BeMaTaC: ein multimodales Map-Task-Dialogkorpus

Oxana Rasskazova & Simon Sauer, Humboldt-Universität zu Berlin

Gesprochene Sprache und Sprachverarbeitung
Workshop des GSCL-Arbeitskreise "Sprache und Dialog"
Pre-Conference-Workshop zur GSCL-Tagung 2013
Darmstadt 23. September 2013

Abstract

BeMaTaC (Berlin Map Task Corpus, Giesel et al. 2013) ist ein Korpus gesprochener Sprache mit Aufnahmen von deutschen Muttersprachlern sowie einem Subkorpus Deutschlernender. Die Dialoge basieren auf einer Map-Task-Aufgabenstellung (Anderson et al. 1991), in welcher die eine Person der anderen eine Route auf einer Karte mit Landmarken (Brinckmann et al. 2008) erklärt. Die Route muss auf diese Weise reproduziert werden – die Karte mit der zeichnenden Hand wird dabei auch auf Video aufgezeichnet. Dieses Setting erlaubt sowohl spontane konversationale Sprache als auch einen thematisch und pragmatisch kontrollierten Kontext.

Das Design (ursprünglich basierend auf HAMATAC, Schmid et al. 2010) reflektiert die ursprüngliche Zielsetzung von BeMaTaC, kontrastive Untersuchungen von gesprochener Lerner- vs. Muttersprache zu ermöglichen. Daneben erlauben vielfältige Annotationen jedoch auch die detaillierte Untersuchung konversationaler Phänomene, wie z.B. von Disfluencies und Reparaturen (Belz & Klapi 2013), Backchanneling, Syntax gesprochener Sprache, räumlicher Konzeptualisierung oder von Interaktionen zwischen Prosodie und Informationsstruktur. Ein besonderer Fokus liegt dabei auf Analysen, welche keine standardsprachlich normierte, sondern die gesprochene Sprache selbst zum Gegenstand haben. Dies soll dazu beitragen, wichtige Rückschlüsse auf linguistische Theorie zu erarbeiten, da diese oftmals auf Nichtstandardvarietäten nicht anwendbar ist.

Ziel dieses Beitrags ist es, BeMaTaC als multimodale Dialogressource sowie die dafür erarbeiteten und verwendeten Annotationsrichtlinien und -werkzeuge vorzustellen. Zur Aufbereitung wurden MAUS (Wesenick & Schiel 1994), Praat (Boersma 2010), EXMARaLDA (Schmidt & Wörner 2009), TreeTagger (Schmid 1994), MaltParser (Nivre et al. 2007) und Arborator (Gerdes 2013) verwendet. Das Korpus verfügt über eine flexible und erweiterbare Multilayer-Standoff-Architektur und ist unter einer Creative-Commons-Lizenz inklusive aller Primärdaten frei verfügbar. Über das SaltNPepper-Konvertierungsframework (Zipser & Romary 2010) sind die Daten in einer Vielzahl von Formaten und Werkzeuge verwendbar. Die browserbasierten Such- und Visualisierungsoberfläche ANNIS (Zeldes et al. 2009) erlaubt mit ihrer flexiblen Knoten-und-Kanten-basierten Abfragesprache und verschiedensten Visualisierungsoptionen einen umfassenden Zugriff auf BeMaTaC.

Referenzen

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.

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

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

Gerdes, Kim. 2013. Arborator. [http://arborator.ilpga.fr]

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. DGfS-CL 2013.

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]

Wesenick, Maria-Barbara & Schiel, Florian. 1994. Applying Speech Verification to a Large Data Base of German to Obtain a Statistical Survey About Rules of Pronunciation. Proceedings of ICSLP 1994, 279-282.

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.