A workflow for creating, analysing, and storing multi-layer corpora: Pepper, Atomic, ANNIS and LAUDATIO

Stephan Druskat[1], Thomas Krause[2], Carolin Odebrecht[2], Florian Zipser[2]
[1]Friedrich Schiller University Jena — [2]Humboldt-Universität zu Berlin

The creation and analysis of corpus linguistic resources can be a costly and error-prone process. Apart from the complexity of the annotation process itself, there are larger technical obstacles to be overcome. Single tools have to be combined in a common workflow, and different formats taken into account. This poster presents a family of well-aligned open source tools which support the conversion, annotation, and analysis of linguistic corpora, as well as securing their long-term accessibility, in a complete workflow.

The interoperability of these tools is guaranteed by the use of a common data model – Salt (Zipser & Romary, 2010) – which, among other things, is used as an intermediate model for the conversion framework Pepper (Zipser et al., 2011). With Pepper, many linguistic formats can be converted into each other, thereby allowing existing data to be included in the workflow. The support for a multitude of linguistic formats allows for the replacement of single components as well as the integration of further tools into the workflow presented here.

The annotation of corpora is carried out in Atomic (Druskat et al., 2014), an extensible annotation platform. Atomic also utilizes Salt – in this case as its concrete data model – and thus allows for theory-neutral annotation which is independent of tagsets and annotation types. By embedding Pepper, it supports a wide variety of source formats for further annotation, as well as target formats for export. Additionally, its plugin-based architecture makes it possible to easily extend the software, e.g., with additional editors, data views, or processing components. For a new annotation type, for instance, a dedicated editor can thus be created and integrated.

At any point in the annotation process, the annotated data can be transferred to the search and visualization tool ANNIS (Krause & Zeldes, 2014) for visualisation and analysis. Conclusions from the analysis can then, for example, also flow back into the annotation process.

When a corpus is ready for publication, it can be released in different formats to a public repository – in the case of historical text corpora, for example, the LAUDATIO-Repository (Odebrecht et al., 2015). Third parties can then download, reference and re-use the data.

References

Druskat, Stephan, Lennart Bierkandt, Volker Gast, Christoph Rzymski & Florian Zipser. 2014. Atomic: an open-source software platform for multi-level corpus annotation. In Josef Ruppert & Gertrud Faaß (eds.), *Proceedings of the 12th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2014)*, 228–234.

Krause, Thomas & Amir Zeldes. 2014. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities.* `http://dx.doi.org/10.1093/llc/fqu057`.

Odebrecht, Carolin, Thomas Krause & Anke Lüdeling. 2015. Austausch von historischen Texten verschiedener Sprachen über das LAUDATIO-Repository. Poster presented at 37. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, 5 March, Leipzig University, Leipzig, Germany.

Zipser, Florian & Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In *Proceedings of the Workshop on Language Resource and Language Technology Standards*, Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta.

Zipser, Florian, Amir Zeldes, Julia Ritz, Laurent Romary & Ulf Leser. 2011. Pepper: Handling a multiverse of formats. Poster presented at 33. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, 24 February, Göttingen University, Göttingen, Germany.