Language Resources and Evaluation (2026) 60:52
https://doi.org/10.1007/s10579-025-09893-6

ORIGINAL PAPER

A new corpus platform for the Texas German Dialect Project

Hans C. Boas[1] · **Thomas Schmidt**[2] · **Margaret Blevins**[1]

Received: 22 February 2025 / Accepted: 3 September 2025

Abstract
Texas German is a contact variety that is the result of dialect mixing of several German dialects brought to Texas from central Europe starting in the 1830s. Since 2001, the Texas German Dialect Project has been assembling a large collection of spoken data documenting this unique variety. The present paper describes how a substantial part of this collection was developed into an annotated corpus and how the corpus is now available through a corpus platform based on the ZuMult technology. We start with an outline of the project's development and its established processes of data collection, transcription, and dissemination. We then explain the process by which the data were cleaned up and enriched with language tagging, orthographic normalization, lemmatization, and part-of-speech tagging. Finally, we illustrate how the new corpus platform makes these annotated data available for systematic browsing and querying. In the outlook, we sketch prospects for future development of the data and for their role in a larger landscape of comparable speech island data.

Keywords Texas German · Corpus platform · Language variation · Language documentation · Spoken language resource · Contact variety

✉ Hans C. Boas
hcb@mail.utexas.edu

Thomas Schmidt
thomas@linguisticbits.de

Margaret Blevins
mblevins@utexas.edu

[1] The University of Texas, Austin, TX, USA

[2] linguisticbits.de, Bingen, Germany

Published online: 24 June 2026

1 Introduction

According to Crystal (2000), half of the world's languages will be extinct by the end of the 21st century. Even though languages have always died out for a number of reasons, this rate of language death is unprecedented in human history and has led linguists to scramble for solutions to the various problems caused by this impending loss of linguistic diversity. To ensure the continued availability of data for a broad range of endangered languages and dialects, documentary linguistics has emerged over the past thirty years as "a fairly independent field of linguistic inquiry and practice that is no longer linked exclusively to the descriptive framework" (Himmelmann, 1998, p. 161). In contrast to descriptive linguistics, which focuses on the analysis of primary data, documentary linguistics concentrates on the collection, processing, and archiving of primary data to help produce a good documentation of a language while speakers still exist.

This increased interest in documentary linguistics over the past 30 years has led to an increasing number of linguists around the world becoming actively involved in documenting and archiving endangered languages and dialects. Some large language documentation projects with substantial long-term institutional support, such as The Language Archive (TLA, Drude et al., 2012), The Native American Languages Collection[1], and the Pacific and Regional Archive for Digital Sources in Endangered Cultures PARADISEC (Thieberger & Harris, 2022), consist of groups of linguists, programmers, and archivists. While some of these projects, such as AILLA (Kung & Sherzer, 2013), focus primarily on building and expanding online archives to house existing linguistic field recordings and their accompanying materials, other projects such as the Endangered Languages Archive (ELAR, Nathan, 2014) and PARADISEC are also conducting linguistic fieldwork and developing computational infrastructure to support linguistic fieldwork and archiving. At a minimum, these archives make language data available through online catalogues where collections are classified according to their metadata, and audiovisual primary data can be downloaded alongside any transcripts or annotations if they exist. If such collections are developed into fully-fledged corpora, they can be provided through corpus platforms with advanced query and browsing capabilities. While this is becoming more and more common for major languages (see, for instance, CLARIN's overview of spoken corpora[2]), it is still rare for collections of endangered languages and dialects (see, however, the Language DOcumentation REference COrpus, Seifart et al. (2024), or the resources for German contact varieties mentioned in Sect. 5).

Most language documentation projects publish the results of their linguistic research and insights into streamlining documentation efforts, but very few provide detailed insights into the evolution of their data resources and of the computational tools available to create, enrich, disseminate, and analyze them. It is undisputed, however, that these "technical" circumstances determine, to a large extent, the oppor-

[1] See https://samnoblemuseum.ou.edu/collections-and-research/native-american-languages/native-american-languages-collections/.

[2] https://www.clarin.eu/resource-families/spoken-corpora.

tunities and limitations of a resource's potential for reuse and archiving (see, e.g., Birds & Simons, 2003).

This paper therefore seeks to fill this gap by discussing the technical evolution of the Texas German Dialect Project (TGDP), which was founded at the University of Texas at Austin (UT Austin) in 2001 as an effort to record and archive Texas German, a critically endangered contact variety of German spoken in Texas since the 1830s (Gilbert, 1972; Boas, 2009). More specifically, we review how the project's data developed over the years, and we assess the various changes it has gone through since it was first reported in the pages of this journal twenty years ago (see Boas, 2006). In this context, we discuss how a substantial part of the TGDP's dialect archive has been transformed into a full-fledged corpus and made available on an online platform for conducting systematic corpus linguistic research, and we present several additional improvements to the original archive platform to ensure accessibility, interoperability, and reusability of the Texas German data.

The remainder of this paper is structured as follows. Section 2 provides some background information about the methodology and aims of the Texas German Dialect Project. Then, it presents an overview of the technical organization and implementation of the Texas German Dialect Archive Online (TGDA Online), which first went online in 2002. The last part of Sect. 2 discusses several technical and linguistic issues that have come up over a period of twenty years and it shows what types of limitations these issues impose on the usability of the archived data for accessibility and linguistic research. Section 3 first summarizes the main insights of Blevins (2022a), who proposes a novel methodology for normalizing Texas German transcription data. Then, we discuss how Blevins' (2022a) approach can be used for streamlining and cleaning up the archived Texas German transcriptions so they can become part of a systematic corpus.[3] Section 4 presents the results of our efforts since 2023 to develop a new platform (ZuMult) for browsing and querying the resulting corpus of interviews with Texas German speakers. Finally, Sect. 5 summarizes the main insights of the paper, discusses the structure and evolution of the TGDP's efforts to make their Texas German data accessible, and makes suggestions for future research.

2 The "original" archive since 2002

In this section, we introduce the project responsible for building the TGDA Online and we outline the digital methods available for accessing its data before the corpus and platform described in the subsequent sections were developed.

2.1 The Texas German Dialect Project (TGDP)

When the first author of this paper, Hans C. Boas, started his new position at the University of Texas in 2001, he noticed that there had been no significant linguistic

[3] In the terminology of Wamprechtshammer et al. (2022, pp. 11ff), this amounts to raising the data's "maturity level" from "Data Maturity Level I – Structured Data Set" to "Data Maturity Level II – Structured Data Set + Structured Data".

research done on Texas German since Glenn Gilbert's pioneering work in the 1960s, which resulted in, among other things, his groundbreaking *Linguistic Atlas of Texas German* (1972). Since the intergenerational transmission of Texas German had effectively stopped by the late 1950s and Texas German was considered an endangered dialect, Boas decided to found the Texas German Dialect Project (TGDP), with the explicit goal to document and archive the remnants of this rapidly eroding German contact variety.[4]

2.1.1 Why is Texas German so interesting?

Texas German is a contact variety that is the result of dialect mixing of several German dialects brought to Texas from central Europe starting in the 1830s. Over the decades of contact with English, Texas German has borrowed many words, phrases, and grammatical constructions from English. In the early 20th century, Texas German was spoken by over 100,000 people across central Texas, primarily in rural areas, but also in larger cities such as Austin and San Antonio. Up until World War I, German enjoyed considerable prestige across central Texas, and the institutional support included German-language schools, churches, singing societies, sports clubs, and newspapers (for an overview, see Nicolini, 2004 and Boas, 2009). With the anti-German attitudes accompanying the entry of the U.S. into World War I, German was effectively pushed out of the public sphere, and it was only spoken at home with family and among friends and neighbors.

During the 1920s and 1930s, Texas German existed in a diglossic relationship with English, the prestige language. The anti-German sentiments during World War II had an additional impact on the prestige of Texas German. As a result, almost all Texas Germans decided to not raise their children with German at home anymore, effectively switching to English. Today, there are almost no speakers of Texas German who learned the dialect at home after the 1950s. By the 1960s, the number of Texas German speakers dropped to about 70,000 and language shift was in full swing. At the beginning of the 21st century, there were an estimated 8–10,000 speakers of Texas German left (Boas, 2018); as of 2026, there exist an estimated 3,000 speakers, most of whom are 80 years and older. Current estimations point to 2035–2040 as the period when Texas German will most likely become extinct (Wilson, 1986; Boas, 2009).

Texas German is linguistically interesting for a variety of reasons. First, in contrast to many other German contact varieties around the world, which are the result of a single German dialect brought to a new location outside of central Europe, Texas German is the result of significant dialect mixing. Second, Texas German, whose beginnings are in the 1830s, is a relatively young contact variety of German in con-

[4] As outlined in Boas (2002), the principles of "Findability" and "Accessibility" and, up to a point, also "Interoperability" and "Reusability" were thus observed long before they became codified in the FAIR principles (Wilkinson et al., 2016). A large part of the later work described in this paper can be understood as stepwise improvements in the FAIRification of Texas German data.

trast to many other German contact varieties in places such as Russia, Romania, Hungary, Pennsylvania, Ohio, and Wisconsin. Third, Texas German is not a focused New World variety of German, according to Trudgill's (2004) theory of new-dialect formation. On the basis of Gilbert's (1972) data as well as more recent data, Boas (2009) demonstrates that Texas German did not go through all four stages of Trudgill's model of new dialect formation, effectively stopping half-way through the third stage in the 1920s, when intergenerational transmission started to deteriorate. Fourth, and perhaps most interesting, Texas German is extremely variable. The degree of inter-speaker and intra-speaker variation is extremely high, and Boas (2009) argues that it is, in fact, impossible to characterize Texas German as a coherent variety, instead suggesting that it should be labeled as a collection of different German varieties spoken in central Texas, possibly even going as far as saying that each speaker of Texas German has his or her own unique linguistic system that sets them apart from other Texas German speakers (for more information, see also Boas, 2018).

2.1.2 Organization and methodology underlying the TGDP

The TGDP's first linguistic fieldwork started in October 2001, after obtaining clearance from the Institutional Review Board at the University of Texas to conduct research with human subjects.[5] This fieldwork resulted in sociolinguistic interviews with five speakers from New Braunfels and Fredericksburg. Linguistic fieldwork continued in the New Braunfels and Fredericksburg areas and in subsequent years covered most of the areas of central Texas in which Texas German was/is spoken. With more students learning about Texas German through linguistics courses at UT Austin, they became interested in conducting linguistic fieldwork themselves. From 2001 to 2004, the focus of linguistic fieldwork was exclusively on sociolinguistic interviews. After that, the types of data elicited became more varied. Since 2005, members of the TGDP have been eliciting a variety of different data from speakers of Texas German. An interview session typically consists of three different parts:

- The first part is a 10-page long biographical questionnaire (in English). Speakers are asked over 100 questions pertaining to their life, language use, and language attitudes. For instance, speakers are asked where they were born, whether they know where their ancestors came from in Europe, which language(s) they spoke with whom growing up, and whether they formally studied German in school. This interview lasts about 25–40 min. These data are elicited for two reasons: (1) to gather a better understanding of a speaker's personal background, including social and family relationships, and (2) to determine a speaker's self reported

[5] Before the start of an interview, each speaker reads and signs a consent form that allows members of the TGDP to record the interview and to process and store the recordings in an online repository for educational and research purposes. The data may be shared for research purposes but may not be used for commercial purposes.

proficiency in Texas German. We are also interested in knowing how they speak English so we can more generally identify the types of speech patterns (such as false starts, hesitations, repairs of sentences, etc.) that are unique to each speaker, but not to the language they speak.

- The second part is an open-ended sociolinguistic interview where Texas German speakers are asked to speak about their lives. Interviewers typically ask their questions in Standard German (where some interviewers are native speakers, while others are not), and interviewees are asked to reply in Texas German (although they may switch to English if they are having difficulties saying what they would like to say in Texas German). Questions for the interviews cover a variety of topics such as personal and family background, the community, family history, jobs, the weather, schooling, cooking, religious practices, etc.[6] We seek to elicit natural conversations in Texas German as far as possible in this setting.[7] This part of the interview usually lasts about 30–60 min.
- The third part is a translation task in which speakers are asked to translate sets of words, phrases, and sentences from English into Texas German. The elicitation lists are based on Eikel (1954), Gilbert (1972), and Guion (1996). These lists were chosen because they were previously used to study Texas German, and in the case of Gilbert and Guion, original research materials are still available, making diachronic studies of Texas German possible. In fact, in 2018, TGDP members located and re-recorded two Texas German speakers who were interviewed by Glenn Gilbert in the 1960s for his *Linguistic Atlas of Texas German* (see Warmuth, 2023). For more details on the interview process, the data management plan, and how the interview data are used for research, teaching, and outreach activities, see Boas et al. (2010); Boas (2021a).[8]

After completing the field recordings, TGDP members segment the audio files into smaller sections for processing. These interview sections are then assigned unique file ID numbers. The narrative interviews are then transcribed (using modified German orthography) and translated word-for-word into English using ELAN (see Boas, 2006 for details). Both of these annotation layers are time-aligned in chunks of typically 2–5 seconds and manually transcribed as span annotations. After transcription and translation of the narrative interviews are completed, they are published on the TGDA Online (*tgdp.org/dialect-archive/*, see Sect. 2.2).

[6] The TGDP has a 10-page list of potential questions interviewers can use during this interview, but they are also given the freedom to come up with their own questions.

[7] These interviews would fall into Himmelmann (1998) 'staged communicative event' category. Staged communicative events are "communicative events that are enacted for the purpose of recording. The important difference between these kinds of communicative events and [observed communicative events] pertains to the fact that staged communicative events are not 'really' communicatively functional, that is, they do not serve any specific communicative purposes other than producing data" (Himmelmann, 1998, p. 185), cf. also the "observer's paradox" (Labov, 1972).

[8] The order in which these interviews are conducted can vary from interview session to interview session. Also, certain interviews are sometimes cut short due to time restrictions, or if, for example, an informant becomes fatigued or noticeably frustrated.

The training of student research assistants requires a considerable amount of time and effort.[9] Between 2001 and 2026, over 120 undergraduate and graduate students have participated as interviewers of Texas German speakers and over 50 students have transcribed and translated interview audio recordings (some students only conducted interviews, other students only transcribed and translated the interviews, while other students were involved in both). Some students worked on interviewing, transcribing, and translating only for a semester, while others decided to stay on for multiple years. In several cases, students became so interested in investigating Texas German that they wrote their undergraduate honors theses (e.g., Rybarski, 2006; Amick, 2020), their M.A. theses (e.g., Bathe, 2005; Whitworth, 2005; Thompson, 2005; Blevins, 2014; Jones, 2022), or even their dissertations (Roesch, 2009; Blevins, 2022a; Warmuth, 2023) on Texas German.[10]

2.1.3 How much data has been collected so far?

The vast majority of the recordings conducted by the Texas German Dialect Project were made between 2002 and 2025. As of February 2026, the Texas German Dialect Archive (TGDA), i.e., the TGDP's entire collection of Texas German audio recordings (both historical and contemporary) includes 850+ Texas German speakers who have been interviewed by the TGDP itself.[11] The range of birth years of the TGDP's participants spans from 1903 to 1975. Table 1 below represents the holdings within the Texas German Dialect Archive from the last ~40 years, which includes the materials that are publicly available via TGDA Online and/or the ZuMult platform

[9] Students were trained by being shown how to use ELAN and by being given a set of project-internal transcription and translation guidelines. Since training was managed by different researchers over the last 25 years, training has slightly differed from researcher to researcher. For example, approximately eight years ago, a document was created that lists several sounds/words that transcribers often have issues with, and this document is shared with all new transcribers. Around the same time, several transcribers also mentioned that it was easier to first work on segmenting the audio (a processing step that comes before transcription) before being asked to transcribe audio, so that they could familiarize themselves with how Texas German sounds before being asked to capture it in detail. Thus in subsequent years, we have tried to ensure that students have the opportunity to segment audio before being asked to transcribe. There is not a common set of test items to transcribe.

[10] Since its inception in 2001, the TGDP has faced a number of significant challenges, including the following: (1) starting a language documentation project from scratch without relevant experience; (2) obtaining adequate funding to support the recording of interviews and for hiring programmers to build the TGDA Online and related digital infrastructure; (3) finding qualified students to help with conducting interviews as well as processing the recordings; (4) receiving adequate institutional recognition during the academic process of promotion and tenure for starting and running a documentary linguistics project from scratch (for the most part, the gold standard remains peer-reviewed publication of monographs and articles); and (5) finding Texas German speakers to interview (this has become especially problematic in the years since the Covid-19 pandemic), as this is the last generation of speakers.

[11] The TGDA also includes a handful of audio recordings from 1989, 1992, and 1994 that were donated to the TGDP and are thus outside the project's general time frame. Information on recording years can be filtered in the ZuMult platform (see Sect. 4). The TGDA also contains additional historical collections of Texas German data, such as the recordings from the 1960s and 1970s from Glenn Gilbert.

Table 1 The audio recording holdings within the Texas German Dialect Archive from the last ~40 years, as of February 2026

Interview Type		Amount	
Open-ended		~945 interviews, ~ 315 h	(~45%)
Translation	Eikel	~160 interviews, ~ 40 h	
	Gilbert	~780 interviews, ~ 195 h	~302 h (~43%)
	Guion	~350 interviews, ~ 58 h	
	Other	~35 interviews, ~ 9 h	
Biographical Questionnaire (usually in English)		~490 interviews, ~ 82 h	(~12%)

(see Sect. 2.2 and 4 below, respectively) as well as the materials that are currently project-internal.[12]

For various reasons, not all of the audio recordings in the TGDA are currently online. About 75 interviews are not online yet because, for example, they have not been segmented. Once an open-ended interview is segmented, each individual segment must first be transcribed and personally identifying information, such as peoples' names, should be bleeped before the segment can be made "public." In addition to this, there are currently 700+open-ended interview segments that have not been transcribed yet. Figure 1 illustrates the different resources that are described in this paper and how they relate to one another. Table 2 (see Sect. 4.1) provides numbers for the portion of the TGDA that is currently publicly available via TGDA Online and/or the ZuMult platform.

2.2 Accessing the Texas German Dialect Archive Online (TGDA Online)

After the initial five speakers from New Braunfels and Fredericksburg were interviewed by the TGDP in the fall of 2001, these interviews were transcribed and translated by three graduate students at UT Austin using the then still relatively new transcription software ELAN (Wittenburg et al., 2006), developed at the Max Planck Institute for Psycholinguistics, The Language Archive in Nijmegen, The Netherlands (see Boas, 2003 and Boas, 2006 for details). In early 2002, two undergraduate students designed and implemented a prototype database and website to make the sound recordings accessible for teaching and research purposes (see Boas, 2002), thus creating the first version of the TGDA Online. Based on feedback from usability studies conducted with students in two linguistics classes at UT Austin, programmers from the Liberal Arts Instructional Technology Services (LAITS) at UT Austin implemented some revisions to the database and website in 2003. Since then, the architecture and data structure of the TGDA Online, i.e., the online portal available at

[12] Texas German speakers were occasionally interviewed multiple times, which can explain how there may be more open-ended interviews than there are Texas German speakers in the TGDA. The time estimates in Table 1 are based on the following: open-ended interview=~20 min; Eikel and Gilbert interviews=~15 min (with some interviews overlapping because multiple people were interviewed at the same time); Guion and other interviews=~10 min; biographical interview=~10 min. These are quite rough estimates, since, for example, open-ended interviews have ranged from 10 min to over 2 h, translation interviews have ranged from 10 min to 1 h, and biographical interviews have ranged from 5 min to an hour. Biographical questionnaire recordings are currently not available online. Table 1 does not include smaller, historical corpora, such as Gilbert's 1960s Texas German collection.

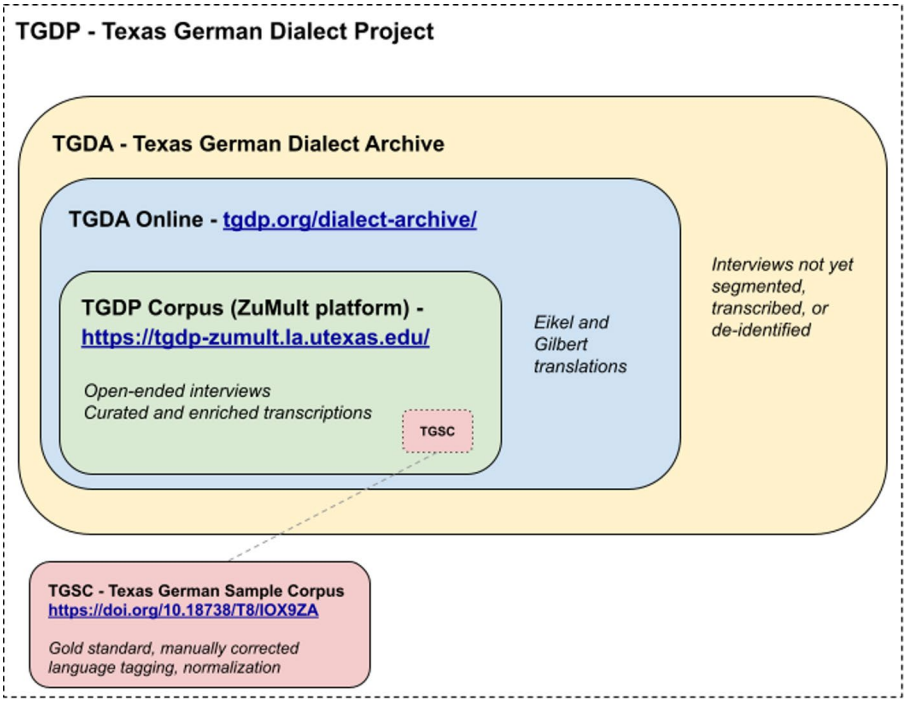

Fig. 1 Resources created by the Texas German Dialect Project

tgdp.org/dialect-archive through which the public can access Texas German recordings and transcripts, have basically remained the same, except for an update in 2015, when LAITS programmers made several upgrades to the database, mainly for security reasons.

Since 2002, users of the TGDA Online have been able to access audio for translation interviews (based on Eikel and Gilbert's translation lists), as well as audio and transcriptions/translations of open-ended interviews.[13] From the homepage of the TGDP (tgdp.org), users click on "dialect archive" and then "enter archive", which leads them to a log-in page where they provide a username and a password to enter the archive.[14] Clicking on "open-ended interviews" leads users to a new page with a map of central Texas that lists all the locations for which open-ended interviews are available. An alphabetical list of locations where interviews have occurred appears under the map, see Fig. 2. Users can click on a location on the map or the name of a location in the list. That brings up a popup window that lists all publicly available interview segments for that location. Each interview segment has a descriptive title summarizing the content of the interview segment (e.g., "Helping parents on the

[13] The audio files of the Texas German translations have not been transcribed. See Boas (2006) for more details on accessing and using the Gilbert and Eikel translation data. Translation data based on Guion's (1996) elicitation list are not currently available online.

[14] Users are required to register for the archive because of regulations regarding the access of research data.

farm" or "Grandfather was half Czech, half Austrian" or "Smoking the pig") and its length in minutes and seconds.

Clicking on the title of an interview segment opens an embedded audio player that plays the recorded segment. Clicking on "view transcript" opens a new window with an embedded audio player, together with a transcription (in modified German orthography) and its word-for-word English translation for that interview section, so that users can read the transcript while listening to the recording. The transcription is displayed in HTML, the audio is in MP3 format. For each interview segment, users also have the option of downloading the higher quality WAV sound file and its transcription as an EAF file so that both can be opened using ELAN to conduct further linguistic analyses of the data (for details, see Boas, 2006).

As of February 2026, the TGDA Online has over 5,000 registered users from around the world, with most users in the United States, Germany, and Austria. Even though the freely available TGDA Online has been a tremendous resource for supporting teaching linguistics and history classes and for providing primary source data for linguistic analyses that formed the basis for theses, articles, and books over the past 25 years, it has had several limitations since it first went online.

First, the TGDA Online only offered limited browsing functionality for sections of interviews. For example, while it was possible to listen to and view audio recordings and transcriptions (as well as their translations), one had to select interviews by first clicking on an interview location. It was not possible, for example, to search for

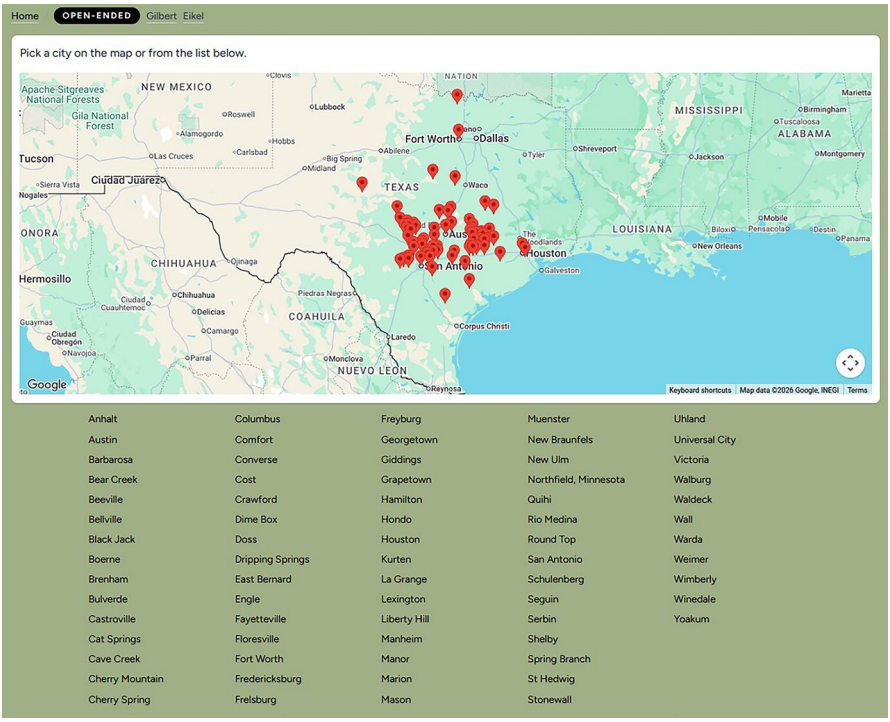

Fig. 2 Map interface for exploring open-ended interviews (available since 2002)

sections relating to a particular topic (e.g., Christmas traditions, how to make sausage, etc.). Second, the site did not offer any systematic way to search the transcripts. Third, there was no way to query speaker metadata. Fourth, the site did not offer any publicly available information on different types of corpus statistics, i.e., total number of speakers, tokens, interviews, etc. Fifth, behind the scenes, the transcriptions contained a lot of variation—both in their file structure, and in the transcriptions themselves—and both kinds of variation made it difficult to conduct large-scale linguistic analyses. This variation was caused by a variety of factors, including different versions of the transcription software ELAN as well as the great number of students transcribing and translating the recordings. Over time, the number and types of inconsistencies grew to a substantial amount. In the sections below, we discuss some of the causes of these variations, as well as how we approached them in our effort to clean up and upgrade the data.

3 From collection to corpus: upgrading the transcription data

In the fall of 2013, Hans C. Boas collaborated with the programmers of LAITS at UT Austin to devise a plan for upgrading the backend of the TGDA Online because most of its technical infrastructure had remained relatively unchanged since 2003. Between 2013 and 2015, LAITS programmers focused primarily on upgrading the database and website to make them compliant with then up-to-date security standards and other important web-based access requirements. In addition, the TGDP's project-internal online administrative and data management interface was revamped to ensure a smoother storage procedure and processing of interview segments.

Since 2016, LAITS programmers started various initiatives to streamline and clean up the many inconsistencies in transcriptions of interviews. They also planned on developing a sophisticated query interface that would allow users to search the transcripts of the interviews in a systematic way. These efforts, however, did not immediately succeed. First, there were significant fluctuations in programming personnel, which made it difficult to achieve progress. Second, when the Covid-19 pandemic broke out in the spring of 2020, almost all resources of LAITS were devoted to implement and support online teaching at UT Austin. Since 2013, Thomas Schmidt (then head of the Archive for Spoken German at the Leibniz Institute for the German Language) had been advising the TGDP in matters of data management and corpus technology. Starting in 2022, this cooperation was intensified by hiring him via linguisticbits.de as a contractor to carry out the tasks described in the following sections. The overall goal was (a) to turn the existing collection of open-ended interviews into a consistent and fully annotated corpus and (b) to make this corpus available for browsing and querying via a dedicated corpus platform. The TGDA Online is kept as a parallel platform for those users not approaching the data as corpus linguists.

3.1 File format cleanup

As was mentioned above, ELAN was chosen early on in the project as the solution for creating time-aligned transcripts of the recordings for the TGDP, so all transcripts are

available in this tool's EAF file format. However, changes in this file format over the years and adaptations to the transcription practices themselves have led to a certain amount of inconsistency in the data, and the fact that ELAN leaves many degrees of freedom in its data model also limits the original transcript's reliability for automatic processing. Before carrying out additional annotations, as described below, the existing transcription data were therefore carefully curated, making sure that tier structures and naming (for tiers, speakers, etc.) follow a common well-defined scheme. This scheme is now also available as a template, making sure that future transcripts will be consistent with the cleaned-up data.

The cleaned-up data could then be further processed automatically, with the likelihood of errors due to inconsistent naming and structures much reduced. As a first step, we converted the ELAN files to the ISO standard for transcriptions of spoken language (ISO, 2016; Schmidt, 2011; Hedeland & Schmidt, 2022), which is based on the guidelines of the Text Encoding Initiative (TEI). In that process, speakers' and interviewers' utterances were tokenized, and English translations and the original transcription text assigned as annotations to the tokenized utterances. Figure 3 provides an example of the resulting XML.

As required by the standard, individual tokens are marked up as <w>(for words) and <pc> (for punctuation) and can be referred to via a unique @xml: id attribute. No changes to the transcribed text are made during tokenization. Thus, according to the TGDP's project-internal guidelines, transcribers are instructed to capitalize all

```
<tei:annotationBlock start="t127" end="t131" who="Interviewer_0001">
   <tei:u xml:id="a55">
      <tei:seg type="contribution" xml:id="d68570e1329">
         <tei:anchor synch="t127"/>
         <tei:w xml:id="a55_w1">AUF</tei:w>
         <tei:pc xml:id="a55_p2">'</tei:pc>
         <tei:w xml:id="a55_w3">m</tei:w>
         <tei:w xml:id="a55_w5">Schiff</tei:w>
         <tei:pc xml:id="a55_p6">.</tei:pc>
         <tei:anchor synch="t131"/>
      </tei:seg>
   </tei:u>
   <tei:spanGrp xml:lang="en" type="translation">
      <tei:span from="a55_w1" to="a55_p6">On the boat.</tei:span>
   </tei:spanGrp>
   <tei:spanGrp type="original">
      <tei:span from="t127" to="t131">AUF 'm Schiff.</tei:span>
   </tei:spanGrp>
</annotationBlock>
```

Fig. 3 Transcript excerpt in ISO format

of the letters of the first word in each functional sentence (i.e., regardless of whether the sentence is actually completed or not). In the English translation however, they are instructed to just capitalize the first letter of the first word of each functional sentence. In this example, that is why 'AUF' is in entirely capital letters, while 'On' is sententially capitalized. Likewise, since the transcriber decided to transcribe the contraction 'auf'm' with an apostrophe, the tokenization divides this into two words and one punctuation token. If the alternative of transcribing this without the apostrophe had been chosen ('aufm'), a single <w> token would have been used in the XML. In normalization (see below), this variation is levelled out.

All subsequent processing described in the following sections is carried out on this representation of the data, with additional annotations added either as attributes to the <w> elements or as additional <spanGrp> elements. In a <spanGrp>, whose @type attribute can be freely defined (so other annotations could be added in an analogous manner), individual span elements contain the actual annotation label (e.g. the English translation "On the boat."), and their extent is given via @from and @to attributes pointing to the @xml: id of the first and last token the annotation refers to.

3.2 Adding and automating token annotation

Previous versions of the data only had the transcription and translation layers (span annotations) as a source for systematic queries. By adding tokenization, language tags, orthographic normalization, lemmatization, and part-of-speech tagging as additional annotation layers, the data lends itself more easily to corpus linguistic analyses. In what follows, we explain how these layers were added.

3.2.1 Language-tagging, orthographic normalization, and the Texas German Sample Corpus

As mentioned above, the original transcripts contained two layers of annotation per speaker: a literary transcription following modified German orthography and a word-for-word English translation, see example (1).[15]

(1) **TRANSCRIPTION**: UND der hat [jeder einer von seine Sohne] aufgesetzt … [in farming and ranching].
TRANSLATION: And he had each one of his sons put-in-charge … in farming and ranching.
(1-55-1-3-a)[16]

[15] The TGDP transcription guidelines also asked transcribers to put square brackets around any non-standard material, such as English words, non-standard grammar, and non-standard word order. Both the transcription and translations are span annotations.

[16] TGDA files that are not part of a smaller, separate corpus (e.g., Gilbert's 1960s audio) have a 4-digit ID number. The first number is the interviewer ID number, the second number is the speaker ID number, the third number is the interview number, and the fourth number is the interview section number. This example is from the third section of interview 1 (an open-ended interview) between speaker 55 and interviewer 1.

While this provided a good foundation for making the materials generally available and accessible, the high amount of orthographic variability in the German transcription made it difficult to conduct in-depth linguistic analyses. For example, a single Texas German audio clip could be transcribed in several ways:

(2) a. Ich wei**ß** nicht, wo **dir** Jung gew**ä**hnlich gehen **duh**.
b. Ich wei**ss** nicht, wo **die ihr** Jung gew**ee**nlich gehen **du**.
'I don't know where her son usually goes.'
Standard German: Ich weiß nicht, wo ihr Junge gewöhnlich hingeht. (Blevins, 2022a, p. 18)

As this orthographic variability is often unpredictable, it could make it challenging for people and computers alike to search the transcripts and find what they are looking for.

This orthographic variability has three main sources: (1) inter- and intra-speaker variation, (2) inter- and intra-transcriber variation, and (3) flexible project-internal transcription guidelines. Inter- and intra-speaker variation is common in any spoken corpus, particularly one in which there are so many donor dialects, as well as contact phenomena at play. This could result in, e.g., one speaker pronouncing *Apfel* 'apple' as [ˈapfəl] while another speaker could pronounce it [ˈapəl], which could lead to two different transcriptions of the same core lexeme, see Table 2.

Compounding this, an additional layer of potential orthographic variation occurs during the transcription phase. Transcribers could interpret and therefore transcribe the same audio clip differently, see Table 3.[17]

Additionally, the TGDP transcription guidelines were designed to be flexible to allow students to quickly learn how to transcribe. These project-internal guidelines were also developed before in-depth transcription systems for contact varieties were

Table 2 Example of orthographic variation in the transcription layer due to inter- and intra-speaker variation

	Pronunciation	Transcription	Standard Language Orthography
Speaker 1	[ˈapfəl]	Apfel	Apfel
Speaker 2	[ˈapəl]	Apel, Appel	

Table 3 Example of orthographic variation in the transcription layer due to inter- and intra-transcriber variation

	Pronunciation	Transcription	Standard Language Orthography
Transcriber 1	[vaːˈʃaɪ̯nlɪç]	wascheinlich	wahrscheinlich
Transcriber 2		vescheilich	

[17] This does not include even more drastic differences, such as transcribers hearing (and therefore transcribing) words differently and/or making mistakes, such as a transcriber transcribing <poodles> instead of <Puter> 'turkeys'.

available.[18] While this has the benefit of helping students not be mired down in the minutia of transcription rules, it has the side effect of introducing additional orthographic variation. For example, in Table 4 below, some transcribers used characters with umlauts such as <ü> while others used the variation <ue>. This is likely due to the fact that most of the people transcribing Texas German audio are American students who do not have an easily accessible ü-key on their keyboard and would need to be told explicitly to use <ü> and not <ue>, which was not included in the transcription guidelines until very recently.

This large amount of orthographic variation within the transcriptions has several consequences: In-depth linguistic research can be time consuming, as it is challenging to identify all of the possible spelling variations for particular lexemes.[19] In addition to that, the (semi-)automatic addition of other annotation layers that are common in modern linguistic spoken corpora, such as part-of-speech and lemma, could have a relatively low accuracy, as many of these tools are often trained on (written) standard-near data such as newspaper texts, and therefore do not have entries for non-standard spellings within their lexicon.[20]

In an effort to make the transcriptions easier to search and manipulate, both manually (by humans) and automatically (by computers),[21] Blevins (2022a) designed an orthographic normalization system for spoken contact language data, using Texas German as a case study.[22] She describes her approach to normalization as follows:

Table 4 Example of orthographic variation in the transcription layer due to flexible project-internal transcription guidelines

	Pronunciation	Transcription	Standard Language Orthography
Transcriber 1		uber	
Transcriber 2	[ˈyːbɐ]	ueber	über
Transcriber 3		über	

[18] There were existing systems for transcribing spoken data, e.g., *Halbinterpretative Arbeitstranskriptionen* (HIAT, Ehlich & Rehbein, 1976) and *Gesprächsanalytisches Transkriptionssystem* (GAT, Selting et al., 1998), which later developed into systems oriented towards computer-assisted transcription (Rehbein et al., 2004; Schmidt et al., 2023). These systems, however, were not originally developed with contact varieties in mind, and spoken contact variety data has unique transcription challenges.

[19] For example, there are at least 10 different orthographic realizations of the lexeme *gehabt* 'had': *gehab*, *ghab*, *gehap*, *ghat*, *gahab*, *gehat*, *gehad*, *gehabt*, *gahabt*, and *ehabt*. For another example, if one did not know that Texas Germans often pronounce *wir* as *mir* and simply searched for < wir > hoping to find all of the instances of first-person plural forms, they would accidentally overlook hundreds of tokens.

[20] When Westpfahl and Schmidt (2013, p. 140) tested using an automatic part-of-speech tagger with literary transcriptions (following cGAT guidelines), there was a high error rate due to there not being lexical entries for the transcribed spoken forms in the tagger.

[21] See Blevins (2022a, pp. 25–26).

[22] This system is based on ten existing, similar systems (Blevins, 2022a, pp. 41–62).

I am primarily guided by:

- Scherrer and [Ljubešić]'s (2016, p. 1) idea of "mapping the variants of what can be identified as the same word to a single representation,"
- Bird et al.'s (2009, p. 108) notion of "identifying non-standard words […] and mapping any such tokens to a special vocabulary,"
- and the desire to map each token to the closest standard language equivalent while preserving the intended part-of-speech [and lemma]. (Blevins, 2022a, p. 38)

It quickly became apparent that normalization decisions—as well as potential later annotation, such as part-of-speech—were necessarily linked to assumptions about a speakers' intended choice of language for a particular lexeme. For example, the following (made-up) sentence is feasible in Texas German:

(3) *mir* ***ham*** ***ham*** *gegessen und* ***die*** ***die*** *gerollt.*
['haṃ] ['hæm] [di:] [daɪ]
'We ate ham and rolled the die.'
Standard German: Wir haben Schinken gegessen und gewürfelt.
(Blevins, 2022a, p. 23)

In this example, although there are seemingly two orthographically identical instances of <ham>, one is a short form of the German verb *haben* 'to have,' while the other is the English noun referring to a pork product. Similarly, the two instances of <die> would necessarily require two different part-of-speech tags (ART and NN or FM respectively[23]). Thus, although "ham" and "ham" as well as "die" and "die" both look orthographically identical in the transcription layer, how they are treated in other layers of annotation, e.g., orthographic normalization and part-of-speech, would differ depending on their language, see Table 5.

It is important to note that the orthographic normalizations as proposed in Blevins (2022a) "primarily reduce variation due to transcription and pronunciation differences but should *not* reduce or 'correct' any non-standard grammatical variation"

Table 5 Example of how language interpretation can disambiguate homographs in the transcription layer

Transcription	mir	ham	ham	gegessen	und	die	die	gerollt
IPA		['haṃ]	['hæm]			[di:]	[daɪ]	
Language	deu	deu	eng	deu	deu	deu	eng	mix:deu.eng
Orthographically normalized transcription	wir	haben	ham	gegessen	und	die	die	gerollt
Part-of-speech	PPER	VAFIN	NN	VVPP	KON	ART	NN	VVPP

[23] Here and in what follows, POS tags are from version 2.0 of the Stuttgart-Tübingen-Tagset (STTS) as described in detail in Westpfahl & Schmidt, 2016. ART = Article, NN = Normal noun, and FM = Foreign material.

(Blevins, 2022a, p. 38). Accordingly, no changes are made to case, gender, number, tense, aspect, etc. For example, in example (4), the definite article *das* is preserved in the normalization layer and not corrected to *dem*, as *das* is an existing standard German definite article, and would therefore lead to the correct part-of-speech and lemma tags without any additional changes.

(4)	**tok**	mit	das	kind	
	lang	deu	deu	deu	
	norm	mit	das	Kind	(TGDP 1-51-2-2-a)

Following the guidelines as laid out in Blevins (2022a) leads to transcriptions that look like example (5) below, with v = the original transcript; tok = the tokenized, lower-cased version of 'v' without punctuation; lang = language; norm = orthographic normalization; trans-deu-utt = utterance translation into standard German; and trans-eng-ww = the original word-for-word translation into English from the TGDA.[24] A summary of Blevins' entire language tagging and orthographic normalization system is available in Blevins (2022a, pp. 688–695).

(5)	v	UND der hat [jeder einer von seine Sohne] aufgesetzt … [in farming and ranching].
	tok	und der hat jeder einer von seine sohne aufgesetzt in farming and ranching
	lang	deu deu deu deu deu deu deu deu deu.txg eng eng eng eng
	norm	und der hat jeder einer von seine Söhne aufgesetzt in farming and ranching
	trans-deu-utt	Und er hat bei jedem seiner Söhne die nötigen Voraussetzungen geschaffen, um in der Land- und Viehwirtschaft selbstständig zu sein.
	trans-eng-ww	And he had each one of his sons put-in-charge … in farming and ranching.

(1–55-1–3-a), Blevins (2022, p. 688)

In an effort to test her proposed language-tagging and normalization guidelines, Blevins constructed and annotated the Texas German Sample Corpus (TGSC). The TGSC is based on a random sampling of the first 600 Texas German speakers interviewed by the TGDP, with the goal of preserving proportionality with respect to birth location (county) and gender.[25] The resulting TGSC contains 13 hours and 35 min of audio, 75,604 tokens (based on the 'norm' layer), and 162 Texas German speakers, spanning 28 birth counties. Using the same estimation numbers as in footnote 12, (open-ended interview = ~20 min), that would mean that the TGSC represented about 6.5% of the audio recordings elicited by the TGDP up to that point. Following the guidelines as outlined in Blevins (2022a), 40% of the transcribed tokens in the TGSC were changed in the normalization layer, and spelling variations decreased by about 25%. The TGSC is available for download in EXMARaLDA's XML-based EXB

[24] The original transcripts and the original word-for-word translations into English (v and trans-eng-ww, respectively) are both written as span annotations in ELAN. In the HTML view of the transcripts on TGDA Online, they are aligned with one other, but not on the token level. It would be possible to refine the ISO/TEI representation of the transcripts to include a token-for-token alignment of transcribed and translated tokens, but this would require another round of consistency checks which we could so far not carry out with our limited resources.

[25] For more information about the target population and sampling procedure, see Blevins (2022a, pp. 664–668, 670–681).

format (Schmidt & Wörner, 2014) at The Texas Data Repository (Blevins, 2022b, https://doi.org/10.18738/T8/IOX9ZA).

According to Blevins (2022a, pp. 686–687),

> In total, the pre-processing and clean-up of all of the files for the TGSC took 15–30 min per file. […] Manually inserting the LANG tags took about 3.3 min per audio minute. Manually inserting the NORM tags took about 3.3 min per audio minute. Manually inserting the TRANS-DEU-UTT layer took about 8.3 min per audio minute. I then did an additional quality check with a native German speaker. Thus, the entire annotation process (including pre-processing and external quality check) took about 60–80 min per audio minute.

3.2.2 Language tagging and normalization

As work on the TGSC has shown, manually adding a language and normalization layer to the transcribed data is very time-consuming. Extending such a manual effort to the entire corpus is unfeasible given the project's limited resources and its focus on getting more recordings transcribed. We therefore explored ways of automating the language tagging and normalization process, using the TGSC as the starting point (a "gold standard") both for informing the automated methods, and for their evaluation.

Automating language tagging and normalization means defining a process which will take transcribed forms as an input and automatically assign them a language tag and a normalized form. The error rate can be calculated by applying the process to a part of the gold standard, comparing automatically assigned forms to the manually assigned ones, and counting as an error all instances where they deviate.

In a first step, we simplified the task by (a) reducing the set of language tags to 4 categories ("deu" for German, "eng" for English, "amb" for ambiguous, and "xxx" for non-lexical forms like hesitation markers), and (b) simplifying the structure of normalizations by allowing only 1:1 relations between transcribed and normalized form (e.g., "gonna" → "going to", whereas previously, contractions like "gonna" → "going" + "to" would be 1:n relations).[26]

We then split the TGSC into a training and an evaluation set. The training set was used to derive a lexicon in which combinations of transcribed form, normalized form, and language tag were recorded with their frequencies. For example, the following entry states that the transcribed form "man" was normalized to German "man" (impersonal pronoun) in 104 cases and to English "man" (male human) in 5 cases.

```
< entry form="man">.
 < n corr="man" lang="deu" freq="104"/>.
 < n corr="man" lang="eng" freq="5"/>.
</entry>.
```

[26] The TGSC used a very fine-grained system of language tags including categories for borrowings from other languages (e.g. Spanish) and word-internal language mixing. These phenomena are rare overall, and the resulting sparsely populated categories would not lend themselves to a statistical approach. Most of these categories were therefore subsumed under the AMB tag for ambiguous forms.

Likewise, the following entry lists different possible German normalizations for the transcribed form “ma”. As a reduced form frequent in casual speech, it is ambiguous and can stand for German “wir” (=personal pronoun ‘we’), “man” (=impersonal pronoun, singular ~ ‘one’), or “mal” (short for adverb ‘einmal’ = ‘once’). It is also conceivable that “ma” could be short for English “Mama” or for a non-standard pronunciation of “my”, although these forms did not figure in the training data. The lexicon entry states that German “wir” is the form most often chosen for normalizing “ma”.

```
< entry form="ma">.
 < n corr="wir" lang="deu" freq="35"/>.
 < n corr="man" lang="deu" freq="32"/>.
 < n corr="mal" lang="deu" freq="10"/>.
</entry>.
```

The automatic process for normalization starts with a lookup in this lexicon. If the transcribed form in question is found, the most frequent normalization and its language tag (including ‘amb’) are chosen. If it is not found, other lexicons are consulted in the following order:

- If the form is found in the normalization lexicon of the FOLK corpus (Schmidt, 2016a/2016b), it is annotated with the most frequent normalized form found in this lexicon and marked as German;
- If the form is found in capitalized form in a lexicon derived from the written DeReKo (Kupietz et al., 2018), it is annotated with this capitalized form and marked as German;
- If the form is found in its non-capitalized form in the DeReKo lexicon, the normalized form will be identical to the transcribed one, and the word is marked as German;
- If the form is found in a lexicon derived from the COCA corpus (Davies, 2008-), the normalized form will be identical to the transcribed one, and the word is marked as English;
- If no lexicon lookup turned up a result, the normalized form will be identical to the transcribed one, and the word is marked as German (since this is the most likely language tag if no other information is available).

A final post-processing step changes “stray” language tags, i.e., single German words within longer sequences of English words or vice versa, to the language tags of the surrounding words.[27] A comparison with the manually annotated TGSC (the “gold standard”) reveals that this simple heuristic leads to an error rate of 6% for the language tagging, and an error rate of 7% for the normalization. Thus, while these automatic annotations are good enough to be used profitably in corpus queries, some

[27] While this may mis-label some potential borrowed words, we found that it accurately labels language more often than not.

methodological caution will be required taking into account that the erroneous normalizations and language tags can lead to false positives or negatives.

3.2.3 POS tagging and lemmatization

Lemmatization and part-of-speech tagging further extend the possibilities of systematic queries to the corpus. While they are standard ingredients for written language corpora and have been successfully transferred to monolingual spoken language close to the standard (e.g., Schmidt, 2016a/2016b), lemmatizing and POS-tagging contact variety data like Texas German involves special challenges.

We used TreeTagger (Schmidt, 1994) to carry out the lemmatization and POS-tagging. We chose the normalization layer as the basis for this task, because, compared to the transcription layer, it has a reduced variety of forms which can be expected to be "recognized" by the TreeTagger parameter files which were trained on standard orthographic forms. For the German tokens, the parameter file developed for tagging the FOLK corpus of spontaneous spoken German in interactions according to the Stuttgart-Tübingen tag set (STTS 2.0, Westpfahl & Schmidt, 2016) was used. In a post-processing step, tokens marked as "eng" on the language layer were all assigned the POS tag "FM" (for: "Fremdsprachliches Material" 'Foreign language material'). In a second run, the English parameter file trained with the PENN tagset (Santorini, 1990) was used, but only lemmas, not POS tags, were written into the data, for tokens that had been language-tagged as "eng" in the previous step. In that way, the POS layer consistently follows the STTS 2.0 tagset guidelines, which is designed for German language material. Everything that is not German (or "amb" – ambiguous) is marked as FM, but the language-sensitive lemmatization will still allow retrieval of words via their base forms.

Evaluating the results of automatic POS tagging against the gold standard of the TGSC yields an error rate of 14.6%. This figure is considerably higher than the 5% reported for the FOLK corpus, but, we think, it is still acceptable for corpus queries in a qualitative approach, i.e. which aim at finding good examples rather than larger full sets of true positives. Using the normalization and language tagging layers as the basis for lemmatization and POS-tagging means that errors from automatic normalization can potentially carry on to the other two layers. We calculated that, if the manually normalized forms were used instead of the automatically normalized ones, the POS tagging error rate would drop to 11.6%.[28]

Figure 4 illustrates the XML markup of the excerpt of Fig 3 after automatic normalization, language tagging, lemmatization and POS tagging.

[28] Between the acceptance of this article and its publication, a second release of the TGDP's data was made available in December 2025. This most recent release included the addition of a second POS layer according to Universal Dependencies POS tags (Nivre et al., 2017). This creates a cross-lingual POS tagging that treats English and German words with the same tag set.

```
<seg type="contribution" xml:id="d68570e1329">
   <anchor synch="ts127"/>
   <w norm="auf" lang="deu" pos="APPR" lemma="auf">AUF</w>
   <pc>'</tei:pc>
   <w norm="dem" lang="deu" pos="NE" lemma="d">m</w>
   <w norm="Schiff" lang="deu" pos="NN" lemma="Schiff">Schiff</w>
   <pc xml:id="a55_p6">.</pc>
   <anchor synch="ts131"/>
</seg>
```

Fig. 4 Transcript excerpt with token annotation in ISO format

4 A new platform for browsing and querying the open-ended portion of the corpus

As mentioned in section 2.2, for years, the Texas German data have been made available to interested researchers, students, and laypersons via the TGDA Online, which is available through the project webpage at http://tgdp.org/dialect-archive/. For registered users, this page offers a map for location-based browsing of the data, including viewing and downloading audio files. The TGDA Online provides access to both the open-ended interviews considered here (including their transcriptions, which can be downloaded as EAF files) as well as the audio from the interviews based on Eikel and Gilbert's translation lists. This portal does not, however, offer ways to query the available data.

A second platform, http://speechislands.org, was set up on a UT Austin server in 2016 in a collaboration between the TGDP and LAITS. It serves as a data management tool for the project but can also be used to browse and query the data. Compared to TGDA Online, this platform offers a more fine-grained and flexible access to the data, and a rudimentary keyword-in-context concordancer, which however, has severe limitations in terms of reliability.[29] Also, this second platform is only available to members of the TGDP as well as a small group of researchers affiliated with the project.

A new platform released in December 2024 (https://tgdp-zumult.la.utexas.edu/) aims to improve the browsing and querying possibilities for the data. It is based on the ZuMult technology, developed as a flexible solution to provide access to multimodal corpora of spoken language (Fandrych et al., 2023, Schmidt 2025, Schmidt/Ferger 2025). It relies on transcripts being available in the ISO standard and can index them for use with the MTAS Query system (Brouwer et al., 2016). In that way, it becomes possible to formulate queries to the data in the CQP query language (Evert et al., 2022)—one of the most widely used query language in corpus linguistics. ZuMult is also used at the Archive for Spoken German (https://zumult.ids-mannheim.de/) where it provides access to, among other things, the corpora of German in Australia,

[29] After the acceptance of this paper, the keyword-in-context concordancer in Speech Islands was discontinued.

German in Namibia, and Unserdeutsch (a.k.a. Rabaul Creole German). Having such different corpora available on a common technical basis facilitates comparable studies across corpora.

ZuMult was thus configured and adapted to work with the Texas German data on a UT Austin server. Besides interfacing the corpus management system of speechislands.org with a process for making the data available in the formats required by the platform, this included:

(1) Creating entry pages for browsing the data tailored to the specific structure of this corpus;

(2) Creating ways of viewing transcripts alongside the underlying recordings, optimized for the way that the data have been transcribed and annotated; and

(3) Building a query interface giving access to the different annotation layers via the CQP query language.

The first official version of the platform has been online since December 2024 at https://tgdp-zumult.la.utexas.edu/. A subsequent release was published in December 2025. For data protection reasons, a free registration is necessary to use the platform. In what follows we illustrate the different platform components in more detail.

4.1 What data are available via ZuMult?

In contrast to TGDA Online, ZuMult follows a versioning protocol to help support replicability of research. In the TGDA Online (described in Sect. 2.2 above), as soon as a new transcript is uploaded and marked as "public" in the back-end, it is available online. This means that the collection of transcripts available via the TGDA Online may differ from week to week. The ZuMult instance, on the other hand, follows a release schedule of 1 or 2 releases a year, and with each new data release, a new version number is given to the data set available through that platform.

At the moment of a new ZuMult release, all of the open-ended interview sections available via the TGDA Online corpus are made available via the ZuMult platform as well, meaning that the set of the interview sections available via TGDA Online and the ZuMult platform are the same. However, as more files get transcribed, they will be made public on TGDA Online before they are made public in the ZuMult plat-

Table 6 The portion of the Texas German Dialect Archive's holdings that are available via TGDA Online and/or ZuMult as of December 2025

Interview Type		Amount	Publicly available?	
			TGDA Online	ZuMult
Open-ended		196 h and 4 min	529 interviews 6,058 interview sections	529 interviews 6,508 interview sections
Translation	Eikel	~37 h	~150 interviews (*not* transcribed)	Not available
	Gilbert	~195 h	~780 interviews (*not* transcribed)	Not available

form. This is because we consider the ZuMult instance to be the basis of linguistic research and should therefore have citable versions that people can refer to in publications to support replicability of research. The TGDA Online interface, however, is intended to be more of a layperson's interface for the public and therefore does not follow strict versioning guidelines.[30]

As can be seen in Table 6, the ZuMult platform currently only includes open-ended data (transcribed), while the TGDA Online contains both open-ended interviews (transcribed) and translation interviews (*not* transcribed). That being said, the TGDA Online has limited search capabilities, no speaker metadata, and only the original transcription and English translation are viewable, while the ZuMult portal, as will be described in Sect. 4.2 through 4.4 below, allows for significantly more sophisticated searches. The TGDP intends to keep both interfaces (TGDA Online and the ZuMult platform) available into the future. While the ZuMult platform currently only includes open-ended interview data, the transcriptions of that data have several additional annotation layers (e.g., token, normalization, language, part-of-speech, lemma), as well as basic speaker metadata, all of which are queryable.

4.2 ZuMult entry pages

To browse and navigate the data for individual interviews, two entry pages are available. The first entry page ("Browse Interviews") lists all interviews by their ID numbers and interview locations and provides filtering possibilities to select interviews by speaker, interviewer, interview location, and year of recording. For each interview, all publicly available sections with their audio and transcripts are displayed. Figure 5 below shows one of the interviews that could be found after filtering for interviews

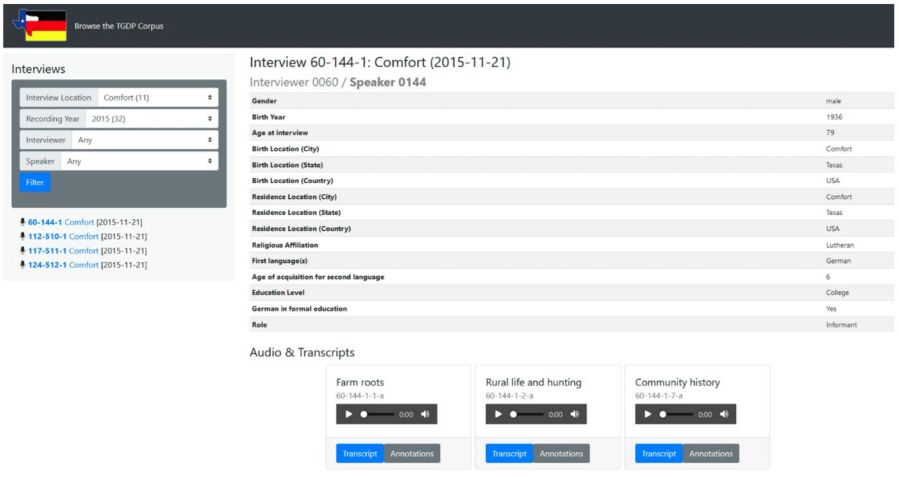

Fig. 5 Entry page "Browse by interview"

[30] Since the acceptance of this paper, this has changed. The TGDA Online now follows the same versioning system as the ZuMult system.

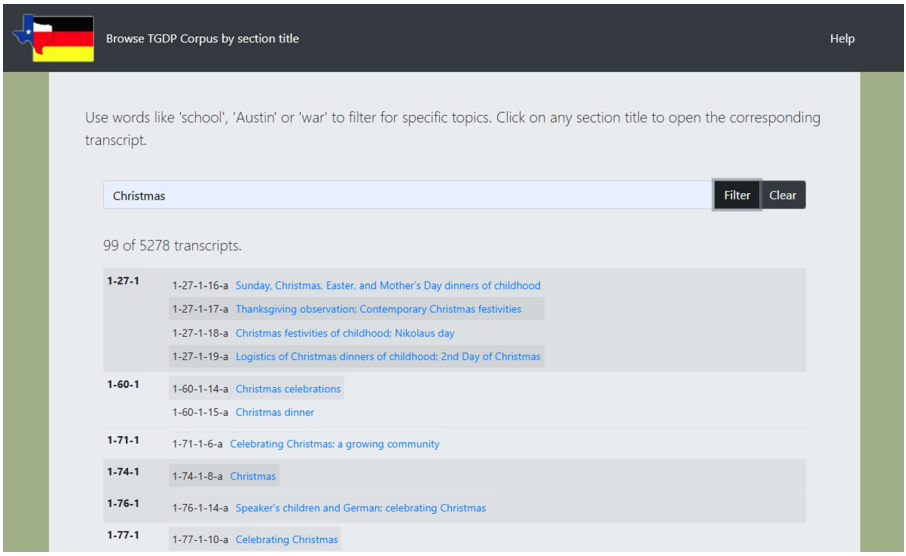

Fig. 6 Entry page "Browse by section title"

conducted in Comfort and recorded in 2015 (see left-side filter panel). The interview in Fig. 5, interview 60–144−1, was conducted in Comfort on November 21st, 2015, and involved interviewer number 0060 and Texas German speaker number 0144. Basic metadata for speaker 0144 is listed at the top of the page, and the interview segments are listed below it.

The second entry page ("Browse by section titles") lists the titles of individual sections. This is a good way of quickly identifying specific topics in the data, for example sections where informants speak about school, church, farming, etc. and thus addresses users who approach the data with a primary interest in the content (eg., cultural studies, oral history). Figure 6 below shows some of the results that occur when one searches for "Christmas".

4.3 Viewing transcripts and annotations

Transcript display plays a crucial role not only for contextualizing corpus linguistic query results, but also for more qualitative approaches to the data. As explained above, transcripts available in the ZuMult platform are richly annotated and aligned with the audio signal. The page for displaying transcripts enables users to navigate this rich information and to interact with transcripts in an explorative manner, adapted to their research interests.

The default view, illustrated in Fig. 7, displays the transcribed contributions in their original form alongside the English translation in the middle part of the screen. Language tagging is visualized by putting English forms in italics and ambiguous

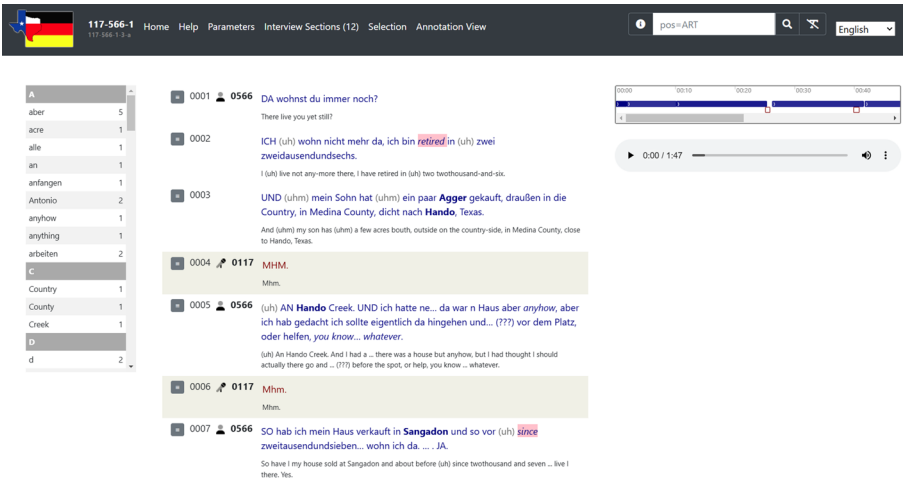

Fig. 7 Transcript visualization

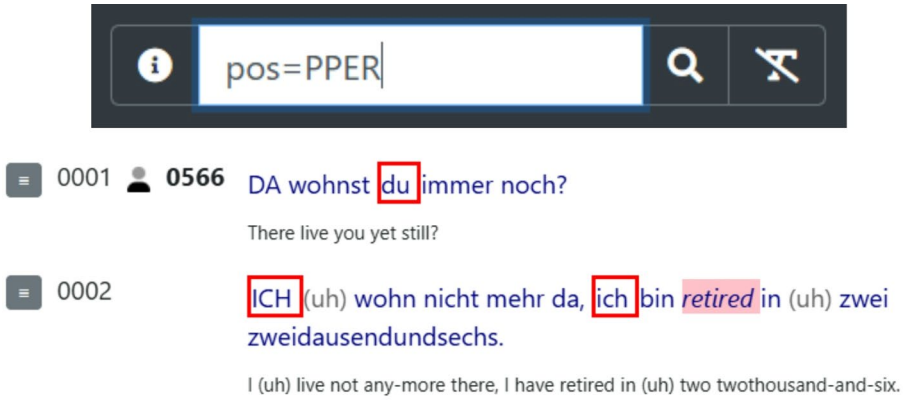

Fig. 8 Search field for the transcript view

forms in bold typeface.[31] Annotated transfer phenomena are highlighted in pink. On the left-hand side, a frequency list presents the lemmas of all words occurring in the transcript. On the right-hand side, a density viewer visualizes in a compact diagram the way in which the interviewer's questions and the informant's answers are distributed. An audio player gives access to the underlying recording.

All components of the view are synchronized, so that double clicking on any place in the transcript will start playback of the corresponding part of the audio, navigation in the audio will scroll the transcript to the corresponding place, clicking on an

[31] "Ambiguous" can mean "language ambiguous" (i.e., it is not clear which language to mark a particular token as) or "form ambiguous" (i.e., it is not clear which normalization form to use) or both. Town names are often marked as ambiguous.

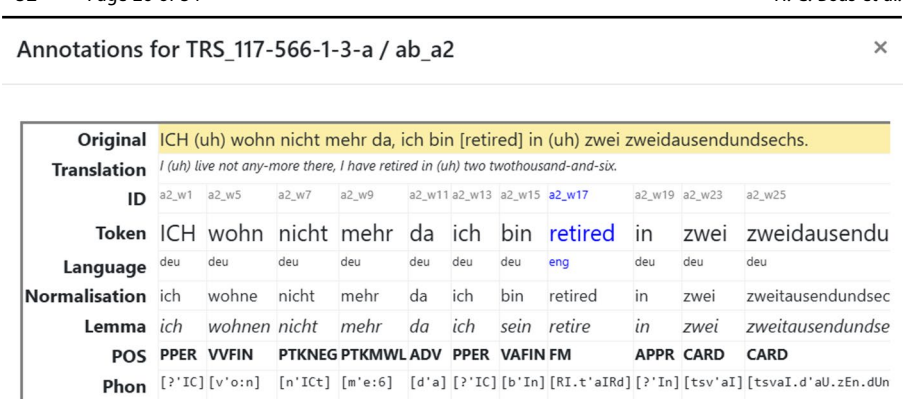

Fig. 9 Annotation view

Fig. 10 Selecting an excerpt

entry in the lemma list will highlight and make visible the places where a form of this lemma occurs, etc.

With the search field in the upper right-hand corner, simple query terms addressing the different annotation layers can be entered to highlight forms in the transcript. For example, the query term in Fig. 8 will highlight all forms which were part-of-speech-tagged as personal pronouns (PPER).

Users can configure the transcript view in different ways, for instance by deciding which annotation level to display (e.g., transcribed or normalized forms), whether or not to visualize deviations between transcribed or normalized form or by slowing down or speeding up audio playback. For each speaker contribution, the entire set of annotation levels can be displayed in a popup window (Fig. 9).

While many analysis steps can be carried out on the transcript display in the browser, it is sometimes necessary to download excerpts onto researchers' computers, for instance in order to carry out acoustic measurements on the audio file in Praat (Boersma, 2014), to add additional annotations in a tool like EXMARaLDA, or simply to integrate an example into a set of slides for a presentation. To achieve this, users can select sections of a transcript by setting a start and an end point (see

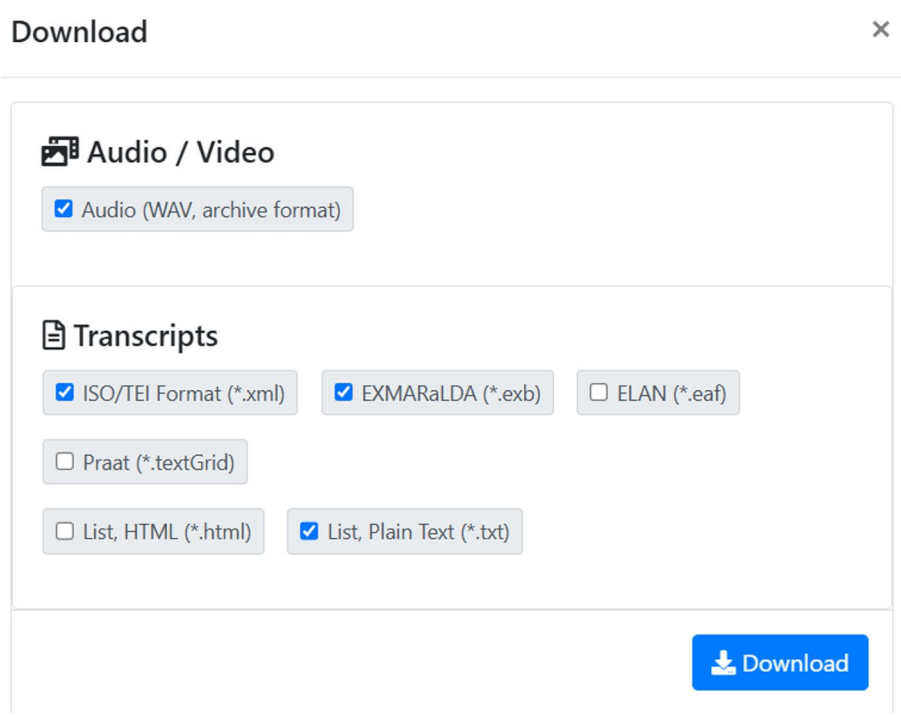

Fig. 11 Downloading data for an excerpt

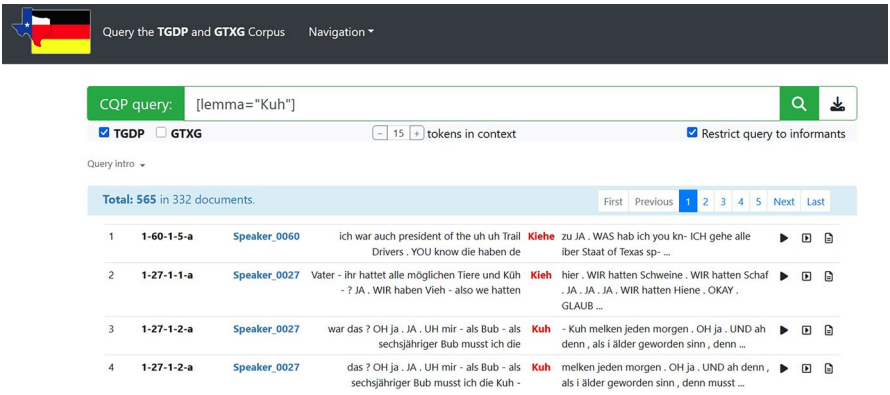

Fig. 12 Query interface with results in KWIC

Fig. 10) and then choose among a set of download options for the transcript and the audio (Fig. 11). Interoperability is thus provided with the most widely used annotation tools Praat, ELAN, and EXMARaLDA.

4.4 Queries

Queries in ZuMult are formulated in CQP, the query language of the Corpus Query Processor (Evert et al. 2022). The query interface of the ZuMult platform is similar to interfaces familiar from written language corpus platforms: A CQP search expression is entered into the search field at the top, and the result of the query is returned as a keyword-in-context (KWIC) concordance which lists all matching parts of the corpus with some amount of preceding and following context. To contextualize an individual query result, the corresponding audio can be accessed directly from the KWIC, metadata about the interview or speaker in question can be displayed in a popup, or the result can be shown in the full context of the transcript as described in the previous section. Figure 12 shows a simple query [lemma= "Kuh"] for the lemmatized form "Kuh" ('cow').

More complex queries can be formulated by concatenation of simple CQP terms and the use of regular expressions. A few short examples will illustrate this on queries that are interesting in the context of a contact variety like Texas German:

(1) Other metadata fields can be included in the CQP queries, for example:

- [norm="hier"] within <Interview_Location="New Braunfels"/> to restrict results to interviews recorded in New Braunfels, or
- [norm="Hochzeit"] within <Sex="male"/> to restrict results to male speakers , or
- [lang="eng"] within <Birth_Year="192."/> to restrict results to speakers born in the 1920s.

(2) Texas German uses a specific quantifier "Masse(n)" ('mass(es)') much more frequently than standard German. To retrieve all instances of this lexical item, a query to the normalization layer can be formulated where a regular expression will make sure that all variations (capitalized vs. lowercase, singular "Masse" vs. plural "Massen") are considered:

[norm="[Mm]ass.+"]

(a)	ne **Masse** von meine Schulfreunde	(32-214-2-4-a)
(b)	die hat da eine **Masse** Land gehabt	(115-663-1-22-a)
(c)	und das hat es **masse** leichter gemacht	(1-547-1-6-a)
(d)	da ist ene **Masse** Deutsch in das Buch	(5-52-1-4-a)
(e)	die hamm **massen** gehabt.	(1-140-1-15-a)

The above selection of query results shows that the item is used with and without a preceding indefinite article (a/b/d vs. c) where the latter case quantifies an adjective rather than a noun. Also, it testifies uses with count nouns (a) as well as non-count nouns (b/d). A full analysis of all 1,539 results would certainly reveal further properties and regularities of the form.

(3) As a variety of German in contact with English, Texas German contains a lot of code switches and other language transfer phenomena (see, e.g., Clyne, 2003; Boas & Pierce, 2011; Dux, 2017). This brings up questions on the morphosyntactic level, such as which German gender is associated with English words preceded by a German article and adjective. The following query retrieves 316 such sequences:

```
[pos="ART"][pos="ADJ."][lang="eng"]
```

(a) ein kleine **Train** (standard German: ein kleine**r** Zug) (115-663-1-7-a)
(b) ein ganze **line** (standard German: ein**e** ganze Linie) (115-663-1-8-a)
(c) ein große **step** (standard German: ein große**r** Schritt) (115-774-1-10-a)
(d) das große **building** (standard German: das große Gebäude) (10-137-1-6-a)
(e) die haupte **reason** (standard German: **der** Hauptgrund) (124-512-1-8-a)
(f) ein verschiedene **Recipe** (standard German: ein verschiedene**s** Rezept) (45-523-1-7-a)
(g) ein große **celebration** (standard German: ein**e** große Feier) (5–52-1-33-a)

The above selection of results shows that the gender article and adjective form is often *not* chosen according to the German equivalent of the English noun. Instead, there seems to be a preference for an unmarked form (i.e. a form with the minimal suffix) for both preceding items. A full analysis could explore this hypothesis further, e.g., by quantifying gender variation for individual items (e.g., [lang="deu"][norm="building"] would find all instances of the noun "building" preceded by a German word), or by considering additional query patterns (e.g., [pos="ART"][lang="eng"] within informant transcripts will find 3,790 German articles followed by an English word without an intervening adjective).[32]

Combining and varying such CQP expressions opens up novel ways of exploring the rich annotations of the Texas German data.

5 Summary, outlook, and perspectives

This paper has presented the Texas German Dialect Project as a long-term project documenting the Texas German dialect in the 21st century. We have shown how the constant evolution of annotation and corpus technology is shaping the project and its workflows for collecting, processing, archiving, and searching data. The ZuMult based platform now available for working with the data is the most recent step in this process. It makes substantial parts of the TGDA data ready for new types of exploration and research approaches, most importantly in a corpus linguistic paradigm. Starting from the current state of the archive with curated and enriched data available in a new platform, we see several concrete prospects for further development.

[32] For a preliminary analysis of gender marking in Texas German, see Boas (2009: 234–236).

First, the body of data will be further expanded: data acquisition is still ongoing, with the aim of documenting more than 1,000 speakers of Texas German altogether. Additional open-ended interviews are being transcribed, which will be made available in future releases. An important desideratum is the inclusion of other data types—most importantly the Gilbert translation tasks—into the corpus. In that context, integrating interviews and elicited data recorded by Gilbert himself in the 1960s can add a diachronic dimension to the Texas German data. We are exploring these prospects in the current phase of our collaboration and expect to be able to publish most of the historical data (as a separate corpus) in the near future.[33]

Second, the annotation processes should be improved and refined. Lower error rates for language tagging and normalization may be achieved by considering additional statistical distributions, such as bi-grams (statistics of word pairs occurring together). With sufficient training data available, we expect the POS tagger might be retrained to reduce errors in POS tagging. Lemmatization still awaits systematic evaluation and insights into possibilities for improvement. Most of these improvements will require considerable manual annotation work and a non-negligible effort for evaluating the results.[34]

Third, we are planning to study how to integrate the full set of speaker metadata into the corpus platform. Based on a 10-page long biographical questionnaire about speakers' backgrounds, language use, language attitudes, speaker attitudes, etc., we will look for ways of making that data available as a part of the corpus platform so that users can search for particular types of speakers based on sociolinguistic variables such as age, gender, religious affiliation, willingness to support the teaching of German in public schools, interest in speaking German with family members, etc. We hope that the addition of these data will enable researchers to conduct even more in-depth research on language contact and language change.

Finally, as part of our long-term vision, we also plan to process recordings from other German contact varieties from Indiana, Pennsylvania, and Wisconsin that we have in our extended collection. Using the same methodology that we developed for Texas German over the past 20+ years, we would like to make these data also available to the research community and to eventually allow a systematic way to conduct comparative analyses of extraterritorial contact varieties of German (Boas, 2021b, 2025). Ideally, this comparative speech islands database could then be linked to other databases of German contact varieties such as the Archive for Spoken German via the Database for Spoken German[35] and the ZuMult instance of the Leibniz Institute for the German Language, which contains data from German in Namibia (Zimmer et al., 2020), Unserdeutsch Creole (Götze et al., 2017; Maitz & Volker, 2017; Lindenfelser, 2022), and Mennonite German in the Americas (Kaufmann et al., 2023) to allow for an even broader comparative approach. We expect this research to help us determine,

[33] Since the acceptance of this paper, the TGDP has published the open-ended portion of Gilbert's 1960s Texas German recordings as a separate corpus available via ZuMult.

[34] Additional annotations at the syntactic level would allow us to conduct dependency parsing as well as (semi-)automatic identification and annotation of grammatical constructions within the frameworks of Construction Grammar as well as its dialect-specific application within construction-based dialectometry (see, e.g., Dunn, 2018).

[35] https://dgd.ids-mannheim.de.

among other things, what types of contact phenomena occur across different contact varieties of German (and why) and what types of contact phenomena are unique to only specific varieties (and why). In addition, this might yield further insights into language contact and change in other contact situations (see, e.g. Clyne, 1981 for German-English contact in Australia; Földes, 2019 for German-Hungarian contact), possibly involving other Germanic languages (see, e.g., Kühl et al., 2019 for Danish in the Americas and Johannessen, 2015 for Norwegian and Swedish).

Acknowledgements We are grateful to Steffen Höeder and Suzan Kung and for their helpful comments on an earlier version of this paper.

Author contributions Each of the three authors wrote different paragraphs and sections of the paper. Each of the authors revised each of the sections as needed. All authors reviewed the manuscript before submission.

Funding Partial financial support was received from the Texas German Endowment Fund at The University of Texas at Austin. The first author, Hans C. Boas, is the founder and director of the Texas German Dialect Project at The University of Texas at Austin.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Amick, A. (2020). *Pitch accent realizations in Texas German*. [Honors Thesis, The University of Texas at Austin].

Bathe, U. (2005). *Plural formation in Texas German*. [Master's thesis, The University of Texas at Austin].

Bird, S. (2003). Seven dimensions of portability for language documentation and description. *Language, 79*(3), 557–582. https://doi.org/10.1353/lan.2003.0149

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. O'Reilly.

Blevins, M. (2014). *Auxiliary* 'tun' *in Texas German*. [Master's thesis, The University of Texas at Austin]. http://hdl.handle.net/2152/32345

Blevins, M. (2022a). *The language-tagging & orthographic normalization of spoken mixed-language data, with a focus on Texas German*. [Doctoral dissertation, The University of Texas at Austin]. https://hdl.handle.net/2152/116703

Blevins, M. (2022b). Texas German Sample Corpus, https://doi.org/10.18738/T8/IOX9ZA Texas Data Repository, V1.

Boas, H. C. (2002). The Texas German Dialect Archive as a tool for analyzing sound change. In P. Austin, H. A. Dry, & P. Wittenburg (Eds.), *Proceedings of the international workshop on resources and tools in field linguistics held in conjunction with the third international conference on language resources and evaluation.*

Boas, H. C. (2003). Tracing dialect death: The Texas German Dialect Project. In J. Larson, & M. Paster (Eds.), *Proceedings of the 28th meeting of the Berkeley Linguistics Society* (pp. 387–398). University of California, Berkeley: Linguistics Department. https://doi.org/10.3765/bls.v28i1.3853

Boas, H. C. (2006). From the field to the web: Implementing best-practice recommendations in documentary linguistics. *Language Resources and Evaluation*, *40*(2), 153–174.

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

Boas, H. C. (2018). Texas. In A. Plewnia & C. M. Riehl (Eds.), *Handbuch der deutschen Sprachminderheiten in Übersee* (pp. 171–192). Narr.

Boas, H. C. (2021a). Zwei Jahrzehnte digitale Dokumentation und Erforschung eines aussterbenden deutschen Auswandererdialekts: Das Texas German Dialect Project (2001–2021). *Zeitschrift für Deutschsprachige Kultur und Literatur*, *30*, 239–268.

Boas, H. C. (2021b). Zur Vergleichbarkeit von Sprachinseldaten: Ein Plädoyer für eine bottom-up Methodologie im Rahmen der Konstruktionsgrammatik und der Frame Semantik. In C. Földes (Ed.), *Kontaktvarietäten des Deutschen im Ausland* (pp. 66–88). Narr Francke Attempto.

Boas, H. C., & Pierce, M. (2011). Lexical developments in Texas German. In M. Putnam (Ed.), *Studies on German language islands* (pp. 129–150). John Benjamins. https://doi.org/10.1075/slcs.123.06bao

Boas, H. C., Pierce, M., Roesch, K., Halder, G., & Weilbacher, H. (2010). The Texas German dialect archive: A multimedia resource for research, teaching, and outreach. *Journal of Germanic Linguistics*, *22*(3), 277–296. https://doi.org/10.1017/s1470542710000036

Boas, H. C. (2025) Towards a systematic methodology for comparing extraterritorial German contact varieties. In R. Szczepaniak, S. Lindenfelser, & A. Prediger (Eds.), *Deutsch als Minderheitensprache in der Welt ,* 9–49. De Gruyter.

Boersma, P. (2014). The use of Praat in corpus research. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 342–360). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571932.013.016

Brouwer, M., Brugman, H., & Kemps-Snijders, M. (2016). MTAS: A Solr/Lucene based multi-tier annotation search solution. *Selected Papers from the CLARIN Annual Conference 2016*, *136*(2), 19–37. https://ep.liu.se/ecp/136/002/ecp17136002.pdf

Clyne, M. (1981). Deutsch als Muttersprache in Australien: Zur Ökologie einer Einwanderersprache. In Zusammenarbeit mit dem Centre for Migrant Studies. Monash University. *Deutsche Sprache in Europa und Übersee* 8. Wiesbaden: Franz Steiner Verlag.

Clyne, M. (2003). *Dynamics of language contact*. Cambridge University Press. https://doi.org/10.1017/cbo9780511606526

Crystal, D. (2000). *Language death*. Cambridge University Press. https://doi.org/10.1017/cbo9781139106856

Davies, M. (2008-). The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/

Drude, S., Broeder, D., Trilsbeek, P., & Wittenburg, P. (2012). The Language Archive: A new hub for language resources. In N. Calzolari (Ed.), *Proceedings of LREC 2012: 8th international conference on language resources and evaluation* (pp. 3264–3267). European Language Resources Association (ELRA).

Dunn, J. (2018). Finding variants for construction-based dialectometry: A corpus-based approach to regional CxGs. *Cognitive Linguistics*, *29*(2), 275–311. https://doi.org/10.1515/cog-2017-0029

Dux, R. (2017). Classifying language contact phenomena: English verbs in Texas German. *Journal of Germanic Linguistics*, *29*(4), 379–430. https://doi.org/10.1017/s1470542717000034

Ehlich, K., & Rehbein, J. (1976). Halbinterpretative arbeitstranskriptionen (HIAT). *Linguistische Berichte*, *45*, 21–41.

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

Evert, S., & Development Team, C. W. B. (2022). The IMS open corpus workbench (CWB): CQP interface and query Language manual. *CWB Version* 3.5.

Fandrych, C., Schmidt, T., Wallner, F., & Wörner, K. (Eds.). (2023). Themenschwerpunkt: Zugänge zu mündlichen Korpora für DaF und DaZ: Das ZuMult-Projekt. *Korpora Deutsch als Fremdsprache (KorDaF)* 3(1).

Földes, C. (2019). Dialekt im Sprachkontakt: Das Ungarndeutsche Zweisprachigkeits- und Sprachkontaktkorpus. In F. Kostrzewa & A. H. Massud (Eds.), *Strukturen und Besonderheiten fremder Sprachen: Unter besonderer Berücksichtigung des Chinesischen, Japanischen, Koreanischen, Arabischen und Ungarndeutschen* (pp. 153–175). Verlag Dr. Kovač.

Gilbert, G. (1972). *Linguistic Atlas of Texas German*. University of Texas Press.

Götze, A., Lindenfelser, S., Lipfert, S., Neumeier, K., König, W., & Maitz, P. (2017). Documenting Unserdeutsch (Rabaul Creole German): A workshop report. In P. Maitz, & C. A. Volker (Eds.), *Language contact in the German colonies: Papua New Guinea and beyond* (=Special issue von Language and Linguistics in Melanesia), (pp. 65–90).

Guion, S. (1996). The death of Texas German in Gillespie County. In P. S. Ureland, & I. Clarkson (Eds.), *Language contact across the North Atlantic: Proceedings of the working group held at University College, Galway, August 29-September 3, 1992 and the University of Gothenburg, August 16–21, 1993*, (pp. 443–463). Niemeyer. https://doi.org/10.1515/9783110929652.443

Hedeland, H., & Schmidt, T. (2022). The TEI-based ISO standard 'Transcription of spoken language' as an exchange format within CLARIN and beyond. In M. Monachini, & M. Eskevich (Eds.), *Selected papers from the CLARIN annual conference 2021* (=Linköping Electronic Conference Proceedings 189) (pp. 34–45). Linköping University Electronic Press. https://doi.org/10.3384/ecp1894

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. *Linguistics, 36*(1), 161–195. https://doi.org/10.1515/ling.1998.36.1.161

ISO (2016). ISO 24624:2016 Language resource management: Transcription of spoken language. https://www.iso.org/standard/37338.html

Johannessen, J. B. (2015). The Corpus of American Norwegian Speech (CANS). In Béata Megyesi (Ed.): *Proceedings of the 20th Nordic Conference of Computational Linguistics*, NODALIDA 2015, May 11–13, 2015, Vilnius, Lithuania. NEALT Proceedings Series 23.

Jones, E. (2022). *Gender roles and language loss: A new perspective from Texas German on language attitudes.* [Master's thesis, The University of Texas at Austin]. https://hdl.handle.net/2152/116033

Kaufmann, G., Gorisch, J., & Schmidt, T. (2023). Das MEND-Korpus im Archiv für Gesprochenes Deutsch: Entstehung, Möglichkeiten, Grenzen. In P. Wolf-Farré, L. Löff, A. Machado, Prediger, & S. Kürschner (Eds.), *Deutsche und weitere germanische Sprachminderheiten in Lateinamerika: Methoden, Grundlagen, Fallstudien* (Vol. 1, pp. 103–147). Lang.

Kühl, K., Petersen, J. H., & Hansen, G. F. (2019). The corpus of American danish: A language resource of spoken immigrant Danish in North and South America. *Language Resources and Evaluation, 54*(3), 831–849. https://doi.org/10.1007/s10579-019-09473-5

Kung, S. S., & Sherzer, J. (2013). The archive of the Indigenous languages of Latin America: An overview. *Oral Tradition*, *28*(2), 379–388. https://journal.oraltradition.org/wp-content/uploads/files/articles/28ii/24_28.2.pdf

Kupietz, M., Lüngen, H., Kamocki, P., & Witt, A. (2018). The German reference corpus DeReKo: New developments – New opportunities. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (Eds.), *Proceedings of the 11th international conference on language resources and evaluation (LREC 2018)* (pp. 4353–4360). European Language Resources Association (ELRA).

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania.

Lindenfelser, S. (2022). *Kreolsprache Unserdeutsch*. De Gruyter.

Maitz, P., & Volker, C. A. (2017). Documenting Unserdeutsch: Reversing colonial amnesia. *Journal of Pidgin and Creole Languages*, *32*(2), 365–397. https://doi.org/10.1075/jpcl.32.2.06mai

Nathan, D. (2014). Access and accessibility at ELAR, an archive for endangered languages documentation. *Language Documentation and Description*, *12*(=Special Issue on Language Documentation and Archiving), 187–208. https://doi.org/10.11647/obp.0032.03

Nicolini, M. (2004). *Deutsch in Texas*. LIT.

Nivre, J., Zeman, D., Ginter, F., & Tyers, F. (2017). Universal Dependencies. *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics: Tutorial abstracts*. Association for Computational Linguistics.

Rehbein, J., Schmidt, T., Meyer, B., Watzke, F., & Herkenrath, A. (2004). Handbuch für das computergestützte Transkribieren nach HIAT. *Working papers in multilingualism*, Series B 56. Universität Hamburg, 2004. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-23681

Roesch, K. (2009). *Texas Alsatian: Henri Castro's legacy*. [Doctoral dissertation, The University of Texas at Austin]. http://hdl.handle.net/2152/19838

Rybarski, J. (2006). *Some developments in the Texas German case system*. [Honors Thesis, The University of Texas at Austin].

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Scherrer, Y., & Ljubešić, N. (2016). Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of international conference on new methods in language processing*.

Schmidt, T. (2011). A TEI-based approach to standardising spoken Language transcription. *Journal of the Text Encoding Initiative*, *1*, 1–25. https://doi.org/10.4000/jtei.142

Schmidt, T. (2016a). Good practices in the compilation of FOLK, the research and teaching corpus of spoken German. In J. M. Kirk, G. Andersen, (Eds.), *Compilation, transcription, markup and annotation of spoken corpora.* (=International Journal of Corpus Linguistics 21(3)) (pp. 396–418). Benjamins.

Schmidt, T. (2025). Représenter et accéder à la parole dans les corpus oraux: diversification et adaptation des méthodes et technologies. In K.-C. Layal, D. Céline, A. Lotfi & G. Annette (Eds.), *Representer la parole*. De Gruyter.

Schmidt, T., & Ferger, A. (2025) Putting things on top of other things. The ZuMult platform for multimodal corpora and its ecosystem. *CLARIN Annual Conference,* Vienna.

Schmidt, T., & Wörner, K. (2014). EXMARaLDA. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 402–419). Oxford University Press.

Schmidt, T., Kupietz, M., & Geyken, A. (2016b). Construction and dissemination of a corpus of spoken interaction: Tools and workflows in the FOLK project. *Journal for Language Technology and Computational Linguistics (JLCL), 31*(1), 127–154. https://doi.org/10.21248/jlcl.31.2016.205

Schmidt, T., Schütte, W., Winterscheid, J., Schürmann, M., Reineke, S., & Schedl, E. (2023). *cGAT: Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2).* https://doi.org/10.14618/chrz-zy56

Seifart, F., Ludger Paschen, & Stave, M. (Eds.). (2024). Language Documentation Reference Corpus (DoReCo) 2.0. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). https://doi.org/10.34847/nkl.7cbfq779

Selting, M., Auer, P., Barden, B., Bergmann, J., Couper-Kuhlen, E., Günthner, S., Meier, C., Quasthoff, U., Schlobinski, P., & Uhmann, S. (1998). Gesprächsanalytisches transkriptionssystem (GAT). *Linguistische Berichte*, *173*, 91–122.

Thieberger, N., & Harris, A. (2022). When your data is my grandparents singing: Digitisation and access for cultural records, the Pacific and regional archive for digital sources in endangered cultures (PARADISEC). *Data Science Journal*. https://doi.org/10.5334/dsj-2022-009

Thompson, J. (2005). *Code-switching in Texas German*. [Master's thesis, The University of Texas at Austin].

Trudgill, P. (2004). *New dialect formation*. Cambridge University Press.

Wamprechtshammer, A., Aznar, J., Arestau, E., Hedeland, H., Isard, A., Khait, I., Lange, H., Majka, N., & Rau, F. (2022). QUEST: Guidelines and specifications for the assessment of audiovisual, annotated language data. *Working papers in corpus linguistics and digital technologies: Analyses and methodology.* https://hal.science/hal-04234971v1

Warmuth, M. (2023). *Phonological convergence and variation in a dying language variety: The case of Texas German*. [Doctoral dissertation, The University of Texas at Austin].

Westpfahl, S., & Schmidt, T. (2013). POS für(s) FOLK–Part of speech tagging des Forschungs- und lehrkorpus gesprochenes Deutsch. *Journal for Language Technology and Computational Linguistics*, *28*(1), 139–156.

Westpfahl, S., & Schmidt, T. (2016). FOLK-Gold: A GOLD standard for part-of-speech tagging of spoken German. In *Proceedings of the tenth conference on international language resources and evaluation (LREC '16).* European Language Resources Association (ELRA).

Whitworth, L. (2005). *Remnants of the German language in Mason County, Texas*. [Master's thesis, The University of Texas at Austin].

Wilkinson, M., Dumontier, M., Aalbersberg, I., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*, Article 160018. https://doi.org/10.1038/sdata.2016.18

Wilson, J. (1986). Texas German and other American immigrant languages: Problems and prospects. In T. Gish & R. Spuler (Eds.), *Eagle in the New World* (pp. 221–240). Texas A&M Press.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.

Zimmer, C., Wiese, H., Simon, H. J., Zappen-Thomson, M., Bracke, Y., Stuhl, B., & Schmidt, T. (2020). Das Korpus Deutsch in Namibia (DNam): Eine Ressource für die Kontakt-, Variations- und Soziolinguistik. *Deutsche Sprache, 48*(1), 210–232. https://doi.org/10.37307/j.1868-775x.2020.03.03