Donnerstag
09.03.2017
12:45 – 13:45
B4 1, Foyer

Agile creation of multi-layer corpora with corpus-tools.org

Stephan Druskat
HU Berlin
druskats@hu-berlin.de

Thomas Krause
HU Berlin
krauseto@hu-berlin.de

Carolin Odebrecht
HU Berlin
odebreca@hu-berlin.de

Agile corpus creation replaces the linear corpus creation process with iterative cycles of query, schema edits, annotation and analysis. We demonstrate corpus-tools.org, a suite of generic tools tailored to the agile creation of multi-layer corpora. It consists of Salt, a graph-based meta model and API for linguistic data; Pepper, a conversion platform; Atomic, an extensible annotation software; ANNIS, a search and visualization architecture for multi-layer corpora. As of now, Atomic lacks search capabilities for agile workflows. ANNIS provides a search system based on annotation graphs, and the ANNIS Query Language (AQL). ANNIS, however, has been optimised for linear workflows, which graphANNIS (`https://git.io/vijrI`), a new C++-based implementation, will change. It will also make ANNIS self-contained, dropping the dependency to a separate database installation. graphANNIS supports a large subset of AQL, aligns its data representation more closely with the Salt model, and provides a Java API. Its encapsulation allows for graphANNIS to be embedded in Atomic, as its search engine. While Atomic will be responsible for storage of corpus data, graphANNIS provides an additional index which is updated whenever a document is changed. For search tasks, Atomic will provide a GUI section for AQL queries. These will be parsed by the ANNIS AQL parser and passed to the graphANNIS search system, which will return the Salt IDs of the matched nodes, in turn used in Atomic to present the results. This setup will provide corpus-tools.org with capabilities for agile multi-layer corpus creation.

CL Poster

References: • Druskat, S.; Krause, T.; Odebrecht, C. (preprint): Agile creation of multilayer corpora with corpus-tools.org. `https://doi.org/10.5281/zenodo.157166`. • Voormann, H.; Gut, U. (2008): Agile corpus creation. *CLLT*, 4(2), 235–251. • Zipser, F.; Romary, L. (2010): A model oriented approach to the mapping of annotation formats using standards. In: *Proceedings of LREC 2010*, Valletta. • Zipser, F.; Zeldes, A.; Ritz, J.; Romary, L.; Leser, U. (2011): Pepper: Handling a multiverse of formats. Poster, *DGfS 2011*, Göttingen. • Druskat, S.; Bierkandt, L.; Gast, V.; Rzymski, C.; Zipser, F. (2014): Atomic: an open-source software platform for multi-level corpus annotation. In: *Proceedings of KONVENS 2014*, 228–234. • Krause, T.; Zeldes, A. (2016): ANNIS3: A new architecture for generic corpus query and visualization. *DSH*, 31(1), 118–139.