Saar-Voice: A Multi-Speaker Saarbrücken Dialect Speech Corpus

Lena S. Oberkircher, Jesujoba O. Alabi, Dietrich Klakow, Jürgen Trouvain
Language Science and Technology (LST), Saarland University,
Saarbrücken, Germany
{s8leober@stud, jalabi@lsv, dietrich.klakow@lsv, trouvain@lst}.uni-saarland.de

Abstract

Natural language processing (NLP) and speech technologies have made significant progress in recent years; however, they remain largely focused on standardized language varieties. Dialects, despite their cultural significance and widespread use, are underrepresented in linguistic resources and computational models, resulting in performance disparities. To address this gap, we introduce **Saar-Voice**, a six-hour speech corpus for the Saarbrücken dialect of German. The dataset was created by first collecting text through digitized books and locally sourced materials. A subset of this text was recorded by nine speakers, and we conducted analyses on both the textual and speech components to assess the dataset's characteristics and quality. We discuss methodological challenges related to orthographic and speaker variation, and explore grapheme-to-phoneme (G2P) conversion. The resulting corpus provides aligned textual and audio representations. This serves as a foundation for future research on dialect-aware text-to-speech (TTS), particularly in low-resource scenarios, including zero-shot and few-shot model adaptation.

Keywords: TTS, German, Saarbrücken, Saarland, Low-resourced Variety

1. Introduction and Background

The field of natural language processing (NLP) has seen substantial advances in recent years, driven by large-scale models and increased computational resources. Nevertheless, research and tools remain heavily focused on standardized language varieties, leaving dialectal varieties underrepresented despite their widespread use and importance to speakers' cultural identity (Blasi et al., 2022). In Germany, for example, more than 40% of the population regularly speaks regional dialects, yet these varieties are often subject to social stereotyping (Adler and Hansen, 2022). This marginalization is reflected in the technological landscape: large language models (LLMs) and speech processing systems frequently struggle with dialectal variation, as they are predominantly trained on standardized language data (Krückl et al., 2025). Consequently, recent years have seen growing efforts to develop language resources for dialectal varieties (Blaschke et al., 2024; Faisal et al., 2024; Blaschke et al., 2025) and to systematically evaluate the performance of existing NLP systems on such data (Bui et al., 2025; Muñoz-Ortiz et al., 2025).

In this paper, we contribute to this line of work by focusing on the Saarbrücken dialect of German spoken in the state of Saarland through the creation of **Saar-Voice**, a six-hour multi-speaker speech corpus.[1] In the following subsections, we provide some basic background about the Saarbrücken dialect, including the main linguistic characteristics of the dialect.

[1]The dataset is available on https://huggingface.co/datasets/UdS-LSV/Saar-Voice.

1.1. Dialect areas

The Saarland is one of Germany's 16 federal states (Bundesländer), located in the southwest of the country and bordering France and Luxembourg. The Saarland has around 1 million inhabitants. The state is, much like the rest of Germany, a linguistically diverse region, with many variants of the standard language, from regiolects to dialects.

A typical problem in linguistics is defining which exact dialect is meant with a given name. In the region of the Saarland, the broad range of regiolects and dialects are loosely referred to as "Saarländisch", but this is not a precise term. On a large scale, "Saarländisch" consists of two different dialect areas: Rhine Franconian and Moselle Franconian (see Figure 1). Both dialect areas are rather diverse on all linguistic levels, mainly phonology, morpho-syntax and lexicon. Looking more closely, there is large variance even within the two regions.

We decided in this paper to consider the Rhine Franconian dialect spoken in and around the capital of and largest city in the Saarland, Saarbrücken, for which a grammar (Steitz, 1981) and a suggested dictionary (Braun and Mangold, 1984) exist, both authored by linguists, as well as a community of active authors (with no background in linguistics).

1.2. Spelling

One big advantage of a standard language is the existence of a standard spelling. For (German) dialects standardized orthographies are missing, which leads to high variance in spelling between

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 12–23
16 May 2026.

Figure 1: The isogloss *dat-das-line* distinguishing areas in which Rhine Franconian dialects ("das") and Moselle Franconian dialects ("dat") are spoken (Drenda, 2008) in Saarland (smaller area in the south-west marked with thin border lines) and Rhineland-Palatinate.

and even within speakers. Often, speakers use spontaneous spelling, orthographically representing the phonology each word as they best see fit. For instance, the exact spelling of "es gebbd" (e. 'there is', 'there are') can vary a lot: "es gebbd, gebd, gäbbd, gäbd, gebbt, gäbt"; or "Zeidung" (e. "newspaper') as "Zeidung, Tseidung, Dsaidung".

Generally, speakers use the letters available in the Standard German alphabet ("a-z, ä, ö, ü" and "ß"). Additionally, depending on the speaker, certain diacritics that are not part of the standard language may be used. For example, "ò" may be used to represent the phoneme /ɔ/ (such as in "dò", "dòò", "do" or "doo", all representing /dɔː/, meaning "there"), and in some cases *á* to represent the phoneme /ɐ/ as in "vákaafe" (/fɐkaːfə/, e. 'sell').

1.3. Differences between dialect and standard language

Were dialects simply accents of the standard language, the approach to creating any speech system would be to simply 'translate' the standard pronunciation to the regional pronunciation (ideally with the intonation of the accent). However, dialects are not just the standard language with regional accents. In fact, they often differ to written standard language on various levels:

- **Orthography and phonology:** Dialectal writing reflects phonological distinctions that are not represented in Standard German orthography. For example, vowel quality and length may be encoded differently, and consonant correspondences do not always align with standard spelling conventions.

- **Morpho-syntax:** Grammatical markers such as case or gender markings may differ between the dialect and standard language, among other morpho-syntactic differences.

  – Standard German: *ein* (indefinite article, masc./neut.) + *schön* (nice) + *Überraschung* (surprise; fem.) → *eine schöne Überraschung*

  – Dialect: *e* (indefinite article, generalized form) + *scheen* (nice) + *lwwerraschung* (surprise) → *e scheeni lwwerraschung*

  Here, both the article system and adjectival inflection differ from Standard German.

- **Lexical variation:** Dialects employ vocabulary items that are absent from Standard German or derived from different etymological sources, e.g., *Gellerieb* ('carrot', from "gelbe Rübe", in Standard German "Karotte" or "Möhre"), *Grumbier* ('potato', from "Grundbirne", in Standard German "Kartoffel"), or *Troddwaa* ('sidewalk', from French "trottoir", in Standard German "Gehsteig" or "Bürgersteig").

An important feature of dialect texts is that they are rarely found in the everyday 'linguistic landscape'. This has not only consequences for feelings of insecurity of spelling among speakers, including inconsistencies of spelling of identical words in the same text, but also for potential readers. Most *speakers* of dialects are not very good *readers* of dialect text, though they are very fluent and competent readers of standard language (but maybe speakers with a regional accent).

1.4. Text sources

There are different sources of dialect-writing, for example:

- web-based: chats, blogs
- books written by (non-professional) dialect authors
- research pieces, e.g. dictionaries, monographs and translations of local dialects
- texts from authors organized in dialect clubs
- public media, e.g. news in local radio stations, glosses in local newspaper

Source	Domain	# Sent	# Words	# U. Words
Printed Books				
An da Saar gefonn	Poetry	698	6,228	1,978
Geschaffd - Geläädbd	Poetry; Prose	394	3,302	1,214
Saa, was de willschd	Prose; Poetry	2,216	19,711	3,147
Was wääs dann isch	Poetry; Prose	2,525	18,445	3,490
Locally Sourced				
(Locally Sourced Texts)	Prose; Folktale; Poetry	2,838	26,789	5,221
Localized Translation				
MASSIVE German	Localized Data	101	805	394
Total		8,772	75,280	11,303

Table 1: Textual corpus statistics of the collected texts for the Saarbrücken dialect.

Depending on selected types of text, the styles can hugely differ. For instance, many dialect writers produce poems, in which the stanzas often rhyme at the end. In contrast, in chats reduced forms of written conversations can be observed which are again quite different to some dialect prose. Thus, the selected text source used for later steps can have a strong bias of style. This is especially important when creating a speech corpus, which may in the end contain biases in pronunciation and intonation.

2. Related Work

German Speech Resources Several resources exist for German spoken language research, but most are limited in dialectal coverage. Existing datasets include unannotated speech collections for speech representation learning, such as VoxLingua107 (Valk and Alumäe, 2020), VoxPopuli (Wang et al., 2021), JesusDramas, WikiTongues, and MMS ULAB v2 (Chen et al., 2024), which are mostly derived from social media, religious texts, or read speech for Standard German. Automatic speech recognition (ASR) datasets include FLEURS (Conneau et al., 2022), and dataset from initiatives like Mozilla Common Voice provide more structured recordings, while TTS datasets such as MLS (Pratap et al., 2020) (also useful for ASR), FLEURS-R (Ma et al., 2024), Thorsten-Voice Dataset (Müller, 2024), and CML-TTS (Oliveira et al., 2023), created from MLS, include high-quality recordings of standard German.

In recent years, some efforts have targeted German dialects for ASR and TTS, covering varieties such as Swiss German (Plüss et al.), Viennese (Schabus et al., 2013; Pucher et al., 2010), non-Vienna Austrian (Pucher et al., 2017), Luxembourgish (Steiner et al., 2017), and three dialect groups spoken in Southeast Germany (Franconian, Bavarian, Alemannic) (Blaschke et al., 2025). However, several dialects, such as the Saarbrücken dialect, remain underrepresented, and to the best of our knowledge only a small corpus is available in Bible MMS (Lux et al., 2024). Furthermore, a recent survey (Blaschke et al., 2024) showed that German speakers are interested in language technologies that can process dialectal (audio) input, and creating and evaluating such systems requires developing speech corpora for underrepresented dialects.

To address these gaps, we introduce **Saar-Voice**, a multi-speaker speech corpus for the Saarbrücken dialect, providing high-quality recordings suitable for training and evaluating modern TTS systems.

Speech Corpora for Dialects and Low-Resource Languages Beyond German-specific efforts, the creation of speech corpora for dialectal and low-resource settings has been widely explored across languages (Zampieri et al., 2020; Ramponi, 2024; Lent et al., 2022; Guellil et al., 2021; Alabi et al., 2025c). Studies on dialectal variants across a range of languages, such as Arabic (Malmasi and Zampieri, 2016; Djanibekov et al., 2025; Talafha et al., 2025), English (Ahamad et al., 2020; Xiao et al., 2023; Olatunji et al., 2023), and African languages (Ahia et al., 2024; Emezue et al., 2024), highlight the importance of accounting for regional and social variation when developing speech datasets, as models trained on standardized varieties often fail to generalize to non-standard speech (Diab, 2016; Aji et al., 2022; Ahia et al., 2024; Alabi et al., 2025b).

In parallel, a substantial body of research has focused on speech data collection for low-resource and underrepresented languages. These settings are typically characterized by limited availability of speakers, scarce linguistic resources, and, in some cases, the absence of standardized orthography (Blaschke et al., 2024). As a result, corpus construction in such contexts often relies on adaptable methodologies, including crowd-sourced data collection, community-driven recording initiatives, and

Speakers	
Total speakers	9
Gender ratio	4F / 5M
Speech	
Recorded sentences	4,871
Total duration (hh:mm:ss)	05:54:48
Avg. duration per sent. (sec)	4.37
Min duration of sent. (sec)	1.59
Max duration of sent. (sec)	14.02

Table 2: Saar-Voice Corpus Statistics.

semi-supervised or lightweight annotation strategies (Emezue et al., 2024; Olatunji et al., 2023).

Despite progress, many speech datasets focus only on common languages and pay little attention to dialect differences. In addition, datasets for low-resource languages often lack consistent annotation or enough speakers, which limits their usefulness for training and evaluating models. To address these challenges, we introduce a carefully curated speech corpus of a German dialect, with nine speakers and about six hours of recordings. The dataset is designed to provide a compact but high-quality resource for studying non-standard German speech and for training and benchmarking speech models, especially TTS, in low-resource dialect settings.

3. Corpus Design

This section describes the methodology used to create the Saar-Voice corpus.

3.1. Text Collection

Digitization of Printed Books For this study, we digitalized four books available in print at the Saarland University library. The books are "An da Saar gefonn: volkstümliche Gedichte in Saarbrücker Mundart" (Jungmann, 1993), "Geschaffd - Geläabd: Mundarttexte" (Fox, 1994), "Saa, was de willschd: Mundart-Kolumnen" (Fox, 2012) and "Was wääs dann isch...?!" (Eckert, 1995). These texts were scanned and digitized using an online OCR software[2]. This OCR software was chosen in particular as it was one of the few softwares that had close to no issue recognizing the special characters "á" and "ò", which caused great problems with other softwares and were usually misclassified as "ä" and "ö", respectively. Remaining errors introduced by the OCR process were removed through manual review by a native speaker of the dialect, while also cross-referencing the original source texts. 5,833 sentences were collected this way, amounting for 66.6% of the full dataset.

[2]https://ocr.ac/de

Locally Sourced Texts 2,838 sentences, accounting for 32.4% of the total data, were collected from internally available texts written by authors from the local community. These texts were already available digitally and only checked for possible spelling errors by a native speaker.

Localized Translations Lastly, we sampled 101 German sentences from the MASSIVE dataset (FitzGerald et al., 2023). These sentences were manually translated with the support of the dictionary of Braun and Mangold (1984). After translation, the data was localized by replacing entities with localized entities, such as locations and personal names. These localized entities were manually inserted to further emphasize dialect-specific variation. Localized location names were taken at random from Braun (1991), as in Example 1. Additionally, numbers were spelled out in dialect orthography rather than given as numerals, as in Example 2, to ensure consistent pronunciation across speakers.

(1) **German Original**: Bitte plane ein Treffen mit Petra in Wiesbaden am Mittag.
German Original (Entities Replaced): Bitte plane ein Treffen mit Anna in Oberbexbach am Mittag.
Translated: Bidde plaan e Dreffe midd Anna in Oberbexbach am Middaach.
Localized: Bidde plaan e Dreffe midd Anna in Owwerbeddschbach am Middaach.

(2) **German Original**: Erinnere mich an Lauras Geburtstag 24 Stunden vorher.
German Original (Entities Replaced): Erinnere mich an Melanies Geburtstag 19 Stunden vorher.
Translated: Erinner misch an Melanies Geburdsdaa 19 Schdunne vòòrhäär.
Localized: Erinner misch an Melanies Geburdsdaa neindsehn Schdunne vòòrhäär.

An overview over text statistics for each of the collected resources can be found in Table 1. This includes an estimation of the domains each of the resources cover, as annotated by a native speaker. Most resources are collections of texts rather than a single running texts. Therefore, for resources that are associated with multiple domains, domains are reported in descending order of frequency.

3.2. Speaker Recruitment

A total of nine participants were recruited using convenience sampling from people known to the research team. The speaker group consisted of four female and five male speakers. Age of participants was collected in categorical ranges. The largest age group was 26-30 years (n=3), followed by 31-35 years (n=2). The remaining participants were distributed across the ranges 18-25, 51-55, 56-60 and 61-65 (n=1 each).

Participants were either native speakers of a Rhine Franconian dialect, or native speakers of a closely related regional dialect with high familiarity and regular, long-term exposure to the target dialect. Self-identified dialect labels included: Saarländisch / Saarländischer Dialekt (*Saarland Dialect*; n=5), Moselfränkisch (*Moselle Franconian*, n=1), Rheinfränkisch (*Rhine Franconian* n=1), Rhein-Moselfränkischer Grenzdialekt (*Rhine-Moselle Franconian Border Dialect*, n=1), and Platt (*local term for the dialect*, n=1).

All speakers reported speaking the dialect daily (n=6) or at least multiple times a week (n=3) in the present time, and the majority (n=7) reports having spoken mainly the dialect during their childhood.

All speakers are also (native) speakers of Standard German and were "alphabetized" with Standard German spelling.

3.3. Recording Setup

Participants were invited to the departmental soundproof recording booth for audio recordings. The texts were presented to the speakers sentence by sentence on a 21.5-inch Full HD monitor. Audio was recorded at a sampling rate of 44.1kHz using a *DAP 2011* microphone, a high-quality directional microphone, using the SpeechRecorder Software (Draxler and Jänsch, 2004). Each recording session took around 1 hour, during which, depending on the participant, 200-300 sentences were recorded in batches of 100.

The speakers controlled their own pace of sentence presentation and recording by mouse clicks on a specified icon. If the speaker felt unsure or had a slip of the tongue s/he had the option for a new recording. Also, the recording supervisor (the first author) had the possibility to ask the speaker to restart the recording of a sentence.

Sentences presented to the speakers were not sampled independently, but derived from a longer, coherent text corpus. The entire textual corpus was concatenated, segmented into sentences and then grouped into chunks of 100 consecutive sentences. For each session, such chunks were randomly selected. This semi-randomized procedure ensured topical and lexical continuity within one session, while still maintaining coverage across the corpus and across domains.

Recording continuous passages like this offers several advantages. Besides keeping the speaker engaged in a cohesive story, it supports a more natural prosody and fluency, leading to consistent pronunciation of recurring and possibly unfamiliar lexical items, and enabling the resulting book subset to be used for long-context TTS experiments.

4. Saar-Voice Corpus

This section presents an overview of the created corpus.

4.1. Data Statistics

Table 2 summarizes the dataset statistics of Saar-Voice. The dataset consists of approximately six hours of speech, comprising 4,871 utterances from nine participants. The average length of the recorded sentence is about four seconds, with minimum and maximum durations of two and 14 seconds, respectively; these are moderately short but of good quality.

4.2. Phoneme Coverage

We applied Epitran (Mortensen et al., 2018) a multilingual grapheme-to-phoneme (G2P) model on the recorded text in Saar-Voice. Our analysis shows that the dataset contains 38 distinct phonemes, which lies in the typical range of phonemes estimated for Standard German (Kohler, 1990) and also the Saarbrücken dialect (Steitz, 1981) and (Braun and Mangold, 1984), if the vowels of German loan words are included. The most common phoneme is /d/, likely due to the voicing of many occurrences of consonants in the dialect that are typically unvoiced in Standard German (e.g., /t/ $\rightarrow$ /d/). It occurs 13,963 times within the speech corpus, and is followed by /n/ (11,284 occurrences), /s/ (10,963), /ʁ/ and /e/ (both 10,547 occurrences).

The least common phoneme is /ø/, occurring only 4 times in the entire speech corpus, followed directly by /œ/ (6 occurrences). Both are realizations of the Umlaut "ö", which is rather rare in the text itself, appearing only 10 times in total. They are followed by /ʏ/ (13 occurrences), /y/ (26 occurrences) and /ʒ/ (64 occurrences). All rounded front vowels /œ, ʏ, y/ can be regarded as phonemes from Standard German loan words.

The Epitran G2P model is multilingual and designed to cover standard German rather than the Saarbrücken dialect. While generally effective for estimating the phoneme coverage of this dialect corpus, the implications of this mismatch are discussed in Section 5.

4.3. Speaker Variability

Analysis of speaker variability shows that mean F0 (pitch) differs by gender. For male speakers, the mean F0 ranges from 115 Hz to 171 Hz (with a median of 107-164 Hz), while for female speakers it ranges from 201 Hz to 251 Hz (with a median of 196-249 Hz). These statistics indicate moderate inter-speaker variability in both pitch and tempo, reflecting natural differences in speech across male

Speaker	Gender	Total audio (hh:mm:ss)	# sentences	Mean F0	Median F0	F0 range (5–95%)	W/s
P01	F	00:23:17	301	203	200	141-278	1.8249
P02	F	01:02:28	1075	251	249	203-312	2.4166
P03	M	01:15:37	899	115	107	80-176	1.7352
P04	M	00:27:56	400	143	140	100-196	2.2056
P05	F	00:27:18	399	213	212	175-256	2.0203
P06	M	00:18:43	300	119	115	84-160	2.3334
P07	M	00:53:21	597	171	164	102-258	1.5016
P08	F	00:44:45	600	201	196	156-265	1.7920
P09	M	00:21:22	300	147	141	101-203	1.8053

Table 3: Speaker-level acoustic statistics. F0 in Hz. W/s = words per second.

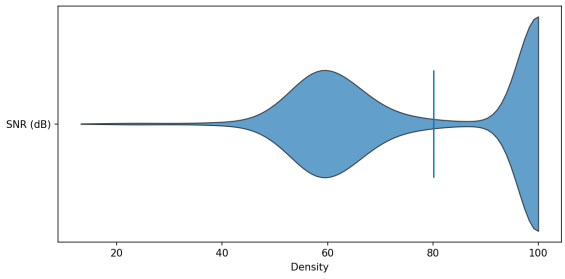

Figure 2: SNR distribution on Saar-Voice.

and female voices. We also observe speech rates ranging from 1.5 to 2.4 words per second across both genders, with values evenly distributed within each gender. Table 3 provides a detailed breakdown for all speakers.

4.4. Audio Quality

We evaluate the signal-to-noise ratio (SNR) for all the recorded audio samples in Saar-Voice using the WADA-SNR (Kim and Stern, 2008) algorithm and plot the distribution in Figure 2. The results show that most samples fall within the clean audio threshold, with SNR values ranging from 20 to 99, confirming the high overall quality of the Saar-Voice dataset.

5. Issues

In this section, we discuss the practical challenges and issues encountered during the corpus creation process.

5.1. Data-Specific Issues

A central challenge of building a corpus for a low-resource dialect like the variant of Rhine Franconian discussed in this project is the absence of standardized orthography for the target dialect. Unlike Standard German, the dialects of the Saarland do not have any official spelling conventions or vocabulary lists. Although dialect dictionaries like Braun and Mangold (1984) can serve as guidance, they do not provide an official, normative reference. Orthographic variation is thus unavoidable and inherent to the dialect. It further affects vocabulary size estimates, as orthographic variants of the same lexical item are counted as distinct types.

Ignoring such variation or attempting to impose a standardized orthography would in many cases even go against what native speakers actually desire: Blaschke et al. (2024) show that 65% of their participating speakers of German dialects explicitly oppose the introduction of a standardized orthography. Our own questionnaire showed no clear consensus regarding standardization. Taken together, these results suggest that there is no stable community-wide preference for standardization, and the more representative survey by Blaschke et al. (2024) indicates the opposite. Any attempt at standardization would risk privileging one subset of speaker preference over others, potentially misrepresenting actual usage practices, and harming user trust.

The choice of OCR software presented an additional technical challenge. Most available systems, such as Python's *tesseract* library, are optimized and effective for Standard German, but showed a substantial difficulty processing dialect-specific characters such as “á” or “ò”, as well as recognizing entire character strings which are not part of Standard German's inventory. This resulted in recognition errors and, even with the final choice of OCR software, required manual corrections, introducing additional processing effort and potential inconsistencies.

Further issues with available technologies arise in the lack of dedicated tokenization tools for the dialect. This led to being able to provide only an approximate vocabulary segmentation by whitespace tokenization. Without dialect-specific morphological or tokenization resources, which are hard to create due to data sparseness, word boundary detection may be imperfect. This in turn affects vocabulary size estimation and lexical frequency calculations.

Beyond orthography issues, the composition of the corpus itself raises questions of linguistic representativeness. The genre of poetry makes up

over one third (35.9%) of the dataset. This may, due to the genre's metrical and rhyming conventions, introduce lexical and especially prosodic and suprasegmental patterns that are unlikely to reflect the natural spoken dialect. More broadly, all textual resources, including prose and localized translations, represents the written register language, which inherently diverges from spontaneous speech. Nonetheless, it accurately represents the distribution of written data within this specific dialectal landscape. Additionally, the dataset lacks spontaneous speech. While this is less critical for the dataset's primary intended use case of TTS synthesis, it may limit its application to downstream tasks such as ASR. Future work should therefore focus on expanding the corpus with transcribed spontaneous speech to improve linguistic representativeness and genre coverage, and broaden its utility.

5.2. Speaker-Specific Issues

Although all speakers were carefully selected as Rhine Franconian speakers, variation within the speaker pool naturally remains inevitable. A variety of self-assessment data from the provided questionnaire indicates high overall proficiency / fluency, but variance appears in dialect usage patterns. Only one speaker reports speaking predominantly exclusively dialect with little to no Standard German influences, while most described their speech as a mixture of dialect and standard language. Some even indicated that they predominantly speak Standard German.

These differences, as well as considering small regional differences in the dialect, suggest that the corpus reflects not a uniform dialect realization, but rather a spectrum of usage patterns. This introduces additional variability for downstream modeling tasks.

5.3. G2P-Specific Issues

Beyond general data limitations, several issues emerged specifically during the G2P conversion process. Representative examples for each of the following issues can be found in Table 4.

Firstly, the most frequent detected IPA symbol in the output is the length mark /ː/, indicating that the model predicts a high number of lengthened vowels. However, through manual inspection, it becomes clear that vowel length is often assigned incorrectly. Words ending in the letter "e", are often transcribed to end in /eː/, which is incorrect. This mistake stands out in particular as the correct transcription would generally be /ə/, just like in Standard German. Also, it appears to misclassify double vowels, which are common in dialectal writing, as a sequence of two lengthened vowels.

Additionally, a common phenomenon in dialectal writing is the occurrence of a doubling of the letter "d", which is rather rare in Standard German and instead often corresponds to the Standard German "tt". This double consonant is misinterpreted by the model as /td/ on several occasions, which reflects an incorrect segmentation.

Lastly, the Epitran (Mortensen et al., 2018) G2P model generated some characters which are impossible in the German phoneme inventory, such as /m̀/, showing expected issues with processing the special characters "á" and "ò".

6. Dialectal TTS Modeling

The primary motivation behind creating Saar-Voice is to enable the development and evaluation of multi-speaker TTS systems. With recent advances in multilingual and multi-speaker (zero-shot) TTS models, such as XTTS (Casanova et al., 2024) and ZMM-TTS (Gong et al., 2024), it is now possible to investigate how well multilingually pretrained systems can generalize to closely related dialectal varieties, even when only limited data are available.

Both models include German in their pretraining, making adaptation to Saarbrücken dialect potentially feasible. However, recent research on low-resource Bildts (a Dutch variety) (Do et al., 2025) observed speaker–language entanglement when adapting ZMM-TTS for unseen speakers (zero-shot), and for seen speakers, there were also multiple cases of mispronunciation (Alabi et al., 2025a). In contrast, Pine (2025) shows that combining Dutch and Bildts and using the StyleTTS (Li et al., 2023) architecture is sufficient for the same task. Hence, our dataset will help to better understand how multilingually pretrained TTS models and architectures generalize to closely related dialectal varieties.

Such evaluation of zero-shot and fine-tuned TTS models using the dataset are a goal of future work.

7. Conclusion

This paper presented the creation of a multi-speaker corpus for a low-resource German dialect. The dataset was constructed using partially OCR-based text extraction and manual normalization, and speech was produced by a variety of speakers. The resulting corpus provides both textual and phonetic representation suitable for speech technology research, with the option to expand the quantity of speech data. The challenges faced in the creation of the corpus reflect broader issues in low-resource dialect modeling. While they may introduce variability, they capture what authentic dialect usage looks like in native speakers.

Word	English	Epitran	Ours
berechne	calculate (Imp.)	bəʀɛxneː	bəʀɛxnə
bidde	please	bɪtdeː	bɪdə
Midde	middle	mɪtdeː	mɪdə
odder	or	ɔtdəʀ	ɔdɐ
buuch	book (Imp.)	buːuːx	buːx
ääner	one	æːæːnəʀ	æːnɐ
òòmens	in the evening	oːòːm̀əns	ɔːmns
Schdigg	piece	ʃdɪçk	ʃdɪg

Table 4: Examples of G2P Errors. Corrections of transcriptions are estimates by a native speaker.

Empirical validation of the dataset through downstream experiments remains an important next step. TTS synthesis experiments using the corpus are left for future work. Through possible zero-shot inference, model fine-tuning and training approaches, further evidence of the dataset's utility for speech technology applications in this low-resource dialect setting may be provided.

Beyond the dataset creation and proposing use cases, we aim to emphasize the necessity of collaborating with the speaker community. Without direct engagement with the dialect writers' scene and the speakers, the amount of texts would have been dramatically reduced. The existence of this corpus is therefore closely tied to the active dialect-writing community.

In dialectal research and Natural Language Processing (NLP) as a whole, methodological decisions should not be guided solely by technical feasibility or research interest, but also by the expectations and preferences of the speaker community itself. Especially in low-resource and non-standard contexts, technological developments interacts directly with questions of representation, identity, language maintenance and culture.

Our questionnaire, following Blaschke et al. (2024), asked speakers directly whether they would welcome more digital applications supporting dialect spoken in the Saarland. The large majority of our participants expressed agreement (strong agreement: n=6; agreement: n=2; neutral: n=1). Similarly, most respondents agree that language technologies can contribute to dialect preservation (strong agreement: n=5; agreement: n=3; neutral: n=1). Despite the limited sample size and biased participant pool, these responses suggest that technological support may be perceived rather positively within the community.

Thus, we consider it an important detail that dialect NLP projects have a contact to the various communities using dialects for getting

- access to written texts and spoken materials,
- recommendations for book authors and relevant sources,
- insights into possible or prominent spelling patterns / conventions and variations,
- feedback on acceptable and desirable technological applications.

Community engagement does and should not only serve as a practical resource for data collection, but also as a means of aligning technological development with actual users.

8. Acknowledgments

We would like to thank all speakers who contributed their voices to this corpus, without whom this resource would not have been possible. Many thanks to the anonymous reviewers for their time and efforts, and for the extensive feedback and questions. Thank you also to our colleague Çağla Kints for carefully proof-reading this paper. Jesujoba Alabi was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

9. Supplementary Materials

9.1. Ethical Considerations

Informed consent. Participants signed a consent from in which they agreed to the process of the recording as well as the anonymized storage, sharing and processing of the recorded data. No personally identifying information of speakers was recorded, and all data is exclusively linked to an anonymized, randomized speaker ID. Participants were also informed that they could withdraw their consent at any time.

Data protection. All recorded data is stored in an anonymized form, only linking to the randomized speaker ID, on secure university servers.

Voluntariness. Participation was completely voluntary and occurred without compensation. No material incentives were offered to participants, and participation had no academic or professional consequences in any way.

Participant well-being. Great care was taken of keeping the duration of recording sessions to

around one hour, though participants were free to end their sessions early if they wished. Breaks were always possible, but taken at least after the recording of 100 sentences. Participants had the option to prepare for the recording, as they received the material to be read during their session digitally around one week in advance.

Community sensitivity. The dataset represents just a subset of the regional dialects of the Saarland. The process of curating the data involved no evaluation of dialectal "correctness", including no aim to standardize or change dialect varieties. The textual content of the dataset reflects the original authors' perspectives and was not modified or filtered to promote or follow any particular set of views.

10. Bibliographical References

Astrid Adler and Karolina Hansen. 2022. Dialekt und Beruf: neue Daten zu Dialekten in Deutschland. Sprache in Zahlen: Folge 7. *Sprachreport*, 38(3):28–33.

Afroz Ahamad, Ankit Anand, and Pranesh Bhargava. 2020. AccentDB: A database of non-native English accents to assist neural speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5351–5358, Marseille, France. European Language Resources Association.

Orevaoghene Ahia, Anuoluwapo Aremu, Diana Abagyan, Hila Gonen, David Ifeoluwa Adelani, Daud Abolade, Noah A. Smith, and Yulia Tsvetkov. 2024. Voices unheard: NLP resources and models for Yorùbá regional dialects. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4392–4409, Miami, Florida, USA. Association for Computational Linguistics.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Jesujoba O. Alabi, Cheng Gong, Erica Cooper, Yu Jiang, Dietrich Klakow, and Junichi Yamagishi. 2025a. Submission from ZMM-TTS for Blizzard Challenge 2025. In *The Blizzard Challenge 2025*, pages 31–36.

Jesujoba O. Alabi, Xuechen Liu, Dietrich Klakow, and Junichi Yamagishi. 2025b. AfriHuBERT: A self-supervised speech representation model for African languages. In *Interspeech 2025*, pages 4023–4027.

Jesujoba Oluwadara Alabi, Michael A. Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025c. Charting the landscape of African NLP: Mapping progress and shaping the road ahead. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27807–27841, Suzhou, China. Association for Computational Linguistics.

Verena Blaschke, Christoph Purschke, Hinrich Schuetze, and Barbara Plank. 2024. What do dialect speakers want? A survey of attitudes towards language technology for German dialects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 823–841, Bangkok, Thailand. Association for Computational Linguistics.

Minh Duc Bui, Carolin Holtermann, Valentin Hofmann, Anne Lauscher, and Katharina von der Wense. 2025. Large language models discriminate against speakers of German dialects. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8212–8240, Suzhou, China. Association for Computational Linguistics.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pages 4978–4982.

Mona Diab. 2016. Processing dialectal Arabic: Exploiting variability and similarity to overcome challenges and discover opportunities. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, page 42, Osaka, Japan. The COLING 2016 Organizing Committee.

Amirbek Djanibekov, Hawau Olamide Toyin, Raghad Alshalan, Abdullah Alatir, and Hanan Aldarmaki. 2025. Dialectal coverage and generalization in Arabic speech recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29490–29502, Vienna, Austria. Association for Computational Linguistics.

Phat Do, Matt Coler, Jelske Dijkstra, Igor Marchenko, Vass Verkhodanova, and Sebastien Le Maguer. 2025. The Blizzard Challenge 2025. In *The Blizzard Challenge 2025*, pages 1–17.

Christoph Draxler and Klaus Jänsch. 2004. SpeechRecorder - A universal platform independent multi-channel audio recording software. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Chris Chinenye Emezue, Ifeoma Okoh, Chinedu Emmanuel Mbonu, Chiamaka Chukwuneke, Daisy Monika Lal, Ignatius Ezeani, Paul Rayson, Ijemma Onwuzulike, Chukwuma Onyebuchi Okeke, Gerald Okey Nweya, Bright Ikechukwu Ogbonna, Chukwuebuka Uchenna Oraegbunam, Esther Chidinma Awo-Ndubuisi, and Akudo Amarachukwu Osuagwu. 2024. The IgboAPI dataset: Empowering Igbo language technologies through multi-dialectal enrichment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15932–15941, Torino, Italia. ELRA and ICCL.

Cheng Gong, Xin Wang, Erica Cooper, Dan Wells, Longbiao Wang, Jianwu Dang, Korin Richmond, and Junichi Yamagishi. 2024. Zmm-tts: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:4036–4051.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5):497–507.

Chanwoo Kim and Richard M. Stern. 2008. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In *Interspeech 2008*, pages 2598–2601.

Klaus Kohler. 1990. German. *Journal of the International Phonetic Association*, 20(1):48–50.

Xaver Maria Krückl, Verena Blaschke, and Barbara Plank. 2025. Improving dialectal slot and intent detection with auxiliary tasks: A multi-dialectal Bavarian case study. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 128–146, Abu Dhabi, UAE. Association for Computational Linguistics.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 19594–19621. Curran Associates, Inc.

Shervin Malmasi and Marcos Zampieri. 2016. Arabic dialect identification in speech transcripts. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 106–113, Osaka, Japan. The COLING 2016 Organizing Committee.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Alberto Muñoz-Ortiz, Verena Blaschke, and Barbara Plank. 2025. Evaluating pixel language models on non-standardized languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6412–6419, Abu Dhabi, UAE. Association for Computational Linguistics.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. 2023. AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Aidan Pine. 2025. The NRCC Submission to the Blizzard Challenge 2025. In *The Blizzard Challenge 2025*, pages 24–30.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

Bashar Talafha, Hawau Olamide Toyin, Peter Sullivan, AbdelRahim A. Elmadany, Abdurrahman Juma, Amirbek Djanibekov, Chiyu Zhang, Hamad Alshehhi, Hanan Aldarmaki, Mustafa

Jarrar, Nizar Habash, and Muhammad Abdul-Mageed. 2025. NADI 2025: The first multidialectal Arabic speech processing shared task. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 720–733, Suzhou, China. Association for Computational Linguistics.

Zedian Xiao, William Held, Yanchen Liu, and Diyi Yang. 2023. Task-agnostic low-rank adapters for unseen English dialects. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7857–7870, Singapore. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

11. Language Resource References

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation. In *Interspeech 2025*, pages 913–917.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Edith Braun. 1991. *Necknamen der Saar und drum herum*. Hempel-Verlag, Lebach.

Edith Braun and Max Mangold. 1984. *Saarbrücker Wörterbuch*. Saarbrücker Druckerei und Verlag.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024. Towards robust speech representation learning for thousands of languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10224, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Georg Drenda. 2008. *Kleiner linksrheinischer Dialektatlas. Sprache in Rheinland-Pfalz und im Saarland*. Franz Steiner Verlag.

Peter Eckert. 1995. *Was wääs dann isch ...?! : hundertzwanzigmal Mundart von der Saar*. KleinstLiteratur aus der edition händmäid. KleinstVerl. Die Kiste, Wadgassen-Differten.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECT-BENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Georg Fox. 1994. *Geschaffd - gelääbd: saarländische Mundarttexte*. cjm-Verlag, Speyer.

Georg Fox. 2012. *Saa, was de willschd!: Mundart-Kolumnen*. PVS-Edition, Heusweiler.

Kurt Jungmann, editor. 1993. *An da Saar gefonn: volkstümliche Gedichte in Saarbrücker Mundart*. Westwind-Verlag, Saarbrücken.

Florian Lux, Sarina Meyer, Lyonel Behringer, Frank Zalkow, Phat Do, Matt Coler, Emanuël A. P. Habets, and Ngoc Thang Vu. 2024. Meta Learning Text-to-Speech Synthesis in over 7000 Languages. In *Interspeech 2024*, pages 4958–4962.

Min Ma, Yuma Koizumi, Shigeki Karita, Heiga Zen, Jason Riesa, Haruko Ishikawa, and Michiel Bacchiani. 2024. FLEURS-R: A Restored Multilingual Speech Corpus for Generation Tasks. In *Interspeech 2024*, pages 1835–1839.

Thorsten Müller. 2024. Tv-44khz-full (revision ff427ec).

Frederico S. Oliveira, Edresson Casanova, Arnaldo Candido Junior, Anderson S. Soares, and Arlindo R. Galvão Filho. 2023. Cml-tts: A multilingual dataset for speech synthesis in low-resource languages. In *Text, Speech, and Dialogue*, pages 188–199, Cham. Springer Nature Switzerland.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. Swiss parliaments corpus, an automatically aligned Swiss German speech to standard German text corpus.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pages 2757–2761.

Michael Pucher, Carina Lozo, and Sylvia Moosmüller. 2017. Phone mapping and prosodic transfer in speech synthesis of similar dialect pairs. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, pages 180–185. TUDpress, Dresden.

Michael Pucher, Friedrich Neubarth, Volker Strom, Sylvia Moosmüller, Gregor Hofer, Christian Kranzler, Gudrun Schuchmann, and Dietmar Schabus. 2010. Resources for speech synthesis of viennese varieties. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Dietmar Schabus, Michael Pucher, and Gregor Hofer. 2013. Joint audiovisual hidden semi-markov model-based speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):336–347.

Ingmar Steiner, Sébastien Le Maguer, Judith Manzoni, Peter Gilles, and Jürgen Trouvain. 2017. Developing new language tools for MaryTTS: the case of Luxembourgish. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, pages 186–192. TUDpress, Dresden.

Lothar Steitz. 1981. *Grammatik der Saarbrücker Mundart*. Saarbrücker Druckerei und Verlag.

Jörgen Valk and Tanel Alumäe. 2020. Voxlingua107: A dataset for spoken language recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.