‚Non-canonical' units in spoken language corpora

Anke Lüdeling
Humboldt-Universität zu Berlin
anke.luedeling@rz.hu-berlin.de

My presentation will deal with the compilation, the annotation, and the analysis of spoken language corpora. I will use the German map task corpus BeMaTaC (Giesel et al. 2013, Sauer & Lüdeling 2013)[1] for illustration.

Most traditional, descriptive, and theoretical grammars describe some kind of idealized (written) language. Most corpus tools (taggers, parsers, etc.) are trained on written language data. Accordingly, the treatment of spoken corpus data poses both conceptual and technical problems.

The first part of this talk will deal with the units (words, phrases, sentences) that occur in spoken language and how they might differ in type and contribution from their counterparts in written language data. I will specifically address the problem of verb-less units for which many grammars provide no analysis even though they are very frequent in spoken data.

In a second part I will talk about the analysis and annotation of disfluencies (filled and unfilled pauses, repetitions, etc., see e.g. Eklund 2004, or Belz & Klapi 2013). Again, disfluencies are a frequent phenomenon that is often not adequately treated in grammatical models.

I will argue that we need a multi-layer model that makes it possible to analyze and annotate phenomena that are characteristic for spoken language (such as verb-less units or disfluencies) in addition to more common 'canonical' units. Deeply annotated corpora that contain all the different kinds of information make complex questions about form and function of spoken units possible.

Referenzen

Belz, Malte & Myriam Klapi (2013). Pauses following Fillers in L1 and L2 German Map Task Dialogues. In: *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech.* Stockholm, Sweden.

[1] BeMaTaC is freely available at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/bematac. While its design is based on the Hamburg Map Task corpus (HAMATAC , see Hedeland & Schmidt 2012), BeMaTaC differs from HAMATC in many respects.

Eklund, Robert (2004) *Disfluency in Swedish Human-Human and Human-Machine Travel Booking Dialogues*. Linköpings Universitet, Dissertation.

Giesel, Linda, Myriam Klapi, Daisy Krüger, Isabelle Nunberger, Oxana Rasskazova & Simon Sauer (2013) *Berlin Map Task Corpus –A Deeply-Annotated Multimodal Map-Task Corpus of Spoken Learner and Native German*. Poster bei der DGfS-CL 2013, Potsdam.

Hedeland, Hanna & Thomas Schmidt (2012) Technological and Methodological Challenges in Creating, Annotating and Sharing a Learner Corpus of Spoken German. In: Schmidt, Thomas & Kai Wörner (eds.) *Multilingual Corpora and Multilingual Corpus Analysis*. John Benjamins, Amsterdam, 25–46.

Sauer, Simon & Anke Lüdeling (2013) BeMaTaC. A Flexible Multi-Layer Spoken Corpus for Contrastive SLA Analysis. Vortrag bei der ICAME 34, Santiago de Compostela, Mai 2013.