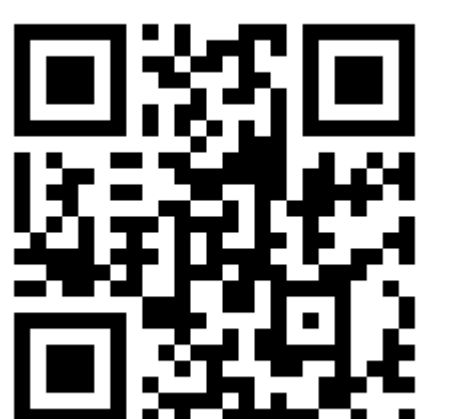


The Texas German Dialect Project Corpus

A Diachronic Resource for Investigating Language Contact

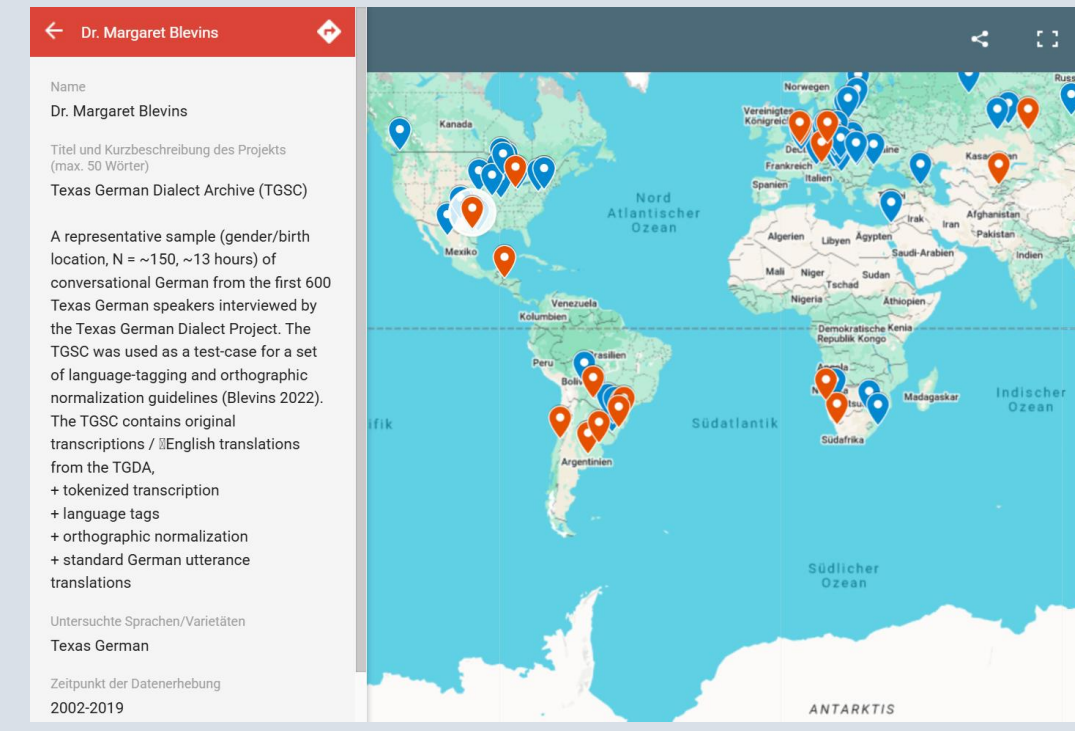


What is Texas German?

Texas German: A Unique Contact Variety

- (1) A mix of at least five different regional dialects brought by German-speaking immigrants Texas beginning in the 1830s
- (2) Relatively young compared to other German contact varieties (e.g. Pennsylvania, Romanian, Hungarian German)
- (3) Unlike other new world contact varieties of German, Texas German did not evolve into a focused new world variety
- (4) One of the most extensively documented extraterritorial varieties of German

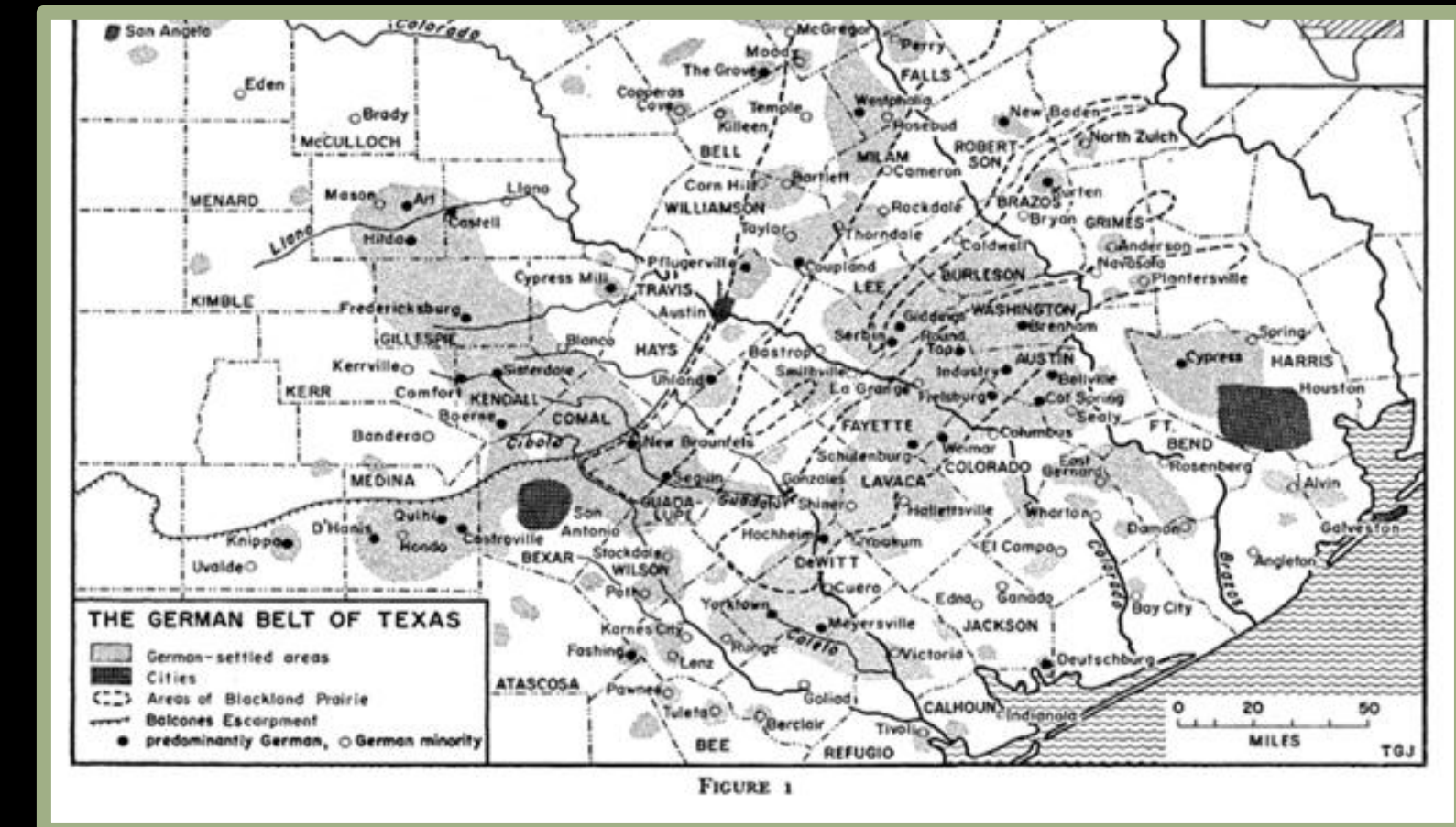
⇒ The Texas German Dialect Project and Archive at the University of Texas at Austin: <https://tgd.org/>



Comparable Contact Varieties with Resources

- Wisconsin German (Eichhoff 1979, Wagener 2002)
- Corpus of American Danish (CaAmDa, Kühl et al. 2017)
- Mennonite German in the Americas (MEND, Kaufmann et al. 2023)
- Australian German (Clyne 1981)
- RUEG Corpus (Schroeder et al. 2024)
- German in Namibia (DNAM, Zimmer et al. 2020)
- Unserdeutsch - Rabaul Creole German (UNSD, Maitz & Volker 2017)
- Russian-German Dialects (RUDI, Berend 2011)

⇒ German Abroad Map at Archive for Spoken German: <https://aqd.ids-mannheim.de/germanabroad.shtml>



The German Belt of Texas - Main target regions of emigration from Germany



Glenn Gilbert: Texas German in the 1960s

- Fieldwork for dissertation *The German Dialect Spoken in Kendall and Gillespie Counties, Texas* (Gilbert, 1963), 2 counties (Gillespie and Kendall), ~40 informants
- 3 types of data: biographical interview, translation task (using different lists of prompts), narrative interview
- Groundwork for *Linguistic Atlas of Texas German* (Gilbert, 1972), 31 counties, ~273 informants)
- Audio recordings, fieldnotes, several boxes of other research materials related to his research on Texas German donated to the TGDP.

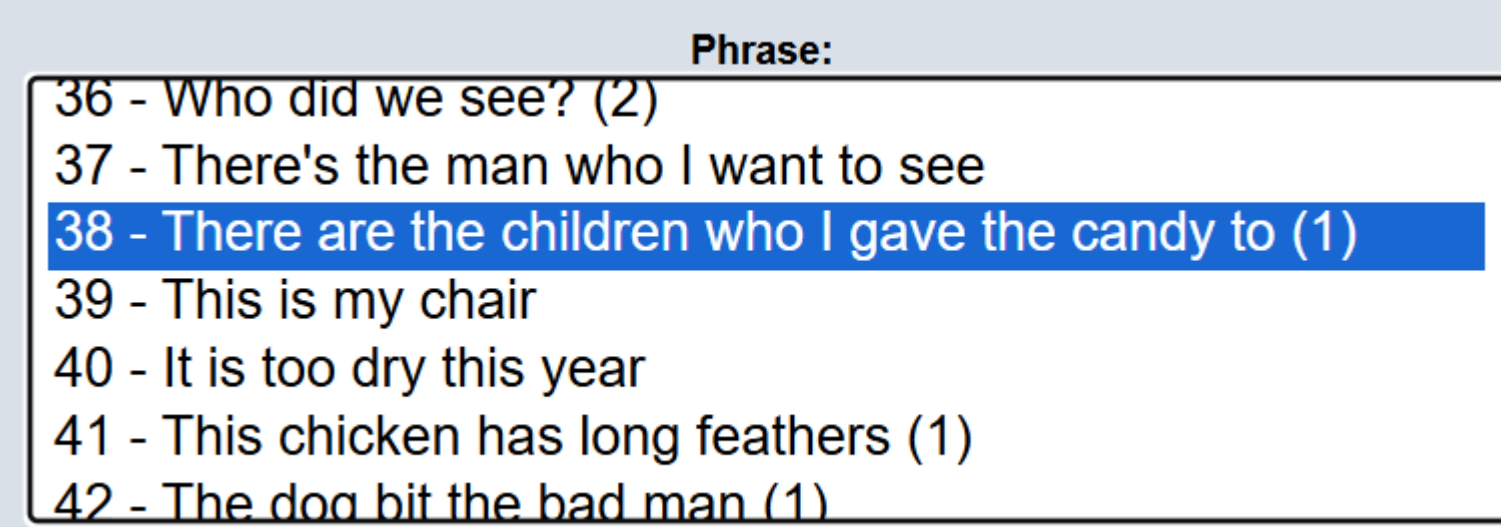
TGDP: Texas German after 2000

- Fieldwork since 2001: Recording the last generation of Texas German speakers
- Hans C. Boas: *The life and death of Texas German*. Duke University Press, Durham 2009, ISBN 978-0-8223-6716-1
- Methodology partly based on Gilbert's work ⇒ direct comparability
- 3 types of data: biographical interview, translation task (using different lists of prompts), narrative interview
- With 1.5 million transcribed tokens, one of the largest resources documenting a German contact variety



Interview Type	Amount
Open-Ended Interviews (OE)	51 interviews, ~ 11 hours
Translation: Gilbert Atlas (GA)	28 interviews, ~ 7 hours

Five other translations lists were also used (37 recordings, ~24.5 hours). Transcription has been completed for all open-ended interviews (published in ZuMult v. 2.1). All GA interviews have been transcribed and will be made public later in 2026. Gilbert also recorded biographical interviews which are still awaiting processing.



The same list of translation prompts was used in both resources.

Interview Type	Amount
Open-Ended Interviews (OE)	925 interviews, ~ 308 hours
Translation: Gilbert Atlas (GA)	810 interviews, ~ 202 hours

Further translation lists used in some TGDP recordings are Eikel (160 interviews, ~ 40h) and Guion (345 interviews, ~ 57 hours). Transcription has been completed and data published on ZuMult (v. 2.1) for 529 recordings, ~200 hours of open ended interviews. The TGDP also collects extensive questionnaire data (spoken and written) for almost all informants.

Data processing and dissemination

- All audio recordings transcribed in **ELAN** (literary transcription)
- Automatic annotation: **Language tagging** and **orthographic normalisation** according to Blevins (2022), **lemmatisation** and **part-of-speech tagging** (STTS 2.0 with TreeTagger for German, UPenn with TreeTagger for English), mapping of language-specific POS to **UD POS tags**, canonical phonetic annotation (via G2P)

Open-ended interviews

- Manual annotation: **English word-by-word translation** and markup of **transfer and contact phenomena**

Translation interviews

- Actual speech in literary transcription ⇒ assigned to **prompt number and prompts** ⇒ mapped onto **Target Hypotheses** ("intended" translation without disfluencies or interactive elements), special conventions for alternatives, replies suggested by the interviewer, implied words, etc.

Prompt #71:	He's running now.
Transcription:	Er lauft. Or rennt in Bauerisch
Target Hypothesis:	Er lauft. +Er rennt.

- Final representation of transcripts in **ISO 24624:2016 Language resource management — Transcription of spoken language** (XML, based on guidelines of the Text Encoding Initiative)
- Extensive **metadata** on speakers and interviews
- Dissemination through a **corpus platform: ZuMult** at UT Austin (Boas et al. 2026, Fandrych et al. 2023): **Browsing, CQP query**

<https://tgd-zumult.la.utexas.edu/>

Texas German Dialect Project (TGDP)
Corpus platform - Version 2.1 (December 2025)

Welcome to the corpus platform of the Texas German Dialect Project (TGDP). This platform offers new ways of browsing and querying transcribed open-ended interviews with speakers of Texas German. You need a user account to access the browsing and query pages. If you do not have an account, you can register for free. A short documentation helps you get started with the platform functionality. To learn more about Texas German and the TGDP, you can visit the TGDP homepage.

Browse by interview
The browsing page is the starting point for exploring the data. It lists all interview data available in the platform together with metadata on the speakers. You can filter by place or year of recording and by speakers and interviewers. Transcripts and their annotations can be viewed with synchronised audio playback and various functionality for closer inspection of transcripts.

Query
Faceted searches on the data (normalized tokens, lemmas (base or class categories), queries CQP languages. Results are (KWIC) from each search in transcript context and

Diachronic Comparison

Example 1: Use of /ɹ/ (voiced alveolar approximant)

- /ɹ/ not present in German phonological inventory of German donor dialects
- /ɹ/ prevalent in American English
- Increasingly Americanized pronunciation of German lexemes due to contact with English?

⇒ search all normalized tokens that contain <r> and listen to their phonetic realization
 ⇒ CQP Query: **[norm="*.r.*"]**
 ⇒ 240,603 hits in all data, can be narrowed down to specific POS or lemmas
 ⇒ Listen to audio in the KWIC concordance

Example 2: English loanwords / Language mixing

- Language mixing occurs, can be expected to increase over time
- Some areas of the vocabulary more affected than others?
- Grammatical environments (dis)favouring language mixing?

⇒ Basic pattern for loanwords: English word in-between two German words
 ⇒ CQP Query: **[lang="deu"] [lang="eng"] [lang="deu"]**
 ⇒ 11,237 hits in contemporary TGDP data (0.75%) vs. 28 hits in historical Gilbert data (0.03%)
 ⇒ Closer inspection in the KWIC concordance (see above)



Thomas Schmidt
linguisticbits.de
thomas@linguisticbits.de

Margaret Blevins
University of Texas, Austin
mblevins@utexas.edu

Hans C. Boas
University of Texas, Austin
hcb@mail.utexas.edu

Glenn Gilbert
University of Southern Illinois
glenngilbert1869@gmail.com



<https://tgd.org/> • <https://tgd-zumult.la.utexas.edu/>