodes or token:
  pos != "ART"
• Text span of token:
  tok != "the"
• Works also on regular expressions:
  pos != /V[MA]FIN/
  tok != /.+ing/
Visualization
B4: T-Codex (, Muspilli, Heliand…)  
(Petrova et al. 2009)

- IS (SFB guidelines)
- Morphosyntactic annotation
- “Interlinear” Latin
MHG Speculum Ecclesiae
(WIP, Hagen Hirschmann/Sonja Linde et al.)

- Syntactically annotated corpus
- Normalization over multiple tokens
- Spans for edition references
- Similarly Monsee Fragments
Falko: Topological Fields, Errors
Parallel Corpora

• Representation of aligned document sets
• Edges to align multiple levels paragraph/sentence/word… (TEI, Romary & Bonhomme 2000)
• Handling of mismatching / circular / incomplete alignment
• Search and visualization
Future Directions: Query Language

• Negation of:
  – Meta data
  – Operators (with and without implied existence)
    #1 !> #2 (does #1 exist?)

• Special operators for subtokens?
What you can do

• Test it! (and report bugs)
• Develop:
  – Write visualizations to handle more kinds (your kinds!) of data
  – Better support for multi-modal data (time alignment/streaming)
  – Write and share importers / exporters for your data
  – …
• Make multi-layer corpora available
• Join the mailing list to stay up to date
Thanks!


http://www.sfb632.uni-potsdam.de/~d1/annis/
http://korpling.german.hu-berlin.de/saltnpepper/
References


References II