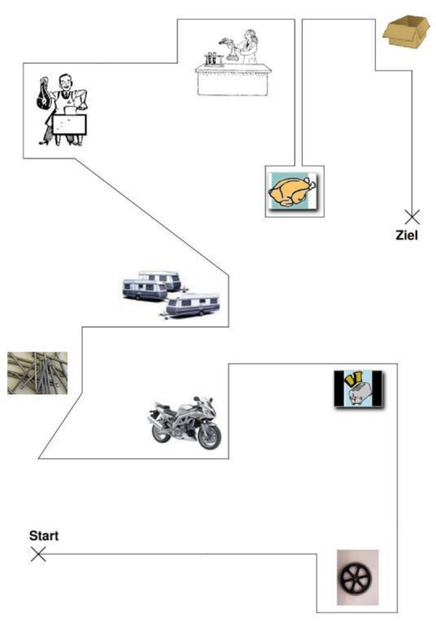
Berlin Map Task Corpus (BeMaTaC)
eine digitale multimodale Ressource für Sprach- und Dialogforschung

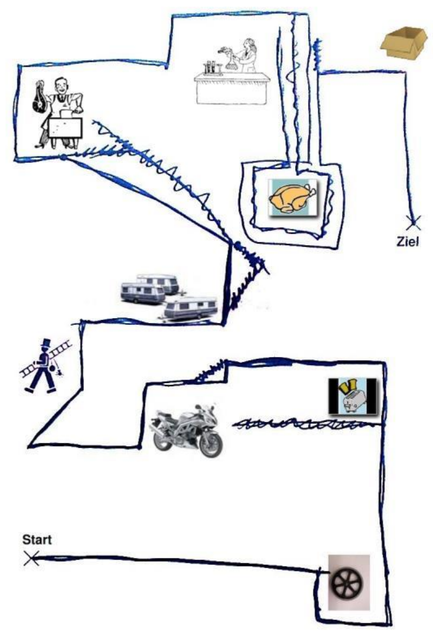
Simon Sauer und Oxana Rasskazova, Humboldt-Universität zu Berlin

Digital Humanities Berlin 28.02.2014 „Grenzen überschreiten – Digitale Geisteswissenschaft heute und morgen“

1. Das Korpus

- Gesprochenes Deutsch
- Subkorpora: Muttersprachler vs. Lerner
- Map Task: eine Person erklärt der anderen eine Route auf einer Karte mit Landmarken, die Route muss auf diese Weise reproduziert werden
- Setting erlaubt sowohl spontane konversationale Sprache als auch einen vergleichbaren kontrollierten Kontext
- Multimodal: Audio und Video der zeichnenden Hand
- Frei verfügbar unter CC-BY 3.0





2. Motivation

- Kontrastive Untersuchungen von Lerner- vs. Muttersprache
- Syntax gesprochener Sprache
- Lexikalische Variation
- Soziale Interaktion im Dialog, z.B.
  - Disfluencies, Hesitationen und Reparaturen
  - Backchanneling und Feedback
  - Räumliche Konzeptualisierung
  - Konvergenz

3. Von Aufnahme bis Veröffentlichung – Aufbau der Ressource

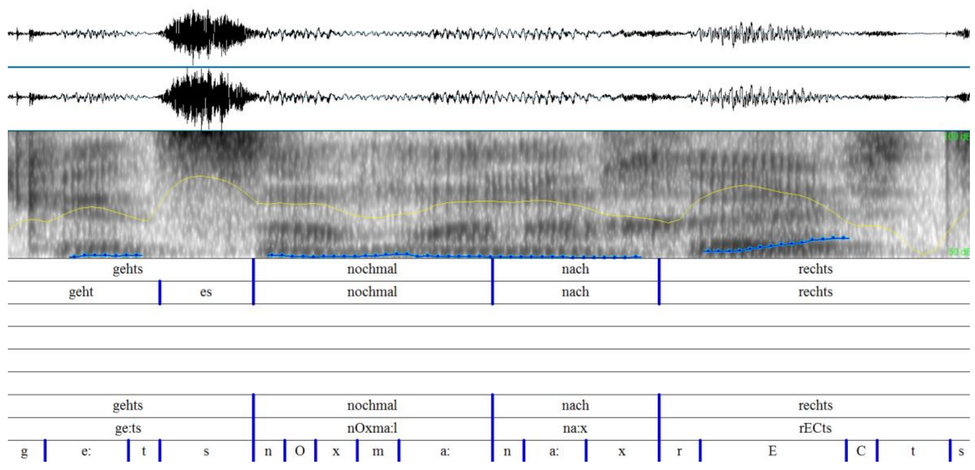
3.1 Aufnahmen

- Ausführliche Metadatenerhebung
  - Alter, Geschlecht, Größe etc.
  - Detaillierte Sprachbiographie
  - Höchster Bildungsabschluss
- Schallisoliertes Phonetiklabor
- Zwei separat platzierte Mikrophone
- Anonymisierung

3.2 Aufbereitung

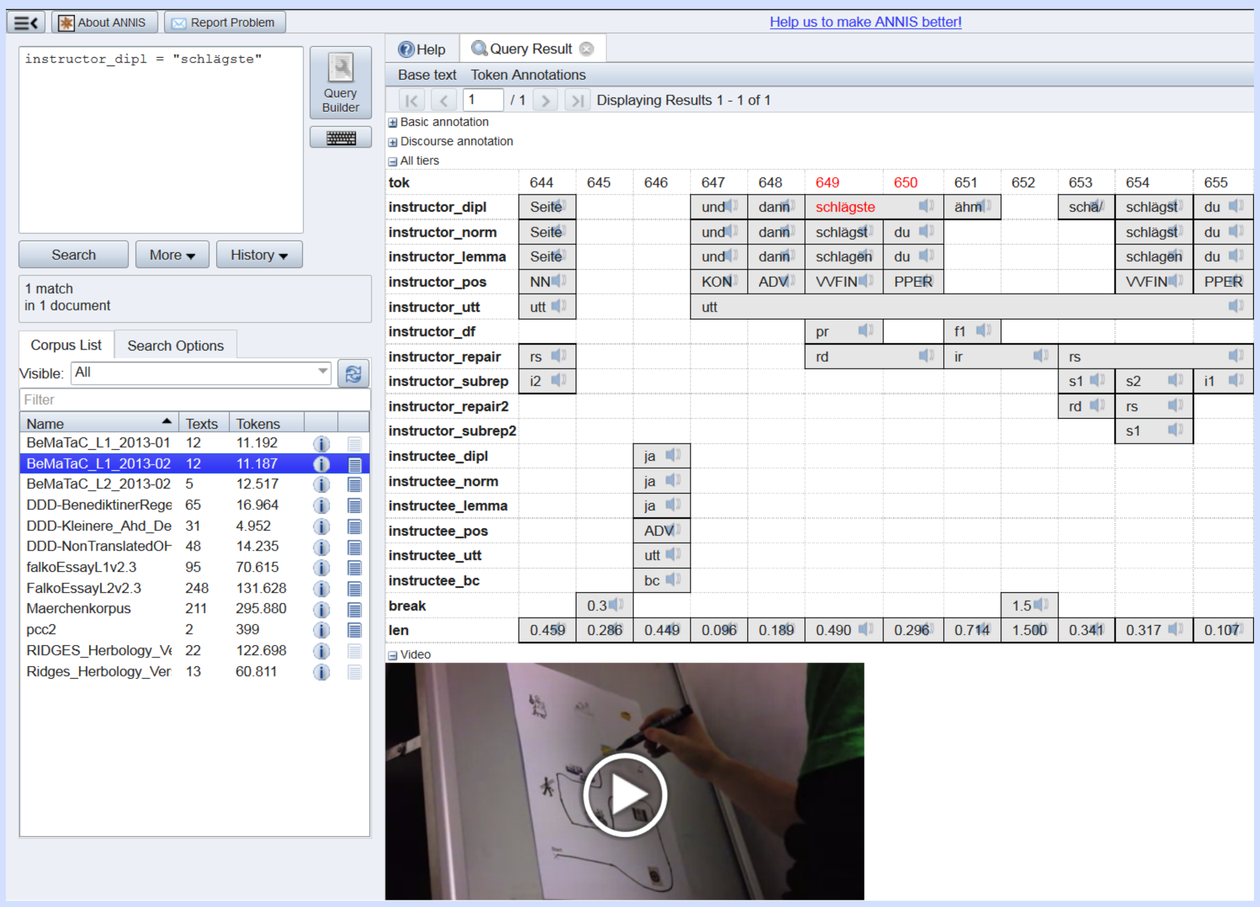
Transkription

- 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



3.3 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
- Abfrage über eine flexible Knoten- und Kanten-basierten Abfragesprache



Annotation

- Automatische Verarbeitung mit TreeTagger
  - Lemmatisierung
  - Wortartenannotation nach dem Stuttgart-Tübingen-TagSet (STTS)
- Annotationssoftware EXMARaLDA
  - Syntaktisch motivierte Äußerungsspannen
  - Disfluencies (Fillers, Wiederholungen, Dehnungen)
  - Reparaturen
  - Backchannelling

4. Institutionelle Anbindung

- Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin
  - Entwicklung, Forschung und Einsatz in Lehre lehrstuhlübergreifend und interdisziplinär mit dem Institut für Informatik
  - Gemeinsame Korpusinfrastruktur im Arbeitsbereich Korpuslinguistik und Morphologie: SaltNPepper und ANNIS
  - Kontrastive Untersuchungen mit den Korpora Falko (geschriebene Lerner- und Muttersprache) und RIDGES (historisches Deutsch)

5. Ausblick

- Erweiterung des Korpus durch zusätzliche dialogorientierte Aufgabenstellungen
- Verbesserung der systematischen Datenaufbereitung
- Verstärkte (Semi-)Automatisierung
- Phonetisch-/phonologische Transkription/Annotation

Referenzen

BeMaTaC: http://u.hu-berlin.de/bematac | **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. | **EXMARaLDA**: T. Schmidt & K. Wörner. 2009. EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics* (19:4), 565–582. | **Falko**: M. Reznicek, A. Lüdeling, C. Krummes, F. Schwantuschke, M. Walter, K. Schmidt, H. Hirschmann, T. Andreas. 2012. Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01. | **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. | **RIDGES**: http://korpling.german.hu-berlin.de/ridges/index_de.html | **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*.