36. Jahrestagung der DGfS – Marburg 5.-7.3.2014 – Postersession der Sektion CL

ANNIS3:
Towards Generic Corpus Search and Visualization

Thomas Krause, Benjamin Weißenfels, Amir Zeldes and Florian Zipser

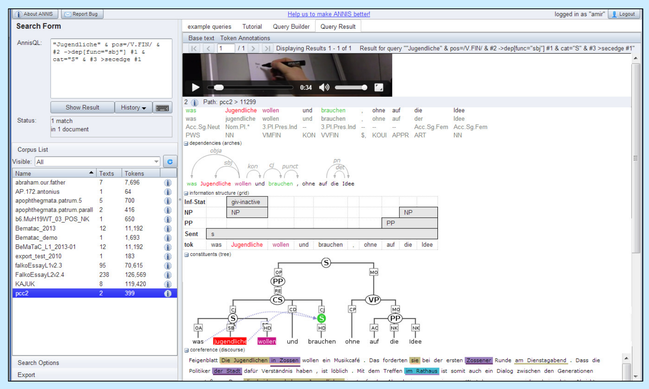
Humboldt-Universität zu Berlin

SFB 632 / D1

LAUDATIO

Challenges

- **Different types of annotation**
  - (semi-)automatic: multiple taggers, constituency/dependency parsers ...
  - manual: coreference, information structure, rhetorical structure ...
- **Different types of corpora**
  - historical – diplomatic and normalized text, manuscript structure
  - multimodal – aligned audio/video streams, multiple overlapping speakers
  - parallel – representing conflicting word and sentence alignment
  - learner language – base text and target hypotheses (Reznicek et al. 2013)
  - multilayer – any and every annotation may repeat or conflict with other structures



- Many corpora violate **assumption of one continuous stream of segments** (multiple languages, speakers, corrected texts...)
  - combinatorial explosion: unrealistic to design tailored system for each type
  - reusing the architecture for unique search and visualization applications
  - Simplifying the query language (**AQL**) to deal with new structures

Unified data model and query language

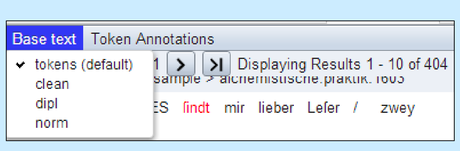
Dealing with multiple source formats

SaltNPepper

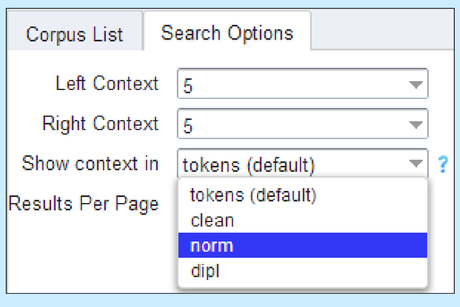
- Annotations come from multiple formats
  - Convert multiple formats with SaltNPepper (Zipser & Romary 2010)
  - Use Salt data model to represent merged information in ANNIS3
  - Reconcile conflicting **segmentations**
- Archive data in PAULA XML (Dipper 2005), a standoff XML format for multi-layer corpora

Segmentations in the ANNIS3 data model

- Deal with multiple alternative base texts: **one segmentation each**
- Any annotation layer can be a segmentation:
  - diplomatic/normalized word forms
  - broad and narrow phonetic transcription
  - data from different speakers



- Segmentations can be selected as:
  - The base text for concordance KWIC views (Key-Word in Context)
  - The unit for defining the desired context (e.g. ±5 normalized word forms)
  - Search criteria for proximity and adjacency in the ANNIS Query Language (**AQL**), using **typed precedence operators**:



  - Search consecutive utterances of a speaker (even if others intervene):

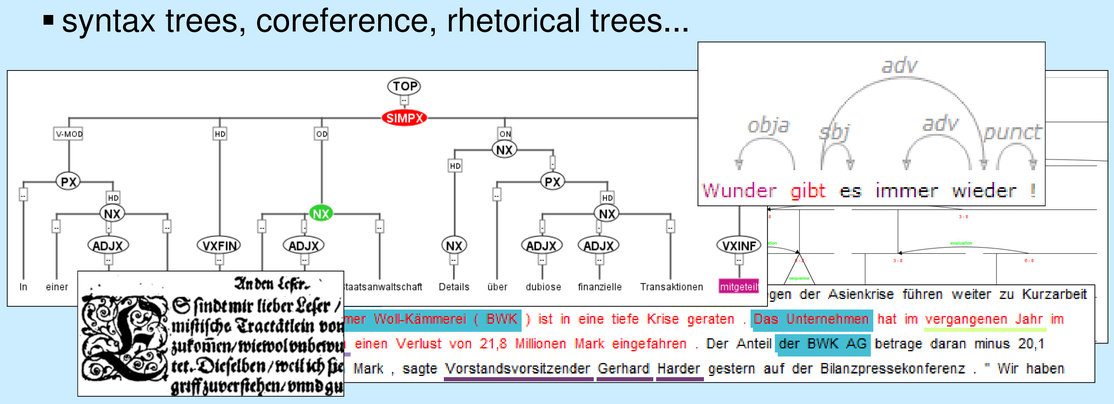
```
"ja" .instructor "ja" //the instructor says ja twice
```

  - Find differently spelled words within 10 diplomatic units in a manuscript:

```
/s.*/ .dipl,1,10 /ſ.*/ &  //words in s- and ſ- in 1-10 dipl
lemma == lemma &          //two identical lemmas
#1 _=_ #3 &               //1st word covers 1st lemma
#2 _=_ #4                 //2nd word covers 2nd lemma
```

Reusable, Configurable Visualizations

- Dedicated visualization are needed for many common data types:
  - syntax trees, coreference, rhetorical trees...



- But many corpora have unique annotations:
  - Different conventions for digital editions of manuscripts
  - Subtitle corpora, film transcripts
  - Information structure...
- Approach: Use **annotation triggered style sheets**
  - Expressiveness of HTML5 with flexibility of CSS3
  - Short development cycles from corpus to visualization
- Implementation as **configuration file** and **CSS file**:

```
tok   span; style="word"       value
spkr  div:spkr; style="spk"    value
lang  span:lang; style="lang"  value
info  t:title; style="info"    value
```

```
.word:after{content: " ";}
div.spk{display: block; padding-top: 6px; padding-bottom:
6px; text-indent: -65px; padding-left: 65px}
div.spk:before{content: attr(speaker) ": "; font-weight: bold}
.lang{color: blue; font-style: italic}
.info:hover{color: red}
```

full scene (annotated dialogue)

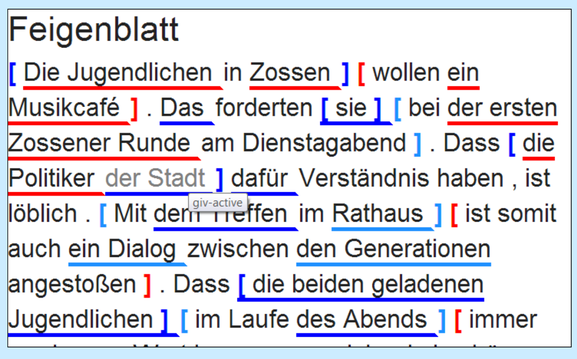
Ibrahim: ka gane abin da nake so ka fahimta ? Yaya za a yi mutan
cewa za sub a da *invitation* nan za su ba *invitation* nan
? *if they are not serious* kawai mu dauki abin nan mu k
kwanaki nawa ? Har man maneji yana neman ya kare n
so ka gane shi ne in ba su ba ni *invitation* din nan yau
kai wani guri a buga mun . *that's it* . To , shi ke nan ina

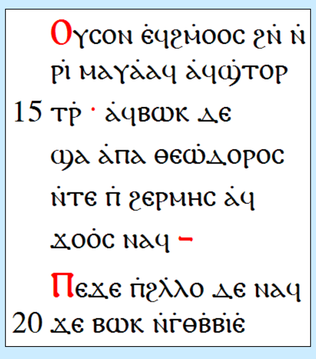
Ummi: Salamu alaikum

Ibrahim: Amin wa alaiku mus salamu

Applications

- A variety of dedicated visualizations can be developed with little code
- Digital manuscript editions for Coptic (Projects KOMeT/SCRIPTORIUM): http://coptic.pacific.edu/
- Visualization of Information Structure in PCC (Stede 2004)





Future directions

- Adding matching javascript files for more interactive visualizers
- Visualizer-triggered searching (click on words, jump between linked results)
- Aggregate visualizers based on results from multiple documents/corpora

References

- Dipper, S. 2005. XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, 39-50.
- Reznicek, M./Lüdeling, A./Hirschmann, H. 2013. Competing target hypotheses in the falko corpus: A flexible multi-layer corpus architecture. In Díaz-Negrillo, A./Ballier, N./Thompson, P. (eds.) Automatic Treatment and Analysis of Learner Corpus Data. Amsterdam: John Benjamins, 101–124.
- Stede, M. 2004. The Potsdam Commentary Corpus. In Webber, B./Byron, D. K. (eds.) Proceeding of the ACL-04 Workshop on Discourse Annotation. Barcelona, Spain, 96–102.
- Zipser, F./Romary, L. 2010. A Model Oriented Approach to the Mapping of Annotation Formats using Standards. *Workshop Language Resource & Language Technology Standards, LREC 2010*. Malta, 7-18.