# 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

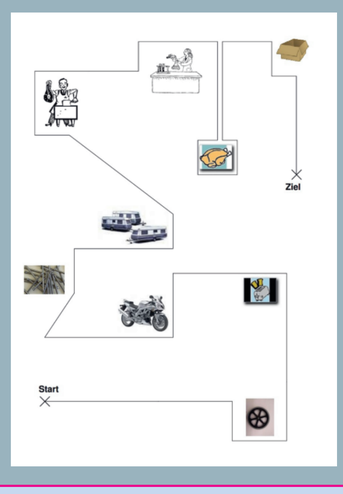
## Motivation

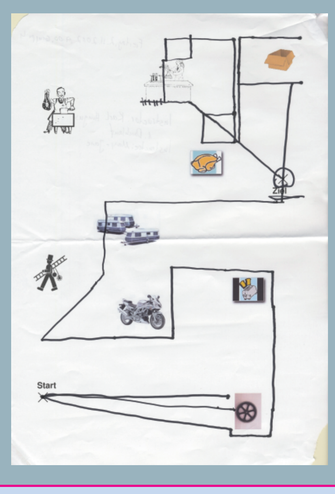
While the analysis of interlanguage is interesting on each particular linguistic level, it is even more valuable to study interactions across several levels. Proper analyses of second language acquisition can only be made by contrasting learner language with native speakers' utterances. The Berlin Map 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
- syncretic studies to learner language
- informal spoken language constructions
- 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

## 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
- 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 IDS Mannheim [3]



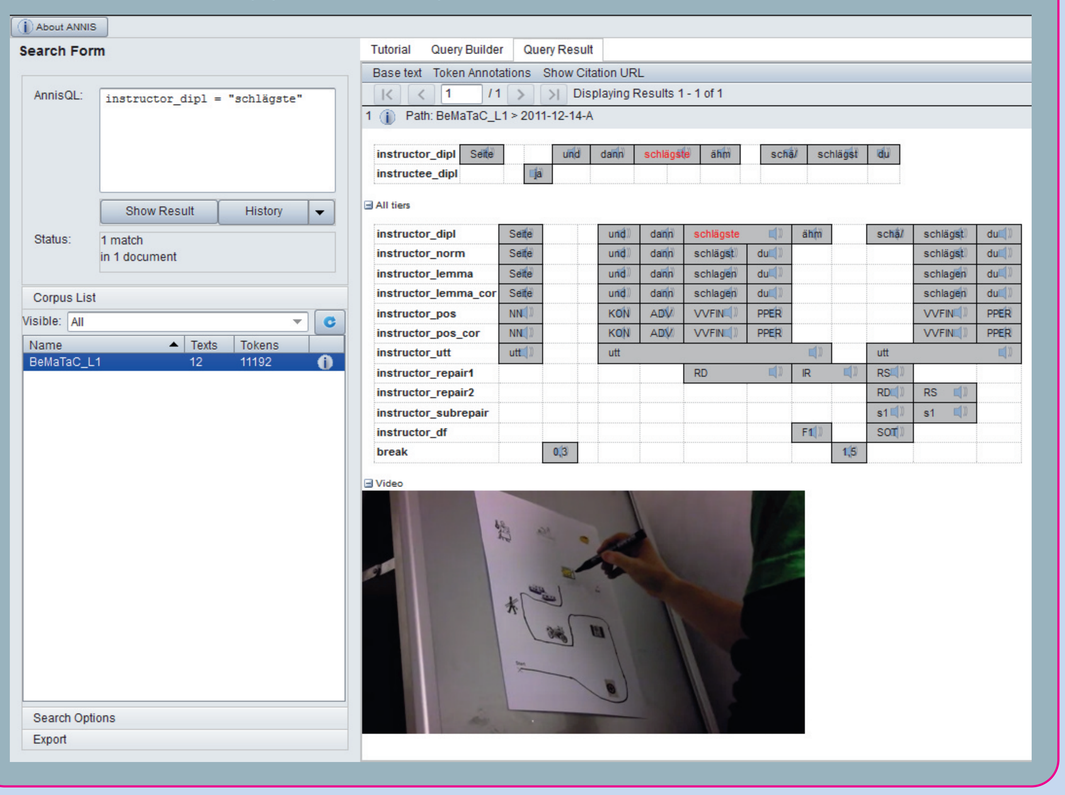




## Metadata

- speaker acquaintance
- dialogue sequence
- native tongue
- foreign languages
- sex
- age
- height
- weight
- handedness
- smoker
- braces
- piercings
- language disorders
- level of education
- location of primary school

## Search & Visualization

- The freely extendable open source SaltNPepper 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:



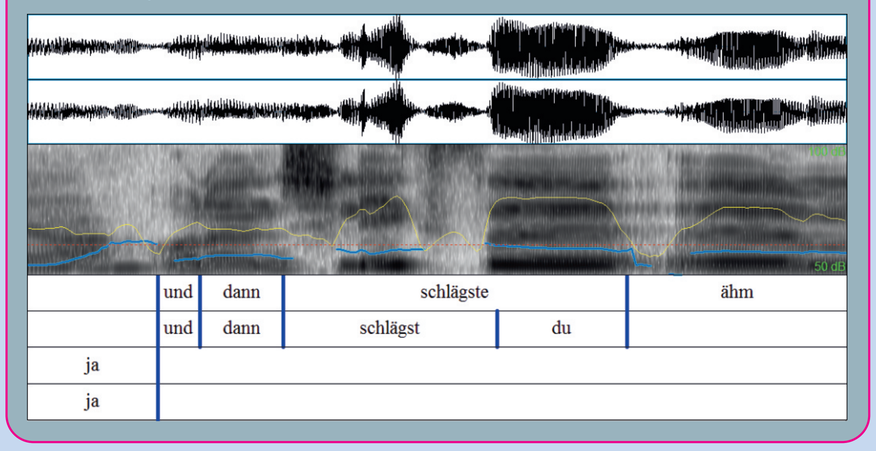

## Data

- 8900 normalized tokens
- 66 mins of audio/video
- 12 dialogues
- 16 native speakers
- 9 female, 7 male
- age 20 to 50

We are currently working on recording further dialogues with learners and extending existing data.

## Transcription & Normalization

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

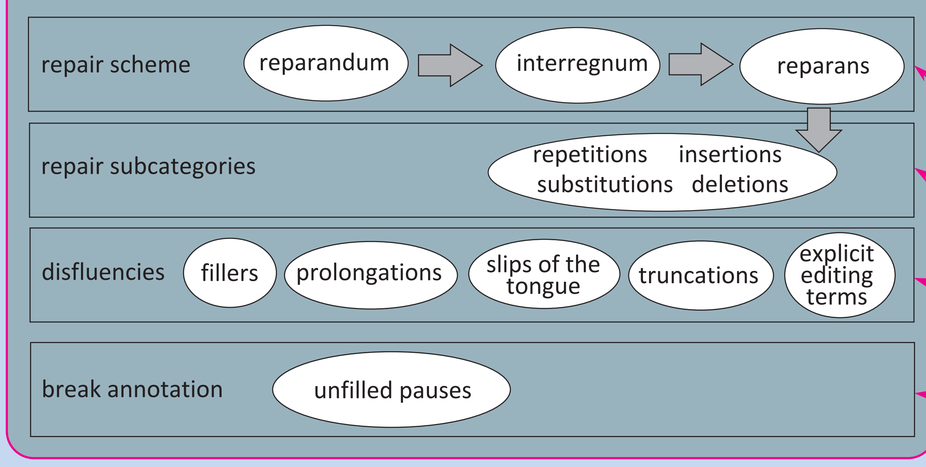



## Lemmatization & POS tagging

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

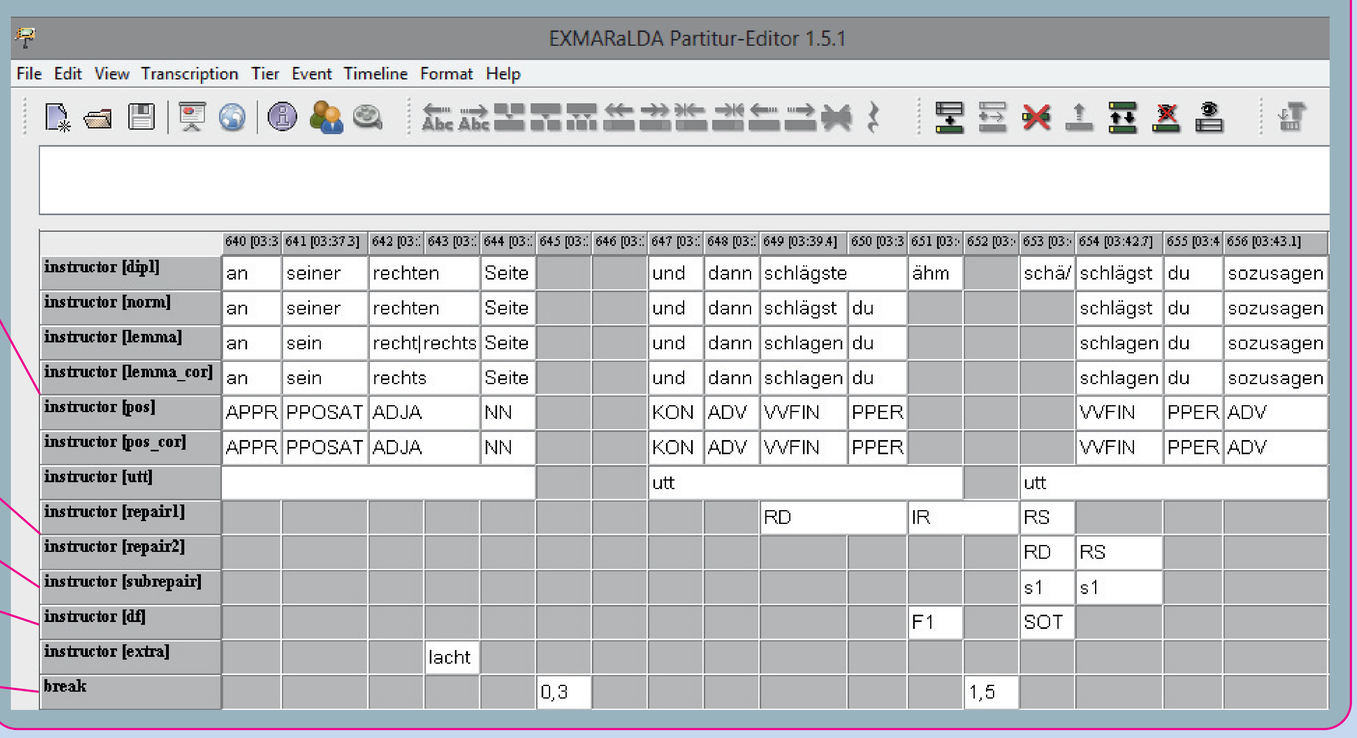
## Disfluencies

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:




## Annotation

- EXMARaLDA: Extensible Markup Language for Discourse Annotation [5]
- flexible multilayer standoff architecture
- freely extendable: new annotation layers can be added at any point




**References:**

**1.** Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, 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. **2.** Thomas Schmidt, Hanna Hedeland, Timm Lehmberg & Kai Wörner. 2010. HAMATAC – The Hamburg MapTask Corpus. **3.** Caren Brinckmann, Stefan Kleiner, Ralf Knöbl, Nina Berend. 2008. German Today: an areally extensive corpus of spoken Standard German. Proceedings 6th International Conference on Language Resources and Evaluation. LREC 2008. **4.** Paul Boersma. 2010. Praat, a system for doing phonetics by computer. Glot International 5 (9/10): 341-345. **5.** Thomas Schmidt & Kai Wörner. 2009. EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. Pragmatics (19:4), 565-582. **6.** Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing. **7.** Anne Schiller, Simone Teufel, Christine Thielen. 1995. Guidelines fuer das Tagging deutscher Textkorpora mit STTS. Technical Report, IMS-CL, University Stuttgart. **8.** Florian Zipser & 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. **9.** Amir Zeldes, Julia Ritz, Anke Lüdeling & Christian Chiarcos. 2009. ANNIS: A Search Tool for Multi-Layer Annotated Corpora. Proceedings of Corpus Linguistics 2009, July, 20-23. **Website:** http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/research/bematac