Berlin Map Task Corpus
A deeply annotated multimodal map-task corpus of spoken learner and native German
Linda Giesel, Myriam Klapi, Daisy Krüger, Isabelle Nunberger, Oxana Rasskazova, Simon Sauer

Motivation
While the analysis of interlanguage is interesting on each particular linguistic level, it is even more valuable to study interactions across several levels. Proper analyses of second language acquisition can only be made by contrasting learner language with native speakers’ utterances. The Berlin Map Task Corpus (BeMaTaC) is a deeply annotated multimodal resource – with both learner and native speakers – enabling contrastive linguistic research in:
- language variation
- interactive task effects on speech rhythm and language adaptability
- syntactic studies to learner language
- informal spoken language construals
- disfluencies, hesitations and repair strategies
- backchanneling and feedback effects
- language architecture and lexical density
- interaction of prosodic features and information structure
- the role of extralinguistic phenomena in discourse and conversation analysis

Search & Visualization
- The freely extensible open source SoftMePepper converter framework [8] makes it possible to use many different dedicated annotation tools on the same data
- BeMaTaC can be accessed using ANNIS [9], an open-source browser-based search and visualization tool for deeply annotated corpora.

Design
- BeMaTaC uses a map-task design, where one speaker (the instructor) instructs another speaker (the instructee) to reproduce a route on a map with landmarks.
- video and audio are recorded with professional microphones in a soundproof environment
- 16 speakers are not able to look at each other
- original map-task design by HCRC [1], corpus design based on HAMATAc [2], maps courtesy of BDS Mannheim [3]

Metadata
- speaker acquisition
dialogue sequence
- native tongue
- foreign languages
- sex
- age
- height
- weight
- handiness
smoker
braces
pincings
language disorders
level of education
location of primary school

Data
- 8900 normalized tokens
- 66 mins of audio/video
- 12 dialogues
- 16 native speakers
- 8 female, 7 male
- age 20 to 50
We are currently working on recording further dialogues with learners and extending existing data.

Lemmatization & POS tagging
- automatic lemmatization and part-of-speech tagging with TreeTagger [6]
- using the STTS Stuttgart-Tübingen-Tagger [7]
- manually corrected lemmas and POS tags on additional tiers

Disfluencies
- Hesitations, repairs and a variety of other disfluency phenomena are important parts of spoken language – both for learners and native speakers. In BeMaTaC they are annotated on 4 different tiers:
  - repair scheme
  - repair subcategories
  - repetitions
  - insertions
  - substitutions
  - deletions
  - proformations
  - prolongations
  - slips of the tongue
  - truncations
  - repair
  - editing
  - terms
  - break annotation
  - unfilled pauses

Transcription & Normalization
- PRAAT: doing phonetics by computer [4]
- loosely orthographic transcription
- orthographic normalization: enables semi-automatic annotation and variant-independent search

Annotation
- flexible multi-layer standoff architecture
freely extendable: new annotation layers can be added at any point.

References: