DGfS-CL

Florian Zipser, INRIA, HU-Berlin
Mario Frank, HU-Berlin
Jakob Schmolling, HU-Berlin

Merging data, the essence of creation of multi layer corpora

The last couple of years have shown an increasing number of multi layer corpora. Such corpora allow the analysis of phenomena spreading through multiple annotation layers, for example corpora like TueBaDZ (see: http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html), PCC (see: http://www.ling.uni-potsdam.de/acl-lab/Forsch/pcc/pcc.html), FALKO (see: https://u.hu-berlin.de/falko) and many other corpora contain annotations on syntactical, rhetorical, information structural and other layers. Often, annotations were created manually, or semi-automatically with different tools like EXMARaLDA (see: http://www.exmaralda.org/), TreeTagger (see: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/) and MMAX2 (see: http://mmax2.sourceforge.net/). These tools are powerful and usable but unfortunately only provide a minimum of interoperability, which impedes the creation of multi layer corpora. Thus, multiple layers of such corpora often had to be merged by hand or by very proprietary scripts implemented for just one use case and therefore could not be reused for other corpora easily.

With this poster we present a tool which merges several layers of annotations into a single multi layer corpus. When creating a multi layer corpus, several analyses base on the same primary data and often also on the same tokenization. For instance, in PCC there are multiple annotation layers covering the same piece of text, as shown in Figure 1.

Figure 1: Figure 1: on the left a dependency annotation of a sentence from the PCC corpus, on the right the same text having information structure annotations. Both visualization are taken from ANNIS (see: http://www.sfb632.uni-potsdam.de/annis/).

We started merging the data on the tokenization level and traversed bottom up, to merge even higher levels of annotation. This concept is implemented in a module for the converter framework Pepper (see: https://u.hu-berlin.de/saltnpepper) with use of the common meta-model Salt. By using Pepper, the merging module is able to handle all formats which can be imported by a Pepper module. Multi layered corpora then can be mapped into a multilayer formats like PAULA (Chiarcos et al. 2008), GrAF (Ide & Suderman 2007) or can be imported into ANNIS.

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J. & Stede, M. (2008),
A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *Traitment automatique des langues* 49, 271-293.

Ide N., Suderman K. (2007).
GrAF: A Graph-based Format for Linguistic Annotations. *Proceedings of the Linguistic Annotation Workshop*, held in conjunction with ACL 2007, Prague, June 28-29, 1-8.