South Tyrolean Dialect-to-Standard Speech Translation: A Resource

Greta H. Franzini, Luca Ducceschi
Institute for Applied Linguistics, Eurac Research
Viale Druso 1, 39100 Bolzano, Italy
{greta.franzini, luca.ducceschi}@eurac.edu

Abstract

This paper presents a developing oral resource for South Tyrolean, a German dialect spoken in Northern Italy. The dialect is ubiquitous in spoken communication but lacks a standardised orthography. In this context, strict transcription into dialect is of limited to no utility to the local community. Instead, there is a distinct and strong demand for technology capable of directly translating spoken dialect into Standard German. To address this specific need, we introduce a dynamic, incrementally growing dataset designed to fine-tune ASR models for the task. Our corpus aggregates diverse sources, including media and research interviews, totalling over 13 hours of aligned audio. We describe a collaborative workflow where community partners contribute audio material in exchange for automated transcriptions, creating a virtuous cycle of data improvement. Additionally, we detail our iterative model fine-tuning strategy, the preprocessing workflow and the resulting improvements in model performance.

Keywords: dialect, German, ASR, language resource, speech corpus

1. Introduction

South Tyrolean, a Southern Bavarian variety spoken by over 90% of the German-speaking population in South Tyrol (Italy) (ASTAT, 2024, p. 1), exists in a state of *medial diglossia* (Auer, 2005, p. 12). While the dialect dominates oral communication, Standard German (Ammon et al., 2016) remains the exclusive medium for formal written domains (Leonardi, 2020). This functional split creates a distinct local demand for NLP: rather than transcriptions into written dialect —a preference mirrored in other Bavarian contexts (Blaschke et al., 2024)— users require technology that translates spoken dialect directly into Standard German. This functional separation between spoken dialect and written standard language means that South Tyrolean remains underrepresented in digitised and annotated speech resources, leaving it a low-resource language despite its high vitality.

Modelling this speech is further challenged by internal fragmentation and language contact. South Tyrol's mountainous geography has fostered distinct varieties (e.g., Puster Valley vs. Vinschgau) with divergent phonology and lexicons, meaning that models calibrated for speakers of the capital, Bozen, often fail in rural side valleys. Furthermore, frequent code-switching with Italian introduces loanwords such as *targa* ('license plate') or *magari* ('perhaps'), which often cause monolingual German models to hallucinate phonetically similar German terms. Together, all of these hurdles necessitate specialised resources for robust ASR in the region.

Contribution. While numerous oral corpora exist for German dialects, with Blaschke et al. (2023) identifying 39 audio resources documenting both Low and High German varieties, no dataset to date has been specifically designed to support dialect-to-standard ASR for South Tyrolean. We address this gap by presenting a growing oral resource developed through an increasingly collaborative, community-driven data collection approach, together with a family of fine-tuned ASR models capable of translating South Tyrolean dialectal speech directly into Standard German (Ducceschi and Franzini, 2025).

To facilitate the interpretation of the qualitative error analysis presented later in the paper, it is worth noting that translation from South Tyrolean dialect into Standard German is not a purely literal mapping task. Differences between spoken dialectal usage and written standard language frequently require structural rephrasing rather than word-for-word rendering. As a result, some model outputs may appear acoustically plausible while nonetheless diverging from the expected Standard German form in terms of morphology, syntax or lexical choice. This distinction is particularly relevant for understanding the error patterns discussed in Section 4.2.

2. Related Work

The development of ASR systems for low-resource languages has gained traction in recent years, yet dialectal varieties remain under-served. The core challenge lies in the data bottleneck: state-of-the-art models like Whisper require thousands of hours of transcribed audio to generalise effectively. While Standard German is supported by vast corpora, its dialectal offshoots often have negligible digital

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 188–194
16 May 2026.

footprints.

Previous work in Germanic dialect-to-standard ASR has concentrated significantly on Swiss German. Initiatives in Switzerland have produced aligned corpora, such as *STT4SG-350* (Plüss et al., 2023), *PASSAGE* (Gerlach et al., 2022), *GermEval 2020 Task 4* (Plüss et al., 2020), *SRF Meteo* (Stadtschnitzer and Schmidt, 2018) or *ArchiMob* (Scherrer et al., 2019). While *ArchiMob* primarily employs a word-aligned normalisation layer that provides Standard German equivalents at the lexical level, these resources collectively facilitate the development of robust models for dialect-to-standard translation. Similar developmental efforts have also been documented for Luxembourgish (Gilles et al., 2023), a related West Germanic language, and Bavarian (Blaschke et al., 2025).

3. Method

3.1. Data Collection and Composition

Our dataset is a dynamic, work-in-progress resource that is incrementally extended as new data becomes available. As Table 1 shows, we aggregate data from a diverse set of sources to ensure coverage of different domains, geographical areas and speech styles.

Crowdsourced Corpus (ALP). We incorporate the South Tyrolean German subset of the AlpiLinK (Alpine Languages in Contact) corpus (Kruijt and Rabanus, 2025). This openly available resource provides crowdsourced oral data across the Alpine regions of Italy (Rabanus et al., 2025). We specifically use the translation task, which features a controlled set of 30 Standard German written sentences orally rendered into dialect by native speakers. This structured repetition provides the model with multiple phonetic realisations of identical lexical content, aiding the mapping between dialectal variation and the Standard German reference.

Subtitled YouTube Videos (IDM). We harvest publicly available video content that already possesses creator-provided subtitles. Specifically, we collected spontaneous speech data from 44 subtitled promotional videos published by *IDM Südtirol*[1]. These videos feature unscripted interviews with speakers of varying ages (excluding children), covering diverse topics from cooking to farming. The videos were downloaded as `MP4` files using the Python *pytube* library[2]. To extract the subtitles, we processed each video frame by enhancing contrast and applying *PaddleOCR* (Kanakaraddi et al., 2024) to the bottom third of the image. OCR errors were primarily due to special character misrecognition (e.g., ß, ä, ü, ö), low video resolution and visual interference. All output was manually corrected to produce clean reference texts. Since the original subtitles are often paraphrased rather than being verbatim transcriptions, the alignment between audio and text is approximate rather than literal.

Language Learning Textbooks (ARB). To ground the model in more structured speech, we use audio materials from dialect teaching resources, which are typically clearer and more grammatically standard than the YouTube data. We used scripted speech data from the publicly available *DaZUgeHÖREN* textbook, designed for learners of South Tyrolean dialect (Gurschler and Tscholl, 2015). The textbook includes audio exercises on CD, featuring dialectal prompts accompanied by printed Standard German translations. Again, the alignment between audio and text is not always exact, as the printed translations occasionally deviate from the spoken content.

Audiovisual Archives (FUM, MEN). We collaborate with cultural institutions and historical research projects to access archival data. As part of a formal agreement with the Office for Film and Media of the Province of Bolzano (FUM), we are transcribing their audiovisual archives, primarily 30+ year old news broadcasts containing both South Tyrolean Standard German and dialect segments, using our best-performing ASR model. Additionally, we have an informal agreement with an Austria-based publishing agency (MEN) producing documentary films on South Tyrolean history. These documentaries feature elderly native speakers with broad, linguistically challenging dialect forms.

Research and Purpose-recorded Interviews (EUR). Beyond external partnerships, we assist other colleagues within our organisation by transcribing their research interviews, thereby incorporating their dialect data into our corpus. Furthermore, we address data gaps through targeted elicitation tasks. We record human translators spontaneously producing spoken dialect translations of Standard German sentences, following a methodology similar to Plüss et al. (2022, p. 2).[3] To enhance the model's performance on local named entities, a frequent stumbling block, we use our company instance of Microsoft Copilot to generate Standard German prompts rich in local terms for these recording sessions.

All audio data is segmented into chunks shorter than 30 seconds, a requirement for effective training with architectures such as Whisper. The files are converted to `WAV` format and resampled to 16 kHz using *ffmpeg* (FFmpeg Developers, 2016), then manually aligned with their respective tran-

[1] https://www.youtube.com/@suedtirol.official

[2] https://pytube.io/en/latest/index.html

[3] Recordings are made in segments of under 30 seconds using a dynamic cardioid Behringer ULTRAVOICE XM8500 vocal microphone.

scriptions in ELAN[4]. A custom Python script subsequently parses the resulting ELAN alignments (EAF) and compiles the final dataset into a single aggregated CSV file for model fine-tuning.

3.2. Collaborative Workflow

Our methodology increasingly centers on a mutually beneficial partnership with internal and external research and memory institutions. Partners provide raw audio archives in exchange for automated Standard German transcripts generated by our latest model. With typical accuracies of 70–80%, partners only correct errors in ELAN rather than transcribing from scratch, significantly reducing manual effort (Russell et al., 2024). Once manually verified, this high-quality data is integrated into the training set for the next fine-tuning cycle. This creates a virtuous cycle where continuous model improvement further decreases partner workload over time.

3.3. Preprocessing and Quality Control

Given that noise or inconsistencies introduced prior to translation may negatively affect downstream performance, we explicitly assessed potential sources of error in the preprocessing pipeline. As mentioned in Section 3.1, all audio was used as retrieved from the original providers and underwent only minimal, standardised processing: conversion to WAV format and resampling to 16 kHz. No signal enhancement, denoising, filtering or speaker normalisation was applied, as our goal was to evaluate translation performance under realistic data conditions rather than optimise acoustic quality.

To avoid segmentation-related artefacts, audio files were manually segmented in ELAN into units shorter than 30 seconds, with segment boundaries placed in silence regions to prevent truncation of words or utterances. Where written translations were missing or derived from noisy sources (e.g. OCR-extracted subtitles), these were —as mentioned in sections 3.1 and 3.2— produced or corrected and subsequently validated. Aside from differences in translation availability and degree of audio–text literalness across sources, the preprocessing pipeline was uniform across all data types.

To characterise the acoustic conditions of the dataset, we assessed its technical quality using *librosa* (McFee et al., 2015). Overall, the corpus exhibits high acoustic quality, with a mean Signal-to-Noise Ratio (SNR) of 40.35 dB and no instances of significant digital clipping ($>$1%). A small subset of the data (26 files from ALP) shows poor SNR ($<$10 dB), reflecting the heterogeneous and sometimes noisy recording conditions typical of real-world material. In addition to this quantitative assessment, we performed qualitative preprocessing checks focussing on segmentation consistency and gross audio integrity. While we did not conduct a systematic or quantitative analysis of audio–text correspondence —since this lies outside the scope of the present work— these checks did not reveal pervasive segmentation errors, severe signal degradation or obvious audio–text inconsistencies that would plausibly account for the translation errors observed in evaluation. Residual variability in recording conditions is therefore treated as an inherent property of the dataset rather than as a preprocessing artefact.

3.4. Dataset Evolution and Tuning Strategy

We employ a rigorous iterative process for dataset construction and model tuning. With every new data increment, the model is fine-tuned to enhance its performance and generalisation capabilities. To ensure the integrity of our resources, we enforce strict consistency checks: each subsequent version of the dataset is verified to be a strict superset of its predecessor.

We maintain an 80/20 train-test split for each version, allowing us to track the model's performance on the specific distribution of each data increment. At the same time, to ensure longitudinal comparability and track performance improvements that are strictly attributable to training data scale rather than changes in test set difficulty, we evaluate all model versions against a common benchmark: the latest test set, v1.3. Currently, the resource comprises over 10 hours of training data and nearly 3 hours of test data.

4. Results

The growth of the dataset and the resulting ASR performance are shown in Table 2. We fine-tuned a separate OpenAI Whisper large-v3 model for each dataset version. We then evaluated all fine-tuned models on the v1.3 test set (1,340 samples, 24,442 tokens, 2h 50m) using both WER and BLEU. While BLEU is typically most robust when calculated against multiple reference translations, we include it here to complement WER in capturing translation quality. This is particularly relevant as our Standard German references occasionally exhibit significant structural or lexical deviations from the original dialectal utterances, moving beyond simple word-for-word mappings. To ensure the metrics reflect semantic and lexical accuracy rather than orthographic or formatting differences, we normalise both the hypothesis and reference texts by

[4]Version 6.4. https://archive.mpi.nl/tla/elan

Table 1: Current overview of corpus (v1.3). The entire audio corpus was segmented into 30-second units, each manually paired with the corresponding written translation.

Source	Type	Speech	Hours	Speakers	Configuration	Age
ARB	transcribed audio	read	47m	3	single-speaker	20-49
ALP	transcribed audio	read, spontaneous	4h 47m	180	single-speaker	10-89
EUR	audio	read, spontaneous	3h 42m	3	multi-speaker	20-49
FUM	video	read, spontaneous	18m	7	multi-speaker	0-79
IDM	subtitled video	spontaneous	1h 03m	86	multi-speaker	20-79
MEN	video	spontaneous	1h 35m	2	multi-speaker	30-99
TOTAL			13h 16m			

Table 2: Training data evolution and ASR performance. All models, including the non-fine-tuned OpenAI Whisper large-v3, are evaluated on the v1.3 test set (1,340 samples, 24,442 tokens, 2h 50m), with v1.2 yielding the best overall performance.

	Training Data				Performance (v1.3)	
Model	**Samples**	**Tokens**	**Avg.L(s)**	**Duration**	**WER ↓**	**BLEU ↑**
OpenAI Whisper large-v3	*n/a (no fine-tuning)*				0.46	44.58
v1.0	4,441	51,474	9.79	6h 39m	0.37	0.52
v1.1	5,048	75,203	11.46	9h 07m	0.27	65.65
v1.2	5,325	87,298	12.26	10h 16m	**0.24**	**69.13**
v1.3	5,363	88,810	12.28	10h 26m	0.24	68.73

lowercasing and removing punctuation prior to evaluation, thus mitigating WER inflation.

4.1. Quantitative Performance

The results shown in Table 2 showcase a clear trajectory of improvement. Notably, already with v1.0 of the dataset —despite its small size— our fine-tuned model obtains decent results. This aligns with the findings of Hollenstein and Aepli (Hollenstein and Aepli, 2014) as well as Samardžić et al. (Samardžić et al., 2015) for Swiss German, who demonstrated that modest amounts of variety-specific data often outperform significantly larger, out-of-domain Standard German corpora in training language processing tools.

The transition from v1.0 to v1.1, which increased training data from roughly 6.5 to 9 hours, yielded the single largest performance gain, reducing WER from 0.37 to 0.27 (a 27% relative improvement). This jump coincides with the introduction of archival interviews and additional purpose-recorded data, which substantially diversified the speaker pool and exposed the model to more varied spontaneous speech. Performance continues to improve with v1.2 (WER 0.24), as broadcast data and further subtitled content bring additional domain coverage. Between v1.2 and v1.3, global WER remains stable at 0.24 while BLEU decreases marginally from 69.13 to 68.73.[5] This plateau is expected: v1.3 adds only 38 new training samples over v1.2, contributing less than 10 additional minutes of audio. Rather than a performance regression, the near-identical scores confirm that the model has effectively saturated on the available data, consistent with the logarithmic relationship between training data size and ASR accuracy observed in large-scale studies (Zhang et al., 2023; Radford et al., 2023). This signals that further gains will require substantive new data increments rather than incremental additions.

4.2. Qualitative Analysis

A qualitative error analysis reveals that the model often confuses inflectional variants. In several instances, it outputs common indicative or finite forms, such as *war* ('was'), *gab* ('there was/were') and *sind* ('are'), instead of the grammatically required subjunctive or infinitive forms, that is *wäre* ('would be'), *gäbe* ('there would be') and *sein* ('to be'). While such distinctions are linguistically subtle, their repeated occurrence suggests systematic difficulties in distinguishing between closely related verbal paradigms. More broadly, the model sometimes fails to disambiguate lexical items that exist in both Standard German and South Tyrolean dialect but differ in meaning. In such cases, it tends to reproduce the acoustically plausible form even when contextually inappropriate. For example, the dialectal form *wert*, intended as a realisation of *wird* ('will'), was transcribed as *wert* ('worth'), although this interpretation rendered the utterance semanti-

[5] BLEU scores are computed using SacreBLEU with signature nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.5.1

cally incoherent in context.

Furthermore, characteristic South Tyrolean forms such as *sem* ('there/at that time'), *homo* ('we have'), *hon* ('I have'), *kimmp* ('comes') and *man* ('mine') are not always rendered accurately, but are generally processed satisfactorily.

The multilingual environment of South Tyrol naturally influences model behaviour. Given the dialect-heavy and translation-oriented training setup, foreign lexical items may be retained, translated or inconsistently transliterated. In one case, the model correctly translated the Italian word *aziendale* ('corporate') as *Unternehmerisch* while in another it incorrectly transliterated the same word as *achtsindale*.

Automated punctuation continues to pose difficulties. The model often struggles to correctly identify sentence boundaries and to insert commas where appropriate. Speaker attribution is likewise challenging. However, this limitation appears to stem primarily from the diarisation component employed, specifically pyannote.audio*pyannote.audio* (Bredin et al., 2019), rather than from the speech recognition model itself.

Finally, the model demonstrates limitations in the recognition of named entities. Place names and personal names are frequently misidentified or inconsistently rendered (e.g., *Potsdn* instead of *Bozen*; *Aldomoro* instead of *Aldo Moro*).

5. Discussion

The collaborative framework successfully addresses the data bottleneck by transforming the reuse and enhancement of audiovisual and research recordings into a driver for technological democratisation. A key finding is the model's ability to maintain semantic coherence across spontaneous, non-linear dialectal utterances, effectively mapping fragmented speech into structured Standard German prose. This functional utility is particularly evident for our partners, for whom the system provides a reliable baseline that significantly accelerates the transcription process, despite the need for human correction.

The results indicate that the model remains robust despite the presence of a subset of messy real-world recordings, as evidenced by the high overall performance despite the 26 files with poor SNR, although certain linguistic nuances remain a challenge. The persistent issues with named entities, as well as specific lexical items and verb forms suggest that future work should focus on data augmentation in these areas.

During the manual revision process, we document these acoustic and linguistic challenges to inform our filtering strategies and provide feedback to partners. This project thus serves as both a technical resource and a capacity-building initiative for the local speaker community.

6. Conclusion and Future Work

This project responds directly to community demands for accessible communication between South Tyrolean dialect and Standard German. By implementing a community-oriented framework, we have established the first dedicated dataset for the region. Our virtuous cycle methodology demonstrates that partnering with institutions can effectively turn various materials into high-quality training data, resulting in a system that is already perceived by users as a practically valuable tool for transcription and documentation.

The resource, however, remains a work in progress. Future work will prioritise filling geographic gaps to ensure the model generalises across South Tyrol's dialectal landscape. We also plan to experiment with larger model architectures, such as NVIDIA Canary, and evaluate whether the integration of Austrian Standard German (e.g., from the *Graz corpus of read and spontaneous speech* (Schuppler et al., 2014)) can further improve robustness.

7. Data Availability

The versioned resource repository is publicly available.[6] While the repository provides comprehensive metadata for all data included in the resource, licensing constraints prevent us from redistributing a substantial portion of the current corpus.

8. Limitations

A primary limitation of this study is that the corpus does not yet provide exhaustive coverage of all South Tyrolean dialectal variants, with certain valleys remaining underrepresented in the current training data. Additionally, while SNR provides a measure of background noise levels, it does not fully capture the acoustic complexity of our corpus. Future iterations will use non-intrusive neural metrics like NISQA (Mittag et al., 2021) to better model the relationship between acoustic quality and translation accuracy.

9. Ethical Considerations

Participants of the EUR corpus consent to the use of their voice and their involvement is entirely voluntary. No metadata or personal identifiable infor-

[6] https://gitlab.inf.unibz.it/commul/speech-to-text/augusta_data/

mation is used in the training or evaluation of the model.

10. Acknowledgements

We extend our sincere thanks to all project partners for their contributions to this work. We are also grateful to the anonymous reviewers for their valuable comments and suggestions, and to Simone Baratella for his technical support in the acoustic assessment of the training data.

11. Bibliographical References

Ulrich Ammon, Hans Bickel, and Alexandra N. Lenz, editors. 2016. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*, 2nd edition. De Gruyter, Berlin.

ASTAT. 2024. Sprachkenntnisse und Sprachgebrauch im Alltag in Südtirol - 2024.

Peter Auer. 2005. *Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations*, pages 7–42. De Gruyter Mouton, Berlin, New York.

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.

Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023. A Survey of Corpora for Germanic Low-Resource Languages and Dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.

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*, page 913–917. ISCA.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2019. pyannote.audio: neural building blocks for speaker diarization.

Luca Ducceschi and Greta H. Franzini. 2025. Speech transcription from South Tyrolean Dialect to Standard German with Whisper. In *Interspeech 2025*, page 5, Rotterdam, The Netherlands.

Ludwig M. Eichinger. 2002. South Tyrol: German and Italian in a Changing World. *Journal of Multilingual and Multicultural Development*, 23(1-2):137–149.

FFmpeg Developers. 2016. ffmpeg tool.

Johanna Gerlach, Jonathan Mutal, and Pierrette Bouillon. 2022. Producing Standard German subtitles for Swiss German TV content. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 37–43, Dublin, Ireland. Association for Computational Linguistics.

Peter Gilles, Léopold Edem Ayité Hillah, and Nina Hosseini Kivanani. 2023. ASRLUX: Automatic Speech Recognition for the low-resource language Luxembourgish. In *Proceedings of the 20th International Congress of Phonetic Sciences*. Guarant International.

Michael Gurschler and Evi Rita Tscholl. 2015. *DaZUgeHÖREN: Südtiroler Dialekt von Jugendlichen für Jugendliche Arbeitsmaterialien zum Südtiroler Dialekt*. Autonome Provinz Bozen - Deutsches Bildungsressort.

Nora Hollenstein and Noëmi Aepli. 2014. Compilation of a Swiss German Dialect Corpus and its Application to PoS Tagging. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 85–94, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Suvarna Kanakaraddi, Ashok Chikaraddi, Shantala Giraddi, Karuna Gull, and Mallanagouda Patil. 2024. Optimized Scene Text Detector and Paddle Optical Character Recognizer Techniques to Extract Text from Images. In *Proceedings of 4th International Conference on Artificial Intelligence and Smart Energy*, pages 218–230, Cham. Springer Nature Switzerland.

Anne Kruijt and Stefan Rabanus. 2025. From VinKo to AlpiLinK: web-based long-term storage and accessibility of information. *Korpus im Text*, 17.

Mara Maya Victoria Leonardi. 2020. "I hardly ever practice the real Standard German." Self-reported language use and language proficiency in South Tyrol (Italy). *Linguistik Online*, 102(2):83–98.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. *SciPy 2015*.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech 2021*, page 2127–2131. ISCA.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A Speech Corpus for All Swiss German Dialect Regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German Speech to Standard German Text Corpus.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2020. GermEval 2020 Task 4: Low-Resource Speech-to-Text. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland. CEUR.

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.

Sam O'Connor Russell, Iona Gessinger, Anna Krason, Gabriella Vigliocco, and Naomi Harte. 2024. What automatic speech recognition can and cannot do for conversational speech transcription. *Research Methods in Applied Linguistics*, 3(3):100163.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2015. Normalising orthographic and dialectal variants for the automatic processing of Swiss German. In *Proceedings of the 4th biennial workshop on less-resourced languages*, pages 294–298. ELRA.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. ArchiMob: Ein multidialektales Korpus schweizerdeutscher Spontansprache. *Linguistik Online*, 98(5):425–454.

Barbara Schuppler, Martin Hagmüller, Juan Andres Morales-Cordovilla, and Hannes Pessentheiner. 2014. GRASS: the Graz corpus of Read And Spontaneous Speech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1465–1470.

Michael Stadtschnitzer and Christoph Schmidt. 2018. Data-Driven Pronunciation Modeling of Swiss German Dialectal Speech for Automatic Speech Recognition. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages.

12. Language Resource References

Rabanus, Stefan and Kruijt, Anne and Alber, Birgit and Bidese, Ermenegildo and Gaeta, Livio and Raimondi, Gianmario. 2025. *AlpiLinK Corpus*. University of Verona and Free University of Bolzano, 1.2.0. PID 10.5281/zenodo.8360169.