REGISTER
Aspects of
Language
in Situation

Documentation and Annotation Guidelines of CoNNAR

Version 1

Megumi Terada[a], Bianca Sell[a], Robert Lange[a], Miriam Müller[a] & Malte Belz[a]

[a]Humboldt-Universität zu Berlin

This documentation describes the data collection process and annotation guidelines of the Corpus of Non-Native Addressee Register (CoNNAR).

Keywords: corpus, register, non-native, video call

1 The corpus

1.1 Summary and description

The Corpus of Non-Native Addressee Register (CoNNAR) is designed to investigate the intra-individual linguistic variation when addressing different interlocutors, namely German native speakers and learners of German as foreign language. The corpus contains two subcorpora: CoNNAR_videocall and CoNNAR_face-to-face.

This documentation focuses on the data collection and annotation process of CoNNAR_videocall which contains 40 conversations between 20 participants and 8 instructed interlocutors (see Section 1.4). The participants, the main interest of our study, repeated the same tasks twice: once with L1 confederates (native speakers of German) and once with L2 confederates (L2 German speakers, L1 English, self-reported German proficiency level on B1, B2 and C1[1]). Table 1 contains a summary of CoNNAR v. 1.

[1]The proficiency level is based on the *Common European Framework of Reference for Languages* (hereafter CEFR, Council of Europe 2001).

BerlinUP Journals

Megumi Terada and Bianca Sell and Robert Lange and Miriam Müller and Malte Belz. 2023. Documentation and Annotation Guidelines of CoNNAR — Version 1. *Register Aspects of Language in Situation (REALIS).* 2(1), pp. 1–32, doi=https://doi.org/10.18452/27898

ISSN 2750-9370

Table 1: Summary of CoNNAR v. 1.

Name	Corpus of Non-Native Addressee Register
Short name	CoNNAR
Version	1
Editors	Prof. Dr. Anke Lüdeling Prof. Dr. Christine Mooshammer Robert Lange, M. A. Bianca Sell, M. A. Megumi Terada, M. A.
Address	Institut für deutsche Sprache und Linguistik Humboldt-Universität zu Berlin Unter den Linden 6 10099 Berlin
Citation of this corpus	Lüdeling, Anke, Christine Mooshammer, Robert Lange, Bianca Sell & Megumi Terada. 2023. Corpus of Non-Native Addressee Register (CoNNAR): Version 1. Berlin: Humboldt-Universität zu Berlin. Medien-Repositorium.
Citation of this documentation	Terada, Megumi & Bianca Sell, Robert Lange, Miriam Müller & Malte Belz. 2023. Documentation and Annotation Guidelines of CoNNAR: Version 1. Register Aspects of Language in Situation (REALIS). 2(6), pp. 1–32, doi=https://doi.org/10.18452/27898
Access	Media repository of HU, https://rs.cms.hu-berlin.de/phon/
Annotators	Malte Belz, Robert Lange, Miriam Müller, Bianca Sell, Megumi Terada
Subcorpora	CoNNAR_videocall
Dialogues	40
Speakers	20 participants, 4 native confederates, and 4 non-native confederates
Tokens	171,511
Duration	34 hours
Language	German
Register	Task-free dialogues (free conversation) & task-based dialogues (Diapix)
Additional material	Word lists
Annotations	Diplomatic transliteration (`TRN`) Tokenisation (`ORT-MAU`) Phonetic transcription (`KAN-MAU`) Phonetic segmentation (`MAU`) Pseudonymization (`pseudo`) Orthographic normalisation (`norm`) Lemmatisation (`lemma`) Part-of-speech-tagging (`pos`) Grammatical information (`gram`) Schwa realisation (`schwareal`) Schwa duration (`schwamau`) Intonation phrases (`IP`) Corner vowels (`vowel`) Request for clear speech (`request`)

1.2 Research rationale

Speakers adapt their linguistic behaviour according to the demands of a communicative situation – they employ different registers (for an overview Lüdeling et al. 2022). One factor influencing register choice is the addressee and their linguistic background (Bell 1984, Hay et al. 1999) with non-native addressees eliciting *non-native addressee register* – NNAR, also called *foreigner talk* or *foreigner-directed speech* (Roche 1998, Bradlow & Bent 2002). For German, NNAR has been mainly investigated for non-native addressees with a low German proficiency, known as *Gastarbeiterdeutsch* 'guest worker German' since the late 1970s (Keim 1978, Hinnenkamp 1982, Roche 1998). However, in these studies it was impossible to differentiate between the addressees' proficiency on the one hand, and a low prestige of their native language and power imbalances on the other hand, possibly leading to a negative bias towards the addressee and influencing linguistic behaviour (cf. Schroedler et al. 2022). In addition, there are (to the best of our knowledge) no spoken corpora available that investigate NNAR in German.

CoNNAR was designed to fill this gap and provide data of participants conversing with non-native addressees with mid to high proficiency[2] in German, and with English as their native language. CoNNAR provides dialogues between German native speakers as a baseline condition for NNAR (see Section 1.3) as well as dialogues in different communicative tasks (see Section 1.6 for a wide variety of research questions).

1.3 General experiment design

In our experiment, we differentiated between the interlocutors as participants and confederates: 20 German native participants went through the experiment design twice – once with a German native (L1) confederate and once with a non-native (L2) confederate. These 20 German native participants were the main experimental targets of our study. This study design was intended to enable an intra-speaker comparison of the participant when talking with L1 and L2 speakers of German. Confederates are instructed interlocutors who participate in the experiment five times – each time with a different participant. There were four L1 confederates and four L2 confederates. Furthermore, participants and confederates were not known to each other prior to the experiment.

The corpus contains four rounds. In each round, an L1 and an L2 confederate conversed with the same five participants. The L1 and L2 confederates were

[2]The term 'mid proficiency' refers to B1/B2 and the 'high proficiency' to the C1 level on the CEFR in this documentation.

matched in age and gender in each round to avoid any bias based on these factors. To reduce the fatigue effect of the speakers, each confederate took part in the experiment five times. In sum, there were 40 sessions with 20 participants (20 sessions with L1 confederates and 20 with L2 confederates).

1.4 Interlocutors

Table 2 shows the metadata of the participants. All participants were native German speakers at an age between 20 and 38 without hearing and speech impairment nor reading and writing disability.

Table 2: Metadata of participants. 'Degree' refers to their most recent degree, 'Region' to the place where they grew up (see abbreviation below), 'Parent1' and 'Parent2' to the place of origin of their parents. 'English' indicates their self-perceived English proficiency level on a scale of 1 (foreign language) to 6 (mother tongue).

ID	Age	Sex	Degree	Region	Parent1	Parent2	English
p01	23	m	baccalaureate	BE	BE	BE	3
p02	30	f	university	BB	DE	DE	4
p03	27	m	university	BE	ST	BB	6
p04	27	f	university	BY	BY	BY	5
p05	33	m	university	BE	BE	BE	3
p06	21	f	baccalaureate	HB/NI	HB	NI	4
p07	32	f	university	MV	MV	TH	4
p08	24	f	university	BY	BY	BW	5
p09	27	m	baccalaureate	BE	BE	MV	2
p10	29	m	university	ST	ST	ST	5
p11	21	f	baccalaureate	HE	PL	PL	5
p12	22	m	baccalaureate	BE	TH	TH	3
p13	38	f	university	NI	NI	NI	5
p14	23	m	baccalaureate	BW	BW	HE	4
p15	24	f	baccalaureate	NI	NI	NI	3
p16	24	f	university	NW	NW	NW	5
p17	20	m	baccalaureate	BB	BB	MV	5
p18	28	f	university	NW	NW	BW	4
p19	22	m	baccalaureate	BE	BE	BE	3
p20	24	m	baccalaureate	MV	MV	MV	3

BB = Brandenburg, BE = Berlin, BW = Baden-Württemberg, BY = Bavaria, HB = Bremen, HE = Hesse, MV = Mecklenburg-Western Pomerania, NI = Lower Saxony, NW = North Rhine-Westphalia, ST = Saxony-Anhalt, TH = Thuringia | DE = Germany, PL = Poland

Table 3 shows the metadata of the confederates. All German native confederates originated from the northern part of Germany and spoke a standard variety

of German without any dialect features. Two L2-confederates were from the USA and two from England. This was due to availability of non-native confederates who matched the experimental conditions (age and proficiency level) and had time to participate in five recordings. For L2 confederates, their self-estimated German proficiency level was provided according to CEFR (Council of Europe 2001).

Table 3: Metadata of confederates

Pseudonym	Age	Sex	German	Degree	Region	Parent1	Parent2
c01	22	f	L1	baccalaureate	BB	BB	NW
c02	23	f	L2 (C1)	baccalaureate	US	US	US
c03	20	m	L1	baccalaureate	TH	TH	TH
c04	23	m	L2 (C1)	baccalaureate	UK	UK	UK
c05	21	f	L1	baccalaureate	BE	SH	BE
c06	24	f	L2 (B2)	university	US	US	US
c07	23	m	L1	baccalaureate	SN	SN	BE
c08	27	m	L2 (B1)	university	UK	UK	UK

BB = Brandenburg, BE = Berlin, NW = North Rhine-Westphalia, SH = Schleswig-Holstein, SN = Saxony, TH = Thuringia | US = United States of America, UK = United Kingdom

1.5 Setup and procedure specific to CoNNAR_videocall

Due to the COVID-19 pandemic, we conducted half of our experiments under different conditions than originally planned. This concerns the sessions with the participants p01 to p20, which constitutes the subcorpus CoNNAR_videocall. The recordings took place between October 2020 and July 2021.

At the beginning of the session, both the participant and the confederate were picked up individually at the entrance to the building. After filling out forms (see Section 1.9), they communicated with each other and with the first experimenter exclusively via the video conferencing software Zoom. The second experimenter was in charge of the recordings.

The participant was seated in a sound-attenuated booth in the lab and the confederate in an office next door. Highly sensitive omnidirectional and cardioid rod microphones were installed in each of the two rooms and connected to the recording computer via audio cable. A fanless tablet was placed in front of the speakers and connected via Zoom and circumaural headphones were connected to the tablets. Figure 1 shows the experimental setup.

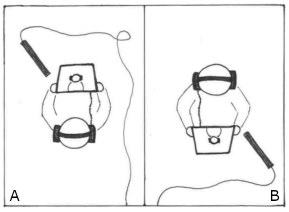

Figure 1: Experimental setup (Miriam Müller 2023, CC-BY 4.0). In the video-call setting (A and B), the speakers were seated in separate rooms and recorded with two rod microphones. Headphones were used to listen to the speech of the other interlocutor connected via Zoom on a tablet.

The audio data was recorded using open source software (Audacity Team 2020). A pre-amplifier was used to filter the audio signal: The gain was set to 0 db, the high-pass filter to 50 Hz, and the output to -6 db. The signals were recorded on separate channels and combined to stereo via an audio interface (Tascam US-2x2), which was connected to the recording computer. In accordance with the data protection regulations, video was not recorded.

The speakers were invited into the shared virtual room via Zoom, then a sound test was performed. They were sent into virtual breakout rooms and counted slowly and audibly from one to five. During this process, the settings on the high-resolution USB audio interface were fine-tuned. There were five tasks: 1) reading the first word list, 2) free conversation, 3) first Diapix task, 4) second Diapix task, and 5) reading the second word list. The recordings started with each speaker reading a word list, again in separate breakout rooms with only one speaker and the experimenter.

In the free conversation task, the speakers chose a topic by themselves after the experimenter suggested a few topics in a verbal instruction. The time limit was set to eight minutes and they were sent to a breakout room without the experimenter. The advantage of setting up breakout rooms was that only the speakers saw each other and the experimenter did not interfere.

Then the interlocutors came back to the Zoom room and were instructed to solve the first of the two Diapix tasks ('spot-the-difference' tasks, see Section 1.6.3). After a short break, the interlocutors completed the second Diapix task, in which the confederate was instructed beforehand to ask their partner to speak more clearly after 3–4 differences were found. Both Diapix tasks were solved in a breakout room without the experimenter.

Finally, both speakers read the second word list in the same setting described for the first word list. The participant filled out a questionnaire after their second

session and was paid (11 Euros per hour). The confederate was paid after the second session and after the fifth session.

The recordings only contain the communication between participants and confederates. The instruction of the experimenter at the beginning and the end of the recordings was excluded. The channels were separated into mono sounds, the left as channel 1 (confederate) and the right as channel 2 (participant). All audio data were stored in an encrypted Veracrypt container on an online cloud of the Humboldt-Universität zu Berlin (HU-Box). The encrypted container was used in compliance with data protection regulations, as the acoustic data included not yet anonymised information.

In short, each session consisted of the following steps:

- filling out of metadata and consent forms prior to the experiment
- arrival of participant and confederate: checking of Covid-19 tests and seating of both interlocutors (5 min)
- filling out Covid-19 contact forms (2 min)
- participant: testing of audio, first reading of the word list[3] (3–5 min)
- confederate: testing of audio, first reading of the word list (3–5 min)
- free conversation (8 min)
- diapix 1 (8 min, participant: version A, confederate: version B)
- diapix 2 (8 min, participant: version A, confederate: version B, clear speech request after finding 3–4 differences)
- participant: second reading of the word list (3–5 min)
- confederate: second reading of the word list (3–5 min)
- participants: filling out of a post-experiment questionnaire after their second session (5 min)
- paying participants after their second session, confederates after their third and fifth session

[3]Every testing of audio as well as readings of word list were carried out in absence of the other interlocutor.

1.6 Materials

The main tasks in this experiment were free conversations and two Diapixes that will be explained in further detail below. Additionally, each speaker had to read a word list before and after the main part of the experiment.

1.6.1 Word lists

Table 4 contains the words as given in the word list. They include all 15 monophthongs in German in the stressed syllable and the two reduced vowels in the last, unstressed syllable. These partly low-frequency words were chosen to avoid pseudowords. One restriction for finding words was that stop consonants needed to surround the stressed vowel in order to make the segmentation easier. To create some variation for the participants and to create some natural variation, the place of articulation of the medial consonant was grouped into three categories: bilabial, alveolar, and velar. The first consonant was allowed to vary freely, so that actual words could be chosen. These restrictions resulted in some low frequency words.

Table 4: Word lists. All words are included in the carrier sentence *Sage X bitte* 'Say X please'. The primary meaning of the following words is given in English below the German.

Place of articulation of the final syllable's onset	bilabial		alveolar		velar	
Ending with [ə]	piepe 'beep.SG.IMP'	[iː]	Güte 'kindness'	[yː]	bücke 'bend.down.1SG'	[ʏ]
	Bube 'boy'	[uː]	Beete 'flower.bed.PL'	[eː]	Böcke 'goat.PL'	[œ]
	Kippe 'cigarette'	[ɪ]	böte 'offer.3SG.SBJV'	[øː]	Pocke 'smallpox.SG'	[ɔ]
	Puppe 'doll'	[ʊ]	Bote 'messenger'	[oː]	Tage 'day.PL'	[aː]
			bäte 'ask.3SG.SBJV'	[ɛː]	packe 'pack.SG.IMP'	[a]
			Kette 'chain'	[ɛ]		
Ending with [ɐ]	Pieper 'beeper'	[iː]	Puder 'powder'	[uː]	Kaper 'caper'	[aː]

1.6.2 Task-free dialogue: free conversation

In the free conversation, the speakers were asked to converse with each other for eight minutes on a topic of their choice. In the verbal instruction, some topics were suggested such as the food at the university cafeteria, ideal vacation, or living in Berlin. To avoid any observer effect, both the participant and the confederate were sent to a breakout room in absence of the experimenters.

1.6.3 Task-based dialogue: Diapix

In the Diapix task, the speakers were supposd to find as many of the 12 differences between the pictures as possible in 8 minutes. The two versions of the same picture (Diapix A and Diapix B) include slight differences and are used to elicit tense vowels (see Figure 2). The Diapixes were translated into German (see Bullock Oliveira & Sell 2022) based on Baker & Hazan (2011). The original version in English can be found in Hazan & Baker (2011).

For the recordings, the following pictures were used: Beach 1–4, Farm 1–4, and Street 1–2 (Bullock Oliveira & Sell 2022). For each session, the pictures were exchanged. Therefore, each speaker never received a picture twice. The speakers were instructed not to show the pictures to each other, only to communicate verbally, and to refrain from taking notes. This task was done twice, Diapix 1 and Diapix 2 respectively.

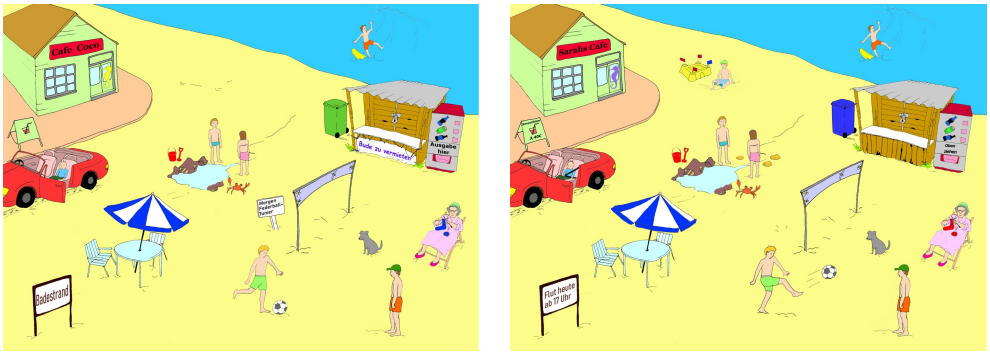

Figure 2: Example of a pair of Diapix: version A on the left and version B on the right side (Bullock Oliveira & Sell 2022).

During the recordings of the second Diapix task, the confederates were instructed to ask for clear speech after finding 3–4 differences (in the middle of the conversation). However, several confederates failed to recall this instruction at the right moment, or failed to recall it altogether (see Table 1.6.3).

Table 5: Table of clarification request during the second Diapix task. If the confederates forgot to ask for clarification or if they used an indirect word, it is marked as such. In the case of a successful request for clarification, it is marked with an 'x'.

Participants	L1 confederates	L2 confederates
Round 1	c01	c02
p01	x	x
p02	x	x
p03	x	x
p04	x	x
p05	x	x
Round 2	c03	c04
p06	forgot	forgot
p07	forgot	x
p08	x	x
p09	x	x
p10	x	x
Round 3	c05	c06
p11	x	forgot
p12	forgot	x
p13	x	x
p14	x	x
p15	x	x
Round 4	c07	c08
p16	x	forgot
p17	x	indirect wording
p18	x	x
p19	x	indirect wording
p20	x	indirect wording

1.7 Data processing

1.7.1 Filename convention

The following convention was used for the filenames: Code for participant, code for confederates, the German language status of confederates (L1 or L2 speakers), proficiency level if the confederate is an L2 speaker (B1, B2, or C1), task (word list 1/2, free dialogue, or Diapix 1/2), and the channel (ch1: confederates, ch2: participants).

For example, p01c02_L2C1_free_ch1.wav stands for the audio signal of the

channel 1 (confederate) of a free conversation between p01 who is a native participant and c02 who is an L2 confederate on C1 level on CEFR.

1.7.2 Audio post-processing and pseudonymization

The recordings only contain the communication between participants and confederates, as the instruction of the experimenter at the beginning and the end of the recordings was excluded. After cutting the audio data, the audio tracks were separated and stored.

In the pseudonymization process, words indicating places of residence/origin and proper names were marked on an annotation tier (cf. Section 3.5). Afterwards, an R script (Belz 2019) was used to pseudonymize the relevant words.

1.7.3 Annotation overview

Table 6 summarises the available annotations for each register and will be described in detail in Section 3. It should be noted that an empty tier `vowel` is included in the free conversation and an empty tier `request` in the Diapix 1 and the free conversation.

Table 6: Summary of annotations in CoNNAR_videocall. All annotation tiers are explained in detail in Section 3.

Annotation tier	Available in	Missing for
`TRN`	Free conversation, Diapix 1+2	
`ORT-MAU`	Free conversation, Diapix 1+2, word lists	
`kan`	Free conversation, Diapix 1+2, word lists	
`mau`	Free conversation, Diapix 1+2, word lists	
`pseudo`	Free conversation	
`norm`	Free conversation, Diapix 1+2	
`lem`	Free conversation, Diapix 1+2	Confederates
`pos`	Free conversation, Diapix 1+2	Confederates
`gram`	Free conversation, Diapix 1+2	Confederates
`schwareal`	Free conversation, Diapix 1+2	Confederates
`schwamau`	Free conversation, Diapix 1+2	Confederates
`IP`	Free conversation, Diapix 1+2	Confederates
`vowel`	Diapix 1+2, word lists	
`request`	Diapix 2	

1.8 Access

The corpus is available via the media repository of Humboldt-Universität zu Berlin (https://rs.cms.hu-berlin.de/phon/). To access the corpora licensed for scientific purposes there, you can use the following options.

- Apply for an account. To do this, send an email to phonetik-labor.german@hu-berlin.de with your name, affiliation, and research purpose.
- Get a link to download. To do this, write an email to phonetik-labor.german@hu-berlin.de with the corpus you are interested in, your name, your affiliation, and your research purpose.

1.9 Consent form

Each speaker agreed to all the terms and conditions of the consent form, which could be ticked off individually. The original wording in German and English translations can be found in the appendix.

1.10 Planned additional subcorpus: CoNNAR_face-to-face

The Corpus of Non-Native Addressee Register (CoNNAR) is divided into two subcorpora: CoNNAR_videocall and CoNNAR_face-to-face. These subcorpora differ in the channel used by the interlocutors to interact with each other during the dialogue recordings – either through the video conference tool Zoom (CoNNAR_videocall) or in a face-to-face setting (CoNNAR_face-to-face). The subcorpus CoNNAR_face-to-face is under construction.

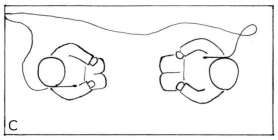

Figure 3: Experimental setup (Miriam Müller 2023, CC-BY 4.0). In the co-presence setting (C), participants were recorded with head-mounted microphones and seated together in a booth at a distance of ca. 1 meter.

2 Annotation scheme

2.1 Word lists

The word lists contain four annotation tiers (cf. Table 7). All tiers are aligned with the audio signal.

Table 7: Annotation tiers of the word lists with their reference to each other and to the acoustic signal (AS).

Tier name	Contains	Reference to
`ORT-MAU` (3.2)	Tokenised transliteration	AS
`KAN-MAU` (3.3)	Phonetic transcription	`ORT-MAU`
`MAU` (3.4)	Phonetic segmentation	`ORT-MAU`, `KAN-MAU`
`vowel` (3.13)	Corner vowels	`ORT-MAU`

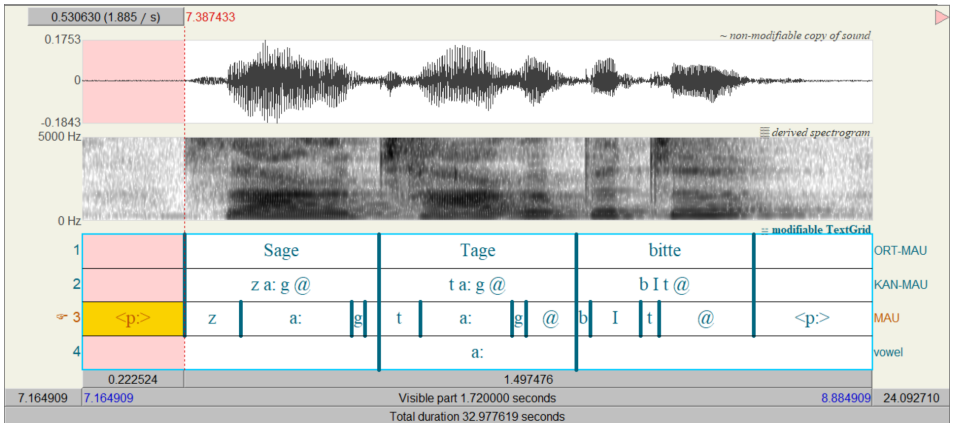

Figure 4: An example of the annotation scheme for the word list (p01c01_L1_lst1_ch1).

2.2 Free conversations

The task-free dialogues (free conversations) contain 12 annotation tiers (cf. Table 8). All tiers are aligned with the audio signal.

Table 8: Annotation tiers of the task-free dialogues (free conversations) with their reference to each other and to the acoustic signal (AS).

Tier name	Contains	Reference to
`TRN` (3.1)	Diplomatic transliteration	AS
`ORT-MAU` (3.2)	Tokenisation	AS
`KAN-MAU` (3.3)	Phonetic transcription	`ORT-MAU`
`MAU` (3.4)	Phonetic segmentation	`ORT-MAU`, `KAN-MAU`
`pseudo` (3.5)	Pseudonymization	`ORT-MAU`
`norm` (3.6)	Orthographic normalisation	`ORT-MAU`
`lem` (3.7)	Lemmatisation	`norm`
`pos` (3.8)	Part-of-speech-tagging	`lem`
`gram` (3.9)	Grammatical information	`norm`
`schwareal` (3.10)	Schwa realisation	`ORT-MAU`, `gram`
`schwamau` (3.11)	Schwa duration	`schwareal`, `MAU`
`IP` (3.12)	Intonation phrase	`schwareal`

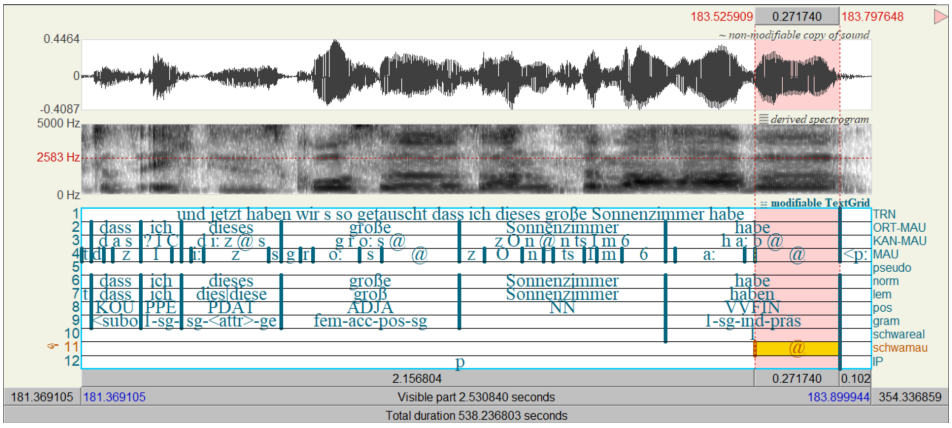

Figure 5: An example of the annotation scheme for the free dialogue (p06c04_L2C1_free_ch2).

2.3 Diapixes

The task-based dialogues (Diapixes) contain 13 annotation tiers (cf. Table 9). All tiers are aligned with the audio signal. The first attempt of the Diapix 1 in session p19c07_L1 had to be restarted after 110 seconds because of a problem with the internet connection. The corpus does not include the interrupted recording.

Table 9: Annotation tiers of the task-based dialogues (Diapixes) and their reference to each other and to the acoustic signal (AS).

Tier name	Contains	Reference to
`TRN` (3.1)	Diplomatic transliteration	AS
`ORT-MAU` (3.2)	Tokenisation	AS
`KAN-MAU` (3.3)	Phonetic transcription	`ORT-MAU`
`MAU` (3.4)	Phonetic segmentation	`ORT-MAU`, `KAN-MAU`
`norm` (3.6)	Orthographic normalisation	`ORT-MAU`
`lem` (3.7)	Lemmatisation	`norm`
`pos` (3.8)	Part-of-speech-tagging	`lem`
`gram` (3.9)	Grammatical information	`norm`
`schwareal` (3.10)	Schwa realisation	`ORT-MAU`, `gram`
`schwamau` (3.11)	Schwa duration	`schwareal`, `MAU`
`IP` (3.12)	Intonation phrase	`schwareal`
`vowel` (3.13)	Corner vowels	`ORT-MAU`
`request` (3.14)	Request for clear speech	`ORT-MAU`

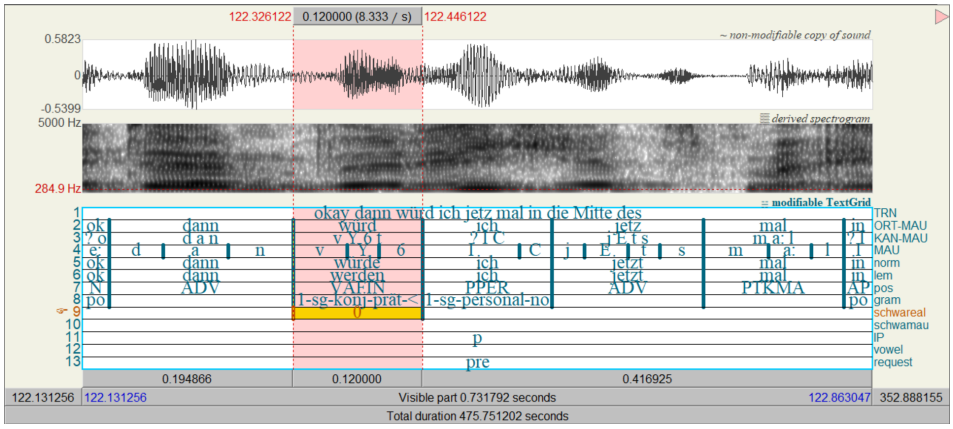

Figure 6: An example of the annotation scheme for the Diapix task (p11c05_L1_pix2_ch2).

3 Annotation tiers

3.1 Tier TRN (Diplomatic transliteration)

Name	TRN
Description	Transliteration as annotation on the acoustic signal
Annotation type	Span annotation on the acoustic signal
Creation	Manually (cf. Section 3.1.1) with CAT (Sauer n.d.), Praat (Boersma & Weenink 2019) and TierTagger (Lange 2023)
Annotation tags	Open set (cf. Section 3.1.2)
Annotators	Robert Lange, Megumi Terada, Bianca Sell
Annotated for	Free conversation, Diapix 1+2

3.1.1 Creation

During the transliteration process, two different methods were used. Because of a compatibility problem with CAT, about the half of the audio files were transliterated using TierTagger (Lange 2023).

Transliteration with CAT

CAT (Sauer n.d.) is a transliteration tool that detects silent pauses in an audio file and segments the audio stream into chunks. These chunks contain only utterances, and when the audio is played, silent pauses are skipped. In transliteration, extralinguistic entities such as laughter or coughing were marked as ⟨usb⟩ (*non-understandable word* or *other human noises*), other noise as ⟨nib⟩ (*non-human noise*). After transliterating the audio file, a TextGrid file was created by CAT, which was manually checked in Praat (Boersma & Weenink 2019).

Transliteration with Praat and TierTagger

As preparation for this, the wav file was segmented into chunks by using the silence annotation of Praat (Boersma & Weenink 2019). Afterwards, the chunked wav files were transliterated by using the TierTagger (Lange 2023), which created a TextGrid file with pre-aligned chunks. Boundaries were checked in Praat as well.

3.1.2 Annotation tags

On the tier `norm`, open annotation tags were used. Diplomatic transliteration was done according to the following guidelines.

- No transliteration of vowel quality, e. g. demonstrative pronoun *das* 'that' even when it was realised as [dəs] instead of [das]
- Transliteration of apocope such as *schwimm* 'swim-PRS.1SG.NOM' for [ʃvɪm] instead of [ʃvɪmə]
- Reduction forms with ambisyllabic consonants: *isses, isser, inner* 'is it', 'is he', 'in the' without token boundaries
- Proclitics: with token boundaries, e. g. *s|gibt* 'there is'
- Epentheses: *darunter* → *dadrunter* 'below'
- Highly frequent forms with reduction:
  - *es* 'it' → *s*
  - *ist* 'is' → *is* (not *s*, to avoid confusion with *es*)
  - *ein* 'a' → *n*
  - *eine* 'a' → *ne*
  - *haben* 'have' → *ham*
  - *haben wir* 'have we' → *ham|wa*
  - *nicht* 'not' → *nich*
- No variation in highly frequent lexical words, e. g. always transcribed as *eigentlich, irgendwie, jetzt* 'actually', 'somehow', 'now'

3.2 Tier `ORT-MAU` (Tokenisation)

Name	`ORT-MAU`
Description	Tokenisation of the transliteration
Annotation type	Span annotation on the acoustic signal
Creation	Semi-automatic (cf. Section 3.2.1) with WebMAUS (Kisler et al. 2017)
Annotation tags	Open set (cf. Section 3.2.2)
Annotators	Megumi Terada, Robert Lange, Miriam Müller, Bianca Sell
Annotated for	Word list 1+2, free conversation, Diapix 1+2

3.2.1 Creation

Boundary correction

Here, the TextGrid files were corrected if chunks included overlong pauses. Boundaries were corrected, where possible, so that utterances of the other speaker were not within the chunks. All redundant interval boundaries added by CAT or Praat were deleted before the alignment with WebMAUS (Kisler et al. 2017). The boundaries include as few pauses as possible. Short pauses (about less than 200 ms) needed to be deleted in order to avoid errors in WebMAUS.

Alignment with the signal

For the alignment between the acoustic signal and the text, the following pipeline in WebMAUS (Kisler et al. 2017) was used with the following settings (the setting remains in default mode if not stated otherwise): Chunkprep → G2P → MAUS[4].

- Language: German (DE)
- Output Encoding: X-SAMPA (ASCII)
- Inter-word silence (MAUS): 7
- KAN tier in TextGrid (MAUS): true
- ORT tier in TextGrid (MAUS): true
- Chunk segmentation (MAUS): true
- Pre-segmentation (MAUS): true

Automatic transliteration with WebMAUS Basic

For word lists the transliteration was created automatically by using WebMAUS Basic (Kisler et al. 2017)[5].

3.2.2 Annotation tags

The tier `ORT-MAU` uses an open tagset which contains tokenised transliteration. The tokenisation was based on the tier `TRN`.

[4] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Pipeline
[5] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic

3.3 Tier KAN-MAU (Phonetic transcription)

Name	KAN-MAU
Description	Canonical phonetic transcription of the tier ORT-MAU
Annotation type	Span annotation on the acoustic signal
Reference	ORT-MAU (Section 3.2)
Creation	Automatic (cf. Section 3.3.1) with WebMaus (Kisler et al. 2017)
Annotation tags	Canonical phonetic form of Section 3.2 (cf. Section 3.3.2)
Annotated for	Word list 1+2, free conversation, Diapix 1+2

3.3.1 Creation

The transliterated TextGrids were transferred to WebMAUS. KAN was created automatically with the Pipeline Chunkprep → G2P → MAUS (cf. Section 3.2).

3.3.2 Annotation tags

The orthographic transliteration on the tier ORT-MAU was translated into canonical phonological form on the tier KAN. Phonetic symbols are encoded in SAMPA (see https://www.phon.ucl.ac.uk/home/sampa/).

3.4 Tier MAU (Phonetic segmentation)

Name	MAU
Description	Phonetic transcription of the tier ORT-MAU
Annotation type	Span annotation on the acoustic signal
Reference	ORT-MAU (Section 3.2), KAN-MAU (Section 3.3)
Creation	Automatic (cf. Section 3.4.1) with WebMaus (Kisler et al. 2017)
Annotation tags	Phonetic form of ORT-MAU in SAMPA (cf. Section 3.4.2)
Annotated for	Word list 1+2, free conversation, Diapix 1+2

3.4.1 Creation

The transliterated TextGrids were transferred to WebMAUS. KAN was created automatically with the Pipeline Chunkprep → G2P → MAUS (cf. Section 3.2).

3.4.2 Annotations tags

On the tier MAU, the actual realisation of the transliterated signal was annotated, which is encoded in SAMPA (see Section 3.3.2). The annotation was done automatically at the segment level and refers to the tier KAN as well as to the acoustic signal. As for Diapixes of participants, this tier was manually corrected in case errors were detected. Alignment of the tier MAU was done with the tiers ORT-MAU and KAN at the outer boundaries.

3.5 Tier pseudo (Pseudonymization)

Name	pseudo
Description	Anonymisation
Annotation type	Span annotation on the acoustic signal
Reference	ORT-MAU (Section 3.2)
Creation	Manual (cf. Section 3.5.1)
Annotation tags	Closed set (cf. Section 3.5.2)
Annotated for	Free conversation

3.5.1 Creation

The annotators listened to all audio data. Proper names and other items which can be used for identifying speakers were annotated. Then the tier pseudo was added to the TextGrid for each channel. After that, a script in https://doi.org/10.18452/20145 (Belz 2019) was applied to replace the speech signal in the annotated intervals with a 200 Hz tone.

3.5.2 Annotation tags

The tier pseudo entails a closed tagset (‘p’ or ‘%’). These were used for proper names and other data that might serve to identify individuals.

3.6 Tier `norm` (Orthographic normalisation)

Name	`norm`
Description	Orthographic normalisation of `ORT-MAU`
Annotation type	Span annotation on the acoustic signal
Reference	`ORT-MAU` (Section 3.2)
Creation	Semi-automatic (cf. Section 3.6.1) with TierTagger (Lange 2023)
Annotation tags	Closed set (cf. Section 3.6.2)
Annotators	Robert Lange, Megumi Terada
Annotated for	Free conversation, Diapix 1+2

3.6.1 Creation

The tier `norm` provides the orthography-based target hypothesis, e. g., the token *isser* 'is he' on the tier `ORT-MAU` was separated into *ist*|*er* on the tier `norm` . The location of the token boundaries are in accordance with the acoustic signal. For breaking off, no goal hypotheses could be formed and the token was replaced by a tag ⟨nn⟩ (*not normalisable*). The tokens on `ORT-MAU` were normalised via an assignment list with TierTagger (Lange 2023) which added the tier `norm` into TextGrid. If a compound contained a silent pause, the token was divided into several parts on the tier `ORT-MAU` and annotated as one token in the tier `norm`. As for Diapix tasks of participants, normalised tokens were manually corrected – words that were not previously normalised were re-normalised manually and then added to the above-mentioned list of words.

3.6.2 Annotation tags

A tag ⟨nn⟩ was given, if the token on the tier `ORT-MAU` could not be orthographically represented. This was the case e. g. for breaking off. ⟨nn⟩ was not assigned to ⟨usb⟩ on the tier `ORT-MAU`. In such cases, there was no token on `norm`.

3.7 Tier `lem` (Lemmatisation)

Name	`lem`
Description	Lemmatisation of the normalised token
Annotation type	Span annotation on the acoustic signal
Reference	`norm` (Section 3.6)
Creation	Semi-Automatic (cf. Section 3.7.1) with TierTagger (Lange 2023)
Annotation tags	Open set (cf. Section 3.7.2)
Annotators	Robert Lange, Bianca Sell, Megumi Terada
Annotated for	Free conversation, Diapix 1+2

3.7.1 Creation

The tier `lem` was created with the TreeTagger and STTS 2.0 parameter file (Schmid 1994), passing the tokens in context (left 5, right 5). Regarding free conversation and Diapix of participants, this tier was manually checked for errors and corrected if necessary. Cliticisation and its components are shown separated by a vertical line. E. g. *im|in+die* 'in the'. Indefinite articles are displayed as *ein|eine*, definite articles as *d|die*. For reflexives *er|es|sie|sich* are used.

3.7.2 Annotation tags

The tier `lem` has an open tagset which consists of lemmatised tokens.

3.8 Tier `pos` (Part-of-speech-tagging)

Name	`pos`
Description	Part-of-speech-tagging of lemmatised tokens
Annotation type	Span annotation on the acoustic signal
Reference	`norm` (Section 3.6)
Creation	Semi-Automatic (cf. Section 3.8.1) with TierTagger (Lange 2023)
Annotation tags	Closed set (cf. Section 3.8.2)
Annotators	Robert Lange, Bianca Sell, Megumi Terada
Annotated for	Free conversation, Diapix 1+2

3.8.1 Creation

The tier `pos` was created with the TreeTagger and STTS 2.0 parameter file (Schmid 1994), passing the tokens in context (left 5, right 5). This tier was manually controlled for free conversation and Diapix of participants as far as noticed. The part-of-speech-tags were annotated with STTS 2.0 (Westpfahl et al. 2017), which is based on the Stuttgart Tubingen Tagset (STTS) (Schiller et al. 1999). Extralinguistic entities or pauses were ignored.

3.8.2 Annotation tags

The part-of-speech was tagged with Stuttgart Tübingen Tagset (STTS) (Schiller et al. 1999) and STTS 2.0 (Westpfahl et al. 2017). They consist of closed tagsets.

3.9 Tier `gram` (Grammatical information)

Name	`gram`
Description	Annotation of grammatical categories
Annotation type	Span annotation on the acoustic signal
Reference	`norm` (Section 3.6)
Creation	Semi-automatic (cf. Section 3.9.1) with TierTagger (Lange 2023)
Annotation tags	Open set (cf. Section 3.9.2)
Annotators	Robert Lange, Bianca Sell, Megumi Terada
Annotated for	Free conversation, Diapix 1+2

3.9.1 Creation

The tier `gram` was created with DEMorphy (Altinok 2018). As DEMorphy offers all possible inflection/declension paradigms, the first entry was always used. Some, but not all, of the participant data was corrected manually. The conjugation of verbs was determined in the linguistic context and was coded using composite tags, orderd by person – number – mode – tense, e. g. *(ich) rannte* '(I) ran': 1-sg-ind-prät.

3.9.2 Annotation tags

Annotation tags	Description
	Composed as follows: Person-Number-Mode-Tense
1 / 2 / 3	Person: 1.P / 2.P / 3.P
sg	Number: Singular
pl	Number: Plural
ind	Mode: Indicative
konj	Mode: Subjunctive (I + II)
imp	Mode: Imperative (number information irrelevant and therefore ignored)
präs	Tense: Present tense
prät	Tense: Preterite
inf	Infinitive (Present tense), e. g. *laufen*
pp1	Present participle (non-attributive), e. g. *laufend*
pp2	Past participle (non-attributive), e. g. *gelaufen*

3.10 Tier `schwareal` (Schwa realisation)

Name	`schwareal`
Description	Annotation on schwa realisation
Annotation type	Span annotation
Reference	`ORT-MAU` (Section 3.2), `gram` (Section 3.9)
Creation	Semi-automatic (cf. Section 3.10.1) with TierTagger (Lange 2023)
Annotation tags	Closed set (cf. Section 3.10.2)
Annotators	Miriam Müller, Bianca Sell, Megumi Terada
Annotated for	Free conversation, Diapix 1+2

3.10.1 Creation

A tier named `schwareal` was added with TierTagger (Lange 2023). The annotation tag was assigned based on Section 3.10.2. This annotation tier was limited to free conversation and Diapix 1 and 2 of participants.

3.10.2 Annotation tags

Annotation tags	Description
1	Yes, schwa is realised
0	No, schwa is not realised
x	Unclear

3.11 Tier `schwamau` (Schwa duration)

Name	`schwamau`
Description	Duration of realised schwa
Annotation type	Span annotation on the acoustic signal
Reference	`MAU` (Section 3.4), `schwareal` (Section 3.10)
Creation	Semi-automatic (cf. Section 3.11.1) with TierTagger (Lange 2023)
Annotation tags	Closed set (cf. Section 3.11.2)
Annotators	Miriam Müller, Bianca Sell, Megumi Terada
Annotated for	Free conversation, Diapix 1+2

3.11.1 Creation

The tier `schwamau` was created with TierTagger (Lange 2023). Based on acoustic signals and annotations on the tier `MAU`, the duration of all realised schwas was annotated. The beginning of the vowel was marked at the rising zero crossing of the first full period in the oscillogram, and the end of the vowel was indicated by the last period before F2 changes from dark to light in the sonogram (at the zero crossing). This annotation tier is limited to free conversation and Diapix of participants.

3.11.2 Annotation tags

The schwa was annotated with '@', when it was realised in the verb ending. Unclear cases, which were to be excluded from the analysis, were annotated with the tag 'x'. Otherwise no tag was given.

3.12 Tier `IP` (Intonation phrase)

Name	`IP`
Description	Intonation phrases
Type of annotation	Span annotation on the acoustic signal
Reference	`ORT-MAU` (Section 3.2)
Creation	Manual (cf. Section 3.12.1)
Annotation tags	Closed set (cf. Section 3.12.2)
Annotators	Bianca Sell, Miriam Müller, Megumi Terada
Annotated for	Free conversation, Diapix 1+2

3.12.1 Creation

An interval tier with the name `IP` was created in Praat. Based on the acoustic signal, intonation phrases were annotated manually. Most researchers distinguish between a superordinate intonation phrases (IP) and subordinate intermediate phrases (ip) (cf. Grice & Baumann 2002). Here, only IPs were annotated. An IP was defined by the maximum extent of a pitch contour perceived as cohesive with at least one ip and at least one nuclear accent. Only IPs containing a possible word-final schwa in verbal inflection suffixes (target tokens as described for `schwareal` in Section 3.10) were annotated up to now. This annotation tier is limited to free conversation and Diapix of participants.

3.12.2 Annotation tags

A tag 'p' was assigned to intonation phrases. This was defined as the maximum extension of a pitch that was perceived as cohesive with at least one accented syllable. The boundary tone was obligatory for this annotation. Unclear cases were marked with an 'x'.

3.13 Tier `vowel` (Corner vowels)

Name	`vowel`
Description	Annotation of corner vowels
Annotation type	Span annotation on the acoustic signal
Reference	`ORT-MAU` (Section 3.2)
Creation	Manual (cf. Section 3.13.1)
Annotation tags	Closed set (cf. Section 3.13.2)
Annotators	Christine Mooshammer, Malte Belz, Bianca Sell, Megumi Terada
Annotated for	Word list 1+2, Diapix 1+2
Empty tier in	Free conversation

3.13.1 Creation

By using a self-written R script, the tier `vowel` was added to annotate corner vowels [aː], [iː], and [uː] in stressed and accented position within content words. Regarding the Diapix tasks, we mainly used the target words in the pictures' written parts (see Section 1.6.3), while some participants read the words less than two times per target vowel. In this case, additional corner vowels in stressed and accented position were annotated manually, avoiding diphthongisation due to r-vocalisation or following vowels. The beginning of each vowel was marked at the rising zero crossing of the first complete period and its end at the zero crossing of F2 fading.

3.13.2 Annotation tags

Annotation tags	Description
a	The annotation of [aː] in stressed and accented position within content words
i	The annotation of [iː] in stressed and accented position within content words
u	The annotation of [uː] in stressed and accented position within content words

3.14 Tier `request` (Request for clear speech)

Name	`Request`
Description	Request for a clearer speech
Type of annotation	Span annotation on the acoustic signal
Reference	`ORT-MAU` (Section 3.2)
Creation	Manual (cf. Section 3.14.1)
Annotation tags	Closed set (cf. Section 3.14.2)
Annotator	Megumi Terada
Annotated for	Diapix 2
Empty tier in	free conversation, Diapix 1

3.14.1 Creation

During the second Diapix task, the confederates were instructed to ask the participants to speak more clearly. On the tier `request`, the start and end time periods of these requests were annotated for both confederates and participants (duplicated tiers).

3.14.2 Annotation tags

Annotation tags	Description
cr	Request to speak clearly
pre	Before the request
med	Between several requests
post	After the request
cr_x	If the intention of the request is not to ask for a clearer speech

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334. The authors would like to thank Anke Lüdeling and Christine Mooshammer for their valuable advice on the design and annotation of the corpus.

References

Altinok, Duygu. 2018. DEMorphy, German Language Morphological Analyzer. http://arxiv.org/abs/1803.00902.

Audacity Team. 2020. *Audacitiy (version 2.4.2).* [Computer program]. https://www.audacityteam.org/.

Baker, Rachel & Valerie Hazan. 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods* 43. 761–770. DOI: 10.3758/s13428-011-0075-y.

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13(2). 145–204.

Belz, Malte. 2019. *Pseudonymisierung von Sprachchunks in Audiodaten mit Praat und R (Pseudonymization of speech chunks in audio data using Praat and R).* Berlin: Humboldt-Universität zu Berlin. DOI: 10.18452/20145.

Boersma, Paul & David Weenink. 2019. *Praat: Doing phonetics by computer.* [Computer program]. Version 6. http://www.praat.org/.

Bradlow, Ann R & Tessa Bent. 2002. The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America* 112(1). 13. DOI: 10.1121/1.1487837.

Bullock Oliveira, Maggie & Bianca Sell. 2022. *PDF and PSD files of DiapixGEtv picture materials – German version adapted to elicit tense vowels.* Zenodo. DOI: 10.5281/zenodo.6510724.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Grice, Martine & Stefan Baumann. 2002. Deutsche Intonation und GToBI. *Linguistische Berichte* (191). 267–298.

Hay, Jennifer, Stefanie Jannedy & Norma Mendoza-Denton. 1999. Oprah and /AY:/: Lexical Frequency, Referee Design and Style. In *Proceedings of the 14th International Congress of Phonetic Sciences*, 1389–1392. San Francisco.

Hazan, Valerie & Rachel Baker. 2011. *High resolution pdf files of DiapixUK picture materials – original English version.* Zenodo. DOI: 10.5281/zenodo.3703202.

Hinnenkamp, Volker. 1982. "Türkisch Mann, Du?": Sprachverhalten von Deutschen gegenüber Gastarbeitern. In *Mehrsprachigkeit in der Stadtregion: Jahrbuch 1981 des Instituts für deutsche Sprache*, 171–193. Schwann.

Keim, Inken. 1978. *Gastarbeiterdeutsch: Untersuchungen zum sprachlichen Verhalten türkischer Gastarbeiter. Pilotstudie*, vol. 41 (Forschungsberichte des Instituts für deutsche Sprache). Tübingen: TBL Verlag Gunter Narr.

Kisler, Thomas, Uwe Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347. DOI: 10.1016/j.csl.2017.01.005.

Lange, Robert. 2023. *TierTagger*. Version 0.9. https://scm.cms.hu-berlin.de/langerob/tiertagger.

Lüdeling, Anke, Artemis Alexiadou, Aria Adli, Karin Donhauser, Malte Dreyer, Markus Egg, Anna Helene Feulner, Natalia Gagarina, Wolfgang Hock, Stefanie Jannedy, Frank Kammerzell, Pia Knoeferle, Thomas Krause, Manfred Krifka, Silvia Kutscher, Beate Lütke, Thomas McFadden, Roland Meyer, Christine Mooshammer, Stefan Müller, Katja Maquate, Muriel Norde, Uli Sauerland, Stephanie Solt, Luka Szucsich, Elisabeth Verhoeven, Richard Waltereit, Anne Wolfsgruber & Lars Erik Zeige. 2022. Register: Language Users' Knowledge of Situational-Functional Variation: Frame text of the First Phase Proposal for the CRC 1412. *Register Aspects of Language in Situation (REALIS)* 1(1). 1–58. DOI: 10.18452/24901.

Roche, Jörg. 1998. Variation in Xenolects (Foreigner Talk). *Sociolinguistica* 12(1). DOI: 10.1515/9783110245172.117.

Sauer, Simon. N.d. *CAT - Chunked Audio Transcription*. Version 3.1.2.

Schiller, Anne, Simone Teufel, Christine Stöckert & Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical Report.* Stuttgart: Institut für Maschinelle Sprachverarbeitung. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the Conference on New Methods in Language Processing*, 44–49. Manchester, UK.

Schroedler, Tobias, Judith Purkarthofer & Katja F Cantone. 2022. The prestige and perceived value of home languages. Insights from an exploratory study on multilingual speakers' own perceptions and experiences of linguistic discrimination. *Journal of Multilingual and Multicultural Development*. 1–18. DOI: 10.1080/01434632.2022.2121402.

Westpfahl, Swantje, Thomas Schmidt, Jasmin Jonietz & Anton Borlinghaus. 2017. *STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS).* Working Paper. Edition: Version 1.1, March 2017. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-60634.

Appendix

Consents on recording setting in English

- I consent to being recorded with a microphone.
- I consent to the use of the video conferencing tool Zoom (without suitability decision and applicable guarantees)
- I agree that my keyboard and/or mouse inputs are recorded.
- I consent to handwritten information being recorded.
- I consent to the individual records collected being linked together.

Consents on recording settings in German (original)

- Ich willige darin ein, dass eine Aufzeichnung mit Mikrofon stattfindet.
- Ich willige darin ein, dass das Videokonferenztool Zoom verwendet wird (ohne Angemessenheitsbeschluss und geeignete Garantien)
- Ich willige darin ein, dass meine Tastatur- und/oder Mauseingaben aufgezeichnet werden.
- Ich willige darin ein, dass handschriftliche Informationen aufgezeichnet werden.
- Ich willige darin ein, dass die einzelnen erhobenen Datensätze miteinander verknüpft werden.

Project

English I agree that my data will be processed as described in the speaker information sheet, listed on a data portal and that my anonymised (biometric[6] and non-biometric) data will be stored in a research data repository. My data will only be processed for the scientific project described above and for no other purpose.

[6]Personal/individual-related information on physical, physiological or behavioral characteristics of a person that enables or confirms the unique identification of that person.

German Ich willige darin ein, dass meine Daten, wie in dem Proband*inneninformationsblatt beschrieben, verarbeitet werden, auf einem Datenportal gelistet und meine anonymisierten (biometrischen und nicht biometrischen) Daten in einem Forschungsdatenrepositorium gespeichert werden. Meine Daten werden nur für das oben beschriebene, wissenschaftliche Projekt und zu keinem anderen Zweck verarbeitet.

Explanation This item allows the storage, processing and use of the data within the project.

Transfer

English Furthermore, I agree that my anonymised (biometric and non-biometric) data will be made available via the research data repository to scientists outside this project for scientific purposes and in compliance with the data processing steps described in the information sheet.

German Darüber hinaus bin ich damit einverstanden, dass über das Forschungsdatenrepositorium meine anonymisierten (biometrischen und nicht biometrischen) Daten Wissenschaftler*innen außerhalb dieses Projektes für wissenschaftliche Zwecke und unter Einhaltung der im Informationsblatt beschriebenen Datenverarbeitungsschritte zur Verfügung gestellt werden.

Explanation This item allows the transfer of data to scientists outside the project for research purposes via repository.

Third parties

English I agree that my anonymised data set is made available to the public for further use outside of this project. This applies exclusively to non-biometric data.

German Ich willige darin ein, dass mein anonymisierter Datensatz außerhalb dieses Projektes zur weiteren Nutzung der Öffentlichkeit zur Verfügung gestellt wird. Dies gilt ausschließlich für nicht biometrische Daten.

Explanation This item allows the publication of anonymised non-biometrical data outside the project.

Anonymisation

English	I was given a code at the beginning of the study and I agree that my records, which are created at different times, may be interconnected by using this code.
German	Ich habe zu Beginn der Untersuchung einen Code ausgehändigt bekommen und willige ein, dass meine Datensätze, die zu unterschiedlichen Zeitpunkten entstehen, mittels dieses Codes miteinander verknüpft werden dürfen.
Explanation	This item allows linking between anonymisation code and recordings.