How to get more out of a corpus: A generic multilayer corpus query interface

Simon Sauer (Humboldt-Universität zu Berlin) & John M. Kirk (Technische Universität Dresden)

Many corpora are made available only in the format(s) used for compiling and annotating. While some formats require corpus users to download and install specific software, others are very generic and can be used with any text editor. Naturally, however, these formats are optimized for the corpus builders' own research interests and might make investigating other research questions very difficult even when the data as such is applicable. Metadata are often not incorporated into the corpus proper and are only available within the corpus's documentation.

Web-based corpus interfaces can alleviate most of these issues. Developing and maintaining a separate interface for each corpus, however, is time-consuming and expensive. Generic interfaces such as CQPweb (1), on the other hand, are often restricted to simple token-based annotations such as part-of-speech tags.

ANNIS (2) is an open source, cross platform, web browser-based search and visualization architecture for complex multilayer linguistic corpora with diverse types of annotation. It is format- and theory-agnostic as all data are modelled as abstract nodes and edges. ANNIS is compatible with a multitude of formats through the SaltNPepper meta model and conversion framework (3). Queries are formulated in a powerful query language, which allows complex queries across different annotation types and even across corpora. Unicode and regular expressions are fully supported. Query hits can be displayed in a variety of independent visualization modules, such as a classic key word in context view, a grid-style view that supports both token-based annotations and annotation spans, a document view where annotations are represented by colour-coded highlighting, arches for dependency relations, trees for syntactic structures and others. Aligned audio and video data can be played back by clicking on any token or annotation.

The underlying multilayer standoff architecture allows for more than one 'basic' text layer and for annotations completely independent from each other. For example, the diachronic corpus RIDGES (4) features a manuscript-near transcription, a further transcription using only contemporary characters and resolving graphical issues such as word-internal line breaks as well as a normalized transcription in today's orthography. Graphical annotations such as paragraphs, pages, headings, or margins are based on the manuscript-near transcription whereas part-of-speech tags refer to the normalized transcription. Metadata can be queried just like any other annotation, so you can easily form ad-hoc subcorpora by limiting your results to for example manuscripts from before 1630.

Besides ANNIS's features in general, this paper will demonstrate its benefits on the concrete example of ICE- and SPICE-Ireland (5), which hitherto were only available in a linear text-based format, with metadata being restricted to the corpus's handbook. In ANNIS, the spoken subcorpus displays speakers on separate layers, so overlaps are graphically represented and much easier to recognize. All annotations are also on separate layers, which significantly improves both legibility of the actual text and the ability to query specific annotations.

References

(1) CQPweb: http://cqpweb.lancs.ac.uk
A. Hardie. 2012. "CQPweb - combining power, flexibility and usability in a corpus analysis tool". In International Journal of Corpus Linguistics. 17 (3): 380–409. Available at: http://www.lancaster.ac.uk/staff/hardiea/cqpweb-paper.pdf

(2) ANNIS: http://annis-tools.org
A. Zeldes, J. Ritz, A. Lüdeling, Ch. Chiarcos. 2009. "ANNIS: A Search Tool for Multi-Layer Annotated Corpora". In M. Mahlberg, V. González-Díaz, C. Smith (Eds.), Proceedings of Corpus Linguistics 2009. Available at: http://edoc.hu-berlin.de/docviews/abstract.php?id=36996

(3) SaltNPepper: http://korpling.german.hu-berlin.de/saltnpepper
F. Zipser, L. Romary. 2010. "A model oriented approach to the mapping of annotation formats using standards". In G. Budin, L. Romary, T. Declerck, P. Wittenburg (Eds.), LREC 2010 Workshop, Proceedings, W4: Language Resource and Language Technology Standards. Paris: ELRA. Available at: http://hal.inria.fr/inria-00527799

(4) RIDGES: http://korpling.german.hu-berlin.de/ridges/index_en.html

(5) ICE- and SPICE-Ireland: http://ice-corpora.net/ice/iceire.htm