The Texas German Dialect Project Corpus as a Diachronic Resource for Investigating Language Contact

Thomas Schmidt[1], **Margaret Blevins**[2], **Hans C. Boas**[2], **Glenn Gilbert**[3]

Department of Germanic Studies, Texas German Dialect Project,
2505 University Avenue, Burdine 336, University of Texas at Austin, Austin, TX 78712-0304
[1]linguisticbits.de, [2]University of Texas at Austin, [3]University of Southern Illinois

thomas@linguisticbits.de, mblevins@utexas.edu, hcb@mail.utexas.edu.
glenngilbert1869@gmail.com

Abstract

The Texas German Dialect Project (TGDP) is a long-standing effort to document the unique variety of German spoken in Texas since the 1840s. For 25 years, the TGDP has built up the freely accessible Texas German Dialect Archive Online (TGDA Online) with recordings and annotations of interviews and language tasks conducted between 2001 and today with some of the last speakers of the variety, which is expected to go extinct within the next 5-10 years. The present paper reports on the most recent addition to the TGDP's online corpus platform— a collection of Texas German data recorded in the 1960s—as well as historical translation elicitations that will be released later in 2026. Both the contemporary and the historical materials follow very similar elicitation methods and are processed using the same pipelines, increasing their comparability. This provides a comparable historical dimension to the resource, enabling diachronic and multidimensional analyses of this endangered variety. These data can help shed light on the dynamics of language contact, dialect contact, and language death.

Keywords: spoken language, variation, corpus, contact variety, Texas German

1. Introduction

Texas German is a unique contact variety for the following reasons: (1) It is a mix of at least five different regional dialects brought by German-speaking immigrants from Europe to central Texas beginning in the 1830s (Boas, 2009). As such, it differs from most other German contact varieties such as Michigan German (Born, 1994), Namibia German (Zimmer et al., 2020), Brazilian German (Altenhofen, (submitted), Kürschner et al., (submitted)), and Hungarian German (Földes, 2002), which can be directly traced back to specific donor dialects from a single region (Boas, 2016). (2) Texas German is, in comparison to other German contact varieties such as Pennsylvania German, Romanian German, and Hungarian German, a relatively young contact variety. (3) Unlike other new world contact varieties of German, Texas German did not evolve into a focused new world variety according to the model of Trudgill (2004). Because intergenerational transmission was severely restricted following World War I, Texas German never went through all three stages of new dialect formation, being "stuck" halfway through the second stage (Boas, 2009).[1] Texas German thus exhibits a very high degree of inter-speaker and intra-speaker variation, making it challenging to provide a clear definition of how Texas German should be characterized linguistically. (4) Texas German is one of the most extensively documented extraterritorial contact varieties of German, making it a particularly fruitful basis for linguistic research.

2. Background

2.1 Similar varieties

Due to research, funding, and time limitations, as well as the availability of the data itself, many corpora of low-resource language varieties contain a potentially limited snapshot of a particular variety (e.g., recordings of natural conversation and elicited translations from a particular 5-year timespan). Or data may be collected as part of a particular research project and then not be made accessible or public. It can therefore be challenging to find accessible data that span several timeframes, genres, and elicitations methods.

There are numerous corpora of (Germanic) heritage languages, all of which have different coverages, timeframes, and accessibility. For example, there are recordings of Wisconsin German (USA) from the 1960s and 2001 (Eichhoff 1979, Wagener 2002), but only the former is currently publicly available. The Corpus of American Danish (CoAmDa) contains recordings from both North America (three sets: 1960s-1980s, early 1990s, and late 1990s) and Argentina (2014-2015). CoAmDa is fully transcribed and marked for language (Kühl et al. 2017). The MEND corpus, a collection of Mennonite German (Kaufmann et al. 2023), contains a large set of translation elicitation

[1] Trudgill's three stages of dialect formation are (1) (first generation): rudimentary leveling, (2) (second generation): variability and mixing, (3) (third generation): emergence of stable, relatively uniform dialect.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 221–229
16 May 2026.

recordings from five different countries from a single time period (1999-2002). A similar situation is true for the RUEG corpus, which is a highly multifaceted collection containing several kinds of data (written and spoken, informal and formal) from a variety of (heritage) language speakers (mono- and multi-lingual speakers of Greek, Russian, German, and Turkish, English, and Kurdish) from multiple locations (Greece, Russia, Germany, Turkey, and the USA) during a ~6 year timespan (Schroeder et al. 2024).[2]

For Texas German, there is a plethora of data available covering many time periods, modes, and genres, including:

- Narrative audio interviews from the 1960s, 1990s, and 2001-today (~500 hours)
- Audio recorded translation tasks from the 1960s, 1990s, and 2001-today (~350 hours)
- Letters and diaries from the mid-1800s through the early 1900s (2,000+ pages)
- Newspapers from the mid-1800s through the mid-1900s (800,000+ pages)
- Reports from the mid-1800s through the early 1900s (1,000+ pages)
- Novels and poetry from the late-1800s through the 1900s

This puts researchers in the unique position to be able to explore the evolution of a heritage contact variety through over 100 years of development from multiple angles.

In what follows, two comparable sets of data are described. Section 2.2 discusses a set of audio recordings of contemporary Texas German (2001-today) and section 3 introduces a set of Texas German recordings from the 1960s. Both sets of data were elicited in similar ways as well as processed and annotated using the same or very similar methods with the goal of making a resource that facilitates diachronic research.

2.2 The TGDP and the TGDA Online

Since 2001, members of the Texas German Dialect Project (TGDP) have interviewed more than 800 speakers of Texas German, recording three different types of data. The first type of data consists of participants' translations of English word lists and sentences taken from Eikel (1954), Gilbert (1972), and Guion (1996) into Texas German. The second type of data consists of sociolinguistic, open-ended interviews, which aim to capture the informants' daily use of Texas German. These interviews seek to elicit casual, relaxed conversations in which speakers produce Texas German spontaneously without being asked to produce specific linguistic structures. The third type of data is a biographical questionnaire that captures relevant extralinguistic information about each speaker, including age, religious affiliation, language use throughout the lifespan, and language, and speaker attitudes.[3]

Audio recordings of participants' translations based on Eikel (1954) and Gilbert (1972)'s wordlists are available at https://tgdp.org/. Data for 750+ speakers from 90+ locations is currently available online.

Audio recordings of open-ended interviews, along with their time-aligned transcripts, are available at https://tgdp-zumult.la.utexas.edu/.[4] The most recent December 2025 release of the data contains 196 hours of audio recordings (1,494,400 tokens, 475 speakers).[5] The transcripts available contain a literary transcription, a tokenized layer, language tags, normalization, part-of-speech tagging (one layer according to STTS 2.0 (Westpfahl & Schmidt, 2016) and another layer for Universal Dependencies POS (Nivre et al. 2020)), lemmatization, a canonical phonetic transcription based on the transcribed forms, and a word-for-word English translation. These transcripts, along with all the annotation layers, can be queried using a CQP query tool.

3. Integrating Gilbert's Historical Recordings

3.1 Background

In the 1960s, Glenn Gilbert recorded several speakers as part of the fieldwork for his dissertation *The German Dialect Spoken in Kendall and Gillespie Counties, Texas* (Gilbert, 1963). Several years later, he interviewed dozens more Texas German speakers, mostly hand-transcribing the interviews in situ, but also audio recording some of them. This fieldwork, in addition to the fieldwork he had already done for his dissertation, laid the groundwork for his *Linguistic Atlas of Texas German* (Gilbert, 1972). In 2017, Gilbert donated all his audio recordings, fieldnotes, and several boxes of other research materials related to his research on Texas German to the TGDP.

These audio recordings create a unique situation: With access to audio recordings from the 1960s and from 2001 to today, it is possible to make

[2] https://www.linguistik.hu-berlin.de/en/institut-en/professuren-en/rueg/rueg-corpus

[3] For details on the interview process, the processing of recordings, the workflow for transcribing and translating the interviews, and how the data are stored, see Boas (2021) and Boas et al. (in press).

[4] This corpus is available via an instance of the ZuMult platform (Fandrych et al., 2023; Boas et al., in press).

[5] Because interviews, audio processing, and transcription are ongoing, only a portion of the interviews recorded by the TGDP are currently available online (~60% of the open-ended interviews, ~90% of the translation interviews).

comparisons between the two sets of data, allowing for a better understanding of how Texas German has evolved over the last 50+ years.[6] This can also help us better understand how languages in contact may evolve over time in general.

3.2 Available Materials

The audio recordings Gilbert donated to the TGDP include sub-collections of other contact varieties, such as Texas Czech, Texas English, and Kansas German. The Texas German portion of his audio collection is referred to as the GTXG corpus ('Gilbert Texas German').[7]

GTXG contains 63 *events* with 98 participants.[8] The term "event" refers to the entirety of recordings made in one informant session (sometimes with one speaker, sometimes with a group of speakers). Each event consists of one or more *speech events*. The term "speech event" refers to one particular type of recorded speech (e.g., translation elicitations, free conversation, etc.). The speech event types are as follows:

- Open-Ended (OE): sociolinguistic, narrative, open-ended interviews in which a non-Texas German speaker talks with a Texas German speaker about their life
- Atlas (GA): A translation elicitation interview in which the interviewer asks the participant to translate a set of English words and phrases into Texas German. This list (~230 prompts) is most similar to the set of words and phrases used to construct Gilbert's *Linguistic Atlas of Texas German* (1972)
- Long (GL): A translation elicitation interview in which the interviewer asks the participant to translate a set of over 1,200 English words and phrases Texas German.[9]
- Wenker sentences (WS): The list of sentences Georg Wenker used to conduct a survey of Germany dialects between 1876 to 1887 (~115 prompts)[10]
- Less frequent translation lists (English into Texas German):
    - Miscellaneous (GM): ~150 prompts
    - Short (GS): ~100 prompts[11]
    - Extra (GE): ~300 prompts
- Biographical interview (BI): Participants' responses to a set of biographical questions, such as when and where they were born.

The amount of data for these eight speech event types is depicted in Table 1 below.

The lengths of the translation lists vary considerably. For example, the "Gilbert Long" (GL) translation list has over 1,200 prompts, while "Gilbert Short" (GS) has a little over 100 prompts. Some translation prompts appear in multiple translation lists.

Interview Type		Amount
Open-Ended (OE)		~ 51 interviews, ~ 11 hours
Translation	Atlas (GA)	~ 28 interviews, ~ 7 hours
	Long (GL)	~ 12 interviews, ~ 21 hours
	Wenker sentences (WS)	~ 16 interviews, ~ 2 hours
	Other (GM, GS, GE)	~ 9 interviews, ~ 1.5 hours
Biographical Interview (BI)		~ 49 interviews, ~ 8 hours

Table 1: GTXG Speech Events

3.3 Data Processing

3.3.1 Processing of Open-Ended Interviews

The GTXG open-ended interviews are segmented into smaller thematic chunks, manually transcribed and translated using the transcription software ELAN[12] and a tokenized layer, language tags, normalization, two layers of part-of-speech tagging (STTS 2.0 and Universal Dependencies POS), lemmatization, and a canonical phonetic transcription based on the transcribed forms are added automatically.[13] Almost all the GTXG open-ended interviews have been transcribed and are available via the ZuMult corpus platform at https://tgdp-zumult.la.utexas.edu/.[14]

[6] There is also one person who was recorded by both Gilbert and the TGDP, allowing for a case study of how a single person's language has changed over time.

[7] Please note that GTXG is referring to the Texas German recordings that Gilbert made in the 1960s and 1970s. In Table 1, "Gilbert" is referring to the collection of recordings the TGDP had made using Gilbert's elicitation list, but these recordings were all made after 2001.

[8] About a quarter of these participants are incidental speakers, i.e., people who were in the room when the primary participant was being interviewed, but who did not speak much themselves.

[9] See Gilbert (1963: 29-63).

[10] See Fleischer (2017) for more information about the Wenker sentences and surveys.

[11] See Gilbert (1963: 78-81).

[12] ELAN (https://archive.mpi.nl/tla/elan) at the Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands (see Brugman, and Russel 2004).

[13] These automatically added annotations are the same layers as are present in the TGDP Corpus annotations of contemporary Texas German and are created using the same pipeline.

[14] The only files that have not been released are the audio files from 2 events (1 in Wendish, 1 in Alsatian), as these files could not be properly quality checked yet.

3.3.2 Processing of Translation Interviews

When building spoken language corpora, focus is often put on more naturalistic speech. Within dialectology studies, however, elicitation lists are also commonly used, but they are often not made searchable in the same way more naturalistic spoken language often is. To make the speech events that were based on elicitation lists more accessible to researchers, the TGDP has developed a system to transcribe them. This transcription system involves several annotation tiers, facilitating different ways to explore the data.

The TGDP is transcribing the elicitation list speech events in phases, starting with the "Gilbert Atlas" speech events (GA). To do this, GA speech events were first segmented using the audio editing software Audacity (Audacity Team 2014). During this phase, students were asked to add labels to all the portions of the audio during which a speaker was providing an answer to a translation elicitation and label that portion of the audio with the appropriate elicitation number. If a student noticed that the interviewer asked for an elicitation which was not on the original list, they were instructed to add it to the list and give it a new number. This was to ensure that the prompt list accurately depicted what the informants were being asked to translate.

Following this step, the translation interviews were processed in two separate ways in parallel. To test the potential of Automatic Speech Recognition (ASR) for economizing on transcription time, the audio files were first processed with Whisper (Radford et al., 2023). For each section marked in Audacity, a suitable prompt of the following kind was generated and sent to the OpenAI platform alongside the corresponding audio.

Hier ist ein Satz in englischer Sprache: „This chicken has long feathers". Können Sie bitte diesen Satz in die deutsche Sprache übersetzen?
[Eng: Here's a sentence in English language: … Can you please translate this sentence into the German language?]

The expectation was that by providing such a highly restricting prompt, useful results could be achieved in spite of the suboptimal audio quality and the non-standard language used by the speakers. At first glance, results looked usable but nowhere near error-free. Students were asked to correct a sample of the results, confirming the findings in Gorisch and Schmidt (2024) that there is hardly a gain in efficiency over fully manual transcription with this type of ASR output. In other words, it took students the same amount of time to transcribe the audio from scratch as to correct a Whisper transcription, and the transcripts that the students constructed tended to keep some of Whisper's normalizations that the TGDP did not want in its transcripts (e.g., standardizing non-standard language, skipping over disfluencies, misunderstanding words, etc.).

Students were therefore asked to transcribe the GA speech events from scratch, focusing only on the portions of the audio that captured participants' responses.[15] These transcriptions were created in ELAN and had four kinds of layers: (1) a transcription tier for each speaker, (2) a prompt tier (PROMPT), (3) a prompt number tier (NO), and (4) a target hypothesis tier (TH). To avoid bias, students were not given the Whisper-generated transcriptions. This resulted in files that looked like Figure 1 below.

In the transcription tier, speakers' utterances were transcribed using modified German orthography to try to capture participant's actual speech with deviations from the norm. The target hypothesis (TH) tier provides an interpretation of the actual speech into the translation the speaker "intended" without any disfluencies or interactive elements of the performed speech. For example, if a speaker said "Uhhm … Geschirr- Geschirrschüssel" as a response to the question "How do you say 'sink' in Texas German?", then the Target Hypothesis would simply be "Geschirrschüssel", i.e., the participants complete "final answer".

Several special characters were used to mark the following special situations (see Table 3 below for a summary). If a participant produced multiple answers, all their answers are listed in the TH layer, separated by a pipe (|). If a word or set of words is "offered" by the interviewer or another speaker and then "accepted" by the participant, each of the "accepted" words are marked with a %. Any words that are implied by the context but not explicitly said by the speaker are marked with a plus sign (+). If the speaker does not provide a translation, then /NO_ANSWER/ is written in the TH tier. Making these distinctions anticipates different ways of analysing the data where, for instance, suggested words will have to be ignored in one approach but will be relevant in another.

[15] This decision was made in an effort to capture the most core data while reducing transcription time. In a future phase, the interviewer's audio could be transcribed as well. This would be helpful for studies interested in e.g., what participants struggle with (for example, by looking at how long it takes a participant to start answering once a prompt is given).

The special symbols can be transferred to explicit annotations in later processing.

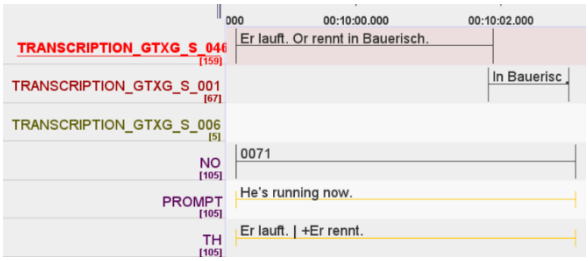

Figure 1: Screenshot of Excerpt from GTXG_E_024_SE_GA Transcript in ELAN.

1. *Er lauft. Or rennt in Bauerisch.*
   he runs or runs in Bauerish
   'He runs. Or 'runs' in Bauerish (a language variety)'

The example in Figure 1 above captures two of these situations. First, the participant offered multiple translations, both of which are captured in the TH layer, separated by a pipe (|). Second, the "er" in "+Er rennt." is marked with a "+" because the *er* 'he' is the implied subject based on the participant's first translation. The speaker does not actually say *er rennt*—he only implies that *er* would be part of his second response.[16]

Target Hypothesis Symbol	Meaning
|	The participant offered multiple translations, and this symbol separates each of them.
%	Someone other than the participant "offered" this token to the participant, and the participant accepted it.
+	The participant did not literally say this token, but it is implied by the surrounding context.
/NO_ANSWER/	The participant does not provide a translation for this prompt.

Table 2: GTXG Target Hypothesis Symbols

As mentioned above, the TGDP is beginning their transcription of elicitation speech events with the "Gilbert Atlas" speech events (GA). The GA portion of the GTXG corpus includes 32 Texas German speakers across 22 events and a list of ~230 translation prompts.[17] There is a total of ~25,000 transcribed tokens and 2,469 recorded target hypotheses.[18] The occurrence of the TH special symbols is summarized in Table 3 below, which can be read as follows: The pipe symbol "|" occurred a total of 68 times in the GA portion of the GTXG corpus. These occurrences spanned 50 different elicitation questions. In other words, there were some elicitation numbers for which multiple speakers provided multiple translations, but generally speaking, participants provided multiple translations for different elicitation prompts. The 68 tokens of the "|" symbol were produced by a total of 20 speakers, with some participants only offering multiple translations once, while other participants produced multiple translations for 6-8 prompts.

This kind of information can aid research on participants' confidence and flexibility. Information about what participants do *not* answer can help shed light on which lexical items and syntactic structures may be more challenging for participants.

Symbol	Tokens	Elicitations	Speakers
|	68	50	20
%	309	80	26
+	193	39	25
/NO_ANSWER/	118	47	24

Table 3: Target Hypothesis Symbols in the GA Portion of the GTXG corpus

3.4 Metadata

The recordings that Gilbert donated to the TGDP did not come with consistent metadata. Therefore, much of the interview/speaker metadata must be gleaned from the audio recordings themselves, as well as relevant publications. There is basic metadata available for ~20 participants (based on Gilbert 1972) and more comprehensive metadata for 7 participants (~30 questions, Gilbert 1963).

4. Example Queries

The availability of (transcribed) Texas German free conversation and translation elicitation audio from both the 1960s and from the last 25 years allows for various kinds of comparative research. On the one hand, one can use the open-ended data to explore how conversational language may

[16] The annotation schema described in this paper for translation elicitations is a novel annotation schema. Guidelines with more examples will be made available when this dataset is made public later this year.

[17] Originally, there were about ~150 prompts in Gilbert's prompt list, including some built in repetitions of prompts. In the process of identifying the individual responses for transcription purposes, another ~80 prompts were identified, often in the form of clarification questions, rephrases of prompts (e.g., 'the cow' vs. 'what do you call the animal that you get milk from'), or small deviations from the original prompts.

[18] Note: In the current iteration of the transcripts, the interviewer's prompt is not transcribed.

have changed within the last 50+ years. On the other hand, one can compare the results from different elicitation methods (conversational data vs. elicited data) to see whether and how results may vary based on elicitation type. Brief examples of both kinds of research are provided below (see also screenshots in the appendix).

4.1 Diachronic Comparison

One way of using the available Texas German audio recordings is to compare how language has (or has not) changed since the 1960s. For example, if one were interested in looking at whether /ɹ/ (a voiced alveolar approximant) has increased in usage since the 1960s, one could search all normalized tokens that contain <r> and listen to their phonetic realization (CQP: `[norm=".*r.*"]`). This could be of interest because /ɹ/ was not present in the German phonological inventory of the German dialects that the majority of German-speaking immigrants who came to Texas spoke, but it is prevalent in American English, and thus could indicate an increasingly Americanized pronunciation of German lexemes due to prolonged contact with English.

Another aspect of language change that could be explored would be to look at the percentage of English loanwords present in Texas German speech. On a basic level, one could conduct the query `[lang="deu"][lang="eng"][lang="deu"]` to find all tokens tagged as English that are surrounded by a token tagged as German on either side.[19] This search finds 11,237 hits in the contemporary TGDP data (i.e., 0.75% of the 1,494,400 total tokens) and 28 tokens in the GTXG corpus (i.e., 0.03% of the 96,309 tokens). Although a more in depth look at the data is necessary, this could indicate that (a) there is a relatively low number of English loanwords and/or nonce borrowings in Texas German, despite it being in contact with English for over 150 years and (b) the number of individual English loanwords in Texas German has increased over the last 50 years.

4.2 Free Conversation vs. Translation Elicitation

Translation elicitations can be a helpful tool in that they ensure that a participant is asked to produce a particular set of lexemes/features/constructions that may not have (frequently) occurred in a less structured conversation.[20] They can also ensure that a set of participants are all asked the same questions, making it easier to compare participants' responses. It is also possible, however, that the prompt itself influences how a participant produces a particular utterance (cf. Bailey et al. 1997, Munro 2022, Vintoniak et al. 2024). In order to get a more thorough picture of a particular variety, it can therefore be helpful to look at data that was collected using multiple elicitation strategies.[21]

For example, one could look at all the translations of prompts "There is the man who I want to see" and "There are the children who I gave the candy to" to get one view of how Texas Germans produce relative pronouns (i.e., how they choose to translate 'who'). Then, one could search for all the relative pronouns in the conversational transcripts and see if speakers used the same strategies within the same contexts (e.g., animate vs. inanimate referents).

5. Conclusion and Outlook

We have introduced the spoken language resources compiled by the Texas German Dialect Project, particularly highlighting the new addition of historical recordings from the 1960s. These data, in addition to the contemporary data from the last 25 years, can facilitate a diachronic exploration of Texas German and support comparative studies.

The open-ended interviews from the historical data (GTXG) were published to the ZuMult platform in December 2025. In 2026, the first phase of the historical elicited translation data will be added (GA speech events within GTXG), and the platform will be extended with functionality to make optimal use of the comparability inherent in this part of the data. In the mid-term, the TGDP will begin processing the corresponding translation data from the (much larger) "modern" corpus in a similar manner with the goal of making it available in ZuMult. In this context, we will evaluate the ASR results in more detail by comparing them to the manual transcriptions currently being created. This should give us a clearer idea if and how ASR can be made fruitful for this type of data.

The historic dimension in the data is currently already made use of in the project "Let the People of the Past Speak! Turning Migrant Letters of the 19th Century into Speech", which aims to reconstruct the language and speech of German immigrants in the 19th century. Comparing the Texas German data from the 21st century to the older recordings is used as a basis to extrapolate

[19] How the search tool is currently configured, queries automatically only search participant transcriptions (i.e., do not include interviewer utterances). Interviewer utterances can, however, be included in searches.

[20] There are many kinds of (linguistic) elicitation strategies, see e.g., Arunachalam (2013), Chambers and Trudgill (2004), and Kainada and Baltazani (2013).

[21] Lenz (2016) is one of many researchers who suggests it is best to use a variety of elicitation methods when investigating a particular set of linguistic phenomena.

certain developments (e.g., dialect levelling, language mixing) to an earlier stage.

6. Availability

The TGDP Corpus data, including the GTXG data, that are available via https://tgdp.org/ and https://tgdp-zumult.la.utexas.edu/ are available for research and teaching purposes. Use of the data for commercial purposes or training of AI requires additional agreements to be negotiated with the TGDP director. More information about user rights and responsibilities are listed at https://tgdp.org/dialect-archive/.

7. Acknowledgments

Over the years, the Texas German Dialect project has profited from financial support from the University of Texas at Austin, the National Endowment for the Humanities, the Texas Council for the Humanities, the Consul General of the Federal Republic of Germany in Houston, Texas, the Alexander von Humboldt Foundation, the Raymond Dickson, Alton C. Allen, and Dillon Anderson Centennial Professorship, the Alkek Foundation, the Texas German Dialect Project Endowment, the Dan L Duncan Foundation, and many private donors who wish to remain anonymous. The project "Let the people of the past speak" is funded by the Volkswagen foundation under grant numbers 9E083 and 9E529. The corpus platforms for the TGDP data were constructed in collaboration with the Liberal Arts Instructional Technology Services at the University of Texas at Austin and linguisticbits.de. We would like to thank all TGDP members and student assistants for their contributions to the project, and, especially, the speakers of Texas German for their willingness to provide their data.

8. Ethics Statement

All participants were informed about the project and the intended use of the recordings. Written consent was obtained for recording, transcription, and inclusion of their data in the corpus. Any directly personally identifying information was removed during transcription and in the audio. When making data available, we are complying with FAIR principles, observing the restrictions resulting from the informed consent. We are also respecting CARE principles by allowing participants to decide whether their recordings are made publicly accessible, and by giving participants the option of having copies of their unedited recordings for their own personal use.

9. Bibliographical References

Altenhofen, C. (submitted). Hunsrückisch in Brazil: The dynamics of a German variety outside of the "Heimat". In Boas, H., Blevins, M., P. Wolf-Farré (Eds.), *German language islands in Latin America*.

Arunachalam, Sudha (2013). Experimental Methods for Linguistics. *Language and Linguistics Compass* 7(4): 221–232, 10.1111/lnc3.12021 Audacity Team (2021).

Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 3.0.0 retrieved March 17th 2021 from https://audacityteam.org/

Bailey, G., Wikle, T., and Tillery, J. (1997). The effects of methods of results in dialectology. *English World-Wide* 18(1), pp. 35-63.

Blevins, M. (2022). *The language-tagging & orthographic normalization of spoken mixed-language data, with a focus on Texas German*. Ph.D. thesis. The University of Texas at Austin.

Boas, H. (2009). *The life and death of Texas German*. Duke University Press, Durham.

Boas, H. (2016). Variation im Texasdeutschen: Implikationen für eine vergleichende Sprachinselforschung. In A. Lenz (Ed.), *German Abroad: Perspektiven der Variationslinguistik, Sprachkontakt- und Mehrsprachigkeitsforschung*, 11–44. Vienna University Press.

Boas, H. (2021). Zwei Jahrzehnte digitale Dokumentation und Erforschung eines aussterbenden deutschen Auswanderer-dialekts: Das Texas German Dialect Project (2001-2021). *Zeitschrift für deutschsprachige Kultur und Literatur* 30, pp. 234–268.

Boas, H., Schmidt, T., & Blevins, M. (in press). A new corpus platform for the Texas German Dialect Project. *Language Resources and Evaluation*.

Born, R. (1994). *Michigan German in Frankenmuth: Variation and change in an East Franconian dialect*. Camden House, Columbia, South Carolina.

Brugman, H. & Russel, A. (2004). Annotating multimedia / multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2004), pp. 2065–2068, Lisbon, ELRA.

Chambers, J. K., and Trudgill, P. (2004). *Dialectology*. Cambridge University Press, Cambridge.

Eichhoff, J. (1979). Deutsche Sprache in Wisconsin. In L. Auburger, H. Kloss, & H. Rupp (Eds.) *Deutsch als Muttersprache in den Vereinigten Staaten, Teil 1 Der Mittelwesten. Deutsche Sprache in Europa und Übersee. Berichte und Forschungen. Vol. 4*. Franz Steiner Verlag, Wiesbaden, pp. 65–75.

Eikel, F. (1954). *The New Braunfels German dialect*. Manuscript. Johns Hopkins University.

ELAN (Version 7.0) [Computer software]. (2025). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan.

Fandrych, C., Schmidt, T., Wallner, F. & Wörner, K. (Eds.) (2023). *Korpora Deutsch als Fremdsprache* 3(1). Themenschwerpunkt:

Zugänge zu mündlichen Korpora für DaF und DaZ: Das ZuMult-Projekt Darmstadt: KorDaF.
Fleischer, J. (2017). Geschichte, Anlage und Durchführung der Fragebogen-Erhebungen von Georg Wenkers 40 Sätzen: Dokumentation, Entdeckungen und Neubewertungen. *Deutsche Dialektgeographie* 123. Hildesheim/Zürich/New York: Olms.
Földes, C. (2002). Kontaktsprache Deutsch: Das Deutsche im Sprachen- und Kulturenkontakt. In U. Hass-Umkehr, W. Kallmeyer, G. Zifonun (Eds.), *Ansichten der deutsche Sprache. Festschrift für Gerhard Stickel zum 65. Geburtstag*, pp. 347-370. Narr, Tübingen.
Gilbert, G. (1963). *The German dialect spoken in Kendall and Gillespie counties, Texas*. Ph.D. thesis, Harvard University.
Gilbert, G. (1972). *Linguistic atlas of Texas German*. University of Texas Press, Austin.
Gorisch, J. & Schmidt, T. (2024). Evaluating workflows for creating orthographic transcripts for oral corpora by transcribing from scratch or correcting ASR-output. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024). ELRA, pp. 6564–6574.
Guion, S. (1996). The death of Texas German in Gillespie County. In P. S. Ureland, I. Clarkson (Eds.), *Language and contact across the North Atlantic*, pp. 443–463. Niemeyer, Tübingen.
Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36, pp. 161–195.
ISO (2016). ISO 24624:2016 Language resource management — Transcription of spoken language.
Kainada, E. and Baltazani, M. (2013). Evaluating methods for eliciting dialectal speech. In *M. Janse, B. D. Joseph, A. Ralli* and *M. Bagriacik* (Eds.), *Proceedings of the 5th International Conference on Modern Greek Dialects and Linguistic Theory, Patras,* pages 101–123.
Kaufmann, G., Gorisch J., Schmidt, T. (2023). Das MEND-Korpus im Archiv für Gesprochenes Deutsch: Entstehung, Möglichkeiten, Grenzen. In P. Wolf-Farré, L. Löff Machado, A. Prediger, & S. Kürschner (Eds.). *Deutsche und weitere germanische Sprachminderheiten in Lateinamerika: Methoden, Grundlagen, Fallstudien*. Peter Lang, Berlin, pages 103–147.
Kühl, K., Heegård Petersen, J., Foget Hansen, G., and Gregersen, F. (2017). CoAmDa: et nyt dansk talesprogskorpus. *Danske Talesprog* 17, pp. 131–60.
Kürschner, S., Habermann, M., Prediger, A., & Pupp Spinassé, K. (submitted). Bohemian varieties in Southern Brazil. In Boas, H., Blevins, M., P. Wolf-Farré (Eds.), *German language islands in Latin America.*
Lenz, A. N. (2016). On eliciting dialect-syntactic data. Comparing direct and indirect methods. In A. Speyer, P. Rauth (Eds.), *Syntax aus Saarbrücker Sicht. Beiträge der SaRDiS-Tagung zur Dialektsyntax* (Zeitschrift für Dialektologie und Linguistik Beihefte. 165), pp. 187-219. Stuttgart: Steiner.
Munro, M. (2022). Variability in L2 vowel production: Different elicitation methods affect individual speakers differently. *Frontiers in Psychology 13.*
Nivre, J. de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4034–4043, Marseille, France.
Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. Doi: 10. 48550/arXiv.2212.04356
Schroeder, C., Lüdeling, A., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Grigoriadou, S., Hartz, R. G., Iefremenko, K., Jahns, E., Katsika, K., Keller, M., Klotz, M., Krause, T., Labrenz, A., Martynova, M., Özsoy, O., Pashkova, T., Pohle, M., and Zürn, N. (2024). RUEG Corpus. 10.5281/zenodo.11234583.
Vintoniak, O., Hnatyyuk, M., Miniailo, R., Turysheva, O., and Kotvytska, V. (2024). Dialectology in modern linguistic research: Theoretical approaches and methods. *Journal of Interdisciplinary Research* 14, pp. 39-44.
Wagener, P. (2002). German dialects in real-time change, *Journal of Germanic Linguistics* 14(3): 271–285.
Westpfahl, S. & Schmidt, T. (2016). FOLK-Gold – A GOLD standard for part-of-speech-tagging of spoken German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). ELRA, pp. 1493-1499.
Zimmer, C., Wiese, H., Simon, H., Zappen-Thomson, M., Leugner, J., Bracke, Y., Stuhl, B., & Schmidt, T. (2020). Das Korpus *Deutsch in Namibia* (*DNam*): Eine Ressource für die Kontakt- Variations- und Soziolinguistik. *Deutsche Sprache* 3, pp. 210-232.

10. Language Resource References

The Texas German Dialect Project. 2001+. *The Texas German Dialect Archive Online.* https://tgdp.org/

The Texas German Dialect Project. 2001+. *The Texas German Dialect Project Corpus*. distributed via ZuMult. https://tgdp-zumult.la.utexas.edu/

Gilbert, G., The Texas German Dialect Project. 1962+. *The Gilbert Texas German Corpus* distributed via ZuMult. https://tgdp-zumult.la.utexas.edu/

11. Appendix

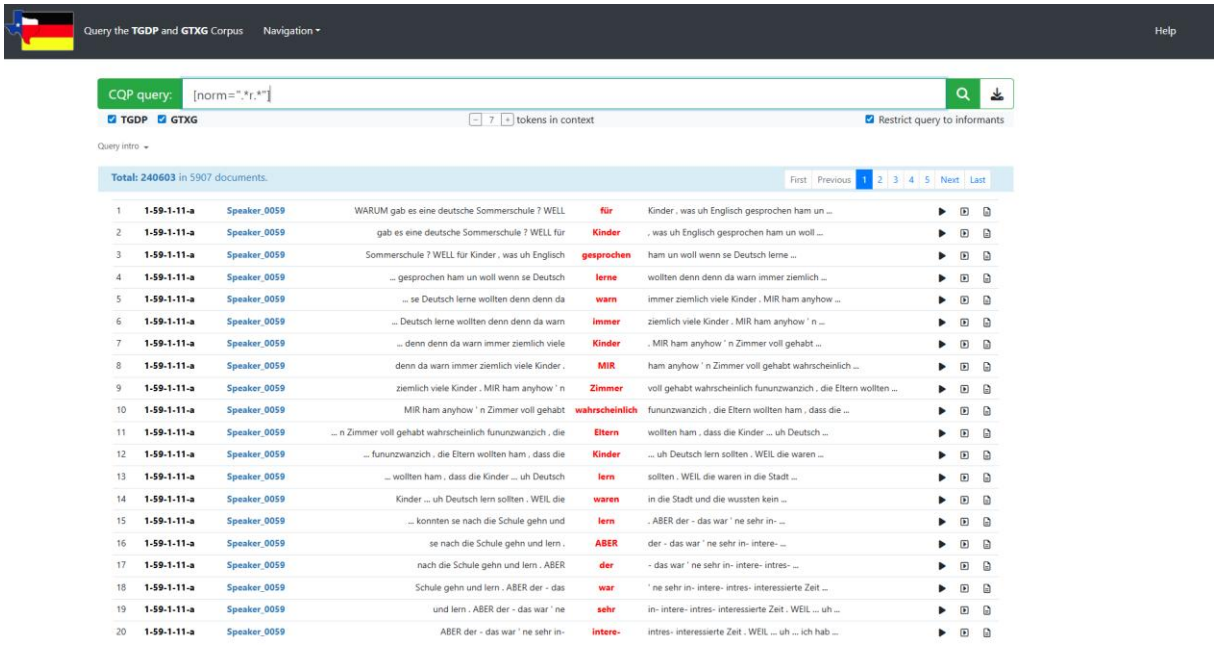

Figure 2: CQP Query `[norm=".*r.*"]` – finding all normalized tokens containing an -r-

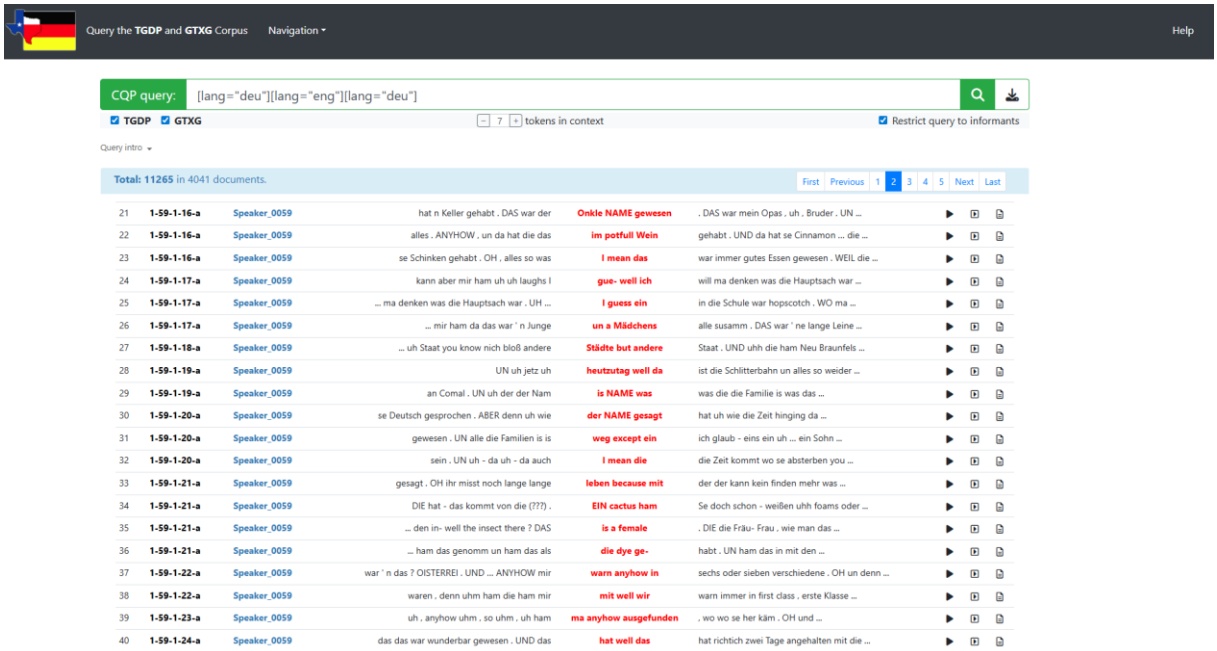

Figure 3: CQP Query `[lang="deu"][lang="eng"][lang="deu"]` – finding all tokens tagged as English that are surrounded by a token tagged as German on either side