Berlin Dialog Corpus (BeDiaCo)
ein multimodales Korpus für Konvergenz- und Dialogforschung

Oxana Rasskazova, Simon Sauer und Christine Mooshammer, Humboldt-Universität zu Berlin
CLARIN-D-Workshop: Sprachdatenbanken – von der Aufnahme zur Publikation

1. Motivation und Ziele

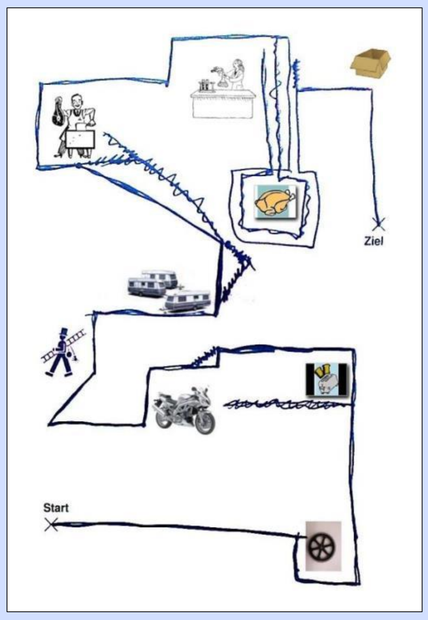
1.1 Phonetische Untersuchungen
- Akustische Analyse für die Konvergenzforschung
  - Messung von Vokaldauer, Vokalqualität usw.
  - Timing zu Turn-Taking
- Vergleich zu artikulatorischem Experiment mit gleichen Aufgaben

1.2 Erweiterung des Berlin Map Task Corpus
- Syntax gesprochener Sprache
- Lexikalische Variation
- Soziale Interaktion im Dialog, z.B.
  - Disfluencies, Häsitationen und Reparaturen
  - Backchanneling und Feedback

1.3 Das Korpus
- Gesprochenes Deutsch
- Besteht aus mehreren Subkorpora je nach Aufgabentyp
- Multimodal: Audio und Video der zeichnenden Hand bei der Map-Task-Aufgabe

2. Korpusdesign

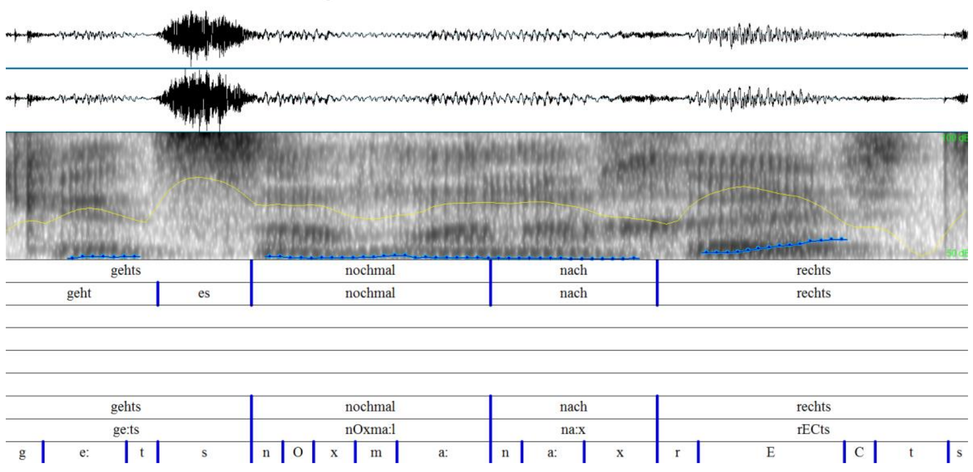
2.1 Aufnahmen
- Ausführliche Metadatenerhebung
  - Alter, Geschlecht, Größe etc.
  - Detaillierte Sprachbiographie
  - Höchster Bildungsabschluss
  - Anonymisierung
  - Persönlichkeitstest
- Schallisolierte Aufnahmekabine
- Aufnahmegerät und zwei separat platzierte Mikrophone bzw. zwei separate Kanäle



2.2 Aufgaben
- Einzelsprecheraufgaben
  - Geschichte vorlesen
  - Lange Wortliste (am Anfang und Ende)
  - Kurze Wortliste (zwischen allen Aufgaben)
- Interaktionsaufgaben
  - Map Task (Route erklären, siehe Abb. links)
  - Diapix-Bilder mit Stimuli (Unterschiede finden, siehe Abb. rechts)
  - Freies Gespräch
  - Synchronsprechen
  - Gegenseitige Imitation

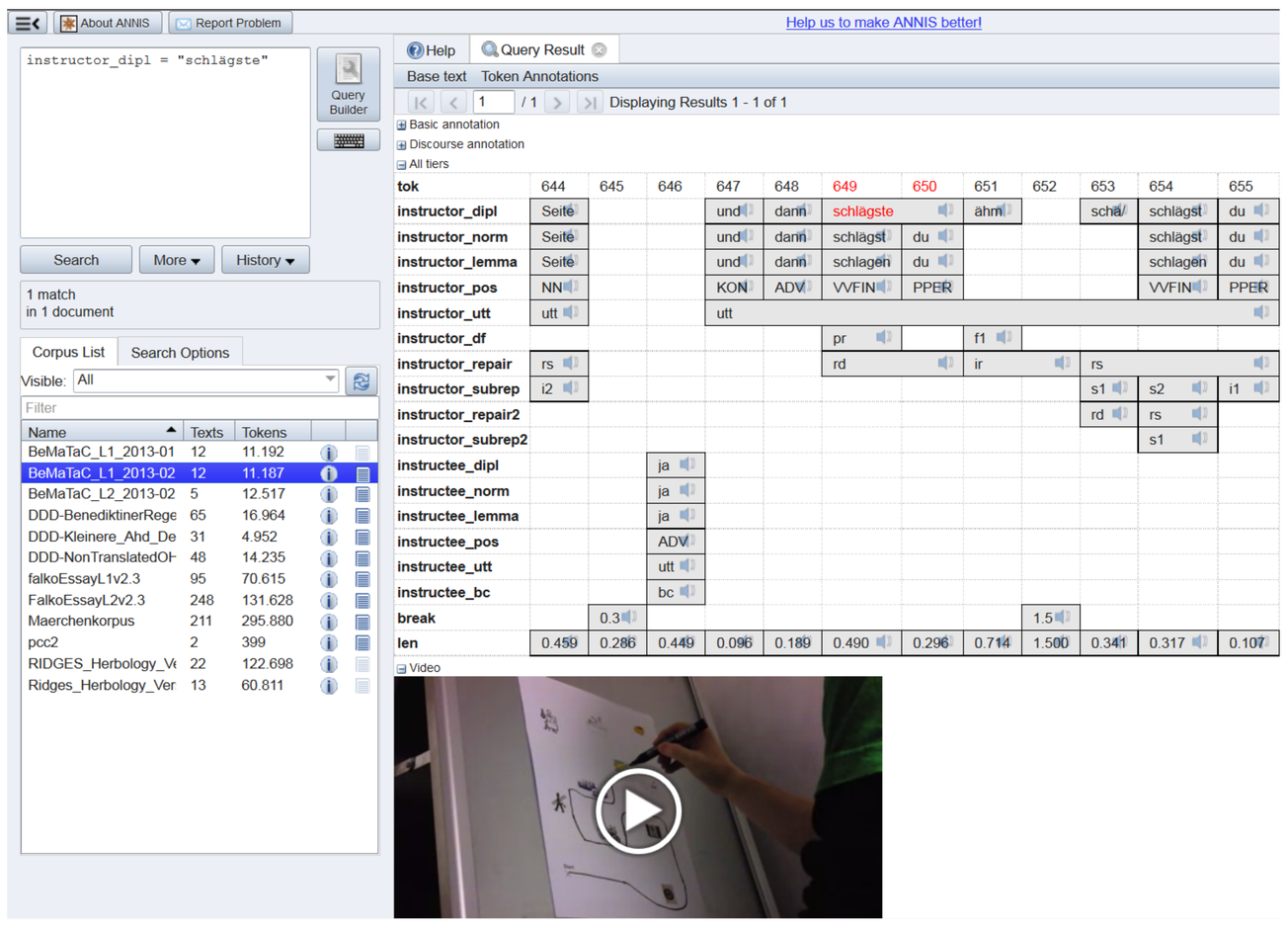
3. Aufbereitung

3.1 Transkription
- Festlegung von Transkriptionsrichtlinien
- Orthographienahe Transkription
- Automatische Verarbeitung mit MAUS
  - Segmentierung auf Laut- und Wortebene
  - Alignierung mit Audio/Video
- Phonetiksoftware Praat
  - Manuelle Korrektur der Alignierung
  - Separate Normalisierung nach amtlicher Rechtschreibung



3.2 Annotation im Rahmen des Berlin Map Task Corpus
- Automatische Verarbeitung mit TreeTagger
  - Lemmatisierung
  - Wortartenannotation nach dem STTS
- Annotationssoftware EXMARaLDA
  - Syntaktisch motivierte Äußerungsspannen
  - Disfluencies (Fillers, Wiederholungen, Dehnungen)
  - Reparaturen
  - Backchanneling

3.3 Veröffentlichung
- Umfassende Dokumentation aller Richtlinien und Vorgehensweisen
- Vereinigung aller Daten mit dem Konvertierungsframework SaltNPepper
- Veröffentlichung aller Daten zum Download
- Suche und Visualisierung in der browserbasierten Oberfläche ANNIS
- Darstellung als mehrere Subkorpora je nach Aufgabenstellung
- Suche über eine flexible knoten- und kantenbasierte Abfragesprache
- Frei verfügbar unter einer CC-BY-Lizenz



Referenzen

ANNIS: A. Zeldes, J. Ritz, A. Lüdeling & C. Chiarcos. 2009. ANNIS: A Search Tool for Multi-Layer Annotated Corpora. *Proceedings of Corpus Linguistics* 2009, July, 20–23. | **BeMaTaC**: http://u.hu-berlin.de/bematac | **DiaPix**: K. J. Van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, A. R. Bradlow. 2010. The Wildcat Corpus of Native- and Foreign-accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profiles. *Language and Speech* 53(4), 510-540. | **EXMARaLDA**: T. Schmidt & K. Wörner. 2009. EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics* (19:4), 565–582. | **Map-Task**: A. H. Anderson, M. Bader, E. Gurman Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson & R. Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech* 34, 351–366. | **MAUS**: F. Schiel, C. Draxler & J. Harrington. 2011. Phonemic Segmentation and Labelling using the MAUS Technique. Workshop *New Tools and Methods for Very-Large-Scale Phonetics Research*. University of Pennsylvania, 2011, January, 28–31. | **Praat**: P. Boersma. 2010. Praat, a system for doing phonetics by computer. *Glot International* 5 (9/10), 341–345. | **SaltNPepper**: F. Zipser & L. 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. | **STTS**: A. Schiller, S. Teufel, C. Stöckert, C. Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). | **TreeTagger**: H. Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*.