‚Nicht kanonische' Einheiten in Korpora gesprochener Sprache

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

Mein Vortrag wird sich mit dem Aufbau, der Aufbereitung und der Auswertung von Korpora gesprochener Sprache beschäftigen. Dabei werde ich das deutsche Map-Task-Korpus BeMaTaC (Giesel et al. 2013, Sauer & Lüdeling 2013)[1] zur Illustration verwenden.

Die meisten Grammatiken (‚traditionelle' deskriptive genauso wie theoretische) setzen eine Art idealisierter (Schrift)sprache voraus. Auch die meisten Korpuswerkzeuge (Tagger, Lemmatisierer, Parser) sind auf schriftsprachlichen Daten trainiert. Bei der Bearbeitung gesprochener Daten müssen daher sowohl technische wie auch grundlegende konzeptuelle Fragen gelöst werden, von denen ich zwei ansprechen möchte.

Im ersten Teil des Vortrags wird besprochen, wie grammatische Kategorien (Wörter, Phrasen, Sätze) in gesprochenen Daten vorkommen und wie sie annotiert werden können. Dabei werde ich besonders auf verblose Einheiten eingehen, die in gesprochener Sprache häufig sind, für die aber in den meisten Grammatiken keine Analyse gegeben werden kann.

Der zweite Teil des Vortrags beschäftigt sich dann mit der Annotation und Auswertung von Disfluencies (gefüllte und ungefüllte Pausen, Wiederholungen, Abbrüche etc., siehe z.B. Eklund 2004). Auch Disfluencies sind häufig, auch sie sind grammatischen Modellen meist nicht integriert. Sie haben viele Funktionen (prozessuale wie die Verringerung von Planungsdruck genauso wie diskursfunktionale wie Turn-Halte-Signale, siehe z. B. Swerts et al. 1995). Erst wenn sowohl grammatische Einheiten als auch disfluencies in einem Korpus annotiert sind, können bestimmte Fragestellungen zur Position von bestimmten Disfluencies in einer Äußerung oder zu Disfluencymustern (Belz & Klapi 2013) ausgewertet werden.

Referenzen

Belz, Malte & Klapi, Myriam (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.

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

[1] Das Korpus ist frei verfügbar unter http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/bematac. Es ist im Design angelehnt an das HaMaTaC (Hedeland & Schmidt 2012), unterscheidet sich aber in vielen Aspekten.

Giesel, Linda, MyriamKlapi, Daisy Krüger, Isabelle Nunberger, OxanaRasskazova, 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 & Schmidt, Thomas (2012) Technological and Methodological Challenges in Creating, Annotating and Sharing a Learner Corpus of Spoken German. In: Schmidt, Thomas & Wörner, Kai (eds.) *Multilingual Corpora and Multilingual Corpus Analysis*. John Benjamins, Amsterdam, 25–46.

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

Swerts, Marc; Wichmann, Anne & Beun, Robert Jan (1995) Filled pauses as markers of discourse structure. In: *Proceedings of ICSLP*, vol. 2,1033–1036.