

Twenty-five years of the Texas German Dialect Project: Methodology, digital resources, and analyses of diachronic and comparative aspects of a contact variety of German

The article discusses how the Texas German Dialect Project (TGDP) has collected, processed, and archived data from its inception in 2001 up until the present day, 25 years later. It also explores how these data have been used to gain new insights into the structure and development of Texas German as well as the dialectology of German-American dialects and language variation and change more broadly. Particular attention is paid to data collection and processing, alongside the evolution of the TGDP's technical infrastructure and organizational framework.

Texas German is a German heritage language present in Texas since the mid-19th century. Despite strong cultural traditions in local communities, English-only laws and anti-German sentiment during the First and Second World Wars led to a stark reduction in German language use in the public domain; likewise, transmission within the family showed a rapid decline. While the first half of the 20th century still saw approximately 160,000 speakers of Texas German (Kloss 1977; Jordan 2004), Boas/Schmidt/Blevins (2026) estimate the current number at no more than 2,000.

Linguistically, Texas German developed from a variety of German dialects, with immigration from regions such as Hesse, the Rhineland, Thuringia, Mecklenburg-Vorpommern, and Lower Saxony (Boas 2009a). Due to the size of the Texas German Belt, where the majority of speakers settled, and the geographic isolation between communities, a singular unified form of Texas German did not develop: Structural regularities indeed exist, but inter-speaker variation seems to be the norm rather than the exception. Larger studies investigating Texas German began in the mid-20th century. Three relevant examples include Eikel (1954), Gilbert (1972), and Boas (2009a).

Founded in 2001 by Hans C. Boas of the University of Texas at Austin, the TGDP is committed to the preservation of Texas German through recorded interviews with the last speakers of this moribund dialect. Conducted with either individuals or small groups, interviewers take speakers through three types of questionnaires: 1) a biographical questionnaire collecting demographic information; 2) a narrative interview in German; and 3) one or more translation lists of words, phrases, and sentences from English into Texas German. These lists, taken from earlier studies (Eikel 1954; Gilbert 1972; Guion 1996), allow for systematic comparison between speakers today, as well as those who participated in the earlier studies. The TGDP interviews are then catalogued, anonymized, segmented, transcribed, and translated. Audio recordings are saved on the server of the Liberal Arts Instructional Technology Services (LAITS) at the University of Texas, and biographical questionnaires are scanned and uploaded to ensure long-term preservation and protection.

The TGDP possesses a wealth of Texas German material, which has been stored in the Texas German Dialect Archive (TGDA), in addition to one collection each of Pennsylvania Dutch and Wisconsin German. Some of these data are freely available through the Texas

German Dialect Archive Online (TGDA Online), which comprises 1) narrative interviews, also known as the TGDP Corpus, 2) translations based on the so-called Gilbert list, and 3) translations based on the so-called Eikel list. While 1) consists of approximately 500 recordings, 2) is made up of 750 recordings for a total of 148 elicitations, and 3) comprises 150 recordings for a total of 190 elicitations. Conceptualized as a resource for teaching and research, the TGDA Online had 6,200 registered users in 2026.

In addition to the TGDA Online, the TGDA also contains the Gilbert Texas German Corpus (GTGXG). This corpus of historical Texas German, compiled by Glenn Gilbert in the 1960s and 1970s, contains approximately 11 hours of narrative interviews and 36 hours of translations. The narrative interviews have recently been made available through the ZuMult corpus platform, where they may be searched alongside the data from the TGDP Corpus. The ZuMult corpus not only contains the original audio to each of its interviews but also offers transcriptions in German and word-for-word translations into English. The release of the translation data is currently in preparation.

Finally, the TGDP also possesses one additional collection of translation data, the bulk of which has been collected through the so-called Guion list, as well as a collection of socio-biographical data, mostly gathered in English. The former accounts for approximately 425 interviews, or 70 hours of recording, while the latter amounts to approximately 500 interviews, or 85 hours of recording. Neither of these has been made publicly available yet.

Over the past 25 years, the narrative interviews in particular have presented the TGDP with a number of challenges. Most importantly, it has long remained impossible to straightforwardly query these data, as each of the more than 50 transcribers followed their own transcription rules. In order to remedy this issue, the project has appealed to the Texas German Sample Corpus (TGSC), as developed by Blevins (2022). The TGSC, which takes the role of an annotation ‘gold standard’, is a 75,000-token corpus accounting for over 13 hours of audio collected from 162 random speakers residing in 28 counties. The corpus aids in (1) orthographically normalizing the interview data and (2) determining the source language for each word according to four levels, namely English, German, ambiguous, or non-lexical. A comparison between the TGDP Corpus, on the one hand, and the TGSC, on the other, results in an average error rate of 7% for the normalization process and 6% for the language categorization process.

Additionally, the interview data have been lemmatized and POS-tagged using TreeTagger (Schmid 1994) following a three-stage system. First, all German tokens were annotated with the help of the Stuttgart-Tübingen-Tagset (STSS 2.0), resulting in an error rate of 14.6% compared to the TGSC. Second, all remaining English, ambiguous, and non-lexical tokens were corrected using the PENN tagset (Santorini 1990). And third, all STSS and PENN tags were mapped onto the Universal Dependencies Tagset, allowing users to run cross-linguistic search queries.

In November 2024, the narrative interviews included in the TGDP Corpus were made accessible via the aforementioned ZuMult corpus platform, an open-source system that was developed between 2018 and 2022 as part of a joint project between the Leibniz Institute for the German Language (IDS), the Herder Institute at the University of Leipzig, and the Hamburg Centre for Language Corpora (HZSK). Since the second release in December 2025, ZuMult also provides access to additional narrative interviews from the

TGDP as well as to the GTXG corpus. As the system also provides access to other corpora of (extraterritorial) German, it is particularly well suited for contrastive analysis.

ZuMult enables interviews to be filtered for interview location, interview year, interviewer, and interviewee. For each interview, the interface displays 1) a list of metadata and 2) thematically labeled interview excerpts, each accompanied with the corresponding audio recordings and transcripts. The interface allows corpus data to be queried using either a simple search targeting the normalized data level or a CQP query. Searches of the latter type may target all annotation levels, including tokens, normalized forms, lemmas, and parts of speech, in addition to enabling queries for language and queries based on specific metadata. The result of a CQP query is displayed as a keyword-in-context concordance (KWIC), with the hit in the center and the relevant context on either side. Finally, individual transcripts may also be studied in detail using an interactive interface equipped with a lemma frequency list, a density viewer providing an overview of the transcript's structure, and an audio player.

Since the founding of the TGDP, research has ranged from the Texas German sound system (Boas et al. 2004; Warmuth 2023), plural formation (Bathe 2005), and case marking (Boas 2005, 2009b) to Texas-Alsatian in Medina County (Roesch 2012), English loan words (Dux 2017, 2020), and aspect markers (Blevins 2018). Other work has focused on the historiography of Texas German research (Pierce/Boas/Gilbert 2019; Pierce/Boas 2019; Pierce/Boas/Gilbert 2021; Boas/Somers/Pierce in the current paper), and on speech islands more comparatively (Saller 2024; Boas 2025; Reher in preparation).

To date, over 300 academic presentations as well as numerous presentations for lay audiences have taken place, and a yearly newsletter keeps the local community engaged with the project. In addition, there have been more than 50 newspaper articles, 12 radio reports, 3 programs produced for tv, and a documentary film to promote public awareness of Texas German. In 2021, the TGDP began collecting, archiving, and digitizing handwritten letters, diaries, and other texts, from which interdisciplinary research has begun to develop (such as Weiss et al. 2025). In 2022, the TGDP also initiated a collaboration with the German Historical Institute in Washington D.C. and their “Migrant Connections” project, which involves the transcription and translation of 19th- and 20th-century handwritten letters sent from Germany to North America. The most recent project, begun in July 2025 with support from the Volkswagen Foundation, reconstructs 1880s Texas German through the use of artificial intelligence. The goal of the project is to create a chatbot that can respond in historically accurate Texas German and discuss events going on at that time.

References

- Bathe, Ulrike (2005): Plural formation in Texas German. Masterarbeit. Austin, TX: University of Texas at Austin.
- Blevins, Margaret (2018): Towards a constructional analysis of the progressive aspect in Texas German. In: Boas, Hans C./Höder, Steffen (eds.): *Constructions in contact: Constructional perspectives on contact phenomena in Germanic languages*. (= *Constructional Approaches to Language* 24). Amsterdam/Philadelphia: Benjamins, pp. 73–114.
- Blevins, Margaret (2022): *The language-tagging & orthographic normalization of spoken mixed-language data, with a focus on Texas German*. Dissertation. Austin, TX: University of Texas at Austin.

- Boas, Hans C. (2005): A dialect in search of its place: The use of Texas German in the public domain. In: Cravens, Craig Stephen/Zersen, David John (eds.): *Transcontinental encounters: Central Europe meets the American heartland*. Austin, TX: Concordia University Press, pp. 78–102.
- Boas, Hans C. (2009a): *The life and death of Texas German*. Durham, NC: Duke University Press.
- Boas, Hans C. (2009b): Case loss in Texas German: The influence of semantic and pragmatic factors. In: Barðdal, Jóhanna/Chelliah, Shobhana L. (eds.): *The role of semantic, pragmatic, and discourse factors in the development of case*. (= *Studies in Language Companion Series* 108). Amsterdam/Philadelphia: Benjamins, pp. 347–373.
- Boas, Hans C. (2025): Towards a systematic methodology for comparing extraterritorial German contact varieties. In Szczepaniak, Renata/Lindenfelser, Siegwalt/Prediger, Angélica (eds.): *Deutsch als Minderheitensprache in der Welt*. (= *Germanistische Linguistik* 337). Berlin/Boston: De Gruyter, pp. 9–49.
- Boas, Hans C./Schmidt, Thomas/Blevins, Margo (2026): A new corpus platform for the Texas German Dialect Project. In: *Language Resources and Evaluation* 60, article 52. DOI: <https://doi.org/10.1007/s10579-025-09893-6>.
- Boas, Hans C./Somers, Joren/Pierce, Marc (current paper): A previously unknown source of 1960s Texas German data. Erscheint in: *Zeitschrift für Dialektologie und Linguistik*.
- Boas, Hans C./Ewing, Karen/Moran, Cheryl/Thompson, Jana (2004): Towards determining the influence of internal and external factors on recent developments in Texas German phonology. In: *University of Pennsylvania Working Papers in Linguistics* 10, 1, pp. 47–59.
- Dux, Ryan (2017): Classifying language contact phenomena: English verbs in Texas German. In: *Journal of Germanic Linguistics* 29, 4, pp. 379–430.
- Dux, Ryan (2020): Code-switching and loan translation in German-American. A cognitive-constructional account. In: *Belgian Journal of Linguistics* 34, pp. 52–65.
- Eikel, Fred (1954): *The New Braunfels German Dialect*. Dissertation. Baltimore, MD: John Hopkins University.
- Gilbert, Glenn G. (1972): *Linguistic atlas of Texas German*. Austin, TX: University of Texas Press.
- Guion, Susan (1996): The death of Texas German in Gillespie County. In: Ureland, P. Sture/Clarkson, Iain (eds.): *Language contact across the North Atlantic*. (= *Linguistische Arbeiten* 359). Tübingen: Niemeyer, pp. 443–463.
- Jordan, Terry G. (1969): The German Settlement of Texas after 1865. In: *The Southwestern Historical Quarterly* 73, 2, pp. 193–212.
- Kloss, Heinz (1977): Die deutsche Schriftsprache bei den Amischen. In: *Deutsch als Muttersprache in Kanada. Berichte zur Gegenwartslage*. (= *Deutsche Sprache in Europa und Übersee* 1). Wiesbaden: Steiner, pp. 97–98.
- Pierce, Marc/Boas, Hans C. (2019): Where was historical linguistics in 1968 and where is it now? In: Boas, Hans C./Pierce, Marc (eds.): *New directions for historical linguistics*. (= *Brill's Studies in Historical Linguistics* 9). Leiden: Brill, pp. 1–41.
- Pierce, Marc/Boas, Hans C./Gilbert, Glenn G. (2019): When is a dissertation not a dissertation? On Eikel 1954. In: *Yearbook of German-American Studies* 53, pp. 187–196.
- Pierce, Marc/Boas, Hans C./Gilbert, Glenn G. (2021): Fred Eikel and the study of Texas German. In: *Yearbook of German-American Studies* 56, pp. 131–148.
- Reher, Caroline (in preparation). *Satzverknüpfung im Kontakt: Ein Vergleich des Deutschen in Chile, Texas und Namibia*. Dissertation. Dortmund: Technische Universität Dortmund.
- Roesch, Karen A. (2012): Language maintenance and language death: The decline of Texas Alsatian. (= *Culture and Language Use: Studies in Anthropological Linguistics* 6) Amsterdam/Philadelphia: Benjamins.

- Saller, Anna (2024): Der Einfluss sprachlicher Vitalität und des Standards auf die Frequenz der ‘tun’-Periphrase in verschiedenen deutschen Sprachminderheiten: Texas, Australien, Namibia und Siebenbürgen. In: Wildfeuer, Alfred/Franz, Sebastian (eds.): *Standardisierung und (Re-)Vitalisierung von kleinen Sprachen*. (= Germanistische Linguistik 55, 1). Hildesheim: Olms, pp. 275–314.
- Santorini, Beatrice (1990): Part-of-speech tagging guidelines for the Penn Treebank Project, Pennsylvania State University. <https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf> (Stand: 27.6.2026).
- Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of the International Conference on New Methods in Language Processing*. Manchester: University of Manchester, pp. 44–49.
- Warmuth, Matthias (2023): *Phonological convergence and variation in a dying language variety: The case of Texas German*. Dissertation. Austin, TX: The University of Texas at Austin.
- Weiss, Jana/Blevins, Margo/Boas, Hans C./Sauer, Verena (2025): Towards a roadmap for exploring past and present Texas German voices: A case study of German-Indigenous relations. In: *Yearbook of German American Studies* 58, pp. 173–214.