SwissGPC v1.0 - The Swiss German Podcasts Corpus

Samuel Stucki, Mark Cieliebak, Jan Deriu
Centre for Artificial Intelligence
Zurich University of Applied Sciences (ZHAW), Winterthur
stku@zhaw.ch, ciel@zhaw.ch, deri@zhaw.ch

Abstract

We present SwissGPC v1.0, the first mid-to-large-scale corpus of spontaneous Swiss German speech, developed to support research in ASR, TTS, dialect identification, and related fields. The dataset consists of links to talk shows and podcasts hosted on Schweizer Radio und Fernsehen and YouTube, which contain approximately 5400 hours of raw audio. After segmentation and weak annotation, nearly 5000 hours of speech were retained, covering the seven major Swiss German dialect regions alongside Standard German.

We describe the corpus construction methodology, including an automated annotation pipeline, and provide statistics on dialect distribution, token counts, and segmentation characteristics. Unlike existing Swiss German speech corpora, which primarily feature controlled speech, this corpus captures natural, spontaneous conversations, making it a valuable resource for real-world speech applications.

Keywords: low-resource ASR dataset, Swiss German dialects, conversational speech corpus

1 Introduction

Swiss German is a family of dialects spoken in Switzerland and belongs to the Alemannic group of German dialects. It differs from Standard German in phonetics, grammar, vocabulary, and syntax. The dialects vary significantly across regions and are collectively spoken by approximately five million people. Unlike many other dialect groups, Swiss German is widely used in both professional and private settings and additionally serves as an expression and representation of a distinct Swiss nationality in the German-speaking part of the country. While it is primarily a spoken language, the rise of informal digital communication has led to an increase in written Swiss German. However, the absence of standardized orthography and its classification as a low-resource language make data collection for Automatic Speech Recognition (ASR) and other speech processing tasks particularly challenging.

There has been growing interest in the past years in researching ASR tasks on Swiss German dialects, which lead to the creation of several corpora such as the Swiss Parliament Corpus (Plüss et al., 2020), SwissDial (Dogan-Schönberger et al., 2021), SDS-200 (Plüss et al., 2022), and STT4SG-350 (Plüss et al., 2023). These corpora contain between 28 and 343 hours of audio and have since enabled various research endeavours (Sicard et al., 2023; Paonessa et al., 2023; Bollinger et al., 2023; Dolev et al., 2024).

However, these corpora are insufficient for data-intensive tasks such as Text-to-Speech (TTS). This paper thus presents the first version of the "Swiss German Podcasts Corpus (SwissGPC v1.0)", the first mid-to-large-scale[1] corpus for Swiss German: It contains links[2] to talk shows and podcasts collected from Schweizer Radio und Fernsehen (SRF) and YouTube (YT). These collected data contain approximately 5400 hours of raw audio, including speech from all dialect regions and Standard German. We utilized the 7 dialect regions outlined in (Plüss et al., 2023) to simplify the dialect classification. Only the source links of the utilized shows are released, as we do not possess the legal rights to distribute the audio or the annotated data of both SRF and YT.

2 Corpus Requirements

Our primary motivation for creating SwissGPC was to train a Zero-Shot Voice Adaptation Text-to-Speech (TTS) system for Swiss German dialects, for which large amounts of high-quality data are

[1] The dataset can be considered large-scale in the context of Swiss German corpora. However, compared to other languages such as English, German, or Mandarin it is still a small- to medium-sized corpus

[2] Due to copyright reasons, we can not provide the audio files, but only the links to the websites.

required. The dataset was thus created with the following goals in mind:

1. The corpus should be sufficiently large with a goal of 4000-5000 hours of primarily Swiss German speech.

2. The corpus must be sufficiently diverse in speakers to provide useful training data for TTS[3].

3. The speech must be recorded with a high-quality recording setup.

4. The corpus should cover a diverse set of topics.

Based on these goals, we decided to collect a large number of dialogues from podcasts that are primarily in Swiss German and to preprocess them to make them applicable for TTS and other speech processing tasks.

3 Data Annotation Pipeline

As outlined in the introduction, we do not have the rights to distribute the audio. We will only publish the links to the podcast sources that comprise the corpus. For SRF podcasts there exists an official API[4], while for YouTube, a third-party tool can be used such as pytube to download the files (specifically the pytubefix fork (JuanBindez, 2025), as the original library is not maintained anymore). Table 1 and 2 list the podcasts and their online source for SRF and YouTube, respectively. All sources combined link, at the time of publication, to 5404 hours of audio.

The data was weakly annotated using an automated pipeline, visualized in Figure 1. First, the raw audio was diarized and segmented on a speaker basis using pyannote (Bredin, 2023). The diarization step only tags actual speech, leading to silent and music segments being implicitly removed. The samples, containing only a single speaker based on the diarization, were cut to be between 2 and 15 seconds long. The time range was chosen to allow diverse sampling of shorter and longer segments, and additionally due to models downstream that required different lengths of audio for transcription or training. This resulted in a reduction of 7.84% from 5404 hours of raw audio to 4979 hours of actual speech with 1.76M unique samples. The segmented audio was then transcribed to Standard German since it has a standardized orthography. The transcription was performed using whisper-v3 (Radford et al., 2022) for its high performance in translating Swiss German speech to Standard German text (Paonessa et al., 2024). Using the approach of (Bolliger and Waldburger, 2024), we applied a wav2vec2 phoneme transcriber (Baevski et al., 2020; Xu et al., 2022). We classified the generated phoneme sequences with a Naïve Bayes n-gram classifier trained on the phonemicized STT4SG-350 corpus (Plüss et al., 2023) and an additional Standard German CommonVoice (Ardila et al., 2020) subset for Dialect Identification (DID). In total 8 different regions were thus used in classification: Basel, Bern, Central CH, Eastern CH, Grisons, Valais, Zurich, and Standard German. Further enrichment processes included the generation of Swiss German text using (Bollinger et al., 2023) and the creation of Mel-Spectrogram of the audio samples using (McFee et al., 2024).

SRF Podcast Name	Length (h)
#SRFglobal	36.97
100 Sekunden Wissen	186.75
Debriefing 404	245.14
Digital Podcast	428.05
Dini Mundart	39.39
Gast am Mittag	33.14
Geek-Sofa	317.28
SRF-Wissen	45.05
Kultur-Talk	55.84
Literaturclub - Zwei mit Buch	31.79
Medientalk	66.46
Pipifax	9.08
Podcast am Pistenrand	18.29
Samstagsrundschau	404.14
Sternstunde Philosophie	159.39
Sternstunde Religion	60.82
Sykora Gisler	152.22
Tagesgespräch	1661.33
Ufwärmrundi	60.98
Vetters Töne	25.42
Wetterfrage	67.68
Wirtschaftswoche	122.30
Wissenschaftsmagazin	393.61
Zivadiliring	50.03
Zytlupe	44.74
Total	4715.87

Table 1: List of SRF podcasts, links to the source, and hours of raw audio.

[3]Note that this will also be very helpful for downstream ASR tasks.

[4]https://developer.srgssr.ch/

YouTube Podcast Name	Length (h)
Auf Bewährung - Leben mit Gefängnis	3.00
Berner Jugendtreff	127.80
Ein Buch Ein Tee	3.73
expectations - geplant und ungeplant kinderfrei	16.84
Fadegrad	49.95
Feel Good Podcast	319.60
Finanz Fabio	58.44
Scho ghört	23.45
Sexologie - Wissen macht Lust	15.41
Über den Bücherrand	14.53
Ungerwegs Daheim	38.67
Wir müssen reden - Public Eye spricht Klartext	17.52
Total	688.93

Table 2: List of YouTube podcasts, links to the source, and hours of raw audio.

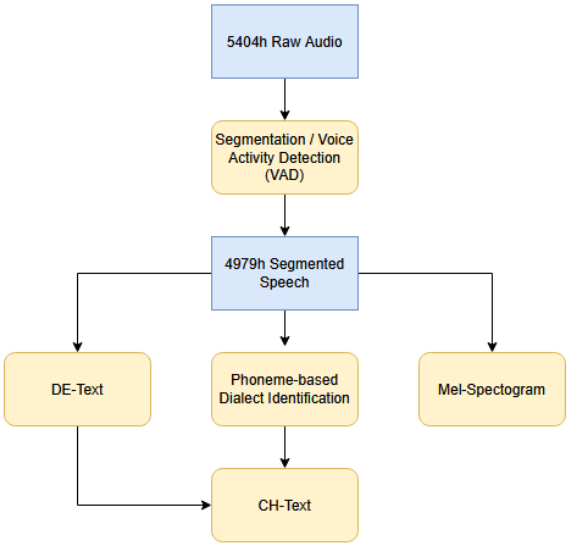

Figure 1: Automated Data Annotation Pipeline

3.1 Pipeline Evaluation

In order to ensure high quality of the automated annotations, we performed an evaluation of the performance of the individual steps of the annotation pipeline. This section presents the evaluation results. The *Zivadiliring* podcast[5] was selected for all evaluations due to its moderate size (approximately 50 hours of raw audio), exclusive use of Swiss German, minimal guest appearances, and the known dialects of its hosts—one from Eastern Switzerland and two from Zurich. These characteristics make it a representative sample of other podcasts.

The diarization was evaluated on a single episode lasting 42 minutes and 38 seconds, using (ela, 2024; Brugman and Russel, 2004) for manual annotation. The diarization pipeline achieved a Diarization Error Rate (DER) of 14.1%, which is comparable to its performance on the AISHELL-4 corpus (Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, 2017), where it reached 12.2%. The Standard German transcription was evaluated by manually transcribing 100 randomly selected audio samples, achieving a Word Error Rate (WER) of 0.30 ± 0.264. Example transcriptions are provided in Table 3. The observed deletions and substitutions can be attributed to the numerous linguistic differences between Swiss German and Standard German. These include the omission of past tense forms, instead preferring the perfect tense, as well as variations in auxiliary verbs, grammatical structures, and the use of Helvetisms or loanwords that either do not exist in Standard German or carry different meanings. Lastly, there is the inherent loss of information when transcribing the audio automatically from Swiss German to Standard German using whisper.

Dialect	Hypothesis	Reference
Eastern CH	Und dann ist quasi die Idee, wenn du als Burning Man gehst, dass du etwas wie einen Provider machst.	Dann ist quasi die Idee auch, dass du, wenn du an das Burning Man gehst, etwas providest.
Zurich	Aber es ist nicht gescheitert. Nein, ich bin ja so hyperemotional. Dann verplatzt es mich und dann bin ich aber wieder ruhig nach vier Sekunden. Aber ich habe dann schon wahrscheinlich ein bisschen umgewettert.	Aber es ist nicht gescheitert an der Wäsche. Nein, ich bin ja schon, ich bin ja so Hyperemotional, dann verplatzt es mich, dann bin ich aber auch wieder ruhig nach vier Sekunden. Aber habe dann wahrscheinlich schon herumgeflucht.
Zurich	Oder was ist er? Weisst du, mit dem Rettchen wüsstest du, über was wir reden. Was ist er gestern gewesen? Was ist er heute?	Oder was ist er? Weisst du damit wir wissen über was wir reden. Was war er gestern? Was ist er heute?

Table 3: Comparison of generated and manual annotated Standard German sentences

The Naïve Bayes classifier from (Bolliger and Waldburger, 2024) was retrained with an additional class for Standard German using the Common-

[5]https://www.srf.ch/audio/zivadiliring

Voice corpus (Ardila et al., 2020). A total of 30 hours of audio was sampled from CommonVoice, ensuring an age and gender distribution similar to that of the phonemicized STT4SG-350 corpus (Plüss et al., 2023), where each dialect region consists of 30 hours of speech. The classifier achieved a macro F1-score of 0.88 across the eight regions. When applied to the *Zivadiliring* episodes, nearly two-thirds of all samples were classified as Zurich and one-third as Eastern Switzerland, aligning with the hosts' origins. Additionally, in an episode where one of the hosts was replaced by a guest from Basel, the classifier correctly identified the samples as Basel.

Lastly, the Swiss German transcription of the same 100 samples used in the Standard German evaluation resulted in a Word Error Rate (WER) of 0.639±0.253. This high error rate is primarily attributed to the lack of a standardized writing system for Swiss German. Example transcriptions are provided in Table 4.

Dialect	Hypothesis	Reference
Zurich	Si känt scho mal din Name, fast. Er isch Content Creator, er isch berüehmt im Internet und er isch super.	Sie kennt scho mal din Name, fast. Er isch Content-Creator, er isch berüehmt im Internet und er isch
Eastern CH	I mein, wa de Onur alles seit. Nur will me zemme wohned isch jetzt nöd de Informationsfluss.	Ich meine, was dä Onur alles seit. Nur will mir zemme wohnet isch ezt do nöd de Informationsfluss.
Zurich	Drum händs so gfunde, ja du bisch irgendwie d'Muetter und denn au irgendwie nöd. Ich glaub, es git nöd die definiert Rolle. Aber ich han so gfunde d'Klaschtante isch no härzig.	Drum hät si d Mueter gfunde, dass si das au nöd gseh hät, dass Klatschstunde no härzig isch.

Table 4: Comparison of generated and manual annotated Swiss German sentences

4 Corpus Statistics

Raw Data. The raw audio is sourced from 25 SRF and 12 YouTube podcasts, comprising 15171 individual episodes with an average length of 1277.28 seconds (21.28 minutes). Episode durations are unevenly distributed, visualized in Figure 2, forming two distinct peaks: one between 100 and 200 seconds and another between 1,600 and 1,800 seconds. The first peak is primarily due to the podcast *100 Sekunden Wissen*, in which hosts provide information about various topics in around 100 seconds. In general, most podcasts produce episodes ranging from 20 to 30 minutes in length, as seen in Figure 4.

The largest podcast is *Tagesgespräch* with 1661.33 hours of raw audio, comprising nearly 31% of the total dataset, clearly visible in Figure 5. On average there are 410 episodes in a podcast, while the median is significantly lower at 104. Outlier episodes ($> 7200s, n = 32$) were typically special episodes, such as yearly recaps, video game playthroughs, or guest interviews. The longest episode in the dataset lasted 13846 seconds (3 hours and 50 minutes), while the shortest was just 19 seconds.

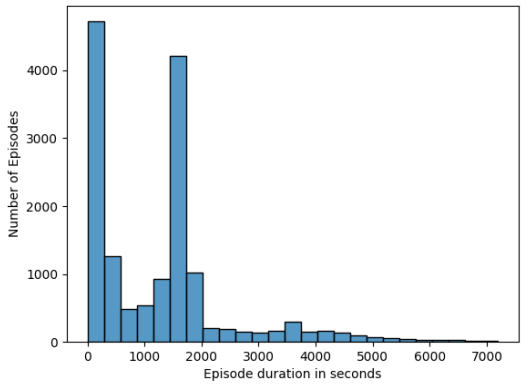

Figure 2: Distribution of episode durations in seconds of all podcasts in the corpus. Outliers ($n = 32$) with length > 7200 seconds are not shown.

Filtering. The filtering step, where we remove, for instance, samples with music only, reduced the data from 5404 hours of audio to 4979 hours of speech, which was segmented into 1.76M samples

Token Counts. After filtering, the data contains 55.85M tokens, calculated using spaCy (Honnibal et al., 2020). The token distribution is shown in 3. Since we did not segment the data on the sentence level, this led to a bimodal distribution of the tokens, visualized in Figure 3, with a large concentration of samples at 15 seconds. Training of a TTS model downstream then led to longer segments being generated better than shorter ones. Future work may improve this. The first peak with token counts of between 7 and 14 could be explained by the characteristics of spontaneous speech in a podcast

setting, in which hosts often interrupt each other in turns or simultaneously, leading to short segments of speech from individual hosts. The second peak with token counts between 40 and 53 can be explained by the generally more information-dense segments in podcasts or shows, where hosts have a monologue telling a story, reading a book or letter, or similar. Additionally, it was found that very short (< 7 tokens) and very large (≥ 65 tokens) samples were often erroneous or incoherent translations by whisper, either due to complex audio or simple mistranslations.

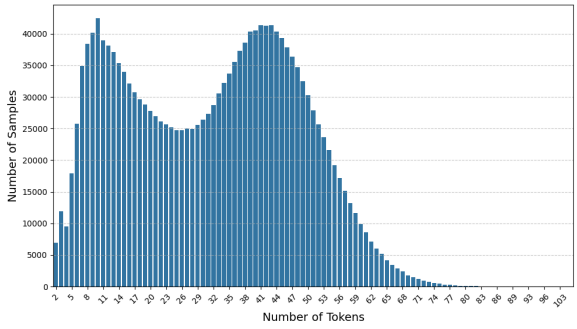

Figure 3: Standard German token distribution of segmented audio samples.

Dialects. At the dialect level, the corpus is highly unbalanced: the two largest regions, Standard German and Zurich, account for 57.53% of audio in the dataset, while the smallest region, Valais, represents only 0.79%. Table 5 provides further insight into the dialect distributions. Additionally, it was observed that Standard German tended to have segments with larger token counts than other dialect regions. This can be attributed to SRF broadcasting more formal and information-dense segments such as science, philosophy, and news programs in Standard German rather than Swiss German. This ensures that all Swiss residents can understand the content regardless of their familiarity with Swiss German dialects. The moderators of these programs are not required to be from Switzerland and could thus originate from Germany, Austria, or another German-speaking region. We hypothesized that the pronunciation of Swiss German speakers using Standard German (i.e., Swiss Standard German) may have a beneficial effect on model training. However, as we are currently unable to distinguish between them, both are grouped under the Standard German label and kept in the dataset.

Region	Samples (K)	Length (h)	% of Dataset	Tokens (M)
Basel	179	460.81	9.25%	5.35
Bern	293	771.38	15.49%	8.98
German	538	1685.72	33.86%	17.23
Grisons	57	151.33	3.04%	1.74
Central CH	121	341.22	6.85%	3.95
Eastern CH	121	350.60	7.04%	4.00
Valais	15	39.46	0.79%	0.43
Zurich	440	1178.58	23.67%	14.13
Total	1767	4979.09	100.00%	55.81

Table 5: Corpus statistics by dialect concerning number of samples, duration, percentage of total duration, and number of tokens.

5 Potential Use Cases

The Swiss German Podcasts Corpus can be a valuable resource for various NLP tasks, particularly for Swiss German. Unlike many existing datasets that focus on scripted or carefully controlled speech, our corpus contains spontaneous, natural, and uncontrolled speech. This makes it particularly useful for real-world applications where speech is often erratic, featuring hesitations, interjections, interruptions, and overlapping speakers. The large size of the corpus and the weak annotation make it particularly useful for weakly supervised learning approaches. An example task where this approach yielded very good results is in *Voice Adaptation for Swiss German dialects* using the XTTSv2 architecture (Casanova et al., 2024). Since the corpus contains a mix of Swiss German and Standard German, it can also serve as an excellent resource for training *Swiss German–to–Standard German machine translation models*. Such models can bridge the gap between spoken dialects and formal written language, enabling better transcription and translation.

6 Corpus Access

SwissGPC v1.0 will be accessible through the Swiss Association for Natural Language Processing (SwissNLP) website. The corpus will include:

1. A comprehensive list of links to all podcasts sourced from SRF and YouTube

2. The code for both downloading any podcast from SRF and YT and the automated annotation pipeline

7 Conclusion

We have presented the Swiss German Podcast Corpus (SwissGPC v1.0), the first mid-to-large-scale Swiss German speech corpus comprising approximately 5400 hours of raw audio (4979 hours of

speech after data cleaning). While the audio can not be released due to licensing concerns, we have provided references to individual podcasts, including an approach for downloading the audio. Additionally, we defined an automated annotation pipeline to weakly label the data for downstream use.

We are convinced that SwissGPC will enable interesting research in the Swiss German speech processing space, and we are excited to see applications utilizing it.

Limitations

The corpus represents a snapshot in time of the selected podcasts. Shows may release new episodes, remove existing ones, change name or location, be discontinued, or be taken offline as a whole. As a result, reproducing the results given here may prove challenging.

SwissGPC v1.0 is highly imbalanced on a dialectal basis, and future work may seek more audio from under-represented regions and add it to the corpus.

The list of podcasts from SRF is not exhaustive, as during the writing of this paper additional podcasts were found that could be utilized. Additionally, it should also be possible to crawl TV shows from SRF, such as *SRF bide lüt*, *Arena*, and more via their website[6] or YouTube channel[7], increasing the size of the corpus further.

References

2024. ELAN (Version 6.8) [Computer software].

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Laura Bolliger and Safiyya Waldburger. 2024. Automatische erkennung schweizerdeutscher dialekte anhand von audiodaten via phonemtranskriptionen. Technical report, ZHAW Zürcher Hochschule für Angewandte Wissenschaften.

Tobias Bollinger, Jan Deriu, and Manfred Vogel. 2023. Text-to-Speech Pipeline for Swiss German – A comparison. *Preprint*, arXiv:2305.19750.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Hennie Brugman and Albert Russel. 2004. Annotating Multi-media/Multi-modal Resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

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.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German. *CoRR*, abs/2103.11401.

Eyal Dolev, Clemens Lutz, and Noëmi Aepli. 2024. Does Whisper Understand Swiss German? An Automatic, Qualitative, and Human Evaluation. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 28–40, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. *Zenodo*.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. AIShell-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline. In *Oriental COCOSDA 2017*, page Submitted.

JuanBindez. 2025. pytubefix: A fork of PyTube with fixes for compatibility issues. Accessed: 2025-03-11.

Brian McFee, Matt McVicar, Daniel Faronbi, Iran Roman, Matan Gover, Stefan Balke, Scott Seyfarth, Ayoub Malek, Colin Raffel, Vincent Lostanlen, Benjamin van Niekirk, Dana Lee, Frank Cwitkowitz, Frank Zalkow, Oriol Nieto, Dan Ellis, Jack Mason, Kyungyun Lee, Bea Steers, Emily Halvachs, Carl Thomé, Fabian Robert-Stöter, Rachel Bittner, Ziyao Wei, Adam Weiss, Eric Battenberg, Keunwoo Choi, Ryuichi Yamamoto, CJ Carr, Alex Metsai, Stefan Sullivan, Pius Friesch, Asmitha Krishnakumar, Shunsuke Hidaka, Steve Kowalik, Fabian Keller, Dan Mazur, Alexandre Chabot-Leclerc, Curtis Hawthorne, Chandrashekhar Ramaprasad,

[6] https://www.srf.ch/play/tv/sendungen
[7] https://www.youtube.com/@srfdoku

Myungchul Keum, Juanita Gomez, Will Monroe, Viktor Andreevitch Morozov, Kian Eliasi, nullmightybofo, Paul Biberstein, N. Dorukhan Sergin, Romain Hennequin, Rimvydas Naktinis, beantowel, Taewoon Kim, Jon Petter Åsen, Joon Lim, Alex Malins, Darío Hereñú, Stef van der Struijk, Lorenz Nickel, Jackie Wu, Zhen Wang, Tim Gates, Matt Vollrath, Andy Sarroff, Xiao-Ming, Alastair Porter, Seth Kranzler, Voodoohop, Mattia Di Gangi, Helmi Jinoz, Connor Guerrero, Abduttayyeb Mazhar, toddrme2178, Zvi Baratz, Anton Kostin, Xinlu Zhuang, Cash TingHin Lo, Pavel Campr, Eric Semeniuc, Monsij Biswal, Shayenne Moura, Paul Brossier, Hojin Lee, and Waldir Pimenta. 2024. librosa/librosa: 0.10.2.post1.

Claudio Paonessa, Yanick Schraner, Jan Deriu, Manuela Hürlimann, Manfred Vogel, and Mark Cieliebak. 2023. Dialect Transfer for Swiss German Speech Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15240–15254, Singapore. Association for Computational Linguistics.

Claudio Paonessa, Vincenzo Timmel, Manfred Vogel, and Daniel Perruchoud. 2024. Whisper fine-tuning for Swiss German: A data perspective. In *Proceedings of the 9th edition of the Swiss Text Analytics Conference*, pages 192–192, Chur, Switzerland. Association for Computational Linguistics.

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. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2020. Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus. *CoRR*, abs/2010.02810.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *Preprint*, arXiv:2212.04356.

Clément Sicard, Victor Gillioz, and Kajetan Pyszkowski. 2023. Spaiche: Extending State-of-the-Art ASR Models to Swiss German Dialects. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 76–83, Neuchatel, Switzerland. Association for Computational Linguistics.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. In *Interspeech 2022*, pages 2113–2117.

A Corpus Statistics

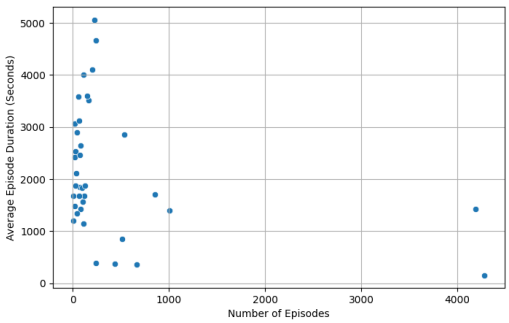

Figure 4: Comparison of total number of episodes in a podcast to the average duration per episode in the podcast.

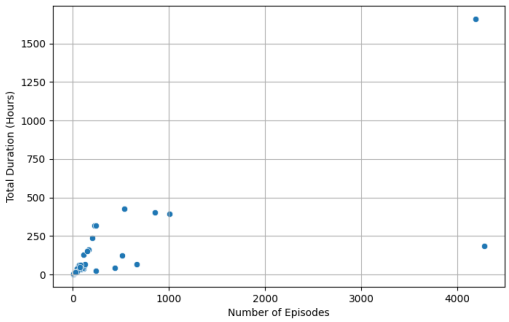

Figure 5: Comparison of the total number of episodes in a podcast to the total duration of all episodes combined in the podcast.

DE-Text	CH-Text	Dialect	Podcast
Die Kosten steigen natürlich dieses Jahr zwischen 6 und 7 Prozent, je nach Leistungserbringer-bereich. Aber letztes Jahr hatte man schon Defizite. Also schon letztes Jahr haben die Prämieneinnahmen die Ausgaben nicht gedeckt. Dieses Jahr wird es noch schlimmer sein. Das nächste	D Kösta stigen natürlich das Johr zwüscha sechs und sieba Prozent, je noch Leistigser-bringerberich. Aber letschts Johr hend d Prämieneinahma d Usgaba nit bedeckt kah.	Grisons	Samstags-rundschau
grossen Deutschschweizer Massenmedien, die noch eine regelmäs-sige Gamekritik gemacht haben. Alle anderen haben das schon viel länger aufgegeben als wir. Und auch bei den Spezialisten, die sich jetzt spezi-fisch für Gamer	Grosse Dütschschwizer Massemedia, wo no e regelmäs-sigi Gameskritik gmacht händ. Alli andere händ das scho vill länger ufgeh wie mir. Und au bi de Spezialiste, wo sich jetzt spezi-fisch für Games.	Zurich	Geek-Sofa
Es war eine Erleichterung, nachdem die UBS angekündigt hat, dass sie auf die staatlichen Garantien verzichtet, die wir im März sprechen mussten. Ohne viel Enthusi-asmus.	Es isch e Erliech-terig gsi, nochdem dUBS aakündigt het, dass sie uf di staatliche Garantie vozichtet hend, wome im März sproche müend.	Eastern CH	Tages-gespräch
nur noch mit Katalysatoren zulassen, so würde man längerfristig den Schadstof-fausstoss massiv beschränken kön-nen.	Nor no met Katalysatore zueloh, so wörd mer längerfrestigi de Schadstof-fusstoss massiv chönne beschränke.	Central CH	100 Sekun-den Wissen

Table 6: Examples of segmented samples in the corpus.