New insights into existing ICE data with a new corpus architecture

Simon Sauer, Humboldt-Universität zu Berlin, sauersim@hu-berlin.de
John M. Kirk, Technische Universität Dresden, jk@etinu.com

Most of the ICE subcorpora available today are provided only in plain text files, marked up according to the standardized ICE guidelines. This format certainly has its benefits in providing a low-tech software-independent solution while at the same time guaranteeing compatibility to all previous ICE releases. However, technological advances have made it possible to parse and annotate large amounts of data (semi-)automatically in a vast variety of different categories. Moreover, digital audio recordings and ever-increasing bandwidths have made it feasible to publish (transcription-aligned) audio and even video data. On the other hand, the more information is put into a corpus, the less legible the original text or transcription becomes, both for human readers and for corpus analysis tools.

The recent ICE Nigeria has opted to make use of several formats, including ELAN files that are time-aligned with the audio recordings. This makes it much easier to look at the data and allows for detailed qualitative analyses. Quantitative analyses, on the other hand, are faced with a number of difficulties. Analyses across several ICE subcorpora require different approaches, as the ICE mark-up has not been applied. Annotations and the time-alignment present in the ELAN files cannot be used with standard corpus analysis tools.

This paper will present the benefits of a generic multilayer corpus architecture that allows multiple, independent annotations of any type (including time-alignment) as well as the inclusion of metadata. At the core of this architecture lies SaltNPepper (1), which consists of a generic meta-model and an extendable converter framework that allows data from a variety of formats and tools, e.g. ELAN or generic XML. The data can then be converted into the format of ANNIS (2), an open source, web browser-based search and visualization interface.

We have converted ICE Ireland (3), spoken and written, as well as the pragmatically and prosodically annotated SPICE Ireland into this format and can therefore present the benefits of this approach by directly comparing it to the standard plain text format (hitherto the only format available for ICE Ireland).

ANNIS provides powerful query capabilities across different (meta)annotation types and even across corpora. This not only allows for complex queries but also for the creation of ad-hoc subcorpora using the metadata, so different groups of documents or speakers can easily be compared or results can be restricted to a specific subset. So far, the only way to filter results was to manually look up every single relevant speaker ID in the corpus handbook.

Query hits can be displayed in a variety of visualizations, including e.g. a grid-style view that displays annotations in separate layers, colour-coded text or syntax trees, as well as downloaded for further (statistical) analysis. Aligned audio/video data can be played back by clicking on any token or annotation. Even spoken subcorpora that can only provide transcriptions benefit greatly by being able to indicate overlaps visually while at the same time showing the transcription proper free from all the mark-up clutter.

References

(1) 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

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