Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3250–3256
Marseille, 20-25 June 2022

SDS-200: A Swiss German Speech to Standard German Text Corpus

Michel Plüss[a], Manuela Hürlimann[b], Marc Cuny[c], Alla Stöckli[c], Nikolaos Kapotis[b], Julia Hartmann[a], Malgorzata Anna Ulasik[b], Christian Scheller[a], Yanick Schraner[a], Amit Jain, Jan Deriu[b], Mark Cieliebak[bc], Manfred Vogel[a]
[a]University of Applied Sciences and Arts Northwestern Switzerland, Windisch
[b]Zurich University of Applied Sciences, Winterthur
[c]SpinningBytes AG, Winterthur
michel.pluess@fhnw.ch

Abstract

We present SDS-200, a corpus of Swiss German dialectal speech with Standard German text translations, annotated with dialect, age, and gender information of the speakers. The dataset allows for training speech translation, dialect recognition, and speech synthesis systems, among others. The data was collected using a web recording tool that is open to the public. Each participant was given a text in Standard German and asked to translate it to their Swiss German dialect before recording it. To increase the corpus quality, recordings were validated by other participants. The data consists of 200 hours of speech by around 4000 different speakers and covers a large part of the Swiss German dialect landscape. We release SDS-200 alongside a baseline speech translation model, which achieves a word error rate (WER) of 30.3 and a BLEU score of 53.1 on the SDS-200 test set. Furthermore, we use SDS-200 to fine-tune a pre-trained XLS-R model, achieving 21.6 WER and 64.0 BLEU.

Keywords: Corpus, Less-Resourced/Endangered Languages, Speech Recognition/Understanding, Speech Resource/Database, Statistical and Machine Learning Methods

1. Introduction

We present Schweizer Dialektsammlung (SDS-200), a corpus of Swiss German dialectal speech with the corresponding Standard German text. The data consists of 200 hours of speech. We make the corpus publicly available [1].

Swiss German is a family of German dialects spoken by around five million people in Switzerland. It differs from Standard German regarding phonetics, vocabulary, morphology, and syntax and is primarily a spoken language. While it is also used in writing, particularly in informal text messages, it lacks a standardized orthography. This leads to difficulties for automated text processing due to spelling ambiguities and huge vocabulary size. Therefore, it is often preferable to work with Standard German text, for which automated processing tools exist in abundance. The main challenge is that Swiss German is not a unified language but a collection of dialects, which sometimes differ significantly in phonetics, grammar, and vocabulary. The immense vocabulary makes it hard to create a Swiss German Automatic Speech Recognition (ASR) system. Due to these reasons, Swiss German is a low-resource language. One way to tackle Swiss German ASR is an end-to-end Swiss German speech to Standard German text approach. This can be viewed as a speech translation (ST) task with similar source and target languages. Training a model for this task requires a substantial amount of data. Unfortunately, not enough public data is available for Swiss German. The largest available corpus, the Swiss Parliaments Corpus (SPC) (Plüss et al., 2021), is limited to the Bernese dialect. However, there are many different dialects in Switzerland, some of which differ substantially from Bernese because the difference between dialects can be significant, especially regarding vocabulary and pronunciation; as many dialects as possible should be part of the training data.

For SDS-200, we created a web recording tool[2] which is open to the public. The idea is that the public can record Standard German sentences in their Swiss German dialect. Other participants then validate the recordings. Almost 4000 different participants from all over Switzerland helped create a high-quality corpus covering a large part of the Swiss German dialect landscape. To cover a wide range of topics and increase vocabulary diversity, we used texts from Swiss newspapers and the German Common Voice corpus. The code of the tool is open source[3].

The remainder of this paper is structured as follows: Related work is discussed in section 2. The data collection process is described in section 3. Corpus preparation and statistics can be found in section 4. In section 5, we describe a baseline model trained on the corpus. Section 6 wraps up the paper and gives directions for future work.

2. Related Work

End-to-end approaches are widely used in deep learning, especially natural language processing (NLP). In the domain of speech translation, suitable corpora are

[1] `https://swissnlp.org/datasets/`

[2] `https://dialektsammlung.ch/de`

[3] `https://github.com/stt4sg/dialektsammlung-public`

scarce. The MuST-C dataset (Di Gangi et al., 2019) provides 400 h of English speech data with sentence-aligned text for eight different languages (German, French, Spanish, Italian, Dutch, Portuguese, Romanian, and Russian). The MuST-C data is collected from TED talks, providing a variety of topics and speakers (male/female, native/non-native speakers). TED talks are manually transcribed and translated, providing a high-quality data source.

Europarl (Iranzo-Sánchez et al., 2020) is another ST corpus with speech and sentence-aligned text for 6 European languages (English, German, French, Spanish, Italian, and Portuguese) containing between 20 and 89 hours of audio for 30 pairs. The sentence alignment is done automatically. Due to the automatic alignment, audio data with low alignment confidence is discarded, and the data quality is lower than manual text alignment. Europarl contains speeches held in the European Parliament.

Four public datasets contain Swiss German audio with transcripts. SPC (Plüss et al., 2021) is the largest corpus with 293 hours of data in the Bernese dialect recorded in the Bernese cantonal parliament. The text and audio are automatically aligned by using commercial Standard German ASR systems, followed by a forced sentence alignment using the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). The ArchiMob dataset (Scherrer et al., 2019) includes 69 hours of Swiss German speech and Swiss German transcript. There are no Standard German transcripts available. The Radio Rottu Oberwallis dataset (Garner et al., 2014) includes 8 hours of speech, 2 of which are provided with Standard German transcripts. Swiss-Dial (Dogan-Schönberger et al., 2021) is a high-quality dataset including eight different Swiss German dialects with roughly 3 hours of audio data per dialect. The sentences are crawled from newspapers and Wikipedia and then manually translated into the selected eight Swiss German dialects. The translated sentences are then recorded sentence by sentence in a studio setting.

SDS-200 combines the strengths of the existing corpora in Swiss German ASR with a large size of 200 hours, Standard German transcripts, and perfect alignment. What makes it unique is the coverage of a large part of the Swiss German dialect landscape and that almost 4000 different speakers made the recordings. We now describe the components in more detail.

3. Data Collection

Our data collection tool is based on the Common Voice platform (Ardila et al., 2020). We adapted the annotation guidelines to the special case of Swiss German. We use the two-step annotation process of the original platform consisting of a recording step and a validation step (see Figures 1 and 2). For the recording step, we presented Standard German sentences from Swiss newspapers, covering diverse topics and Switzerland-specific named entities, and texts from the German Common Voice corpus to the participants. They were then asked to translate each sentence into their Swiss German dialect and record it. For the validation step, the participants were presented with a sentence-recording pair and asked if the recording contained an accurate Swiss German translation of the Standard German sentence.

The goal was to create a corpus with as many hours and as much dialect and topic diversity as possible. We worked extensively with the Swiss media to reach as many people as possible. To enhance the engagement, we organized two contests on our platform. The leaderboard contest awarded prices to the participants with the most recordings, factoring in the quality of their translations. The *Clash of Cantons* contest was a competition between the 26 Swiss cantons.

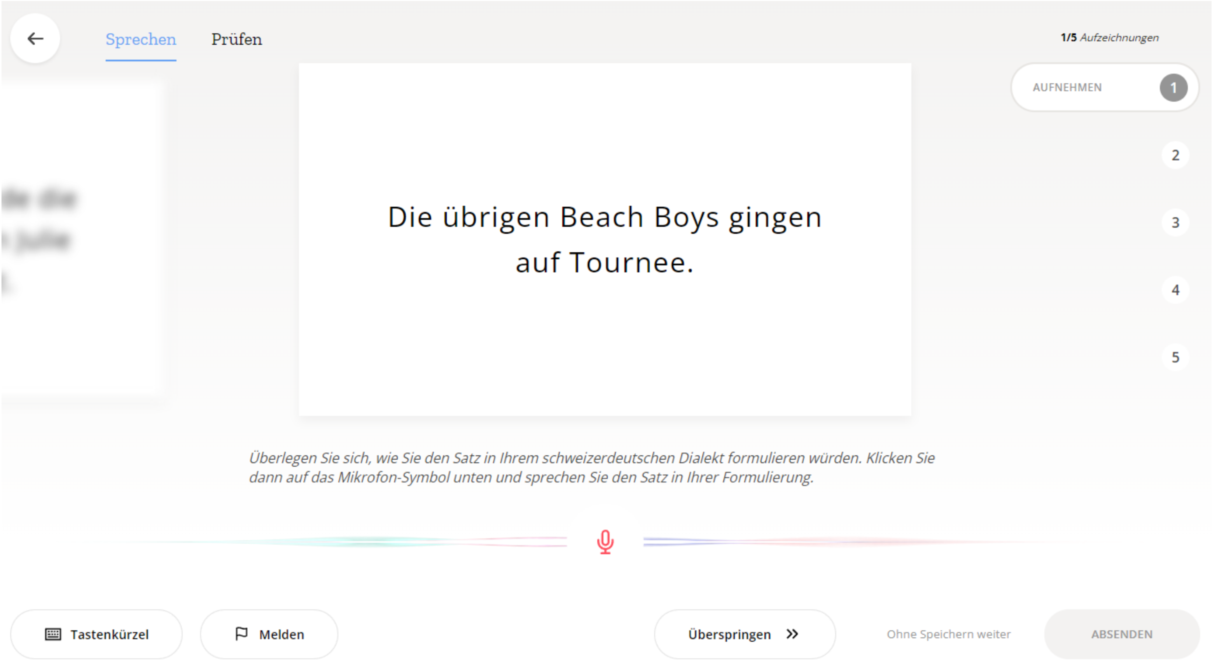
3.1. Sentence Selection

The sentences used for the recordings were derived from Swiss newspapers and the German dataset of Common Voice. We used newspaper articles from all categories from the past five years. As the speakers' task consisted of translating the sentences from Standard German to Swiss German, not just reading them, we expected the speakers' cognitive effort to be larger, hence the error probability to be higher. Keeping this in mind, we carefully selected sentences to ensure lexical diversity and reduce sentence complexity. To this end, we selected only sentences between 5 and 12 tokens long. We applied the following filtering criteria:

- Exclude sentences containing tokens that occur less than 1000 times per billion words. We use the Exquisite Corpus[4] to compute the word frequencies.
- Exclude sentences with a large number of rare words having an average word frequency below 10'000 per billion words.
- We removed sentences with dates and numbers with more than three digits. This is to reduce inconsistencies in how speakers read or translate the prompts.
- Sentences containing citations, e-mail addresses, hashtags, and phrases in brackets are also removed.
- We kept only complete sentences. We used simple heuristics to remove incomplete sentences. For instance, each sentence begins with an uppercase letter or a digit, and a sentence should contain at least one noun, pronoun, or proper noun and one verb.

The final set of prompts contains 1'267'195 sentences. Our tool samples newspaper sentences in 80% of cases, and in 20% of cases, it samples from the German Common Voice pool.

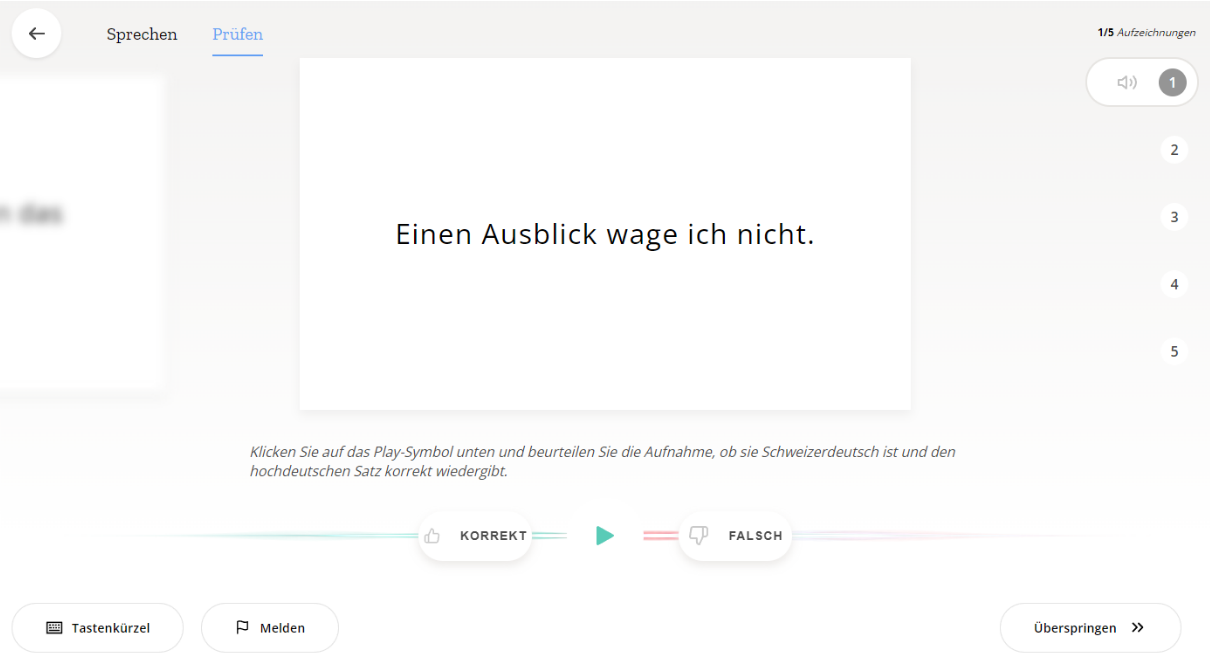
[4] https://github.com/LuminosoInsight/exquisite-corpus



Die übrigen Beach Boys gingen auf Tournee.

Überlegen Sie sich, wie Sie den Satz in Ihrem schweizerdeutschen Dialekt formulieren würden. Klicken Sie dann auf das Mikrofon-Symbol unten und sprechen Sie den Satz in Ihrer Formulierung.

Figure 1: Recording step in our tool. ”*Die übrigen Beach Boys gingen auf Tournee.*” is the sentence to be recorded.



Einen Ausblick wage ich nicht.

Klicken Sie auf das Play-Symbol unten und beurteilen Sie die Aufnahme, ob sie Schweizerdeutsch ist und den hochdeutschen Satz korrekt wiedergibt.

Figure 2: Validation step in our tool. ”*Einen Ausblick wage ich nicht.*” is the Standard German sentence. The recording must be played and then judged as correct (”*Korrekt*”) or wrong / inaccurate (”*Falsch*”).

3.2. Recording Tool

We made two adaptions to the original Common Voice (Ardila et al., 2020) platform. First, we added the possibility for the participants to specify the zip code of origin of their dialect[5]. This allows us to investigate dialects in different granularity levels: coarse dialect regions, cantons, fine-grained dialect regions, and even individual municipalities. Additional demographic information such as age and gender selection is already

[5] The origin of a participant’s dialect could for example be the place where he or she grew up and / or went to school. The specified zip code is not to be confused with the current place of residence, which would not allow reliable inference of a participant’s dialect.

Split	Hours	Sentences	Speakers
train (raw)	188.9	144′468	3428
train (filtered)	178.3	135′271	3247
validation	5.2	3638	288
test	5.4	3636	281

Table 1: Data splits of the Dialektsammlung corpus.

available in Common Voice. Second, we adapt the annotation guidelines to cover the special case of Swiss German. The annotation is performed in two steps: a recording step and a validation step.

Step 1: Recording. During the recording step, depicted in Figure 1, the participant is shown a Standard German sentence and asked to translate it to Swiss German speech. Sentences are recorded in packages of 5 and can be skipped or reported if necessary. One crucial point for our Swiss German speech to Standard German text use case is the inherent translation step the participant has to do before recording. As an example, the participant is presented with the following Standard German sentence: "*Robben verstand dies wie viele andere Spieler nicht.*". The participant should then think about how he or she would formulate this sentence in his or her Swiss German dialect, e.g. "*De Robben het das wie vieli anderi Spieler nid verstande.*", before actually recording the Swiss German version. This can include vocabulary as well as grammar changes, such as changing the past tense from Standard German "*verstand*" to Swiss German "*het (...) verstande*", which is necessary because the imperfect tense does not exist in Swiss German, where the perfect tense is used instead. We display an explanation popup with examples before the first recording to make this clear to participants. We also display a short explanation below the sentence to be recorded (see Figure 1).

Step 2: Validation. Figure 2 depicts the validation function. Participants are asked to listen to other recordings and judge whether the recording contains an accurate Swiss German translation of the Standard German sentence. Recordings are again validated in packages of 5 and can be reported or skipped if necessary. Similar to the recording function, we display a detailed explanation with examples of wrong (e.g. recording is in Standard German rather than Swiss German) or inaccurate (e.g. wrong tense) translations when a participant visits the validation page for the first time.

3.3. Collection Process

To reach as many people as possible, we collaborated with a range of national and local newspapers, television networks, and radio stations. In addition, four well-known Swiss comedians agreed to record a short video supporting the project and share it on their social media accounts, some of them reaching more than 100’000 followers.

To keep the participants motivated, we organized two contests, the leaderboard contest and the *Clash of Cantons*.

Leaderboard. The leaderboard contest was a competition between all registered participants. For each participant, we computed a score based on the number of recordings, the number of validations given, and the number of positive validations received. The top ten of the leaderboard were awarded attractive Switzerland-themed prizes. Furthermore, the participant with the highest recording quality (lowest rejection rate) was awarded a special prize.

Clash of Cantons. The *Clash of Cantons* was a competition between the 26 Swiss cantons. The idea was to spark a competition between the cantons and for participants to "fight" for their respective canton. The winning canton was picked according to its number of recordings, weighted by their average quality, normalized by the population of the canton.

The data of the corpus described here was collected over seven months, with 58 % of recordings made during the 38 days where the two contests were held. The current version contains 200 hours of raw speech data in MP3 format with a sampling rate of 32 kHz.

4. Corpus Preparation and Data Statistics

4.1. Data filtering

Crowd-sourced data needs filtering to ensure high data quality. We used the public validation process to filter bad samples such as empty, truncated, or silent recordings and wrong translations.

Of all recorded data, 33% have been validated, and of these samples, 88% have been accepted. To also use a large amount of unvalidated samples, we allow unvalidated samples as well under the following conditions:

- The speaker *has some* validated recordings and more than 80% of the validated clips are accepted.
- The speaker *has no* validated recordings and the duration is within 2 to 12 seconds.

We found that we were able to filter out many clips with recording problems (e.g., empty recordings) with the second rule. Since the added unvalidated data likely contains some invalid samples, they will need to be filtered further as more clips are validated. We also provide the unfiltered train data so that corpus users can compile their own filter rules.

4.2. Corpus Structure

We provide randomly generated train, validation, and test splits, ensuring that each speaker is part of only one split. The target size of the validation and test splits is 5.3 hours each. Table 1 shows the number of hours, sentences, and speakers of each split. To ensure optimal quality, validation and test splits only contain validated samples. Furthermore, to obtain balanced sets

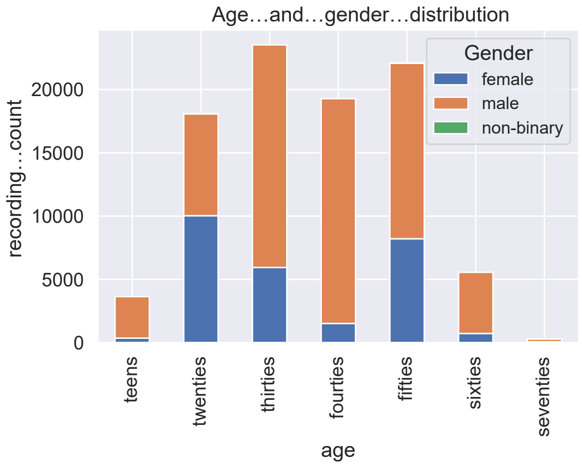


Figure 3: Number of utterances per speaker's age group and gender.

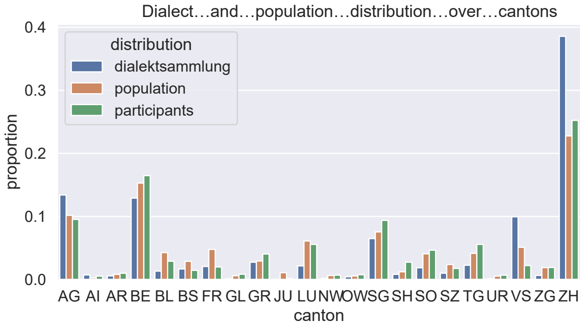


Figure 4: Canton distribution in the dataset compared with the relative population and the relative number of unique speakers per respective canton. Only cantons where Swiss German is spoken are shown.

and a larger variety of speakers, we only allow speakers with 5 to 200 recorded sentences to be part of either validation or test splits.

4.3. Data Statistics

On average, an utterance is 4.8 seconds long with a standard deviation of 1.3 seconds. The shortest and longest utterances are 2 and 11.2 seconds long, respectively. In Figure 5 we display the utterance length distribution.

By crowdsourcing the data, we obtain a diverse set of speakers regarding age, gender, and dialect. In total, the filtered SDS-200 contains 142'545 utterances with 138'553 unique sentences. The vocabulary consists of 41'289 German words. Out of 3816 speakers, 8% are male, 6% are female, 86% did not reveal their gender, and 4 participants are non-binary. In terms of utterances, 19% of utterances are voiced by females, 46% by males, and 35% of unknown gender. On average, each participant recorded 37 utterances with a standard deviation of 364 utterances. The participant with the most speech donations recorded 13'333 utterances. In Figure 3 we display the age and gender distribution over the recorded utterances. In Figure 4 we show the distribution over the number of recordings for each canton and compare them with the population of the respective cantons[6] and the proportion of unique speakers. The collected dialects follow the dialect distribution in Switzerland closely, with some exceptions. For Appenzell Innerrhoden, we have four times more utterances than the relative population. Wallis and Zürich have almost twice as many utterances. In the canton Wallis, one speaker recorded 10'368 out of 11'739 samples. The cantons Baselland, Glarus, Jura, Luzern, Nidwalden, Uri, and Zug are underrepresented in the SDS-200.

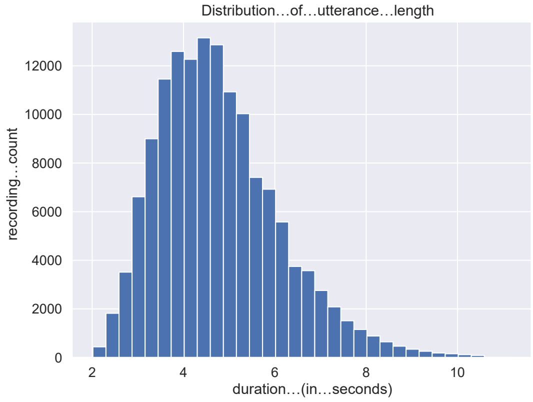


Figure 5: Distribution of utterance lengths in the SDS-200.

5. Baseline

We conducted experiments to demonstrate the use of the SDS-200 corpus for speech translation. We further evaluated how the corpus can be combined with the SPC (Plüss et al., 2021). Finally, we assessed how large-scale pre-training on unlabeled speech data can improve the performance by finetuning XLS-R Wav2vec models (Babu et al., 2021) on the SDS-200 train set.

Transformer Baseline. We employed Transformer (Vaswani et al., 2017) based models implemented in the FAIRSEQ S2T library (Ott et al., 2019; Wang et al., 2020) as our baselines. These models consist of a two-layer convolutional subsampler followed by a Transformer network with 12 encoder layers and six decoder layers. For the Transformer network, we employed eight attention heads, an embedding dimension size of 512, and a dropout rate of 0.15. We used the default model hyper-parameters and learning rate schedules provided by the library without any task-specific tuning. We evaluated the model performance when training on SDS-200 alone as well as the combination of SDS-200 and the SPC. After training, we

[6] We use the canton information as an indicator for the dialect.

Model	Train data	Model parameters	WER		BLEU	
			valid	test	valid	test
Transformer	SDS-200	72M	31.3	30.3	52.1	53.1
Transformer	SDS-200+SPC	72M	24.9	24.7	60.9	61.0
XLS-R (0.3B)	SDS-200	317M	27.2	26.9	54.9	54.6
XLS-R (1B)	SDS-200	965M	21.7	21.6	63.9	64.0

Table 2: Performance of the Transformer Baseline and XLS-R Wav2Vec models finetuned on the SDS-200 train set. We report Word Error Rate (WER) and BLEU scores obtained from evaluating on the SDS-200 valid and test splits.

averaged the weights of the ten checkpoints with the lowest validation loss to obtain the final model.

XLS-R fine-tuning. For the Wav2vec experiments, we employed XLS-R models (Babu et al., 2021) that were pre-trained on 436K hours of unlabeled speech data covering more than 128 languages and are publicly available[7]. Importantly, Swiss German was not part of the training data. Of the available pre-trained models, we evaluated XLS-R (0.3B) and XLS-R (1B), whereas the number in braces denotes the number of model parameters. XLS-R Wav2vec models consist of a convolutional feature encoder, followed by a stack of transformer blocks. Details of the architecture configurations can be found in (Babu et al., 2021). For the fine-tuning on the SDS-200 corpus, we followed the procedure and hyper-parameters described by the authors.

Results. The results of our experiments are shown in Table 2. Both additional labeled training data and large-scale self-supervised pre-training on unlabeled speech data lead to performance improvements. The strong performance of XLS-R (0.3B) highlights the benefits of latter in low-resource settings, even if the target language was not available during pre-training. Notably, for all our experiments, we did not use any external language model.

6. Conclusion

In this work, we presented SDS-200, a speech translation dataset for Swiss German speech to Standard German text. The main characteristics of this corpus are the large variety of Swiss German dialects that are covered and the large number of speakers that contributed to the data collection. The baseline achieved 30.3 WER score, and 53.1 BLEU score on the SDS-200 test set. The current version contains around 200 hours of speech.

Our goal is to increase the size of the corpus in the future, which will allow for even better performance. We plan to find new ways to engage the public, for instance, by adding gamification components to keep the engagement high. The current version is publicly available.

7. Acknowledgements

First and foremost, we would like to thank all participants for their contribution to this corpus. Furthermore, we thank Tamedia for providing the newspaper texts. We also thank Claudio Zuccolini, Frölein da Capo, Mike Müller, Renato Kaiser, Simon Enzler, and Sina for their public relations efforts.

8. Bibliographical References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv*, abs/2111.09296.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D., and Pino, J. (2020). fairseq S2T: Fast Speech-to-Text Modeling with fairseq. In *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*.

9. Language Resource References

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., and Turchi, M. (2019). MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the*

[7] `https://github.com/pytorch/fairseq/tree/main/examples/wav2vec/xlsr`

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2012–2017, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dogan-Schönberger, P., Mäder, J., and Hofmann, T. (2021). SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German.

Garner, P. N., Imseng, D., and Meyer, T. (2014). Automatic Speech Recognition and Translation of a Swiss German Dialect: Walliserdeutsch. In *Proceedings of Interspeech*, Singapore, September.

Iranzo-Sánchez, J., Silvestre-Cerdà, J. A., Jorge, J., Roselló, N., Giménez, A., Sanchis, A., Civera, J., and Juan, A. (2020). Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Plüss, M., Neukom, L., Scheller, C., and Vogel, M. (2021). Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus. In *Swiss Text Analytics Conference 2021*, Proceedings of the Swiss Text Analytics Conference 2021.

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