# Towards a Best Practice for sharing and reusing learner corpora: the Hamburg Map Task Corpus

Hanna Hedeland, Timm Lehmberg,
Thomas Schmidt, Kai Wörner

hzsk hamburger zentrum für sprachkorpora

# Overview

- A corpus for a different purpose
- The Hamburg MapTask Corpus
- What accounts for reusability
- Future steps

# A corpus for a different purpose

# Building a corpus to...

...investigate linguistic phenomena

- Information Structure
- Acquistion of Syntax
- Language/Dialectal Variation

# Building a corpus to...

...investigate linguistic phenomena (and maybe share it)

- Potsdam Commentary Corpus (Stede 2004)
- DUFDE* (Meisel 1994)
- SiN** (Elmentaler et al. 2006)

*Deutsch und Französisch - Doppelter Erstspracherwerb
**Sprachvariation in Norddeutschland

# Building a corpus to...

...build a corpus and share it?

- Reference Corpus for British English
- Dictionary of the German Language
- Corpus of contemporary spoken Dutch

# Building a corpus to...

...build a corpus and share it?

- BNC
- DWDS (Geyken 2007)
- CGN (http://lands.let.kun.nl/cgn/ehome.htm)

# Building a corpus to...

...share it and investigate aspects of reusability

# The Hamburg MapTask Corpus

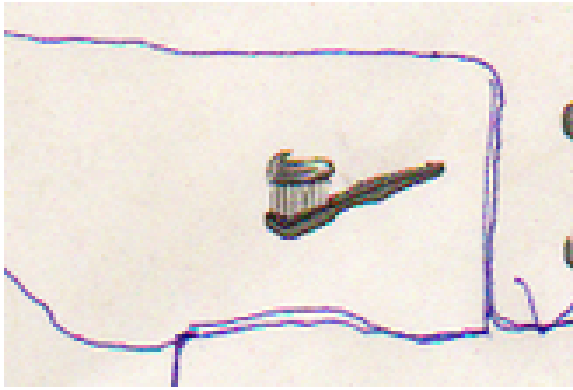

hzsk hamburger zentrum für sprachkorpora

# The Hamburg MapTask Corpus

Hamburg MapTask Corpus

http://www.exmaralda.org/files/maptask

## Hamburg MapTas



The Hamburg MapTask Corpus (HAMATAC)

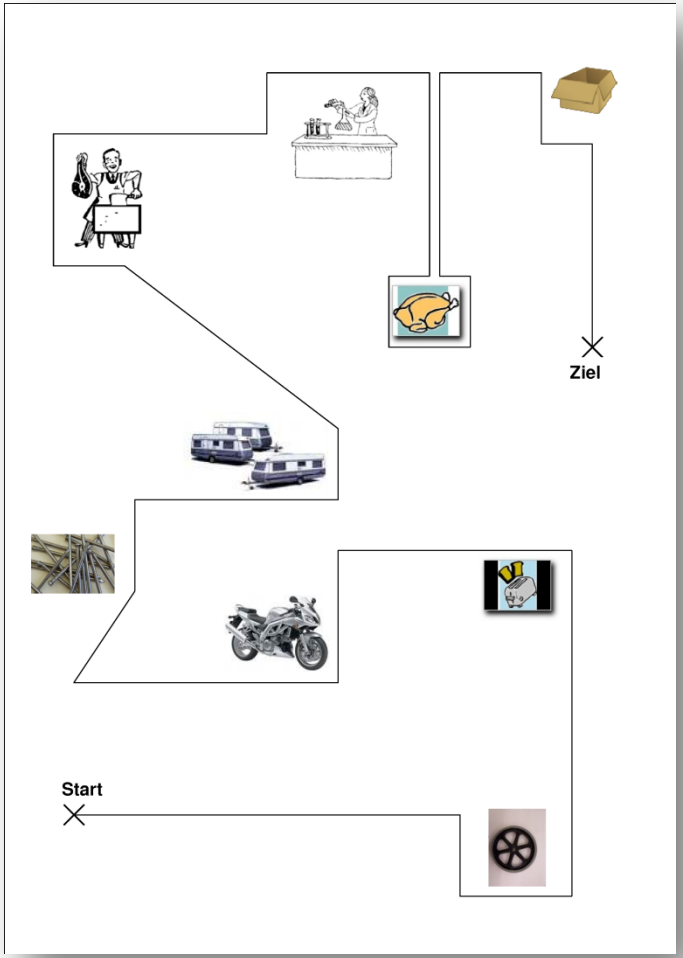
hzsk hamburger zentrum für sprachkorpora

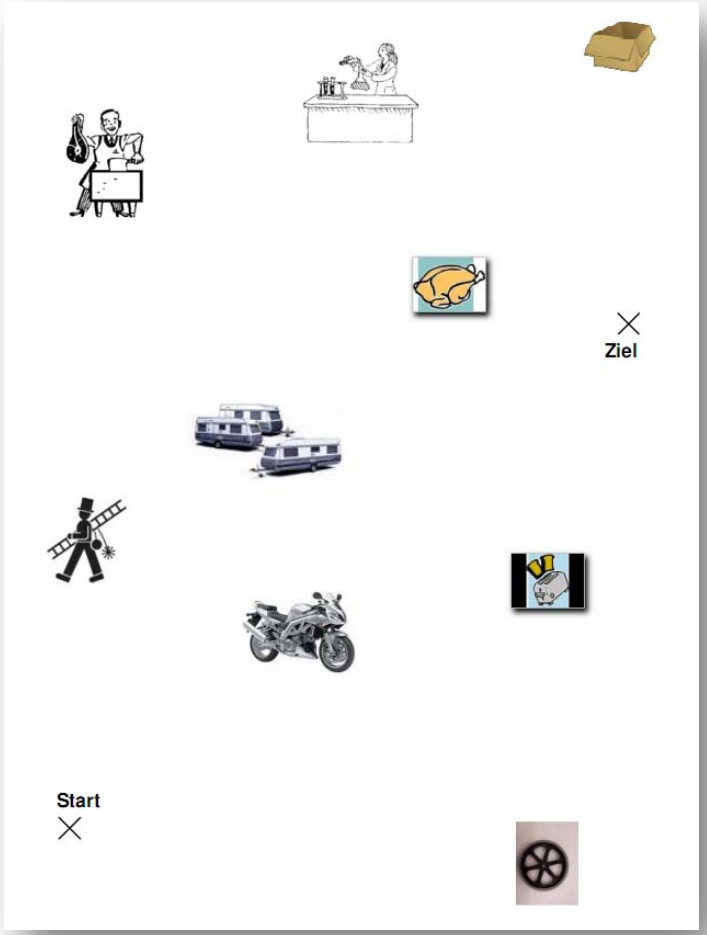
# Design

- Map Task (Edinburgh description):

  "The Map Task is a cooperative task involving two participants. The two speakers sit opposite one another and each has a map which the other cannot see. One speaker -- designated the **Instruction Giver** -- has a route marked on her map; the other speaker – the **Instruction Follower** -- has no route. The speakers are told that their goal is to reproduce the Instruction Giver's route on the Instruction Follower's map. The maps are not identical and the speakers are told this explicitly at the beginning of their first session. It is, however, up to them to discover how the two maps differ."
- Maps used from the "Deutsch Heute" Map Task Corpus

# Design



# Metadata

Universität Hamburg

SFB 538 „MEHRSPRACHIGKEIT"
TEILPROJEKT Z2
DR. THOMAS SCHMIDT
MAX-BRAUER-ALLEE 60
22765 HAMBURG

**Metadaten für die Map-Task**

1) Kennen sich die Map-Task-Teilnehmer?

ja ☐ nein ☐

2) Alter:

3) Geschlecht:

4) Seit wann leben Sie in Deutschland?

5) In welcher Region in Deutschland haben Sie Deutsch gelernt?

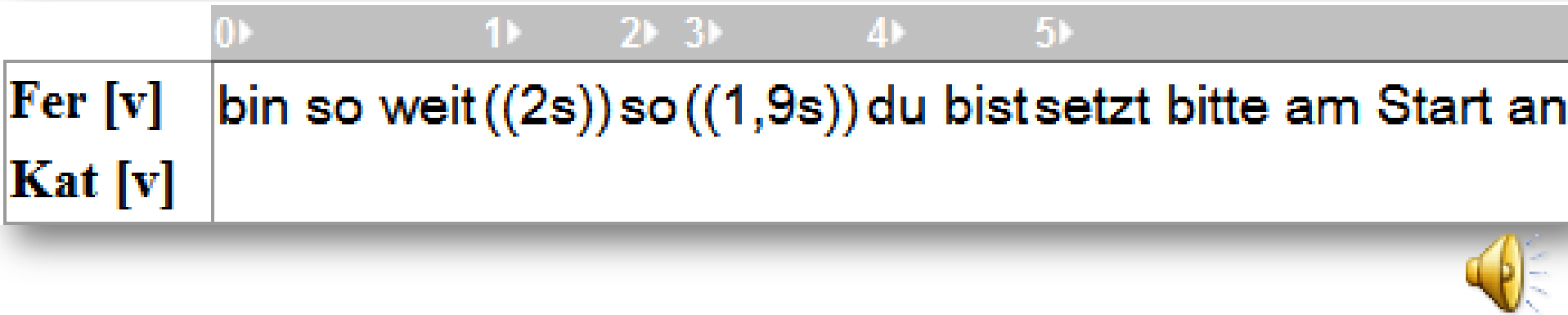
6) Welche Sprachen sprechen Sie seit wann und wie oft in Ihrem Alltag?

| Sprache / wann erlernt | | | | | |
|---|---|---|---|---|---|
| Spreche ich im Alltag ausschließlich | | | | | |
| hauptsächlich | | | | | |
| etwa zur Hälfte | | | | | |
| kaum | | | | | |

hzsk hamburger zentrum für sprachkorpora

# Transcription conventions

- Orthographically transcribed, no punctuation, Caps for nouns and proper names, but not at the beginning of turns.
- Only strong deviations are transcribed in „literary transcription“
- Non-phonological productions in double-round-brackets
- Cut-off word marked with a slash



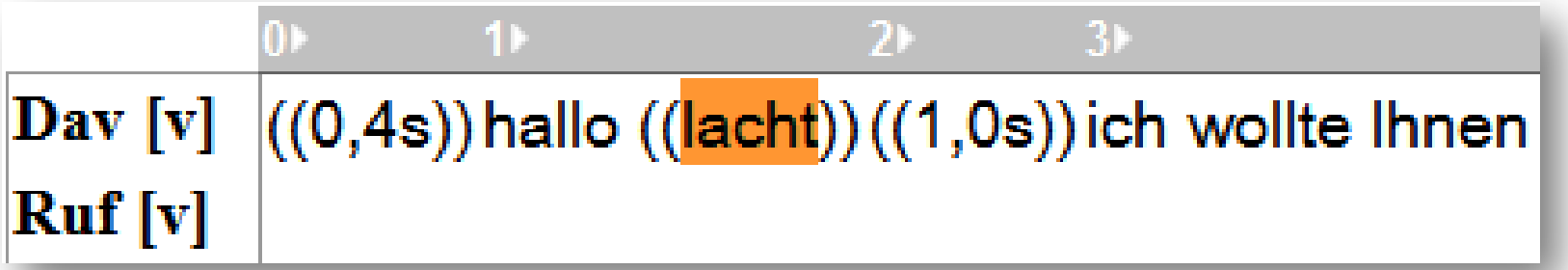
# Transcription conventions

- Orthographically transcribed, no punctuation, Caps for nouns and proper names, but not at the beginning of turns.
- Only strong deviations are transcribed in „literary transcription“
- Non-phonological productions in double-round-brackets
- Cut-off word marked with a slash

| |
|---|
| von hier aus ((0,2s)) einderthalb Zentimeter |

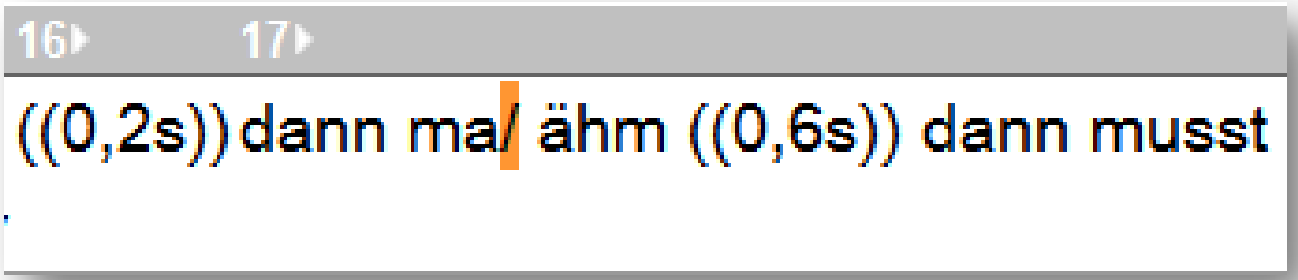
# Transcription conventions

- Orthographically transcribed, no punctuation, Caps for nouns and proper names, but not at the beginning of turns.
- Only strong deviations are transcribed in „literary transcription“
- Non-phonological productions in double-round-brackets
- Cut-off word marked with a slash



# Transcription conventions

- Orthographically transcribed, no punctuation, Caps for nouns and proper names, but not at the beginning of turns.
- Only strong deviations are transcribed in „literary transcription“
- Non-phonological productions in double-round-brackets
- Cut-off word marked with a slash



# Status of the Corpus

| Version | Release Date | Changes |
|---|---|---|
| 0.1 | 16 September 2010 | first version<br>orthographic transcriptions only |

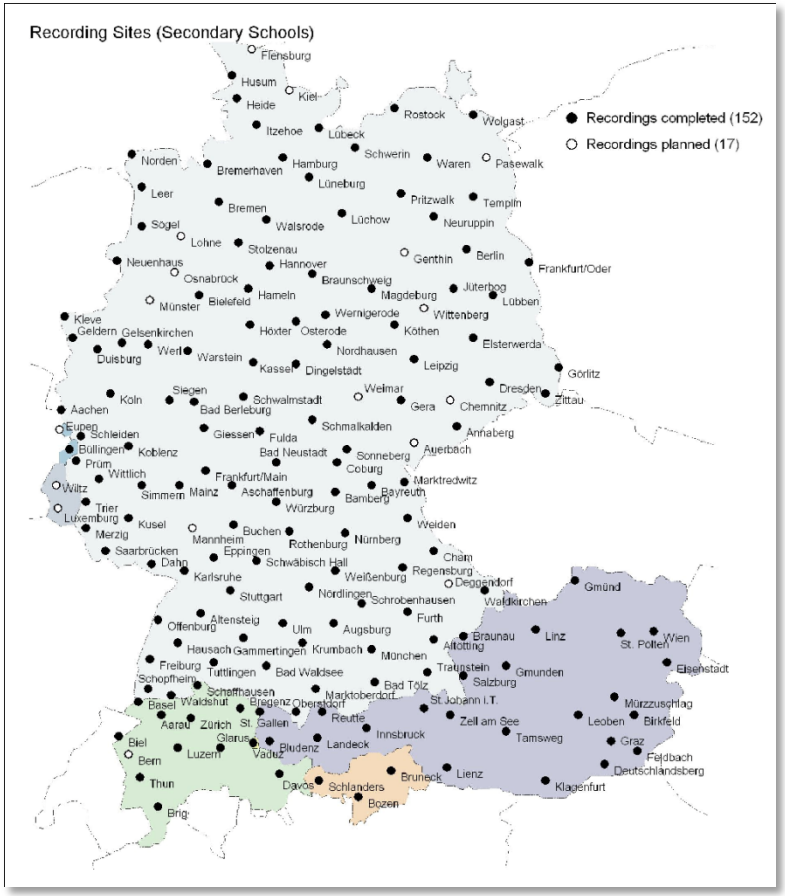
- Double-checked
- No annotations (yet)
- Released on the WWW

# What accounts for reusability

- Comparability and reproducibility of corpus design
- Data protection issues
- Theory-(in)dependence and quality of annotations
- Interoperability and standardisation of formats

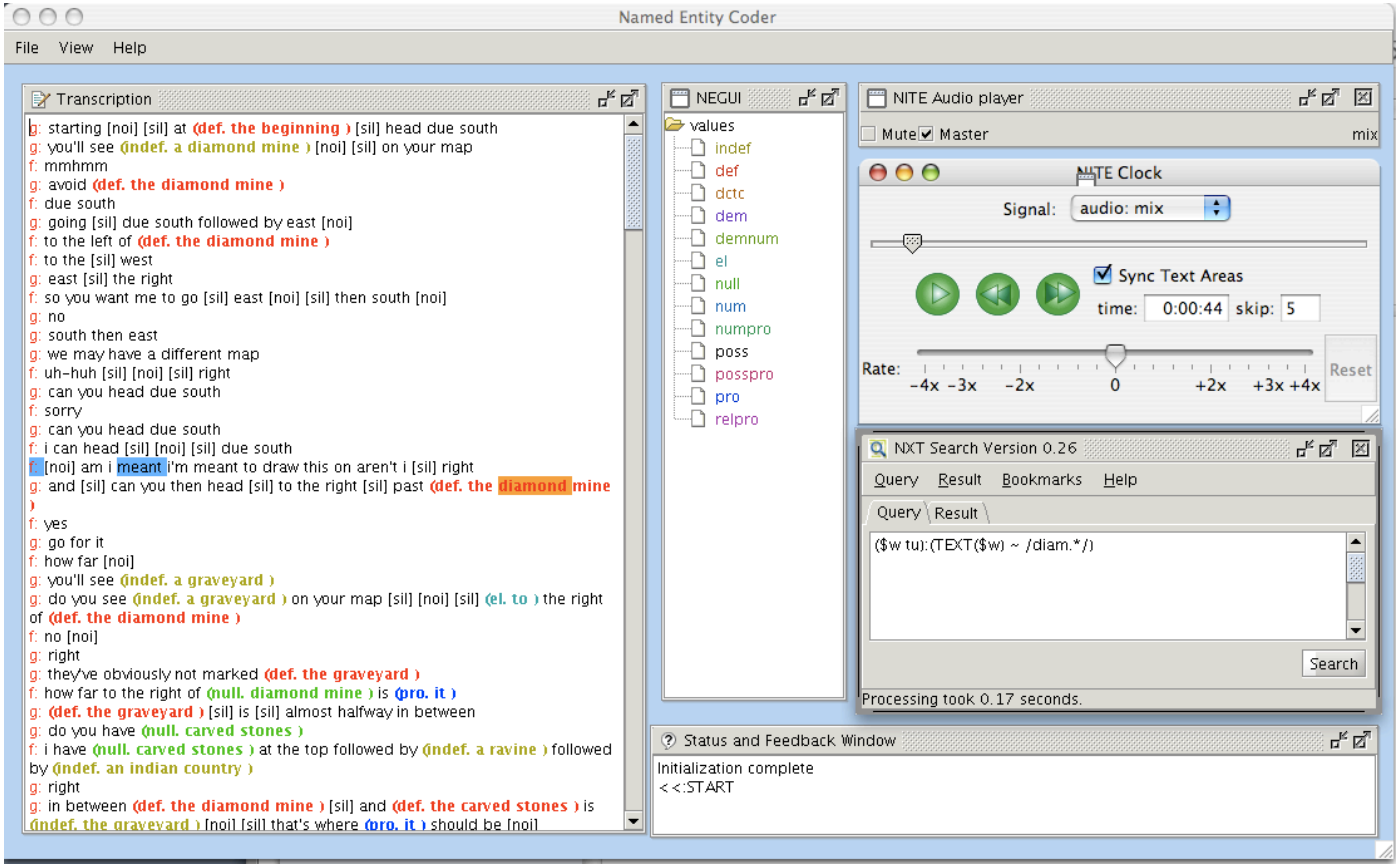
# Comparability and reproducibility of corpus design

„Deutsch Heute“ Map Task Corpus (Brinckmann et al. 2008)



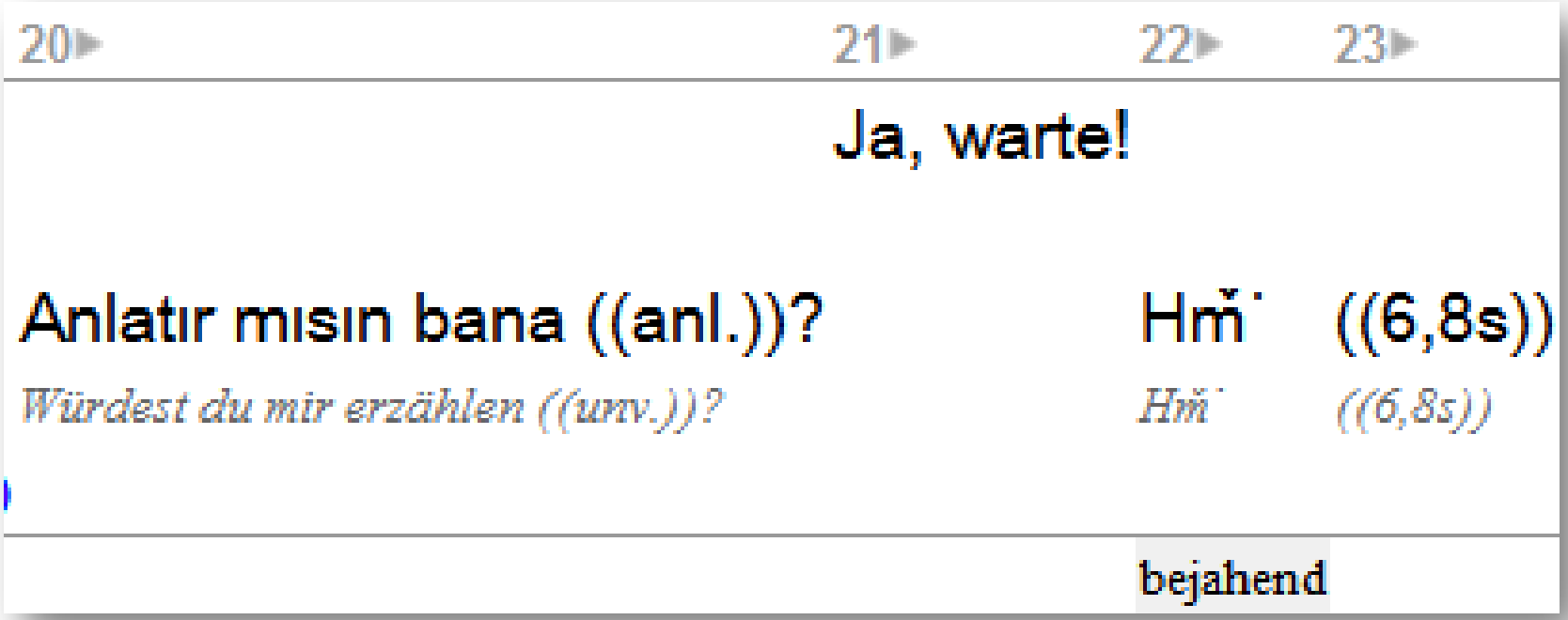
# Comparability and reproducibility of corpus design

HCRC Map Task Corpus Edinburgh (Anderson et al. 1991)



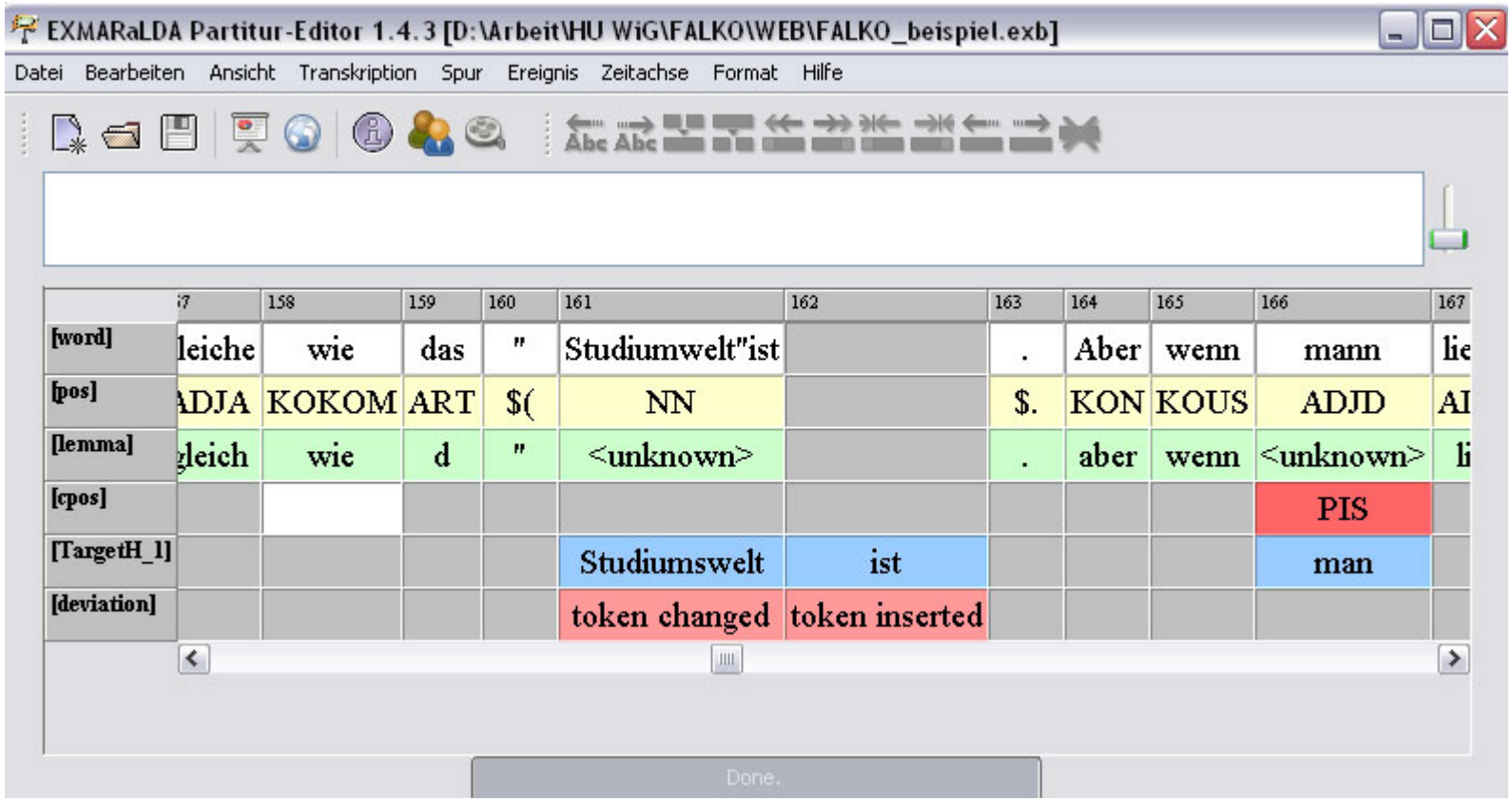
# Comparability and reproducibility of corpus design

SFB 538 (http://www.exmaralda.org/corpora/sfbkorpora.html)



# Comparability and reproducibility of corpus design

FALKO (Lüdeling et al. 2008)



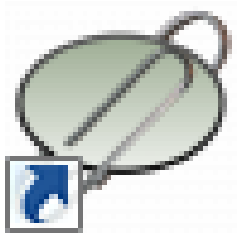

| | 157 | 158 | 159 | 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| [word] | leiche | wie | das | " | Studiumwelt"ist | | . | Aber | wenn | mann | lie |
| [pos] | ADJA | KOKOM | ART | $( | NN | | $. | KON | KOUS | ADJD | AI |
| [lemma] | gleich | wie | d | " | <unknown> | | . | aber | wenn | <unknown> | li |
| [cpos] | | | | | | | | | | PIS | |
| [TargetH_1] | | | | | Studiumswelt | ist | | | | man | |
| [deviation] | | | | | token changed | token inserted | | | | | |

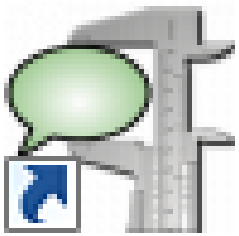
# Comparability and reproducibility of corpus design

**Documentation**

The following documentation is available:

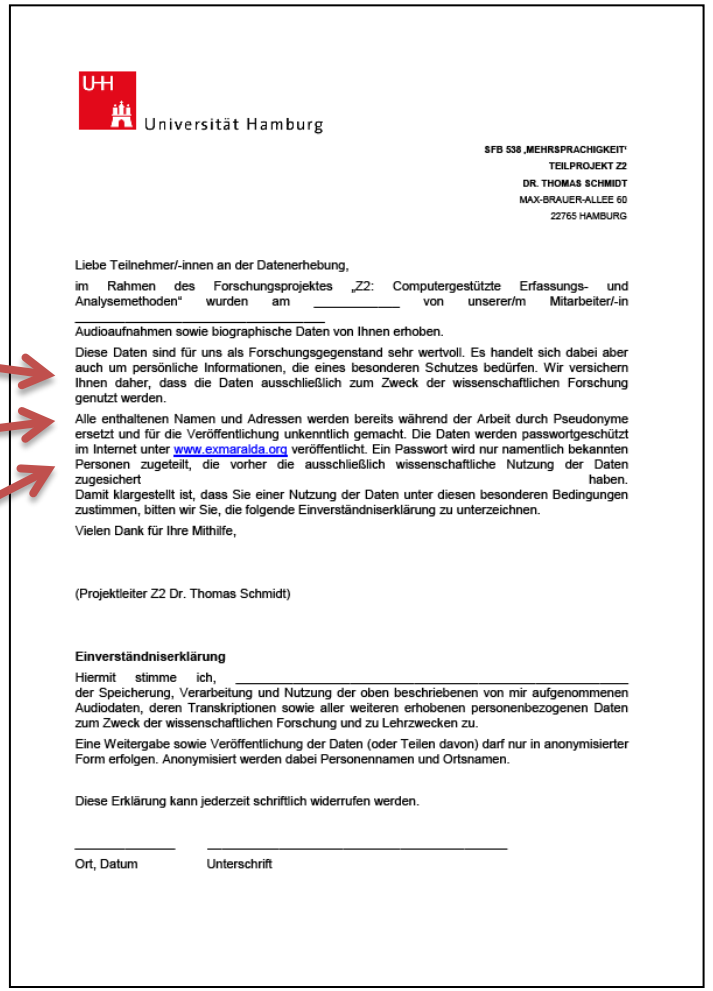
- A PDF document explaining the design and structure of the corpus.
- The maps used in the experiment: Map 1 [with path] [without path] / Map 2 [with path] [without path]
- A HIAT transcription manual explaining the conventions used for orthographic transcription in the corpus
- A PDF document explaining online and offline use of EXMARaLDA corpora




# Data protection issues

- Only for scientific research
- Anonymisation
- Limited number of users



Universität Hamburg

SFB 538 „MEHRSPRACHIGKEIT"
TEILPROJEKT Z2
DR. THOMAS SCHMIDT
MAX-BRAUER-ALLEE 60
22765 HAMBURG

Liebe Teilnehmer/-innen an der Datenerhebung,

im Rahmen des Forschungsprojektes „Z2: Computergestützte Erfassungs- und Analysemethoden" wurden am _____ von unserer/m Mitarbeiter/-in _____ Audioaufnahmen sowie biographische Daten von Ihnen erhoben.

Diese Daten sind für uns als Forschungsgegenstand sehr wertvoll. Es handelt sich dabei aber auch um persönliche Informationen, die eines besonderen Schutzes bedürfen. Wir versichern Ihnen daher, dass die Daten ausschließlich zum Zweck der wissenschaftlichen Forschung genutzt werden.

Alle enthaltenen Namen und Adressen werden bereits während der Arbeit durch Pseudonyme ersetzt und für die Veröffentlichung unkenntlich gemacht. Die Daten werden passwortgeschützt im Internet unter www.exmaralda.org veröffentlicht. Ein Passwort wird nur namentlich bekannten Personen zugeteilt, die vorher die ausschließlich wissenschaftliche Nutzung der Daten zugesichert haben.
Damit klargestellt ist, dass Sie einer Nutzung der Daten unter diesen besonderen Bedingungen zustimmen, bitten wir Sie, die folgende Einverständniserklärung zu unterzeichnen.

Vielen Dank für Ihre Mithilfe,

(Projektleiter Z2 Dr. Thomas Schmidt)

**Einverständniserklärung**

Hiermit stimme ich, _____
der Speicherung, Verarbeitung und Nutzung der oben beschriebenen von mir aufgenommenen Audiodaten, deren Transkriptionen sowie aller weiteren erhobenen personenbezogenen Daten zum Zweck der wissenschaftlichen Forschung und zu Lehrzwecken zu.

Eine Weitergabe sowie Veröffentlichung der Daten (oder Teilen davon) darf nur in anonymisierter Form erfolgen. Anonymisiert werden dabei Personennamen und Ortsnamen.

Diese Erklärung kann jederzeit schriftlich widerrufen werden.

_____ _____
Ort, Datum Unterschrift

# Theory-(in)dependence and quality of annotations

- Strong dependence on theory makes reusability difficult (e.g. adding „missing morphology“ in transcriptions)
- HaMaTaC uses a „less theory dependent“ way of transcription (e.g. transcription according to HIAT conventions, but no marking of utterances)
- „More theory dependent“ annotations will be added in a stand-off manner, so the user can decide whether to use them or not

# Interoperability and standardization of formats

**Map:** Shirin.BMP

**Recordings (4.653 minutes):** MT_270110_Shirin.mp3 MT_270110_Shirin.wav

**Transcription Shirin_Zhi_Zhi**
EXMARaLDA: [Transcription] [Segmented]
Visualisation: [Partitur] [RTF] [PDF] [XML] [Utterances] [Words]
Export: [TEI] [AG] [EAF] [Praat] [Chat] [FOLKER]

# Interoperability and standardization of formats

- Possible Platforms to publish the corpus
  - Talkbank
  - ANNIS
  - MPI

  - (IDS)
  - CLARIN

# Future improvements

- Plans for annotating the corpus for:
  - Disfluency phenomena
  - Errors (based on FALKO's error-annotation)
    - Morphosyntax
    - Lexicon
    - Phonology
    - Syntax

# References


- Brinckmann, Caren/Kleiner, Stefan/Knöbl, Ralf/Berend, Nina (2008): German Today: an areally extensive corpus of spoken Standard German. In: Proceedings 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakesch, Marokko.

- Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. and Weinert, R. (1991). The HCRC Map Task Corpus. Language and Speech,34, pp. 351-366.

- Isard, Amy (2001): An XML Architecture for the HCRC Map Task Corpus. In: Kühnlein, P.; Rieser, H. & Zeevat, H. (ed.): Proceedings of BI DIALOG 2001, Bielefeld.

- Lüdeling, Anke / Doolittle, Seanna / Hirschmann, Hagen / Schmidt, Karin / Walter, Maik (2008): Das Lernerkorpus Falko. In: Deutsch als Fremdsprache 2(2008), 67-73.

- Stede, M. (2004). The Potsdam Commentary Corpus. In Webber, Bonnie, Byron, Donna K. (Hrsg.): Association for Computational Linguistics (ACL) 2004 Workshop on Discourse Annotation, S. 96-102. Barcelona, Spain.

# References ctd.


- Meisel, Jürgen (Hrsg.) (1994): Bilingual First Language Acquistion: French and German Grammatical Development (Language Acquisition & Language Disorders). Amsterdam: John Benjamins
- Elmentaler, Michael/Gessinger, Joachim/Macha, Jürgen/Rosenberg, Peter/Schröder, Ingrid/Wirrer, Jan: Sprachvariation in Norddeutschland. Ein Projekt zur Analyse des sprachlichen Wandels in Norddeutschland. In: Osnabrücker Beiträge zur Sprachtheorie (OBST) 71 (2006) 159-178
- Geyken, Alexander (2007): The DWDS corpus: A reference corpus for the German language of the 20th century, in: Idioms and Collocations: Corpus-based Linguistic, Lexicographic Studies, Continuum Press