DGfS-CL

Thomas Krause[1], Benjamin Weißenfels[1], Amir Zeldes[1] and Florian Zipser[1,2]
1. Humboldt-Universität zu Berlin, 2. INRIA

ANNIS3: Towards Generic Corpus Search and Visualization

This poster showcases the latest developments in the generic representation of digital corpora using ANNIS (Zeldes et al. 2009), a browser-based, open source corpus query architecture. Recent advances in computational linguistics tools and corpus formats have led to a rapid growth of multi-layer corpus projects, integrating data from fully manual annotation (e.g. coreference, information structure), fully-automatic tools like taggers and parsers for morphology and syntax, as well as semi-automatic annotations combining the two. Adding to these the special requirements of different corpora, such as highly detailed historical corpora with diplomatic and normalized text, multimodal corpora with aligned A/V streams, or parallel corpora with aligned multilingual texts, quickly leads to a combinatorial explosion: each corpus requires unique search and visualization capabilities, and the overhead of designing the query system core and encoding formats is repeated countless times before research can progress.

With ANNIS3, we attempt to move one step closer to a generic solution for the corpus search and representation problem. We model primary linguistic data (transcriptions from any number of simultaneous speakers, aligned multilingual texts, diplomatic and normalized levels…) as nodes in a graph and designate specific layers of information as segmentation layers (cf. Krause et al. 2012). These are treated as "word forms" or "tokens", though there can be any number of such layers and they may overlap freely and be used to define adjacency in queries, token-distance between search elements or query hit context. Above and below these segmentations, we represent any and all annotation types as a multi-DAG, an annotation graph which may contain as many subgraphs as needed, including cycles, as long as each type of annotation is free of cycles within itself.

The problem of visualizing heterogeneous data is approached from two directions using an extensible plugin-based system. The system offers dedicated, highly optimized visualizations for some common data-types, such as constituent and dependency trees, coreference, annotation grids, aligned pdf and multimedia plugins and more. A new module in ANNIS3 constructs annotation-triggered HTML/CSS on the fly, beginning and ending HTML tags depending on the scope of annotation nodes and filling the attributes and styles of such elements with values from the annotation model. Figure 1 below demonstrates some of the applications of this flexible architecture, which will be shown in the demo.

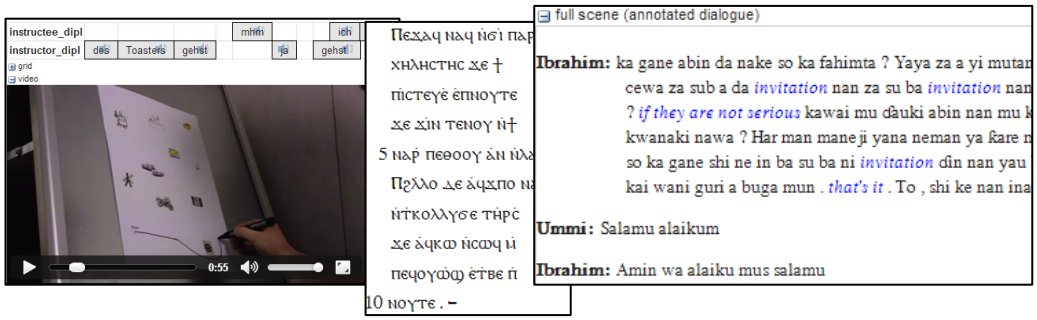

Figure 1. Visualizing dialogue data, historical manuscripts and film subtitles in ANNIS.

References

Krause, T./Lüdeling, A./Odebrecht, C./Zeldes, A. 2012. Multiple tokenizations in a diachronic corpus. In: *Exploring Ancient Languages through* Corpora. Oslo, Norway.

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, UK.