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

Hanna Hedeland, Timm Lehmberg,
Thomas Schmidt, Kai Wörner

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
- Acquisition 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>)

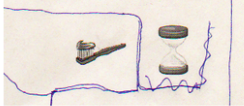
Building a corpus to...

...share it and investigate aspects of reusability

The Hamburg MapTask Corpus



Hamburg MapTask Corpus



The Hamburg MapTask Corpus (HAMATAC) is a spoken language corpus documenting the performance of 24 L2 learners of German in a map task. HAMATAC was recorded and transcribed in [project Z2](#) at the [Research Centre on Multilingualism](#). The current version 0.1 contains orthographic transcriptions of the recordings. We plan to add further annotations in future versions.

Terms of use

By using the HAMATAC corpus, you agree to:

- use it for non-commercial research and teaching purposes only
- not redistribute it to third parties
- cite the following source in any published work which is based on the corpus:

Thomas Schmidt, Hanna Hedeland, Timm Lehberg & Kai Wörner (2010): HAMATAC - The Hamburg MapTask Corpus. [<http://www.exmaralda.org/files/HAMATAC.pdf>]

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

Online data

The following data can be viewed online

- A [corpus overview](#) which links to all transcriptions, recordings, visualization and export documents
- A [corpus statistics](#) organised by communications
- A [corpus statistics](#) organised by speakers
- A [word list](#) for the whole corpus

Downloadable data

The following data can be downloaded for offline use:

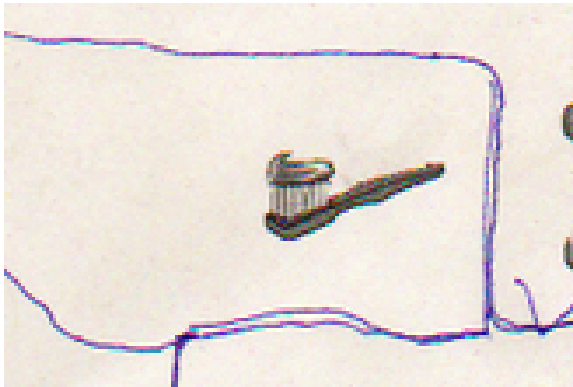
- A [zip archive](#) with all data in EXMARALDA formats (basic transcriptions, segmented transcriptions, Coma file)
- A [zip archive](#) with transcriptions in FOLKER format
- A [zip archive](#) with transcriptions in ELAN (*.eaf) format
- A [zip archive](#) with transcriptions in TEI format
- A [zip archive](#) with transcriptions in Praat TextGrid format

Audio recordings in MP3 and/or WAV format can be downloaded separately from the [corpus overview](#).

The Hamburg MapTask Corpus



Hamburg MapTask



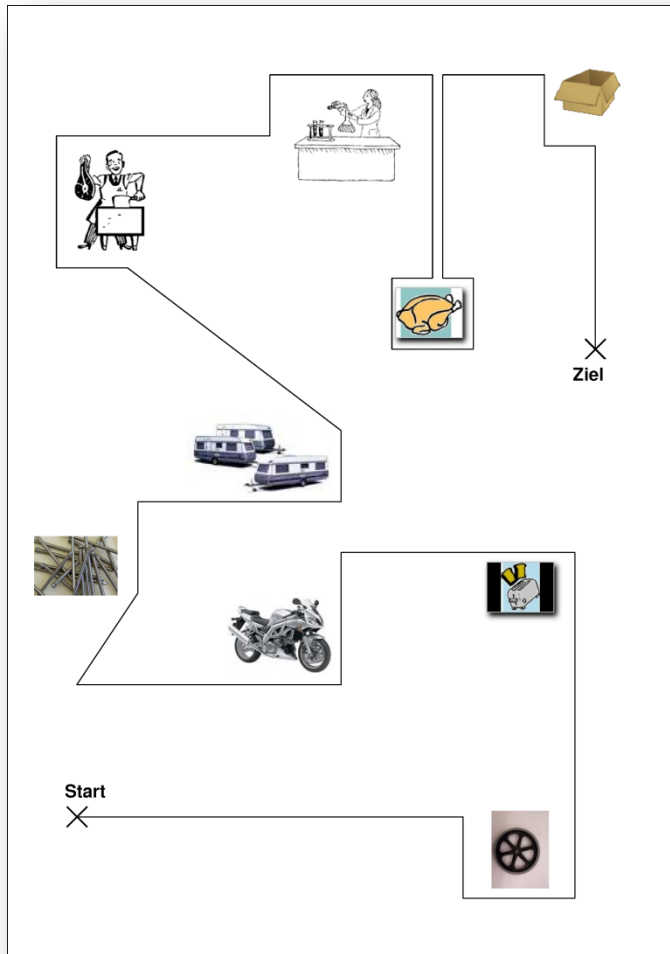
The Hamburg MapTask Corpus (HAMATAC) documenting the performance of 24 L2 learners

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					

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

	0▶	1▶	2▶	3▶	4▶	5▶
Fer [v]	bin so weit ((2s)) so ((1,9s)) du bist setzt bitte am Start an					
Kat [v]						



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

	0▶	1▶	2▶	3▶
Dav [v]	((0,4s))	hallo	((lacht))	((1,0s)) ich wollte Ihnen
Ruf [v]				



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

16▶ 17▶
((0,2s)) dann ma/ ähm ((0,6s)) dann musst



Status of the Corpus

Version	Release Date	Changes
0.1	16 September 2010	first version orthographic transcriptions only

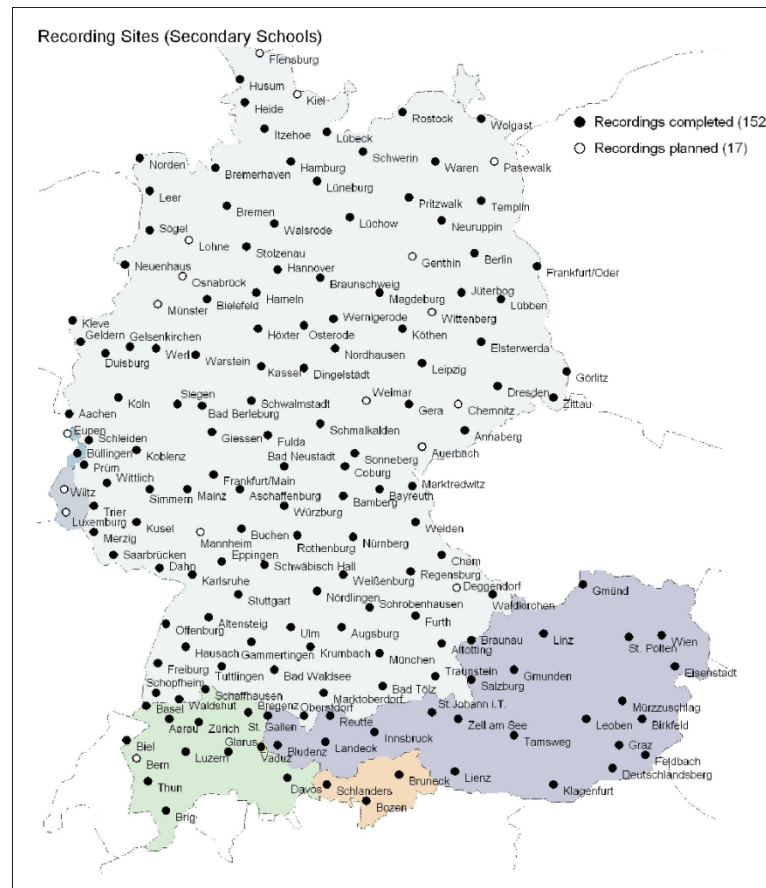
- Double-checked
- No annotations (yet)
- Released on the WWW

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)

The screenshot displays the Named Entity Coder (NEGUI) software interface. The main window is titled "Named Entity Coder" and contains several panes:

- Transcription:** A list of dialogue lines with named entities highlighted in red and green. For example, "starting [noi] [sil] at (def. the beginning) [sil] head due south".
- NEGUI:** A tree view showing the hierarchy of named entity classes, including "values", "indef", "def", "dctc", "dem", "demnum", "el", "null", "num", "numpro", "poss", "posspro", "pro", and "relpro".
- NITE Audio player:** A window for playing audio files, showing "Signal: audio: mix" and playback controls like play, stop, and volume.
- NITE Clock:** A window for time management, showing "time: 0:00:44" and "skip: 5".
- NXT Search Version 0.26:** A search interface with a query field containing the regex "(\$w tu):(TEXT(\$w) ~ /diam.*)".
- Status and Feedback Window:** A small window at the bottom showing "Initialization complete" and "<<:START".

Comparability and reproducibility of corpus design

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

20▶	21▶	22▶	23▶
	Ja, warte!		
Anlatır mısın bana ((anl.))?		Hm̃' ((6,8s))	
<i>Würdest du mir erzählen ((unv.))?</i>		<i>Hm̃' ((6,8s))</i>	
		bejahend	

Comparability and reproducibility of corpus design

FALKO (Lüdeling et al. 2008)



The screenshot shows the EXMARaLDA Partitur-Editor 1.4.3 interface. The main window displays a table with columns for word indices (157-167) and rows for linguistic annotations. The table content is as follows:

	157	158	159	160	161	162	163	164	165	166	167
[word]	gleiche	wie	das	"	Studiumswelt"	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					

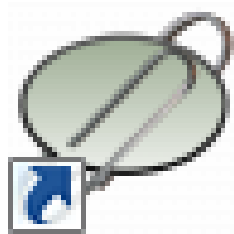
Below the table, there are navigation arrows and a 'Done.' button.

Comparability and **reproducibility** of corpus design

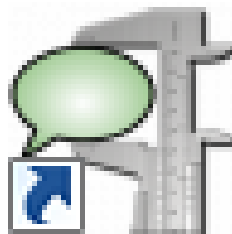
Documentation

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



Coma




EXAKT



Partitur-Editor

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
Hiernit 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: [Partiture] [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), 6773.
- 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 Acquisition: 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