German Dialects Across Situations, Generations, and Regions: The REDE corpus as an Oral Resource for NLP

Hanna Fischer, Alfred Lameli
Research Center Deutscher Sprachatlas, Marburg University
Pilgrimstein 16, 35032 Marburg, Germany
{hanna.fischer, lameli}@uni-marburg.de

Abstract

Recent advances in speech and language technologies increasingly rely on large and diverse corpora that represent linguistic variation across dialect regions, communicative situations, and social speaker characteristics. While substantial resources are available for Standard German, comparable spoken corpora for German dialects have so far been largely lacking, limiting the development and evaluation of dialect-sensitive NLP systems. The REDE corpus addresses this gap by providing a methodologically uniform collection of spoken German for 148 locations that systematically covers all major dialect areas in Germany. It comprises contemporary recordings collected in multiple elicitation and interaction settings, capturing variation across speaking styles, situational contexts, and speaker generations. With more than 1,500 hours of speech and rich metadata on regional and social dimensions, the REDE corpus constitutes a large-scale oral resource suitable for both linguistic research and NLP applications. This paper presents the design, structure, and methodological foundations of the corpus and discusses its relevance for current speech technology requirements.

Keywords: corpus development, speech recordings, German, dialects, regional accents, REDE corpus, Wenker sentences

1. Introduction

The current state of the art in natural language processing (NLP) and regional language research—particularly with regard to spoken language data—is characterized by a marked discrepancy between rapid advances in applications for standard languages and persistent challenges in the acquisition, modeling, and processing of regional varieties (Netzwerk Regionale Sprache und Künstliche Intelligenz, 2026). Transformer-based models such as BERT (Devlin et al., 2019) and self-supervised speech representations such as wav2vec 2.0 (Baevski et al., 2020) have dramatically improved performance on Standard German tasks, and generative systems such as GPT (OpenAI et al., 2024) have achieved widespread popularity. This development is attributable to the availability of extensive corpora of both spoken and written Standard German (Kupietz and Schmidt, 2018; Deppermann et al., 2023).

Within this context, automatic dialect identification (ADI) emerges as a central task. ADI not only enables new insights into the characteristics, distribution, and evolution of dialects, but also provides the necessary foundation for deploying dialect-sensitive AI systems—such as automatic speech recognition (ASR) or advanced generative language models—by systematically accounting for dialectal variation (Yan and Vaseghi, 2002). However, for dialects and minority languages, such functional language models are still largely lacking. The primary reason is the absence of suitable training data: for most regional varieties, there are insufficient spoken and written resources available.

This paper focuses on a corpus of regional varieties of German (dialects, regional accents), where this problem also applies. While ASR systems for Standard German—such as the Whisper model (Radford et al., 2023)—and speech synthesis technologies like the Eleven Multilingual v2 model (ElevenLabs[1]) have already achieved high levels of accuracy, comparable systems for dialects and other regional varieties remain either largely unavailable or perform significantly worse (Gilles et al., 2023; Dolev et al., 2024; Fischer et al., 2025).

What is therefore required is a corpus that meets the technical requirements of NLP methods while adequately representing regional diversity. Such a corpus must satisfy several requirements: (i) regionally diversified coverage of different dialect areas; (ii) it must include contemporary recordings that reflect present-day linguistic realities; (iii) recordings must be collected using a uniform methodology to ensure comparability and high recording quality; (iv) the corpus should encompass different speaking styles and modes to capture real-world variation—both in terms of dialect intensity (regional accents vs. local dialects) and experimental settings (e.g., controlled tasks vs. free speech); (v) to reflect social differentiation, the corpus should include speakers from different generations; (vi) the availability of transcriptions (orthographic, phonetic, or conversation-analytic) is required for a range of

[1] https://elevenlabs.io

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 144–152
16 May 2026.

analytic approaches; (vii) legal requirements for usability of the data must be met and the recordings must be technically prepared and accessible in open standard file formats.

The REDE corpus is the first nationwide, methodologically uniform spoken language resource for German dialects that meets all seven of these requirements. The aim of this paper is to present the design of the corpus, its structure, and relevance for linguistic research and NLP.

2. The REDE corpus

2.1. Context and Objectives

The REDE corpus was developed within the still ongoing research project Regionalsprache.de (REDE) (Schmidt et al., 2020–). One of the project's core objectives is the first systematic collection, analysis, and documentation of contemporary regional languages of German. The concept of *regional languages* encompasses all speech forms within the variation space between local dialects and regionally accented Standard German (Kehrein, 2019b; Schmidt and Herrgen, 2011).

2.2. Data Collection

The speech recordings were collected between 2008 and 2015 following a uniform methodological protocol. Data collection was conducted by twelve trained fieldworkers at a total of 148 locations across Germany. Location selection aimed at balanced regional coverage while also prioritizing sites that functioned as emergency call centers, as the core speaker group consisted of police officers engaged in emergency call handling. (For an overview of all locations and their assignment to the dialect areas, see the maps in in Figure 1 and Figure 2. For a tabular overview of the number of locations, speakers, and recordings per dialect area, see Table 1.

Recordings were made in stereo using high-quality microphones and recorders (microphone: *SONY ECM-MS957*, recorder: *MARANTZ professional PMD661 handheld solid state recorder*) at a sampling rate of 44.1 kHz and a bit depth of 16 bit. All sessions were conducted in quiet, controlled environments (e.g., offices, meeting rooms) at the respective police stations or fieldwork locations.

2.3. Speakers

Speech data were collected from three speaker groups differing in age and occupation (see Table 2). Group 2 (G2) constitutes the core group: male middle-aged police officers with comparable educational backgrounds and professional profiles,

Figure 1: Location network of the REDE corpus (see REDE-Infothek (Lipfert, 2024b) for location codes and Figure 2 for color coding).

who also participated as speakers in REDE's predecessor project, the so-called DiGS project[2].

In addition, speakers from two comparison groups were recorded that differ from the police officers with respect to age, educational background, and profession (see Table 2). For all REDE speaker groups, the selection criterion of local rootedness—at least second, ideally third generation—applied. In total, recordings were made from 780 speakers (all male; see Section 5 for a discussion of this limitation and its implications for NLP applications).

2.4. Recording Methodology

The recordings were conducted on site in direct interaction with the speakers. Each speaker was recorded in up to five recording situations that vary with respect to speech mode (translation, reading, conversation) and intended variety (local dialect vs. Standard German). In addition to three controlled settings, two free speech settings were included (see Table 3). The recording situations can be characterized as follows.

In the WS-D setting, speakers were asked to translate standard-language Wenker sentences[3]

[2]The DiGS project was conducted in cooperation with the German Federal Criminal Police Office (BKA).

[3]Wenker sentences are a standardized set of 40 elicitation sentences that have been widely used in dialect sur-

Dialect macro regions	Dialect regions (with transition zones)	Locations	Speakers	Recordings
Low German				
	Northern Low German	13	70	343
	N. Low German-Mecklenb.-West Pom.	1	5	25
	Mecklenburgish-West Pomeranian	9	43	212
	Mecklenb.-West Pom.-Brandenburgish	1	4	20
	Central Pomeranian	1	6	30
	Westphalian	8	49	240
	Eastphalian	4	23	112
	Eastphalian-Brandenburgish	1	6	28
	Eastphalian-Northern Low German	1	5	25
	Brandenburgish	6	27	132
	Brandenburgish-South Markish	2	10	50
Subtotal Low German		47	248	1217
West German				
	Low Franconian	2	9	45
	Ripuarian-Low Franconian	3	16	78
	Ripuarian	5	25	117
	Ripuarian-Moselle Franconian	2	11	55
	Moselle Franconian	7	42	203
	Moselle Franconian-Rhine Franconian	2	11	52
	Moselle Franconian-Central Hessian	2	9	45
Subtotal West German		23	123	595
Central German				
	Berlin	1	4	20
	North Upper-Saxon-Low Markish	2	8	40
	Upper Saxon	4	19	78
	Thuringian	4	21	104
	Thuringian-Upper Saxon	5	23	112
	Silesian	1	6	24
	North Hessian	3	15	74
	East Hessian	1	12	59
	Central Hessian	5	30	140
	Central Hessian-North Hessian	1	4	20
	Central Hessian-Rhine Franconian	1	8	37
	Rhine Franconian	9	46	226
	Rhine Franc.-Mosl. Franconian	1	4	19
	Rhine Franc.-Mosl. Franc.-Central Hessian	1	5	25
	Rhine Franc.-Low Alemannic	1	5	24
Subtotal Centr. German		40	210	1002
Upper German				
	Low Alemannic	2	8	40
	Central Alemannic	3	13	65
	Swabian	9	48	234
	Swabian-East Franconian	1	6	28
	High Alemannic	1	5	25
	High Alemannic-Low Alemannic	1	4	20
	East Franconian	8	42	203
	East Franconian-North Bavarian	1	5	25
	North Bavarian	3	14	70
	North Bavarian-Central Bavarian	1	6	30
	Central Bavarian-North Bavarian	1	7	33
	Central Bavarian	5	29	144
	Central Bavarian-Swabian-South Bavarian	1	7	34
	Bavarian-Swabian	1	5	22
Subtotal Upper German		38	199	973
Total		**148**	**780**	**3787**

Table 1: Overview of locations, speakers, and recordings by dialect area; dialect macro regions following Lameli (2013), subdivided according to Wiesinger (1983).

Group	Age range	Occupation	N
Group 1 (G1)	60+	manual occupations (e.g., crafts, agriculture)	260
Group 2 (G2)	42–59	police officers	326
Group 3 (G3)	17–26	upper secondary students and university students	194
Total			780

Table 2: Speakers

Figure 2: Classification of German dialects (Fischer and Lameli (forthcoming); Low German dialects are coloured in blue, Central German dialects in yellow, Upper German dialects in orange, and Western German dialects in green. Non-Germanic varieties, e.g., Sorbian, are not depicted on the map).

Abbreviation	Setting description
WS-D	Translation of Wenker sentences into the local dialect
WS-S	Translation of Wenker sentences into Standard German
NoSo	Reading task
I	Guided interview
FR	Conversation with a friend

Table 3: Recording settings

read aloud by the fieldworker into the local dialect. In WS-S, speakers listened to dialect recordings of Wenker sentences made prior to the survey that represented their own dialect region and translated them into Standard German. In the NoSo reading task, speakers read aloud Aesop's fable *Der Nordwind und die Sonne* provided in written form.

The interview (I) was conducted using a semi-structured guideline and recorded. Topics included speakers' language-biographical background, individual language use, language attitudes, and self-assessment of linguistic competence. In the friend conversation (FR) setting, a free conversation between the speaker and self-selected acquaintances from the same dialect region was recorded. The conversational partners were also recorded as part of the corpus; their speech is excluded from annotations. In addition, sociodemographic data were collected, and a questionnaire on linguistic background was completed.

Not all settings could be realized for every speaker, for example, because speakers did not understand the dialectal stimulus recordings in the WS-S translation setting. However, it was ensured that for each location and each speaker generation, at least one speaker was recorded in all settings.

2.5. Corpus Size and Structure

In total, the REDE corpus comprises 3,787 recordings with an overall duration of 1,542 hours. This corpus size enables fine-grained linguistic analyses and makes the resource attractive for NLP applications. Table 4 provides an overview of the number and total duration of recordings by setting. On average, interviews constitute the longest recordings (ca. 56 minutes), whereas the reading task lasts approximately one minute. Free speech settings (I and FR) together account for 72% of the total recording time, providing ample naturalistic data for language modeling and ASR. An overview on the number of available recordings per dialect area is provided by Table 1.

Setting	N	Total duration (hh:mm:ss)	Avg. duration (hh:mm:ss)
NoSo	769	12:15:34	00:00:57
WS-S	766	214:27:25	00:16:48
WS-D	770	212:21:41	00:16:33
I	765	711:09:29	00:55:47
FR	717	420:07:23	00:35:09
Total	3,787	1,570:21:32	00:24:53

Table 4: Number and duration of recordings by setting

veys since their original deployment in Georg Wenker's *Sprachatlas des Deutschen Reichs* (1889–1923) (see Lameli, 2014; Fleischer, 2017).

2.6. Data Processing and Documentation

The recordings are available as WAV files, organized by location, speaker, and recording setting. The speech recordings were further processed in specific subsamples. Orthographic and phonetic transcriptions are available as TextGrids for recordings from 86 locations—each including speakers from all three generations and all five situational contexts. For a further subsample of 60 locations, measures of dialect intensity were additionally performed and published within the REDE SprachGIS. Additional orthographic transcriptions exist for 1,000-word excerpts from 224 recordings from the FR and I settings for 56 locations. Phonetic transcriptions follow IPA conventions and are available at the phone, syllable and word level.

Furthermore, for FR of 56 young speakers (G3) at 48 locations, conversation-analytic transcriptions following the GAT2 conventions (Selting et al., 2009) were produced (Fischer et al., 2023).

Rich metadata are available for all recordings and speakers, including information on location, dialect area, language-biographical background, self-assessed dialect competence, and dialect use.

All documentation related to data collection and processing is provided in the REDE Infothek. Templates such as recording protocols and the linguistic background questionnaire, as well as data sets, are published in the LinguRep repository. A selection of results is available on the platform *Regionalakzente in Deutschland* (Pheiff and Pistor, 2023).

Currently, the audio recordings are available upon request and following the conclusion of a cooperation agreement. Publication in the LinguRep repository under a free license is in preparation.

3. The REDE corpus as a Research Resource

3.1. Suitability for Variationist Linguistics

Based on the REDE recordings, a wide range of qualitative and quantitative studies have investigated speakers' situational variation behavior. A particular focus has been placed on phonetic and phonological variation in individual regions of the language area covered by REDE (e.g., Kehrein, 2012, 2019a; Rocholl, 2015; Keil, 2017; Vorberger, 2019; Bohnert-Kraus, 2020; Stiel, 2020; Limper, 2024; Lameli, 2025). These studies employ measures of dialect intensity (measured as phonetic difference from codified spoken standard variety (Herrgen et al., 2001; Lipfert, 2024a; Lameli, forthcoming)) as well as qualitative analyses and statistical methods. Dialectal prosody has been examined by Pistor (2022, 2025a,b), while phonetic studies on vowel space (Kleen et al., 2024) and on pre-boundary lengthening (Spina and Lameli, 2024; Lameli and Spina, 2025; Spina and Lameli, 2025) demonstrate that the acoustic quality of the recordings is sufficient for fine-grained phonetic analyses. Morphological and morphosyntactic variation has been investigated by Fischer (2022) and Fischer and Rabanus (2023).

Taken together, these studies validate the acoustic and linguistic quality of the REDE corpus across multiple levels of analysis and confirm its suitability for systematic variationist linguistics.

3.2. Suitability for NLP Applications

Automatic dialect identification. The suitability of the REDE corpus for NLP applications has already been demonstrated in several studies. The recordings constitute a central data resource in the project *AnDy: Automatische Analyse der Dynamik dialektalen Sprechens mit Methoden der Künstlichen Intelligenz*. The project addresses both spatial and register-based classification of German dialects (see, e.g., Fischbach et al., 2025a,b) and evaluates speech processing methods such as speaker diarization (Fischbach, 2024).

Automatic speech recognition. Further NLP applications have drawn on REDE subsets: Bystrich (2023), for instance, used Bavarian Wenker sentences from REDE to evaluate ASR models. His work demonstrates that large corpora with orthographically transcribed data enriched with dialectal syntactic, lexical, and morphological features are essential for effectively combining standard language models with dialect-specific models. A similar objective was pursued by Lindner (2025), who adapted a standard ASR model to the Bavarian dialect using transfer learning on Wenker sentences recordings from three REDE locations. The REDE corpus has also been used in forensic applications: Siewert (2023) trained an ASR model on REDE conversational recordings and achieved highly accurate speaker localization.

Voice cloning and speaker localisation. In addition, Fischer et al. (2025) demonstrate that the REDE corpus can be used for modern voice cloning technologies, including speech synthesis systems such as the Eleven Multilingual v2 model (ElevenLabs), which are capable of replicating individual speaker voices with high fidelity (see also Gilles et al., 2023; Dolev et al., 2024).

4. Discussion

The REDE corpus is, to date, the only spoken language resource for German dialects that combines

nationwide coverage (locations across all dialect macro regions, i.e., Low German vs. High German), situational breadth (five recording settings), generational stratification, and a uniform collection protocol, yielding 1,542 hours with rich sociolinguistic metadata. While methodologically uniform corpora with dialect recordings do exist, these were collected in the 1950s and 1960s, for example, the Zwirner corpus (Zwirner, 1962) and the GDR corpus (Ehlers, 2022). Other corpora are restricted to specific dialect areas and therefore do not permit analyses of cross-dialect continua or convergence zones, for example, SwissCoco for Switzerland (Cieliebak et al., 2025), Schnëssen for Luxembourg (Entringer et al., 2021), DiÖ for Austria (Lenz, 2018), and SiN for northern Germany (Elmentaler et al., 2015). By contrast, regionally diversified corpora such as Deutsch heute corpus (Kleiner, 2015) focus exclusively on regionally accented reading pronunciation, while others, such as the Pfeffer corpus (Pfeffer, 1975), target regiolectal speech and exclude base dialects.

Through its methodological setup, the REDE corpus provides a curated, machine-readable resource for developing and evaluating AI systems. Its annotated structure supports both linguistic research and the development of language-aware AI systems that are responsive to regional variation. This is particularly relevant for accessible communication, human–machine interaction, and regionally adaptive voice technologies.

Furthermore, the REDE corpus complements existing resources such as DIALECTBENCH (Faisal et al., 2024) by its only spoken modality and also by providing evaluation data for underrepresented varieties and contributing to more inclusive benchmarking practices. It enables computational modeling of dialect continua and convergence zones, thereby opening new avenues for sociolinguistic and AI-driven analyses.

Beyond research applications, the REDE corpus is also available for dialect revitalization and preservation as well as for dialect studies and teaching. Central to these applications is the integration of the recordings into the linguistic geographic information system REDE SprachGIS, which provides cartographic access. In addition, REDE SprachGIS offers analyses of the dialect–standard continua for 60 REDE locations, each represented by one speaker from a given speaker group. These visualizations make regional language variation accessible for educational purposes (Pheiff et al., 2019; Schmidt, 2017).

5. Limitations

The REDE corpus has several limitations worth noting. First, coverage is restricted to Germany, excluding Austrian and Swiss dialect areas, though the methodologically parallel DiÖ corpus (Lenz, 2018) offers a complementary resource for Bavarian and Alemannic varieties. At present, there are no plans to extend the REDE corpus to additional German-speaking regions. Second, all speakers are male, reflecting the occupational gender bias of the core speaker group; this limits the transferability of NLP models trained on REDE data to female speech. Third, the recordings date from 2008–2015 and may not fully reflect current spoken varieties given ongoing dialect leveling. Finally, data annotation remains incomplete in parts and is being continuously expanded; access to the recordings is currently granted upon request and after conclusion of a cooperation agreement, with open publication in the LinguRep repository in preparation.

6. Acknowledgements

This research is supported by the *Academy of Sciences and Literature Mainz* (grant REDE 0404), the *Federal Ministry of Research, Technology and Space* (BMFTR) (grant AnDy 16DKWN007), and the Research Center *Deutscher Sprachatlas* of Marburg University. We are grateful to Lisa Dücker, Georg Oberdorfer, and Salome Lipfert for providing us with statistical overviews and information regarding the REDE data. We would also like to thank two anonymous reviewers for their valuable comments.

7. Bibliographical References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Mirja Bohnert-Kraus. 2020. *Regionalsprachliche Spektren im Mittelalemannischen*. Olms, Hildesheim.

Tobias Bystrich. 2023. Data-driven and rule-based approaches to improving Bavarian speech recognition. Bachelor's thesis, Universität Bonn and Fraunhofer-Institut IAIS.

Mark Cieliebak, Jonathan Gerber, and Manuela Hürlimann. 2025. SwissCoco2025 - the Swiss corpora collection 2025. In *Proceedings of the 10th edition of the Swiss Text Analytics Conference*, pages 133–148, Winterthur, Switzerland. Association for Computational Linguistics.

Arnulf Deppermann, Christian Fandrych, Marc Kupietz, and Thomas Schmidt, editors. 2023. *Korpora in der germanistischen Sprachwissenschaft*. De Gruyter, Berlin, Boston.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Eyal Liron Dolev, Clemens Fidel Lutz, and Noëmi Aepli. 2024. Does Whisper understand Swiss German? An automatic, qualitative, and human evaluation.

Klaas-Hinrich Ehlers. 2022. Die Tonbandaufnahmen der deutschen Mundarten im Kontext der (niederdeutschen) Dialektologie der DDR. IDSopen: Online-only Publikationen des Leibniz-Instituts für Deutsche Sprache, No. 3.

Michael Elmentaler, Joachim Gessinger, Jens Lanwer, Peter Rosenberg, Ingrid Schröder, and Jan Wirrer. 2015. Sprachvariation in Norddeutschland (SiN). In Roland Kehrein, Alfred Lameli, and Stefan Rabanus, editors, *Regionale Variation des Deutschen: Projekte und Perspektiven*, pages 397–424. De Gruyter, Berlin and Munich and Boston.

Nathalie Entringer, Peter Gilles, Sara Martin, and Christoph Purschke. 2021. Schnëssen. surveying language dynamics in luxembourgish with a mobile research app. *Linguistics Vanguard*, 7(s1).

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: A NLP Benchmark for Dialects, Varieties, and Closely-Related Languages.

Lea Fischbach. 2024. A Comparative Analysis of Speaker Diarization Models: Creating a Dataset for German Dialectal Speech. In *Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024)*, pages 43–51, Bangkok. Association for Computational Linguistics.

Lea Fischbach, Akbar Karimi, Carolin Kleen, Alfred Lameli, and Lucie Flek. 2025a. Improving Low-Resource Dialect Classification Using Retrieval-based Voice Conversion. In *Proceedings of Interspeech 2025*, pages 2780–2784.

Lea Fischbach, Caroline Kleen, Lucie Flek, and Alfred Lameli. 2025b. Does Preprocessing Matter? An Analysis of Acoustic Feature Importance in Deep Learning for Dialect Classification. In *Proceedings of the 25th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 159–169, Tartu. University of Tartu Library.

Hanna Fischer. 2022. *Tempus und Regionalsprache. Eine gebrauchslinguistische Studie*. Olms, Hildesheim.

Hanna Fischer, Brigitte Ganswindt, and Georg Oberdorfer. 2023. Die regionalsprachlichen Tonkorpora des Forschungszentrums Deutscher Sprachatlas. In Marc Kupietz and Thomas Schmidt, editors, *Neue Entwicklungen in der Korpuslandschaft der Germanistik. Beiträge zur IDS-Methodenmesse 2022*, pages 189–202. Narr Francke Attempto, Tübingen.

Hanna Fischer and Alfred Lameli. forthcoming. Regional varieties of German. In Joshua Bousquette and Simon Pickl, editors, *The Oxford Handbook of the German Language*. Oxford University Press, Oxford.

Hanna Fischer, Alfred Lameli, Martha Schubert, and Ingo Siegert. 2025. Cloning dialects: Recreating and localizing dialectal voices. In *2025 IEEE International Professional Communication Conference (ProComm)*, pages 358–367.

Hanna Fischer and Stefan Rabanus. 2023. Zwischen dialektalem Hintergrund und standardsprachlicher Norm: verbalmorphologische Variation in standardintendierter Sprechweise. In Hanna Fischer and Stefan Rabanus, editors, *Morphologische und syntaktische Variation in den deutschen Regionalsprachen: Impulse für die Erforschung der sprachlichen Vertikale*. Olms, Hildesheim.

Jürg Fleischer. 2017. *Geschichte, Anlage und Durchführung der Fragebogen-Erhebungen von Georg Wenkers 40 Sätzen: Dokumentation, Entdeckungen und Neubewertungen*. Olms, Hildesheim and Zürich and New York.

Peter Gilles, Nina Hosseini Kivanani, and Léopold Edem Ayité Hillah. 2023. LUX-ASR: Building an ASR system for the Luxembourgish language. In *Proceedings - 2022 IEEE Spoken Language Technology Workshop (SLT)*.

Joachim Herrgen, Alfred Lameli, Stefan Rabanus, and Jürgen Erich Schmidt. 2001. Dialektalität als phonetische Distanz. Ein Verfahren zur Messung standarddivergenter Sprechformen. Marburg University.

Roland Kehrein. 2012. *Regionalsprachliche Spektren im Raum – Zur linguistischen Struktur der Vertikale*. Steiner, Stuttgart.

Roland Kehrein. 2019a. 5. Areale Variation im Deutschen „vertikal“. In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Volume 4 Deutsch*, pages 121–158. De Gruyter Mouton, Berlin, Boston.

Roland Kehrein. 2019b. Vertical language change in germany: Dialects, regiolects, and standard german. In Stanley D. Brunn and Roland Kehrein,

editors, *Handbook of the Changing World Language Map*, pages 1–13. Springer International Publishing, Cham.

Carsten Keil. 2017. *Der VokalJäger. Eine phonetisch-algorithmische Methode zur Vokaluntersuchung. Exemplarisch angewendet auf historische Tondokumente der Frankfurter Stadtmundart*. Olms, Hildesheim.

Caroline Kleen, Marina Frank, and Alfred Lameli. 2024. Vertical differences in German Vowel Space Areas. In *Proceedings of the 5th International Symposium on Applied Phonetics (ISAPh 2024)*, pages 43–48.

Stefan Kleiner. 2015. „Deutsch heute“ und der Atlas zur Aussprache des deutschen Gebrauchsstandards. In Roland Kehrein, Alfred Lameli, and Stefan Rabanus, editors, *Regionale Variation des Deutschen: Projekte und Perspektiven*, pages 489–518. De Gruyter, Berlin and Munich and Boston.

Marc Kupietz and Thomas Schmidt, editors. 2018. *Korpuslinguistik*. De Gruyter, Berlin, Boston.

Alfred Lameli. 2013. *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. De Gruyter, Berlin and Boston.

Alfred Lameli. 2014. *Erläuterungen und Erschließungsmittel zu Georg Wenkers Schriften*. Olms, Hildesheim and New York and Zürich.

Alfred Lameli. 2025. Gesprochenes Deutsch in den Regionen. Eine Standortbestimmung für die Bundesrepublik Deutschland. In Nadine Proske, Thilo Weber, Monika Dannerer, and Arnulf Deppermann, editors, *Gesprochenes Deutsch. Struktur, Variation, Interaktion*, pages 51–79. De Gruyter, Berlin and Boston.

Alfred Lameli. forthcoming. Evaluating phonetically weighted and unweighted distance measures in dialectometry.

Alfred Lameli and Nadja Spina. 2025. The Signs of Time: The Indexical Value of Pre-boundary Lengthening. In Toke Hoffmeister, Christina Kauschke, and Mathias Scharinger, editors, *Repräsentationen aus linguistischer und interdisziplinärer Perspektive*, pages 357–381. Olms, Baden-Baden.

Alexandra N. Lenz. 2018. The special research programme: German in austria: Variation – contact – perception. *Sociolinguistica*, 32(1):269–278.

Juliane Limper. 2024. *Regionalsprachliche Spektren im Bairischen*. Steiner, Stuttgart.

Moritz Lindner. 2025. Anpassung eines Standard-Hochdeutsch-ASR-Modells an den bayerischen Dialekt mittels Transfer Learning. Master's thesis, Hochschule für Ökonomie & Management München.

Salome Lipfert. 2024a. REDE-Infothek: Phonetische Abstandsmessung (PAM). In Jürgen Erich Schmidt, Joachim Herrgen, Roland Kehrein, Alfred Lameli, and Hanna Fischer, editors, *Regionalsprache.de (REDE). Forschungsplattform zu den modernen Regionalsprachen des Deutschen*. Forschungszentrum Deutscher Sprachatlas, Marburg.

Salome Lipfert. 2024b. REDE-Infothek: REDE-Orte. In Jürgen Erich Schmidt, Joachim Herrgen, Roland Kehrein, Alfred Lameli, and Hanna Fischer, editors, *Regionalsprache.de (REDE). Forschungsplattform zu den modernen Regionalsprachen des Deutschen*. Forschungszentrum Deutscher Sprachatlas, Marburg.

Netzwerk Regionale Sprache und Künstliche Intelligenz. 2026. Regionale Sprache und Künstliche Intelligenz im Zeitalter der digitalen Transformation. *Zeitschrift für Dialektologie und Linguistik*, (Online First):1–27.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal ..., and Barret Zoph. 2024. Gpt-4 technical report.

J. Alan Pfeffer. 1975. *Grunddeutsch: Erarbeitung und Wertung dreier deutscher Korpora. Ein Bericht aus dem Institut for Basic German, Pittsburgh*. Narr, Tübingen.

Jeffrey Pheiff and Tillmann Pistor. 2023. Des reine Hochdeutsch bring ich net zamm. Lautliche Merkmale des Regionalakzents im Ostfränkischen. *Wiener Linguistische Gazette*, 94:35–95.

Jeffrey Pheiff, Tillmann Pistor, and Anna Wolańska. 2019. Zur Verwendung der Forschungsplattform Regionalsprache.de (REDE) bei der Vermittlung arealer Sprachvariation in den Bereichen Deutsch als Fremdsprache und Deutsch als Zweitsprache. *Linguistik Online*, 94(1):101–137.

Tillmann Pistor. 2022. *Universelle Intonationsmuster. Ein empirischer Nachweis konstanter prosodischer Strukturen in Regionalsprachen des Deutschen und darüber hinaus*. Steiner, Stuttgart.

Tillmann Pistor. 2025a. Phonetische Merkmale regionalsprachlicher Prosodie in Deutschland. *Zeitschrift für Dialektologie und Linguistik*, 92(3):332–373.

Tillmann Pistor. 2025b. So. Weischt? Zur Interaktion von Prosodie, syntagmatischer Position und lexikalischer Semantik von Diskursmarkern im Niederalemannischen. In Susanne Oberholzer and Noemi Adam-Graf, editors, *Dialekt in Gesellschaft und Schule. Variation und Wandel in Gebrauch und Wahrnehmung des Alemannischen*, pages 277–289. Steiner, Stuttgart.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Josephine Marie Rocholl. 2015. *Ostmitteldeutsch – eine moderne Regionalsprache? Eine Untersuchung zu Konstanz und Wandel im thüringisch-obersächsischen Sprachraum*. Olms, Hildesheim.

Jürgen Erich Schmidt. 2017. Vom traditionellen Dialekt zu den modernen deutschen Regionalsprachen. In Deutsche Akademie für Sprache und Dichtung and Union der deutschen Akademien der Wissenschaften, editors, *Vielfalt und Einheit der deutschen Sprache. Zweiter Bericht zur Lage der deutschen Sprache*, pages 105–143. Stauffenburg, Tübingen.

Jürgen Erich Schmidt and Joachim Herrgen. 2011. *Sprachdynamik: eine Einführung in die moderne Regionalsprachenforschung*. Schmitt, Berlin.

Jürgen Erich Schmidt, Joachim Herrgen, Roland Kehrein, Alfred Lameli, and Hanna Fischer. 2020–. Regionalsprache.de. Forschungsplattform zu den modernen Regionalsprachen des Deutschen.

Margret Selting, Peter Auer, Dagmar Barth-Weingarten, ..., and Susanne Uhmann. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion*, 10:353–402.

Marlon Siewert. 2023. Automatic Speaker Recognition and Dialectal Variation. Master's thesis, Marburg University.

Nadja Spina and Alfred Lameli. 2024. Regional variation in pre-boundary lengthening from a horizontal and vertical perspective: Evidence from German dialect- and standard-targeted speech. In *Proceedings of the Conference on Speech Prosody 2024*, pages 990–994.

Nadja Spina and Alfred Lameli. 2025. Slow or Long? Generational Variation of Pre-boundary Lengthening in Different Dialect Areas of German. In *Proceedings of the Third International Conference on Tone and Intonation*, pages 60–64.

Rico Stiel. 2020. *Phonemwandel im gesprochenen Standard. Dynamik des //-Phonems im Deutschen*. Ph.D. thesis, Marburg University.

Lars Vorberger. 2019. *Regionalsprache in Hessen. Eine Untersuchung zu Sprachvariation und Sprachwandel im mittleren und südlichen Hessen*. Steiner, Stuttgart.

Peter Wiesinger. 1983. Die Einteilung der deutschen Dialekte. In Werner Besch, Ulrich Knoop, Wolfgang Putschke, and Herbert Ernst Wiegand, editors, *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung*, volume 1.2 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 807–900. De Gruyter, Berlin and New York.

Qin Yan and Saeed Vaseghi. 2002. A comparative analysis of uk and us english accents in recognition and synthesis. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–413–I–416.

Eberhard Zwirner. 1962. *Deutsches Spracharchiv 1932–1962. Geschichte, Aufgaben und Gliederung, Bibliographie*. Aschendorff, Münster (Westfalen).