ICAME 34
Pre-conference workshop “Compilation and annotation of spoken corpora: Towards best practice?”
Abstract

BeMaTaC: A Flexible Multilayer Spoken Dialogue Corpus for Contrastive SLA Analyses

Simon Sauer & Anke Lüdeling (Humboldt-Universität zu Berlin)

This paper describes the construction of a deeply annotated multimodal corpus of spoken learner German and its native speaker reference corpus. While the analysis of interlanguage (Selinker 1972) is interesting on each particular linguistic level, it is even more valuable to study interactions across several levels. How do learners use prosody in connection with information structure? How do disfluencies tie in with lexical density or syntactic complexity? In order to make studies like this possible, we constructed a corpus which (a) enables maximally flexible annotation and analysis and (b) conforms to sustainable and well-described formatting standards.

The Berlin Map Task Corpus (BeMaTaC) uses a map task design (Anderson et al. 1991) where one speaker instructs another speaker to reproduce a route on a map with landmarks. This setting allows for spontaneous dialogue in a controlled context. The drawing hand of the instructee is recorded on video. BeMaTaC builds on ideas developed for the Hamburg Map Task Corpus HaMaTaC (Schmid et al. 2010), which, however, is not consistently tokenized and documented and is therefore unsuitable for quantitative studies or further annotation. Our transcription follows a loosely orthographic scheme and is tokenized. All data is stored in an extensible and flexible multilayer standoff architecture, which allows the addition of annotation layers at any point. The freely extendable open source SaltNPepper converter framework (Zipser & Romary 2010) makes it possible to use many different dedicated annotation tools on the same data. Transcriptions and phonetic/phonological annotations are created and carefully aligned with Praat (Boersma 2010), followed by automatic lemmatization and POS-tagging (Schmid 1994). Token-based and span annotations such as disfluency tags are added using EXMARaLDA (Schmidt & Wörner 2009), while dependency structures are annotated using the MaltParser (Nivre et al. 2007).

BeMaTaC can be accessed using ANNIS (Zeldes et al. 2009), an open-source browser-based search and visualization tool for deeply annotated corpora. Using a highly flexible node-and-edge-based query language and permitting queries over multiple corpora, ANNIS can visualize data like simple token or span annotations but also syntax trees or pointing relations. BeMaTaC’s annotations and its audio and video recordings are time-aligned, permitting the playback of the corresponding sequence with a simple click on a token.

References

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Docherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson & Regina Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech* 34, 351-366.

BeMaTaC. [http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/bematac], freely available under the Creative Commons licence.

Boersma, Paul. 2010. Praat, a system for doing phonetics by computer. *Glot International* 5 (9/10): 341-345.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov & Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13 (2), 95-135.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*.

Schmidt, Thomas & Kai Wörner. 2009. EXMARaLDA - Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics* (19:4), 565-582.

Schmidt, Thomas, Hanna Hedeland, Timm Lehmberg & Kai Wörner. 2010. HAMATAC - The Hamburg MapTask Corpus. [http://www.exmaralda.org/files/HAMATAC.pdf]

Selinker, Larry. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching* 10/3, 31-54.

Zeldes, Amir, Julia Ritz, Anke Lüdeling & Christian Chiarcos. 2009. ANNIS: A Search Tool for Multi-Layer Annotated Corpora. *Proceedings of Corpus Linguistics* 2009, July 20-23.

Zipser, Florian & Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. *Proceedings of the Workshop on Language Resource and Language Technology Standards*, LREC 2010.