Systematic Normalization of Spoken Mixed-Language, Mixed-Dialect Data

Margaret Blevins
The University of Texas at Austin
mblevins@utexas.edu

Abstract

Literary transcriptions of spoken language often deviate from standard, written language. These variations can lead to higher than desirable error rates in NLP processing. This is particularly the case for spoken data of low resource varieties, including dialects and contact varieties of higher resource languages. This paper outlines a proposal for the systematic dialect-to-standard normalization of spoken language from language contact and dialect contact situations. This proposed system is then validated and evaluated on the Texas German Sample Corpus (approx. 13 hours), a set of audio and transcripts of Texas German conversations. Texas German is an umbrella term for a set of a heritage varieties of German spoken in Texas, USA that descend from multiple German dialects and that have been in contact with English for 150+ years. The proposed normalization system, along with the accompanying language-tagging system, can act as a starting point for other projects interested in normalizing their mixed variety data.

Keywords: dialect-to-standard normalization, language contact, Texas German

1. Introduction

Transcriptions of spoken language often use modified orthography ("literary transcription", "eye dialect") to more faithfully represent the actual phonetic form of words. In that way, they deviate from standard, written language. At the same time, many NLP tools are trained using standard(-near) texts. When such tools are used on transcriptions of spoken language, it can lead to higher than desirable error rates (see the OOV 'out-of-vocabulary' problem).[1] A similar situation can occur when using NLP tools trained on high resource standard languages on low resource dialects of that language. These issues multiply when working with spoken mixed-language, mixed-dialect data. Creating a normalized annotation layer for transcripts of spoken, mixed-language data can provide automatic NLP tools with more standard-like data, thereby increasing their accuracy.

This paper outlines an approach to normalization for spoken language contact data, mapping non-standard forms to standard language equivalents whenever possible. This system:

- provides a standardized, flexible system to normalize (non-standard, mixed-language, mixed-dialect) spoken data,
- puts normalization decisions on an explicit, linguistically grounded basis, thereby making normalization decisions transparent and reproducible, and
- is publicly accessible.

The importance of this latter point is often overlooked. Although (pre-)processing steps have many implications for searchability and downstream processing and knowing how earlier annotation decisions are made helps researchers use data more accurately and effectively, the decisions made during these steps are often not made public.

As a validation in a real world corpus compilation scenario, the guidelines proposed here are applied to a set of approx. 13 hours of Texas German conversational interview transcripts. Texas German is an umbrella term for a set of heritage contact varieties of German in Texas, USA. As such, it contains a great deal of dialectal variation and contact phenomena, which makes it a useful test for many normalization challenges. The purpose of this study is to develop and test a set of normalization guidelines by manually annotating a set of data and then use those annotations to (a) create a pipeline for automating normalization for this dataset and (b) use the manually annotated data as a gold standard to calculate error rates of the automated process.[2]

[1] E.g., Westpfahl and Schmidt (2013, 140) demonstrate that using an automatic pos-tagger on literary transcriptions leads to a high error rate because the transcriptions of spoken forms do not have lexicon entries in the tagger.

[2] Manually inserting the LANG tags (see section 3) took about 3.3 minutes per audio minute. Manually inserting the NORM tags (see section 5) took about 3.3 minutes per audio minute. Thus, this process took about 7 minutes per audio minute. See conclusion for error rates.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 58–69
16 May 2026.

2. Background

2.1. Texas German

In the mid-1800s, German-speaking immigrants from several dialect regions in central Europe settled in central and south-central Texas (i.e., the "Texas German Belt", see Jordan 1977, 3). Over the last 150+ years, these immigrants and their descendants have come into contact with each other and with other people in the area who speak other languages such as American (Southern) English and (Mexican) Spanish. This resulted in a multitude of (mutually intelligible) contact varieties throughout the Texas German Belt.

Texas German contains a great deal of variation throughout the linguistic spectrum and within its literary transcripts. Section 2.2 below briefly demonstrates several kinds of variation that can appear in transcripts of Texas German audio.[3]

2.2. Variation Within Transcripts

Variation within literary transcripts of spoken language-contact data has three primary sources: the speaker(s), the transcriber(s), and the annotation system(s). Brief examples of each of these within Texas German are provided below.

2.2.1. Inter- and Intra-Speaker Variation

Inter- and intra-speaker[4] variation can occur at all levels of speech and can have dialectical, language contact, or other roots, e.g.,

1. Phonological variation, e.g.,
   [**t**iːɣ̊] vs. [**d**iːɣ̊] 'animal' (SG[5]: [tiːɣ̊])
2. Morphological variation, e.g.,
   er*zählen* vs. ***ver****zählen* 'to tell (e.g., a story)' (SG: *erzählen*)
3. Lexical variation, e.g.,
   Magenschmerzen vs. *Leibweh* vs. *Panzweh* 'stomach ache'
4. Morphosyntactic variation, e.g.,
   (a) case:
   mit ***sie***-NOM/ACC vs. *mit* ***ihr***-DAT 'with her' (SG: *mit ihr*)
   (b) possessive:
   meine Frau ihre Mutter (lit. 'my wife her mother') vs. *die Mutter meiner Frau* (lit. 'the mother my-DAT wife') 'my wife's mother' (SG: *die Mutter meiner Frau*)

Code-mixing and other transfer phenomena can also occur at a variety of levels of speech, e.g.,

5. Foreign language lexeme(s)
   (a) *... wo die* ***fence*** *gebaut ist*
   (TGDP 1-51-1-27-a)[6]
   'where the fence was built / is'
   (SG: *... wo der Zaun gebaut wurde / ist*)
   (b) *sie haben gefragt ob sie — **how do you say 'spies'** — ob sie spies waren*
   'they asked whether they — how do you say 'spies' — whether they were spies'
   (SG: *sie haben gefragt ob sie — wie sagt man 'spies' — ob sie Spione waren*)
6. Mixing of morphemes from different languages
   (a) *ge****jump****t* 'jumped' (SG: *gesprungen*)
   (b) ***fence****n* 'fences' (SG: *Zäune*)
7. Phrasal calques
   ich kann nicht für sicher sagen
   I can not for sure say

   'I can't say for sure.'
   (SG: *ich kann nicht mit Sicherheit sagen*)

2.2.2. Inter- and Intra-Transcriber Variation

Once speakers produce spoken language, how a transcriber depicts that language can add another layer of variation within transcripts. A single transcriber may transcribe something inconsistently, or there may be variation between different transcribers, e.g.,

8. Use of special characters:
   *gro****ß*** vs. *gro****ss*** 'big' (SG: *groß*)
9. Use of abbreviations and/or punctuation:
   St. *John****'s*** vs. ***Saint*** *John****s***
10. Location and inclusion of word boundaries:
    nennses *"links"* vs. ***nen se s*** *"links"*
    (*nennen sie es "links"*, 'they call it "links"',

[3] For the purposes of this study, I use literary transcripts from the Texas German Dialect Project Corpus (TGDP Corpus). This corpus is part of the larger Texas German Dialect Project (TGDP, tgdp.org).

[4] Intra-speaker variation refers to variation within a single speaker's speech—a single person may vary, e.g., how they say a particular word throughout a conversation. Inter-speaker variation refers to variation between different people, e.g., person A and person B may use different lexemes for a particular object.

[5] SG = Standard German

[6] This is an interview segment ID number from the TGDP Corpus. It can be read as follows: the first number is the interviewer ID, the second number is the speaker ID, the third number is the interview number, and the fourth number is the interview segment number. Thus, this is the twenty-seventh segment of interview number 1 that occurred between interviewer 1 and speaker 51.

lit. 'call they it "links"')
(SG: *nennt man sie Verbindungen*)

11. Difference in identification/interpretation:
[fo:g̊] being transcribed as *for* (English) vs.
vor (German)

12. Transcriber misunderstanding:
poodles (English) vs. *Puter* 'turkeys'

Inter-speaker and inter-transcriber variation are present in any transcription of spoken language. Transcriber variation can be reduced by having detailed transcription guidelines and transcription quality checks.[7]

2.2.3. Inter-Corpus and Transcription System Variation

Additionally, there is inter-corpus variation with respect to transcription systems. There are many systems for transcribing spoken language, e.g., cGAT (Schmidt et al., 2023), HIAT (Rehbein et al., 2004), and Jefferson (Jefferson, 2004). Example (13) shows how a reference to the Texas-based grocery store H-E-B could be transcribed according to two transcription systems:

13. Transcriptions referring to H-E-B

    (a) cGAT (Schmidt et al., 2023, 33)
    ‹aitch› ‹ey› ‹bee›

    (b) TGDP project-internal guidelines
    ‹H-E-B› (or) ‹HEB› (or) ‹H.E.B.›

The two transcriptions above differ both in terms of orthography (i.e., which characters are used to represent each letter of the alphabet) and in terms of how many tokens they contain (three tokens vs. one token, respectively).

While researchers may want to minimize some variation (e.g., transcriber variation), they may want to embrace other kinds of variation to achieve a project's goals (e.g., dialectal variation). Deciding what to (not) include within transcriptions and annotations is an integral step of designing and building a corpus.

[7]The TGDP Corpus transcriptions are particularly variable due to (a) the amount of speaker variation, (b) the project's relatively limited transcription guidelines, and (c) the fact that L1, L2, and L3 German speakers transcribe the data, none of whom are native speakers of Texas German and all of whom have varying degrees of familiarity with German dialects. For example, the German word *wahrscheinlich* 'probably' is spelled in at least 10 different ways: *wahrscheinlich, verscheinlich, washeinlich, wahrscheinlih, wahrscheinlich, vescheinlich, vescheilich, warscheinlich, warscheinlih*, and *werscheinlich*. See also Bailey et al. (2005) for a discussion of transcriber effects on phonetic transcription.

2.3. Definitions of Normalization

As mentioned in the introduction, this paper proposes a systematic normalization system. Normalization is a common pre-processing step in NLP which systematically decreases variation within data, thereby making downstream processing more accurate.

CLARIN describes normalization as "the process of transforming parts of a text into a single canonical form" (CLARIN). However, which canonical form to choose (and why) differs between researchers and research communities.[8]

Within the NLP community, normalizing often means something like reducing all characters to their lower-case equivalents, **stemming**, and/or **lemmatization** (Bird et al., 2009, 107–108).[9] Another normalization task can be "identifying non-standard words, including numbers, abbreviations, and dates, and mapping any such tokens to a special vocabulary" (Bird et al., 2009, 108).[10] For example, ‹July 4, 2020›, ‹4th of July, 2020›, and ‹Fourth of July, 2020› could all be normalized to ‹2020-07-04›. Text normalization can also refer to "the task of transforming lexically variant words to their canonical forms" (Mehmood et al., 2020).[11]

Similarly, in the computer-mediated communication (CMC) community, the term text normalization has been used to mean "the challenge of discovering the English [i.e., the relevant standard language] words corresponding to the unusually-spelled words used in social-media messages and posts" (Jahjah et al., 2016, 180).[12]

Within the Text-to-Speech research community (TTS),[13] text normalization means "convert[ing] a written representation of a text into a representation of how that text is to be read aloud" (Yolchuyeva et al., 2018, 589), e.g., changing ‹$5› to ‹five dollars› (Baumann et al., 2019).

The term normalization has also been used within the historical corpus community, albeit with differing meanings. Bollmann et al.'s (2012) corpus of Early New High German texts contains both

[8]Bigi (2014) provides a brief overview of previous approaches to text normalization (a task that is generally thought to be language-dependent and task-dependent) and attempts to make a text normalization approach that is as little language- and task-dependent as possible.

[9]Also sometimes referred to as **text normalization**. Both stemming and lemmatization have the goal "to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form" (Manning et al., 2009, 32). For information, see Manning et al. (2009, 32–34).

[10]See also Sproat et al. (2001, 289), Bigi (2014, 515).

[11]See also **lexical normalization** (Toska, 2020).

[12]See also Kuparinen (2023).

[13]Often linked to the Automatic Speech Recognition (ASR) community.

a normalization and a **modernization** annotation layer. The normalization layer "stay[s] close to the original and, e.g., keep[s] historical inflection even if it violates modern morpho-syntactic constraints" while the modernization layer "adjust[s] inflection to [fit] modern constraints" (Bollmann et al., 2012, 342).[14] There are several corpora that refer to these kinds of annotation processes as **regularization** or regularized spelling, e.g., the EarlyPrint project and the British National Corpus (BNC).[15] Alternatively, Jurish et al. (2013) describe **canonicalization** as "assigning an extant equivalent to each word of the input text and deferring application analysis to these canonical cognates" (see also Jurish (2010, 2011); Gotscharek et al. (2009); Reffle et al. (2008); Makarov and Clematide (2020)).

A similar approach, **dialect-to-standard normalization**, i.e., "transforming dialectal text into a standard variety while maintaining as much of the original meaning as possible," has been used to help lower-resource varieties benefit from the tools built for higher-resource standardized varieties (Dimakis et al. (2025, 1); see also Joshi et al. (2025)).[16]

It suffices to say that the term normalization does not have a single, consistent definition across corpora and different sub-fields of linguistics. For the purposes of this paper, I use the term normalization to refer to mapping literary transcription forms to their standard-orthographic equivalent (e.g., *zwohundert* → *zweihundert* 'two hundred').[17] This definition does not state, however, that the single representation needs to be a standard, *grammatical* language form, see section 5 for more information.

[14]The RIDGES (Register in Diachronic German Science) corpus (Lüdeling et al., 2020) has similar annotation layers but calls them by different names: RIDGE's 'clean' annotation layer appears to be comparable to Bollman et al.'s (2012) 'normalization' layer, and RIDGE's 'normalization' layer appears to be similar to Bollman et al.'s (2012) 'modernization' layer.

[15]See Burnard (2007, 324) for information about regularized spelling in the BNC.

[16]For more information about dialect-to-standard normalization, see e.g., Kuparinen et al. (2023), Partanen et al. (2019).

[17]The term **normorthographic** is also used in similar corpora, such as the SiNProjekt (Sprachvariation in Norddeutschland, 'Language Variation in Northern Germany') and Regionalsprache.de (REDE) (Kehrein and Vorberger, 2018, 141–142). As Rehbein and Schalowski (2013, 208) note, "[a]dding an orthographic normalisation to the transcription might be seen as a 'poor man's target hypothesis.'"

2.4. Previous Corpora

The guidelines outlined in sections 3-5 below are based on the annotation guidelines for 10 non-standard varieties of spoken German, five of which are extraterritorial contact varieties: the Texas German Dialect Project (TGDP) (Boas et al., 2010); the *Namdeutsch* 'Namibia German' corpus (DNam) (Zimmer et al., 2020); the *Unserdeutsch* (Rabaul Creole German) corpus (Maitz et al., 2016; Maitz and Volker, 2017; Götze et al., 2017); the *Russlanddeutsche Dialektdatenbank* 'Russian German Dialect Database' (RuDiDat) (Berend and Frick, 2021); and the *Digitales Portal: Ungarndeutsches Zweisprachigkeits- und Sprachkontaktkorpus Projekt* (UZSK) 'Digital portal: Bilingual and language contact corpus of German as a minority language in Hungary' (Földes, 2016a,b,c).

The other five corpora are spoken varieties of (non-standard) German within DACH: the *Forschungs- und Lehrkorpus Gesprochenes Deutsch* (FOLK), a general corpus of spoken German (Kaiser, 2018; Schmidt, 2016, 2018; Deppermann and Hartung, 2011); the *KiezDeutsch* corpus (KiDKo), a corpus of speech from relatively mono- and multi-ethnic communities in Berlin (Wiese et al., 2012; Rehbein et al., 2004); the Berlin Map Task corpus (BeMaTaC), a learner corpus (Sauer and Lüdeling, 2016); ArchiMob, a dialectal, Swiss German corpus (Scherrer et al., 2019a,b); and a code-switching corpus of Turkish-German (SAGT) (Çetinoglu, 2017).[18]

3. Language Tagging

The first step of normalizing language contact data is to language tag the data. This is because the language assigned to a token has implications for how it should be normalized. For example, if the token <ham> (i.e., a common spoken form of German *haben* 'to have') were tagged as German, then it should normalized to <haben> (i.e., the standard German form) but if the same token were tagged as English, it should be normalized as <ham> (meaning 'pork meat product'). This language interpretation also has implications for other annotations such as part-of-speech, with <haben> being tagged as a verb and <ham> being tagged as a noun.

The LANG annotations in the system proposed in this paper are token annotations in which the annotator records which language(s) they interpret a particular token to be.[19] The purpose of the language layer is to make annotators' interpretations and assumptions transparent. For example, this

[18]See language resources section for corpora URLs.

[19]The language tags are based on the ISO 639-2 guidelines (ISO).

fictional Texas German sentence would be tagged as follows:

14. **tok** mir ham die die gerollt
    lang deu deu deu eng mix.deu+eng

Example (14) means *Wir haben den Würfel geworfen* 'We rolled the die.' These language tags are based on each token's phonological form, morphological form, part-of-speech, semantics, and (sometimes) surrounding tokens.

As a first step, if a token's (phonological[20] and morphological) form, meaning, and part-of-speech all correlate to a lexical entry in a standard reference work, then it should be tagged with a standard language tag, such as `deu` (German) or `eng` (English), see Figure 1.

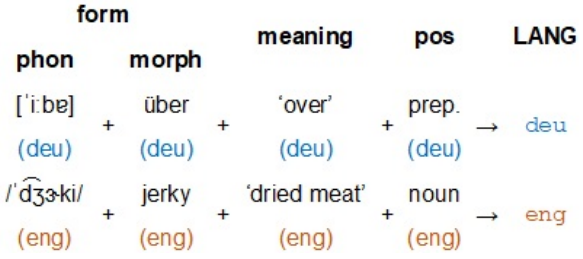

Figure 1: Depiction of scenarios leading to standard language tags.

When a token's form, meaning, and/or part-of-speech do *not* correlate to a lexical entry in a standard reference work, the token should be language-tagged with something other than a standard language tag. This can occur for several reasons, briefly illustrated below.

If a token has a form that is made up of German morphemes but there is no lexical entry for that token's intended meaning and/or part-of-speech, then it should be tagged as `deu.txg`. See e.g., Figure 2.[21]

If there is morphological/lexical material from multiple languages present in a single token, that token should be marked as `mix.LANG+LANG`. For example, the token *gemoved* 'moved' would be tagged as `mix.deu+eng` because it is comprised of a mixture of German *ge-* and English *moved*. The language tags are written in alphabetical order

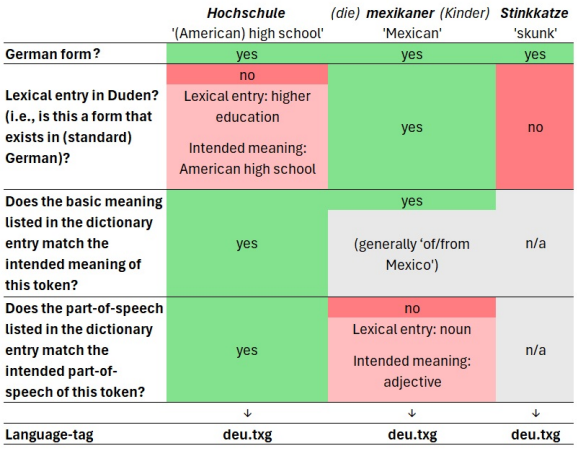

Figure 2: Different `deu.txg` scenarios.

to keep them consistent throughout the corpus.[22]

The language-tag `xxx` is used in situations in which it is unclear which language a particular token is. For example, all material marked as unintelligible (indicated by a (???) in the TOK layer) would be language-tagged as `xxx`.[23]

If a token was incomplete (e.g., the speaker cuts themselves off mid-word), but the presumed language of the intended token was relatively unambiguous, then the token should be language-tagged with the appropriate language, albeit with a hash tag (#) immediately preceded the language token to mark it as an interpretation of an incomplete utterance.

An asterisk is used to indicate that there are multiple possible interpretations for a given token. Each interpretation is marked with an asterisk. For example, the token *mädchens* could have been interpreted as an entirely German token, albeit with a nonstandard German plural marker (the standard German plural would be the null-plural *die Mädchen*), or it could have been interpreted as a German noun (*Mädchen*) + the English plural marker *-s*. The former interpretation would lead to a `deu` language-tag, while the later interpretation would lead to a `mix.deu+eng` language-tag. Both of these interpretations are included in the language-

[20] The phonology does not need to be a perfect match, but it should be similar.

[21] If one wanted to annotate for several dialects under the umbrella of a standard language, one could use different subtags, e.g., `deu.bai` for Bavarian or `deu.pla` for Low German. The main thing is to systematically and consistently use a set of language tags throughout an entire dataset. Distinguishing between related dialects is its own challenge, see e.g., Heeringa (2004).

[22] Pronouncing English words with a German "accent" does not trigger the use of the `mix` tag.

[23] In the version of these guidelines described in Blevins (2022), the `xxx` tag includes three subtags for material that is consistently language-ambiguous: `xxx.hes` (filled pauses / vocalized hesitations), `xxx.exc` (the exclamation *oh*), and `xxx.bcl` (backchanneling, such as *uh-huh* to indicate agreement). This was done to expedite the manual annotation process. Including labels to common non-lexical utterances such as `xxx.hes` helped improve subsequent pos-tagging.

tag layer, e.g.,[24]

15. **tok** mädchens
    lang *deu *mix.deu+eng

For a summary of the language tags and symbols within the LANG layer, see Table 1.[25]

4. Ambiguous Situations withing Language Tagging

I approach potentially ambiguous tokens in one of three ways: create a system to disambiguate them, always mark them as ambiguous, or always mark them in the same way, although they are ambiguous. These approaches are discussed in more detail below.

4.1. Disambiguating Cognates and Homographs Using Phonology

One way of disambiguating potentially language-ambiguous tokens is to rely on phonology. This is particularly useful if the pronunciation of the tokens is question is quite distinct. For example, German *Musik* [muˈziːk] has second syllable stress while English *music* /ˈmjuzɪk/ has first syllable stress, making them usually easy to differentiate.

4.2. 4.2 Cognates and Homographs with Similar Phonology

For cognates that are more similar than those described in 4.1, there are three options: (1) always mark such tokens as ambiguous, (2) always choose the same language for a particular token, or (3) have a system for deciding whether to mark something as a single language or ambiguous between multiple languages. All three of these approaches are taken in different situations within this proposed annotation system.

[24]The asterisk was used in a similar way in the normalization layer to indicate ambiguity.

[25]The standard language reference works used are as follows: German = *Duden* (online, www.duden.de); English = Merriam-Webster online (www.merriam-webser.com); Spanish = *Diccionario del español de México* 'Dictionary of Mexican Spanish' (DEM) (https://dem.colmex.mx//); Czech = (1) *Slovník spisovného jazyka českého* 'The Dictionary of the Written Czech Language' (2011), https://lexiko.ujc.cas.cz/texts/ssjc.html, (2) ORAL2013 portion of the Czech National Corpus (www.korpus.cz) (Křen, 2015; Válková et al., 2012), and (3) *Deutsch-böhmisches Wörterbuch* (Dobrovský et al., 1802-1821); Wendish = (1) *Prawopisny słownik hornjoserbskeje rěče: Hornjoserbsko-němski słownik* 'German-Upper Sorbian Dictionary' (Völkel, 2005), (2) *Niedersorbisch-deutsches Wörterbuch* 'Lower Sorbian-German Dictionary' (Starosta, 1999) (niedersorbisch.de).

4.2.1. Tokens That are Always Marked as Ambiguous

For this particular language pair (German-English), the following situations were always marked as language ambiguous.

Situation 1: A speaker pluralizes a German noun with an *-s* plural marker (for German nouns that use plural markers other than *-s* in standard German), e.g., *Tante**s*** 'aunts' (SG: *Tante**n***), *Baum**s*** 'trees' (SG: *B**ä**um**e***). This is because the *-s* plural marker is present in both German and English. Therefore, if it occurs in an "unexpected" place, it is unclear if it is a non-standard German plural ending or a borrowing of an English plural ending.

Situation 2: Tokens for which it is unclear whether a speaker is using an English loan that has been phonologically adapted to reflect German-like pronunciation, or whether they are using a term that was in the original Texas German donor dialects, and that that term has gone through semantic shift (e.g., *Grad* '(school) grade' (SG: *Klasse*), *Trubel* 'trouble' (SG: *Ärger*).

Situation 3: When it is ambiguous whether a speaker omitted a final schwa or whether they pronounced an English word according to German phonology, e.g., *Seit* 'side' (SG: *Seite*).

4.2.2. Ambiguous Tokens That are Always Tagged as a Single Language

There are two particularly common tokens that are always language-tagged as `deu` although they could technically be considered ambiguous.

First, Texas German speakers commonly say *denn* meaning 'then' (standard German *dann*). It is ambiguous whether *denn* should have the language tag `deu` (implying that it is a dialectal, phonological variant of *dann*, leading to the normalization form *dann*), `deu.txg` (implying a that it is German form with a non-standard meaning, which would lead to the normalization form *denn*), or `eng` (implying that it is a dialectal, phonological form of English *then*, leading to the normalization form *then*). Thus, if one wanted to capture all three of these potential interpretations, one would language-tag (and normalize) *denn* 'then' as `*deu *deu.txg *eng` (leading to the normalization forms ***dann* *denn* and ***then*, respectively). However, in the proposed system, *denn* 'then' is language-tagged as `deu` and normalized as <dann>. This is done because *denn* is an attested form of standard German *dann* 'then' in continental German dialects (e.g., in Cologne[26]).

[26]https://dat-portal.lvr.de/woerterbuch/D – "denn": "Häufig für 'dann': "Gestern wa ich ers bein Aazt, un denn wa ich noch auf en Maakt. Wenn schon, denn schon! (nicht kleckern, sondern klotzen!) Ich kauf mir

Tag	Explanation	Example
`deu`	German	*Kuh, gespielt*
`deu.txg`	German material ... (a) with non-standard semantics, or (b) that does not appear in *Duden*	(a) *Luftschiff* 'airplane', *Hochschule* '(American) high school' (b)*Stinkkatze* 'skunk', *mitaus* 'without'
`eng`	English	*well, baseball*
`spa`	Spanish	*leche* 'milk'
`ces`	Czech	*pivo* 'bier'
`wen`	Wendish	*wótnic* 'grandmother'
`mix.lang + lang` (e.g., `mix.deu+eng`)	Mixture of morphological / lexical material from multiple languages	*Hickorybaum* 'hickory tree', *gemovt* 'moved', *schlipperich* 'slippery'
`* + lang` (e.g., `*deu`)	Ambiguous language	*in* 'in', *Mädchens* 'girls'
`# + lang` (e.g., `#deu`)	Unknown but (partially) reconstructable material	*ausgezeich*
`xxx`	Unintelligible material, material for which the intended language is indeterminate	(???), *g*

Table 1: List of available language tags

A similar approach is taken with the token *is* 'is'. If a speaker pronounces *is* as [ɪs] (not US-American English /ɪz/) and if the token is otherwise surrounded by German tokens, then it is marked as `deu` and not as ambiguous.

16. **tok** sie is schnell
lang deu deu deu
norm sie ist schnell
'she is fast'

Similar to *denn* 'then,' this is done because *is* [ɪs] is an attested colloquial pronunciation of standard German *ist* [ɪst] and practically speaking, this saves time for the annotator, as this is a common occurrence in the corpus. Otherwise, it is treated like *in*, see section 4.2.3.

4.2.3. Tokens for Which There is a Disambiguation System

Certain tokens have a great deal of overlap with regards to orthographical form, meaning, and pronunciation in English and German. One common example of that is the preposition *in*.

Instead of always tagging *in* as `deu`, `eng`, or `*deu *eng`, the language tag is reliant on the language of the tokens to the immediate left and right of the in token. If *in* is surrounded by German tokens, it is marked as `deu` (e.g., (17)). If it is surrounded by English tokens, it is tagged as `eng` (e.g., (18)). If a German token is on one side, and an English token on the other, it is tagged as ambiguous (e.g., (19)). If it is surrounded by German tokens but used in a way that is not available in standard German, it is tagged as `deu.txg` (e.g., (20)).

17. **tok** irgendwo in die gegend
lang deu deu deu deu
'somewhere in the area'
(TGDP 1-51-1-9-a)

18. **tok** fighting in the civil war
lang eng eng eng eng eng
(TGDP 1-51-1-23-a)

19. **tok** nicht mehr in saloon
lang deu deu *deu *eng eng
'[she can] no longer [go] into the saloon'
(TGDP 10-171-3-24-a)

20. **tok** was das in deutsch ist
lang deu deu deu.txg deu deu
'... what that is in German'
(TGDP 9-122-5-1-13-a)

5. Normalization

5.1. Definition

The NORM ('normalization') layer is a token annotation that exists in addition to the above-mentioned layers. I use the term *normalization* to mean associating transcribed utterances to a set of consistent, systematic representations, the majority of which follow standard language orthography. This can be seen as a kind of dialect-to-standard normalization. I am primarily guided by Scherrer and Samardžić (2016, 1) idea of "mapping the variants of what can be identified as the

jetzt nen Daimler, wenn schon, denn schon!"

same word to a single representation," Bird et al.'s (2009, 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 lemma and part-of-speech.

The normalizations in this proposed system are token-based reductions of phonological and/or orthographic variation. No changes are made to grammar (e.g., case, gender), (lexical / morphological) word choice, word order, part-of-speech, or lemma. No tokens are inserted or removed, but contractions can be split into their constituent parts.

For example, normalization can disambiguate homographs that occur due to Texas German pronunciation patterns, e.g., both *Tür* 'door' and *Tier* 'animal' being pronounced (and transcribed) as ‹tier› due to the presence of an unrounded vowel where standard German has a rounded vowel. In the process of normalization, ‹tier› 'animal' would be normalized to ‹Tier› and ‹tier› 'door' should be normalized to ‹Tür›, thereby disambiguating these instances.

These normalizations are not a translation into (more) standard German, nor are they a "correction" of non-standard grammatical variation. For example, *Schrippe* 'roll/bun' (Berlin dialect) would not be translated to the more ubiquitous, standard term *Brötchen*, as these are two entirely different lexemes. Also, if a speaker says *mit er*-NOM or *mit ihn*-ACC 'with him' although standard German would require the dative *ihm* for 'him' following the preposition *mit* 'with', *mit er* and *mit ihn* would be normalized to *mit er* and *mit ihn* (i.e., no changes would be made to their forms) because it would be inaccurate to claim that *er* and *ihn* are phonological or orthographic forms of *ihm*.[27]

There are several main principles that these normalization guidelines follow:

1. Normalization forms are based on the presumptions implied by the language tag each token is given. (e.g., see ‹ham› discussion in section 3).

2. If a token is marked with a standard language language-tag (e.g., `eng` or `deu`), then the orthographic form in the normalization layer should be an existing standard language reference work.[28]

3. The presumed part-of-speech and lemma of the token in the TOK layer should also be the same as the part-of-speech and lemma for the token in the NORM layer

4. In order to achieve this form (as described in points 2 and 3), phonological and orthographic changes are permissible, while lexical and grammatical changes are not.

5. Tokens language-tagged as something other than a standard language have different rules, see section 5.2.

5.2. Non-Standard Situations

Tokens tagged as `mix` or `deu.txg` were morphologically normalized, i.e., each morphological segment of a token was normalized in its own right, based on each morpheme's presumed source language if there was one, see Table 2.

LANG	transcribed form	NORM
`deu.txg`	‹Stinkkatz› *stink* 'stink' (deu) + *Katze* 'cat' (deu)	‹Stinkktze› 'skunk'
`mix.deu+eng`	‹Futtershtore› (*Futter* 'feed' (deu) + *shtore* (eng) 'store'	‹Futterstore› 'feed store'
	‹fencen› *fence* 'fence' (eng) + *n* (plural) (deu)	‹fencen› 'fences'
	‹schlipperich› *slipper-* (eng) + *-ig* 'y' (deu)	‹slipperig› 'slippery'

Table 2: Normalization of tokens with non-standard tags

If a token was incomplete (e.g., the speaker cuts themselves off mid-word), but the presumed intended word was relatively unambiguous (e.g., due to a later mention of the same word) then the normalized token was preceded by a pound sign (`#`) to mark it as a reconstruction of an incomplete utterance. Portions (syllables) of the reconstructed utterance that were not produced by the speaker

[27] This differs from, say, an L2 corpus of German in which annotators may want to annotate speaker "errors" / target hypotheses. Although Texas German shares many similarities with L2 German, it is not the goal of this annotation to make their utterances conform to standard German grammar.

[28] For each standard language tag, a standard reference work was chosen to guide the spelling of each normalized form. The reference work(s) a given project uses should be explicitly stated. See footnote 25 for a list of the standard reference works used in this study. While this is possible for (varieties of) standardized languages, a slightly different approach may be necessary when normalizing a (variety of a) language that does not have a standardized orthography. For example, if one wanted to normalize a corpus of Pennsylvania German, it would be helpful to choose a particular dictionary / spelling convention to guide the normalization process and make those decisions public.

originally were included in square brackets, e.g., <#ausgezeich[net] > 'excel[lent]'.

Backchanneling utterances contained a great deal of orthographic variation. In order to systematize the forms, each meaning of a backchanneling signal was given a single form. For example, all affirmative signals such as <hmhm>, <uh-huh>, and <mhmm> were normalized to the form <hmhm> (see Blevins, 2022, 627).

6. Texas German Sample Corpus

In order to develop and test the effectiveness of the language-tagging and normalization guidelines as described above, I manually annotated a set of Texas German conversational audio (the Texas German Sample Corpus, TGSC). The audio and initial transcripts of the TGSC come from the Texas German Dialect Project Corpus, a collection of Texas German interviews from 2001-today (see Boas et al. (in press)).

The TGSC is collection of randomly selected audio from 150 randomly sampled speakers that is representative of the TGDP's first 600 Texas German speakers who:

- were recorded by the TGDP itself (i.e., not recordings that were subsequently donated to the project),
- do *not* speak a particular unique sub-dialect of Texas German (e.g., Texas Alsatian),
- were born in Texas and still live in Texas,
- whose ancestors came to Texas before 1917, and
- grew up speaking Texas German and learned Texas German at home through relatives.

The TGSC is proportionally representative based on the extra-linguistic variables sex and birthplace (county) and is available for download at https://doi.org/10.18738/T8/IOX9ZA.

7. Conclusion

The guidelines outlined here are presented in order to highlight some of the challenges that occur when normalizing spoken language-contact data and proposes possible solutions to those challenges. Admittedly, teasing apart two related languages (here: German and English) may have different challenges than other contact situations (e.g., a tonal language in contact with an atonal language, or a language in contact with another language that uses a different writing system, etc.). These guidelines are therefore not intended as a one-size-fits all solution; they provide a resource for other projects to use as a starting point, and/or to react against rather than starting from scratch. It is a step towards transparency of pre-processing decisions, thereby facilitating downstream processing and the accuracy of future research using these materials.

After the TGSC was constructed, it was used to train an automatic tagger to add LANG and NORM annotations to the larger TGDP Corpus. Following the process described in Boas et al. (in press), this led to a language tagging error rate of 6% and a normalization error rate of 7%. One of the additional benefits of the TGSC itself is that it is a step towards developing a comprehensive dictionary of (dialectal) terms that do not exist in any of the related standard language dictionaries (cf. Dimakis et al., 2025).

8. Acknowledgments

The author would like to thank Thomas Schmidt for his valuable insights and support; Hans C. Boas, the founder and director of the Texas German Dialect Project, for his support and for providing access to the data; and the hundreds of Texas German speakers who have generously agreed to be interviewed by the TGDP.

9. Availability

The Texas German Sample Corpus is available for download at https://doi.org/10.18738/T8/IOX9ZA and may be used for non-commercial and non-AI/LLM purposes. See https://tgdp.org/dialect-archive/ for more information regarding user rights and responsibilities.

10. Ethics Statement

All participants were informed about the project and the intended use of the recordings. Written consent was obtained for recording, transcription, and inclusion of their data in the corpus. Directly personally identifying information was removed during transcription. When making the data available, I am complying with FAIR principles, observing the restrictions resulting from the informed consent.

11. Bibliographical References

G. Bailey, J. Tillery, and G. Andreas. 2005. Some effects of transcribers on data in dialectology. *American Speech*, 80(1):3–21.

T. Baumann, A. Köhn, and F. Hennig. 2019. The Spoken Wikipedia Corpus: Harvesting, alignment and an application to hyperlistening. *Language Resources and Evaluation*, 53(2):303–329.

N. Berend and E. Frick. 2021. Russlanddeutsche Dialekte online: Dokumentation, Präsentation und Recherche deutscher Auslandsvarietäten im Internet am Beispiel des Russlanddeutschen. In C. Földes, editor, *Kontaktvarietäten des Deutschen im Ausland*. Narr Francke Attempto.

B. Bigi. 2014. A multilingual text normalization approach. In *Human language technology challenges for computer science and linguistics, 8387*, pages 515–526.

S. Bird, E. Klein, and E. Loper. 2009. *Natural language processing with Python*, 1st edition. O'Reilly.

M. Blevins. 2022. *The Language-Tagging and Orthographic Normalization of Spoken Mixed-Language Data, with a Focus on Texas German*. Ph.D. thesis, The University of Texas at Austin.

H.C. Boas, M. Pierce, K.A. Roesch, G. Halder, and H. Weilbacher. 2010. The Texas German Dialect Archive: A multimedia resource for research, teaching, and outreach. *Journal of Germanic Linguistics*, 22(3):277–296.

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

M. Bollmann, S. Dipper, J. Krasselt, and F. Petran. 2012. Manual and semi-automatic normalization of historical spelling: Case studies from Early New High German. In *Proceedings of KONVENS 2012, The 11th Conference on Natural Language Processing (LThist 2012 workshop)*, Vienna, Austria.

L. Burnard, editor. 2007. *Reference Guide for the British National Corpus (XML Edition)*.

CLARIN. *Tools for normalization*.

Ö. Çetinoglu. 2017. A code-switching corpus of Turkish-German conversations. In *Proceedings of the 11th Linguistic Annotation Workshop*, Valencia, Spain.

A. Deppermann and M. Hartung. 2011. Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des 'Forschungs- und Lehrkorpus Gesprochenes Deutsch' (FOLK) am Institut für Deutsche Sprache (Mannheim). In M. Müller E. Felder and F.Vogel, editors, *Korpuspragmatik: Thematische Korpora als Basis diskurslinguistischer Analysen*, pages 414–450. De Gruyter.

A. Dimakis, J. Pavlopoulos, and A. Anastasopoulos. 2025. Dialect normalization using large language models and morphological rules. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 23696–23714.

J. Dobrovský, S. Leška, A. J. Puchmajer, and V. Hanka. 1802-1821. *Deutsch-böhmisches Wörterbuch*. In der Herrlischen Buchandlung, Prag.

C. Földes. 2016a. Diskurse im Wirkungsraum von Zweisprachigkeit: Werkstattbericht aus einem Forschungs- und Dokumentationsprojekt. In M. Dus, R. Kolodziej, and T. Rojek, editors, *Wort–Text–Diskurs*, page 321–336. Peter Lang GmbH Internationaler Verlag der Wissenschaften.

C. Földes. 2016b. Ungarndeutsche Sprachvariation und Mehrsprachigkeit: Ein Korpusprojekt auf der Basis von empirischer Feldforschung und Online-Sprachdokumentation. *Sprachtheorie und germanistische Linguistik*, 26(2):167–190.

C. Földes. 2016c. Ungarndeutsches Zweisprachigkeits- und Sprachkontaktkorpus: Konzept, Design und Inhalte. *Zeitschrift für interkulturelle Germanistik*, 7(1):167–181.

A. Gotscharek, U. Reffle, C. Ringlstetter, and K. U. Schulz. 2009. On lexical resources for digitization of historical documents. In *Proceedings of the 9th ACM Symposium on Document Engineering*, pages 193–200, New York. Association for Computing Machinery.

A. Götze, S. Lindenfelser, S. Lipfert, K. Neumeier, W. König, and P. Maitz. 2017. Documenting Unserdeutsch (Rabaul Creole German): A workshop report. *Journal of the Linguistic Society of Papua New Guinea*, pages 65–90.

W. J. Heeringa. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. thesis, University of Groningen.

ISO. 2016. *ISO 24624:2016 Language resource management — Transcription of spoken language*.

V. Jahjah, R. Khoury, and L. Lamontagne. 2016. Word normalization using phonetic signatures. In *Proceedings of the 29th Canadian Conference on Artificial Intelligence*, pages 180–185.

G. Jefferson. 2004. Glossary of transcript symbols with an introduction. In G. H. Lerner, editor, *Conversation Analysis: Studies from the First Generation*, pages 13–31. John Benjamins, Amsterdam and Philadelphia.

T. G. Jordan. 1977. The German element in Texas: An overview. *Rice University Studies*, 63(3):1–11.

A. Joshi, R. Dabre, D. Kanojia, Z. Li, H. Zhan, G. Haffari, and D. Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, 57(6):1–37.

B. Jurish. 2010. Comparing canonicalizations of historical German text. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, page 72–77, Uppsala, Sweden. Association for Computational Linguistics.

B. Jurish. 2011. *Finite-state canonicalization techniques for historical German*. Ph.D. thesis, Universität Potsdam.

B. Jurish, M. Drotschmann, and H. Ast. 2013. Constructing a canonicalized corpus of historical german by text alignment. In P. Bennett, M. Durrell, S. Scheible, and R. J. Whitt, editors, *New Methods in Historical Corpora*, pages 221–234. Narr.

J. Kaiser. 2018. Zur Stratifikation des FOLK-Korpus: Konzeption und Strategien. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion*, 19:515–552.

R. Kehrein and L. Vorberger. 2018. Dialekt- und Variationskorpora. In M. Kupietz and T. Schmidt, editors, *Korpuslinguistik*, pages 125–150. De Gruyter.

M. Křen. 2015. Recent developments in the Czech National Corpus. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 1–4.

O. Kuparinen. 2023. Murrevikko – a dialectally annotated and normalized dataset of Finnish tweets. In *Proceedings of the 10th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 31–39, Dubrovnik. Association for Computational Linguistics.

O. Kuparinen, A. Miletić, and Y. Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828, Singapore. Association for Computational Linguistics.

A. Lüdeling, C. Odebrecht, Schnelle G. Krause, T., and C. Fischer. 2020. Ridges herbology.

P. Maitz, W. König, and C. A. Volker. 2016. Unserdeutsch (Rabaul Creole German): Dokumentation einer stark gefährdeten Kreolsprache in Papua-Neuguinea. *Zeitschrift für germanistische Linguistik*, 44(1):93–96.

P. Maitz and C. A. Volker. 2017. Documenting Unserdeutsch: Reversing colonial amnesia. *Journal of Pidgin and Creole Languages*, 32(2):365–397.

P. Makarov and S. Clematide. 2020. Semi-supervised contextual historical text normalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7284–7295.

C. D. Manning, P. Raghavan, and H. Schütze. 2009. *An introduction to information retrieval*. Cambridge UP.

K. Mehmood, D. Essam, K. Shafi, and M. K. Malik. 2020. An unsupervised lexical normalization for Roam Hindi and Urdu sentiment analysis. *Information Processing & Management*, 57(6).

N. Partanen, M. Hämäläinen, and K. Alnajjar. 2019. Dialect text normalization to normative standard Finnish. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT)*, pages 141–146, Dubrovnik, Croatia.

U. Reffle, C. Ringlstetter, A. Gotscharek, and K. U. Schulz. 2008. Successfully detecting and correcting false friends using channel profiles. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(3):17–22.

I. Rehbein and S. Schalowski. 2013. STTS goes Kiez: Experiments on annotating and tagging urban youth language. *Journal for Language Technology and Computational Linguistics*, 28(1).

J. Rehbein, T. Schmidt, B. Meyer, F. Watzke, and A. Herkenrath. 2004. *Handbuch für das computergestützte Transkribieren nach HIAT*, volume 56 of *Arbeiten zur Mehrsprachigkeit*.

S. Sauer and A. Lüdeling. 2016. Flexible multilayer spoken dialogue corpora. *International Journal of Corpus Linguistics, Special issue: Complation, transcription, markup and annotation of spoken corpora*, 21(3):419–438.

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

Y. Scherrer, T. Samardžić, and E. Glaser. 2019a. ArchiMob: Ein multidialektales Korpus schweizerdeutscher Spontansprache. *Linguistik Online*, 98(5):425–454.

Y. Scherrer, T. Samardžić, and E. Glaser. 2019b. Digitising Swiss German: How to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53(4):735–769.

T. Schmidt. 2016. Construction and dissemination of a corpus of spoken interaction: Tools and workflows in the FOLK project. In M. Kupietz and A. Geyken, editors, *Corpus Linguistic Software Tools*, pages 117–144.

T. Schmidt. 2018. Gesprächskorpora. In M. Kupietz and T. Schmidt, editors, *Korpuslinguistik*, pages 209–230. De Gruyter, Berlin and Boston.

T. Schmidt, W. Schütte, J. Winterscheid, M. Schürmann, S. Reineke, and E. Schedl. 2023. *cGAT: Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2)*.

R. Sproat, A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 2001. Normalization of non-standard words. *Computer Speech and Language*, 15(3):287–333.

M. Starosta. 1999. *Niedersorbisch-deutsches Wörterbuch*. Domowina-Verlag, Bautzen.

M. Toska. 2020. A rule-based normalization system for Greek noisy user-generated text. Master's thesis, Uppsala University.

L. Válková, M. Waclawičová, and M. Křen. 2012. Balanced data repository of spontaneous spoken Czech. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 3345–3349, Istanbul. European Language Resources Association.

P. Völkel, editor. 2005. *Prawopisny słownik: Hornjoserbsko-němski słownik / Obersorbisch-deutsches Wörterbuch*, 5th edition. Domowina-Verlag, Bautzen.

S. Westpfahl and T. Schmidt. 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.

H. Wiese, U. Freywald, S. Schalowski, and K. Mayr. 2012. Das KiezDeutsch-Korpus: Spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. *Deutsche Sprache*, 40:97–123.

S. Yolchuyeva, G. Németh, and B. Gyires-Tóth. 2018. Text normalization with convolutional neural networks. *International Journal of Speech Technology*, 21(4).

C. Zimmer, H. Wiese, H. J. Simon, M. Zappen-Thomson, Y. Bracke, B. Stuhl, and T. Schmidt. 2020. Das Korpus Deutsch in Namibia (DNam): Eine Ressource für die Kontakt-, Variations- und Soziolinguistik. *Deutsche Sprache*, 3:210–232.

12. Language Resource References

BNC Consortium. 2007. *The British National Corpus, XML Edition*. Oxford Text Archive.

Deppermann, A. and Hartung, M. and Schmidt, T. and Reineke, S. 2025. *Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)*.

Frick, E. 2019. *Russlanddeutsche Dialektdatenbank (RuDiDat)*.

Földes, C. 2019. *Digitales Portal: Ungarndeutsches Zweisprachigkeits- und Sprachkontaktkorpus Projekt (UZSK)* .

Maitz, P. and König, W. 2022. *Unserdeutsch Rabaul Creole German (UNSD)*. Archiv für Gesprochenes Deutsch (AGD), IDS Mannheim.

Margaret Blevins. 2022. *Texas German Sample Corpus*.

Mueller, M. and Loewenstein, J. *Early Print Project*.

Texas German Dialect Project. 2025. *Texas German Dialect Project Corpus Platform*.

Sauer, S. 2016. *The Berlin Map Task Corpus (BeMaTaC)*.

Scherrer, Y. and Samardžić, T. and Glaser, E. 2019. *ArchiMob*.

Simon, H. and Wiese, H. 2019. *Deutsch in Namibia (DNam)*. Archiv für Gesprochenes Deutsch (AGD), IDS Mannheim.

Wiese, H. and Freywald, U. and Rehbein, I. and Schalowski, S. 2015. *Das KiezDeutsch-Korpus*.

Çetinoğlu, Ö. 2004. *SAGT: Computational Structural Analysis of German-Turkish Code-Switching*.