Swiss German Speech Translation and the Curse of Multidialectality

Martin Bär
University of Malta, University of the Basque Country
martin.bar.22@um.edu.mt

Andrea De Marco
Department of Artificial Intelligence
University of Malta
andrea.demarco@um.edu.mt

Gorka Labaka
HiTZ Center
University of the Basque Country
gorka.labaka@ehu.eus

Abstract

In many languages, non-standardized varieties make the development of NLP models challenging. This paper explores various fine-tuning techniques and data setups for training Swiss German to Standard German speech-to-text translation models. While fine-tuning on all available Swiss German data yields the best results, ASR pre-training lowers performance by 1.48 BLEU points, and jointly training on Swiss and Standard German data reduces it by 2.29 BLEU. Our dialect transfer experiments suggest that an equivalent of the *Curse of Multilinguality* (Conneau et al., 2020) exists in dialectal speech processing, as training on multiple dialects jointly tends to decrease single-dialect performance. However, introducing small amounts of dialectal variability can improve the performance for low-resource dialects.

1 Introduction

Swiss German (*Schweizerdeutsch*) is considered one of the most distinct and lively varieties of German with unique features on the phonological, morphological, syntactic and lexical levels[1]. It is a continuum of mostly High Alemannic German dialects in Switzerland, spoken by more than 5 million people. Swiss German is used extensively in everyday situations, including spoken communication, text messaging, local and national TV programs, and even regional parliaments. Standard German (*Hochdeutsch*) is used for formal and institutionalized forms of communication (Christen et al., 2020). This coexistence of two varieties with clearly separated use cases in a single speaker group has been described as *diglossia* by several researchers (Ferguson, 1959; Ender and Kaiser, 2009; Russ, 1990).

Swiss German dialects can vary significantly within Switzerland, sometimes even leading to difficulties in understanding between Swiss German speakers from distant regions (Christen, 2010). Due to particularities on all linguistic levels, Swiss dialects are hard to understand for many German speakers outside of Switzerland (Ender and Kaiser, 2009) and German learners who are primarily familiar with Standard German (Schlatter, 2024). This makes the need for systems that can translate from Swiss German speech to Standard German text apparent. It could facilitate the integration of non-Swiss-German speakers into Swiss society by enabling them to understand local TV programs, radio shows, dialectal voice messages, and conversations between their co-workers. Furthermore, dialectal speech translation can help preserve dialectal varieties and make language technologies more accessible to dialect speakers, contributing to the development of fair and equitable technologies (Joshi et al., 2024). In a study by Blaschke et al. (2024), 61% of respondents were in favor of systems that can translate dialect speech to Standard German text. This highlights the demand for dialectal translation systems beyond academic interests.

In the case of Swiss German, Automatic Speech Recognition (ASR) and Speech Translation (ST) are closely related. As Swiss German does not have any standardized written form and all of its speakers understand Standard German (Ender and Kaiser, 2009), it seems natural to prioritize Swiss German speech to Standard German text ST instead of Swiss German speech to Swiss German text ASR. Although there are works about the latter (Garner et al., 2014; Scherrer et al., 2019), ST is the subject of most research (Khosravani et al., 2021b,a; Paonessa et al., 2023; Sicard et al., 2023; Mutal et al., 2023) and was one of the shared tasks at the Swiss Text Analytics Conference[2] in 2021

[1]For a complete list of Swiss German particularities, we refer the reader to Russ (1990) and Christen (2019).

[2]https://www.swisstext.org/

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 165–179
July 31 - August 1, 2025

(Plüss et al., 2021) and 2022 (Plüss et al., 2023b).

Although the area is being actively researched, SwissText 2022 (Plüss et al., 2023b) has demonstrated that the problem is far from being solved. None of the participating teams were able to outperform the baseline model, a simple Transformer fine-tuned on three datasets. Later works achieved improvements over this baseline by using more data and experimenting with fine-tuning pre-trained models (Sicard et al., 2023; Plüss et al., 2023a,b). However, they did not explore further pre-training, nor did they utilize all the available data for Swiss German, or employ Standard German data. Paonessa et al. (2023) showed that one of the main challenges is that Swiss German ST needs to handle a considerable amount of dialectal variability. They found that some dialects benefit from positive transfer from related dialects, whereas others negatively influence overall performance. It remains unclear, however, how many dialects can be used together to improve performance and when performance starts to degrade. Here, we expect a breaking point as observed for the Curse of Multilinguality (Conneau et al., 2020) even for the closely related Swiss dialects. Furthermore, we don't know how small amounts of dialectal variability affect performance.

We aim to close these research gaps by:

1. Exploring fine-tuning and pre-training to improve performance for Swiss German ST and determine the usefulness of Standard German data.

2. Investigating whether there is a *Curse of Multidialectality* for Swiss German.

3. Observing how small amounts of dialectal variability affect the performance of Swiss German ST models.

2 Multidialectal Speech Processing

Joshi et al. (2024) highlight that variability within dialects of a language is one of the biggest challenges for dialectal NLP. This issue, referred to as *multidialectality* in the present work, has already been investigated in speech processing. ASR systems are often only trained on standard accents, making them perform poorly on other dialects of the same language (Sanabria et al., 2023; Parsons et al., 2023). Yadavalli et al. (2022) find that a model trained on multiple Telugu dialects jointly performs worse than a system trained on each dialect separately, indicating negative transfer. Similar issues have been observed for Japanese (Imaizumi et al., 2020), Chinese (Ding et al., 2024), Tibetan (Zhao et al., 2019), Flemish/Dutch (Herygers et al., 2023), Armenian (Arthur et al., 2024), and Arabic (Nasr et al., 2023; Ali et al., 2021).

Researchers have proposed various techniques to mitigate performance drops due to multidialectality, with a primary focus on Automatic Speech Recognition (ASR). Using pre-trained models has been found to outperform monolingual training from scratch (Arthur et al., 2024; Luo et al., 2021). Imaizumi et al. (2022) suggest dialect-aware ASR modeling by simultaneously performing dialect identification and ASR for Japanese dialects, Dan et al. (2022), Das et al. (2021), and Yadavalli et al. (2022) apply similar multi-task training approaches to Chinese, English, and Telugu. Using the standard and dialectal varieties together during training has been found to increase performance for Tunisian Arabic (Messaoudi et al., 2021), for multiple other Arabic dialects (Chowdhury et al., 2021), and for Thai when combined with curriculum learning[3] Suwanbandit et al. (2023).

3 Swiss German ST

For German, research in dialectal speech processing is scarce. Wepner (2021) calls for adapting ASR systems to Austrian German as they observe a performance discrepancy between German Standard German and Austrian Standard German. Similarly, Baum et al. (2010) find an increase of 24.8% in WER when evaluating a German ASR system on dialectal utterances, and Wirth and Peinl (2022) see the need to include dialectal varieties in German ASR datasets. Paonessa et al. (2023) find that the multidialectal nature of Swiss German, briefly described in the introduction, is one of the main challenges for Swiss German ST. They observe positive and negative transfer between dialects, mainly depending on their overall similarity as determined by Scherrer and Stoeckle (2016).

Swiss German ST is actively researched, and many datasets have been released in the past years[4].

[3] This is a multi-stage training approach where a model is trained on increasingly complex tasks (Bengio et al., 2009).

[4] This is not the case for other German dialects. ASR datasets have been released for Upper-Saxon (Herms et al., 2016), Austrian German (Schuppler et al., 2014), and the Southern Bavarian dialect De Zahrar (Gulli et al., 2024). However, we did not find any freely available datasets or other research on ST for these dialects, nor the widely spoken Bavar-

Table 1 lists these datasets and their abbreviations. STT and SDS were both collected by crowd-sourcing with a web recording tool, similar to the Common Voice datasets (Ardila et al., 2020). They contain Standard German sentences that participants were asked to translate into their dialect and record. SPC was automatically compiled from audio recordings of the Bernese cantonal parliament. These were automatically aligned with their Standard German transcriptions. Similarly, GRZH contains speech from the Zurich parliament. It does, however, not include transcriptions. AM is the only dataset we found that contains dialectal transcriptions. It was compiled by segmenting interviews that were conducted and transcribed in Swiss German.

Abbr.	Dataset	Total h	Train h	Cantons	T
STT	STT4SG-350 (Plüss et al., 2023a)	343	239	17	StG
SDS	SDS-200 (Plüss et al., 2022)	200	50	21	StG
SPC	Swiss Parliaments Corpus (Plüss et al., 2020)	293	217	N/S	StG
SDial	SwissDial (Dogan-Schönberger et al., 2021)	36	36	8	StG
GRZH	Gemeinderat Zürich Corpus (Plüss et al., 2021)	1208	1208	N/S	-
AM	ArchiMob (Samardzic et al., 2016)	80	0	14	SwG
-	Total data with Standard German labels	872	542	-	StG

Table 1: Swiss German speech datasets. *Total h* and *Train h* show the number of hours and the hours used in our experiments, respectively.
Abbreviations for the T (Transcriptions) column: *StG* = Standard German, *SwG* = Swiss German.

Early work on Swiss German to Standard German ST has focused on single dialects and pipeline systems (Garner et al., 2014), as ST data was scarce. However, Khosravani et al. (2021a) emphasize that the lack of a standard orthography and limited resources make it difficult to train cascade systems, making end-to-end architectures dominate the Swiss German ST area (Nigmatulina et al., 2020; Büchi et al., 2020; Sicard et al., 2023; Plüss et al., 2023a).

Current state-of-the-art models for Swiss German ST mostly follow the pre-train and fine-tune paradigm. Plüss et al. (2023a) fine-tune an XLS-R 1B model on the STT dataset and achieve state-of-the-art performance on the SDS, STT, and Swiss-Text2021 test sets (69.6 BLEU, 74.7 BLEU, and 66 BLEU, respectively). Sicard et al. (2023) find that Whisper exhibits strong zero-shot capabilities for Swiss German, outperforming the previously mentioned model on the SPC test set. Paonessa et al. (2023) trained three small models on the STT data, with XLS-R 0.3B outperforming Whisper S and a Transformer model trained from scratch. These findings make it difficult to determine which architecture is the most suitable for Swiss German ST. Furthermore, recent pre-trained multilingual models, such as SeamlessM4T (Communication et al., 2023) and AudioPaLM (Rubenstein et al., 2023), have not yet been evaluated for this task.

ian, Swabian, and Alsatian dialects.

4 Data and Models

In this section, we detail the models and datasets used for our speech-to-text translation experiments for Swiss German. The methodology used for the experiments will be described in Section 5 and 6.

4.1 Data and Dialects

Swiss German datasets were briefly introduced in Section 3. Table 1 summarizes them, and Table 2 lists the Standard German datasets we used for our fine-tuning experiments. For Standard German, we randomly sampled 180 hours from each dataset to obtain a total of 540 hours, the same amount we used for Swiss German. Initial experiments showed that this yielded better performance for Swiss German. To track model performance during training, we use validation splits of Swiss German (STT, SDS, SPC, GRZH) and Standard German (CV) datasets. The SPC and GRZH validation sets are not official splits and were created by randomly sampling 10% and 1% of their training data, respectively.

Abbr.	Dataset	Total h	Train h (long)	Train h
CV	Common Voice v17.0 (Ardila et al., 2020)	1423	933	180
MLS	Multilingual Librispeech (Pratap et al., 2020)	1995	1966	180
VP	VoxPopuli (Wang et al., 2021a)	282	264	180
-	Total data with Standard German labels	3700	3163	540

Table 2: Standard German ASR datasets. *Train h* shows the hours of speech used in our final experiments.

For the dialect transfer experiments, we only use the STT dataset because it is the largest available dataset that contains dialect region labels for every utterance. The SDS and SwissDial datasets also include dialect information, but the regions differ from the STT regions, limiting their usefulness for dialect experiments. Figure 1 shows all the regions from STT: *Basel* (BS), *Bern* (BE), *Central Switzerland* (CS), *Eastern Switzerland* (ES), *Grisons* (GR), *Valais* (VS), *Zurich* (ZH).

Test sets We use the test splits of STT, SDS, SPC, as well as the test sets of the SwissText 2021 (Plüss et al., 2021) and SwissText 2022 (Plüss et al., 2023b) shared tasks for model evaluation. To track

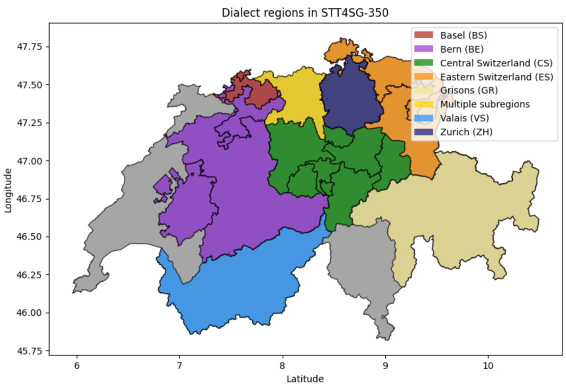


Figure 1: Dialect regions (from Paonessa et al., 2023).

the performance of our systems in Standard German ASR, we use the test split of CV. In addition to evaluating the STT test set per dialect, we provide the average performance over all datasets (including and excluding CV, denoted as $\varnothing$ and $\varnothing_{noCV}$, respectively) to be able to compare the models' robustness across different domains.

Data pre-processing All audios were resampled to a sampling rate of 16,000 Hz, the rate accepted by XLS-R. Similar to (Plüss et al., 2023a), all transcripts were normalized to only contain letters of the English alphabet (a-z), numbers, and the German umlauts *ä*, *ö*, *ü*. We use the unidecode package[5] to transform all other characters with accents or other special characters to ASCII. Then we remove all the non-alphanumerical characters (including punctuation) and lowercase the transcripts. We apply this normalization to all transcripts and translations used for training and evaluating our models. SPC was filtered to only include samples longer than 2 seconds and shorter than 15.5 seconds.

4.2 Models

We use XLS-R (Babu et al., 2021) as the base model for all our experiments. Its architecture is based on wav2vec 2.0 (Baevski et al., 2020), which is designed to learn high-quality speech representation through self-supervised learning, similar to masked language modeling in BERT (Devlin et al., 2019).

XLS-R is a multilingual version of wav2vec 2.0 and was pre-trained on 128 languages using 436,000 hours of unlabeled data for one million updates. In this way, the model learned powerful speech representations in several languages, similar to what happens for multilingual text models such as mBERT (Pires et al., 2019; Wu and Dredze, 2019; Tanti et al., 2021). Through fine-tuning, these representations can later be leveraged for downstream tasks across multiple domains and languages.

We train all of our models with Fairseq (Wang et al., 2020) and use the official checkpoints of XLS-R 300M and 1B (after the pre-training) as a starting point. We add a randomly initialized linear layer on top of the network and freeze the Transformer part of the network for the first 10,000 updates, similar to (Baevski et al., 2020). For generating the transcriptions, we use CTC decoding because Paonessa et al. (2023) found that it yields better results for Swiss German ST than seq2seq decoding. Additionally, we add a 5-gram language model (LM)[6] for decoding (LM fusion decoding) as this was shown to improve results, especially in low-resource contexts (Baevski et al., 2020; Babu et al., 2021). All results reported in this paper are achieved by applying LM fusion when applying CTC decoding.

5 Fine-Tuning Experiments

To improve the state-of-the-art of Swiss German ST and investigate whether using data from a closely related language (Standard German) is beneficial for ST performance, we conduct a series of experiments. Experiments 1-4 focus on different fine-tuning strategies and data setups, while Experiment 5 involves continued pre-training of XLS-R. All experiments aim to improve overall Swiss German translation performance and train robust models that perform well across different data domains.

5.1 Overview and Setup

Table 3 is an overview of all the fine-tuning experiments. Experiment 1 recreates the baseline model from Plüss et al. (2023a). In Experiment 2, we extend the fine-tuning data to all available Swiss German ST datasets to investigate how the additional variance introduced through these datasets affects performance on STT and/or specific dialect regions.

In Experiments 3 and 4, we use a multi-stage fine-tuning approach[7]. This has been shown to

[5] https://github.com/avian2/unidecode

[6] Similarly to (Plüss et al., 2023a), the LM was trained with kenlm (https://kheafield.com/code/kenlm/) on 100M Standard German sentences. Details are in Appendix A.

[7] In some works (e.g., (Suwanbandit et al., 2023)), this is also referred to as curriculum learning.

improve performance on low-resource tasks in MT (Imankulova et al., 2019; Luo et al., 2019), ASR (Medeiros et al., 2023; Deng et al., 2023; Yang et al., 2022), and ST (Kesiraju et al., 2023; Stoian et al., 2020; Wang et al., 2021b). Experiment 3 applies ASR pre-training (Kesiraju et al., 2023; Stoian et al., 2020) on Standard German data in the first step. Then, the resulting model is fine-tuned on the Swiss German ST data. In Experiment 4, we shuffle equal parts of the Standard German and Swiss German datasets together and fine-tune the model on all of them jointly in the first step. Then, we again fine-tune the resulting model on the Swiss German ST data.

In Experiment 5, we explore further pre-training on unlabeled Swiss German data. This is also called continued pre-training or language-specific pre-training and has been shown to improve downstream ASR performance (Bartelds et al., 2023; Nowakowski et al., 2023; Paraskevopoulos et al., 2024; Huang and Mak, 2023). XLS-R's pre-training data does not include any Swiss German, and the model might benefit even more from further pre-training on Swiss German data. Due to computational limitations, we do not use the labeled Swiss data for continued pre-training. However, we use it to fine-tune the resulting model in a second step.

Training Configuration We use the same hyperparameters as (Plüss et al., 2023a), who base theirs on (Babu et al., 2021). The only difference is that we use 1 GPU (NVIDIA A100 with 80 GB of memory) for training instead of 4. We tried to make up for this by using 4x the gradient accumulation steps but initial experiments showed that the performance gains were not worth the increased training time. The hyperparameters are listed in Table 8 in Appendix B.

Evaluation After fine-tuning, we generate predictions for the test sets described in Section 4.1 and evaluate the best model of the training run by BLEU and WER[8]. As Swiss German ST is more of a translation task, we use BLEU for the primary evaluations. The BLEU score is computed with SacreBLEU[9] (Post, 2018) on the references that were normalized as described in section 4.1. For the per-dialect results, we calculate the BLEU score using the entire corpus of the respective dialect. To calculate WER, we use the jiwer package[10].

As fine-tuning our models is resource-intensive, we are not able to conduct multiple training runs with different random seeds to determine if the differences between models are statistically significant. Instead, we use bootstrapping resampling to calculate system BLEU scores, as proposed in Koehn (2004) and implemented by SacreBLEU. This allows us to calculate confidence intervals and the statistical significance of BLEU score differences.

5.2 Results

Table 4 summarizes the results of the fine-tuning experiments. Using all available labeled data to fine-tune XLS-R proved to be the most effective approach, yielding the best overall model. While our model did not outperform the previously published baselines on each test set individually (see Figure 4 in Appendix C), we achieved the best average performance ($\varnothing_{noCV}$) across all test sets. This is most likely because the test set domains are very different, and we can assume that the domain-specific data resulted in some interference with the other domains.

Experiments 3 and 4 demonstrated that using Standard German data does not improve Swiss German dialect translation performance. Neither the ASR pre-training nor mixing Standard German and Swiss German data during fine-tuning improved the results for Swiss German. However, the Standard German data helped improve performance on the Common Voice dataset, adding 39.9 to the BLEU score when comparing the model only trained on Swiss German data (*AllSwiss*) and the model trained on a mixture of Swiss and Standard German data (*Joint_ft*). Nevertheless, the average Swiss German performance dropped by 2.29 BLEU for this setup. We observed this drop when the ratio of Swiss German and Standard German data was kept equal, and when 7 times more Standard German was used. We suspect that there were no improvements over *AllSwiss*, because the model is incapable of learning Standard German ASR and Swiss German ST simultaneously without any additional task separation, resulting in interference of the Standard German data.

Further pre-training the XLS-R on Swiss German speech from the GRZH corpus did not improve

[8] This is usually done in Swiss German ST, see Plüss et al. (2023a, 2021, 2023b); Sicard et al. (2023)

[9] Version 2.4.0

[10] https://jitsi.github.io/jiwer/

No.	Name	Description	Fine-tuned from	Fine-tuning data	Total hours
1	Baseline	Baseline replication from Plüss et al. (2023a)	XLS-R 1B	STT	239
2	AllSwiss	Fine-tune XLS-R on all available labeled data for SwG ST	XLS-R 1B	STT, SPC, SDS, SDial	542
3.1	ASR	Fine-tune model for StG ASR	XLS-R 1B	CV, MLS, VP	542
3.2	ASR_ft	Fine-tune StG ASR model on SwG ST data	3.1 ASR	STT, SPC, SDS, SDial	542
4.1	Joint	Jointly fine-tune on shuffled StG ASR and SwG ST data	XLS-R 1B	CV, MLS, VP, STT, SPC, SDS, SDial	1084
4.2	Joint_ft	Fine-tune jointly trained model on SwG ST data	4.1 Joint	STT, SPC, SDS, SDial	542
5.1	SwSSL	Continued pre-training on unlabeled SwG data	XLS-R 1B	GRZH	1208
5.2	SwSSL_ft	Fine-tune SwG pre-trained model on SwG ST data	SwSSL	STT, SPC, SDS, SDial	542

Table 3: Overview of fine-tuning experiments. *StG* = Standard German, *SwG* = Swiss German.

Test set	BLEU								WER							
	STT4SG	Baseline	AllSwiss	ASR	ASR_ft	Joint	Joint_ft	SwSSL_ft	STT4SG	Baseline	AllSwiss	ASR	ASR_ft	Joint	Joint_ft	SwSSL_ft
STT	**74.7**	71.9	72.2	9.6	70.2	68.9	69.4	70.9	**14.0**	15.9	15.6	73.9	16.8	17.7	17.5	16.4
SDS	**69.6**	66.8	67.2	6.6	65.2	63.0	63.5	66.3	**18.2**	19.9	19.6	78.7	20.9	22.5	22.2	20.3
SPC	54.9	52.8	**61.3**	7.3	60.2	60.2	60.5	60.7	30.2	32.4	**24.4**	79.8	25.6	25.6	25.4	24.8
ST21	**66.0**	62.4	64.7	10.1	64.1	62.5	62.7	62.9	**20.7**	22.9	21.4	73.6	21.7	22.6	22.4	22.7
ST22	-	73.7	**73.9**	11.8	72.4	71.5	71.8	73.2	-	14.7	**14.3**	69.6	15.6	15.9	15.7	15.1
$\varnothing_{noCV}$	66.3	65.5	**67.9**	9.1	66.4	65.2	65.6	66.8	20.8	21.2	**19.1**	75.1	20.1	20.9	20.6	19.9
CV	-	35.7	37.7	**84.9**	46.5	78.8	77.6	33.8	-	45.8	44.3	**8.6**	36.6	12.6	13.3	48.7
$\varnothing$	-	60.5	62.9	21.7	63.1	67.5	**67.6**	61.3	-	25.3	23.3	64	22.9	19.5	**19.4**	24.7

Table 4: Results of the baseline from Plüss et al. (2023a) and our experiments. Best results for each dataset are bold.

fine-tuning results either. We conjecture that this is due to low data quality and overfitting to the Zurich dialect, which was the only dialect in the dataset. Performance might benefit from (1) audio pre-processing or cleaning, and (2) adding more dialects to the unlabeled pre-training dataset.

Figure 2 shows the per-dialect results of the models. Comparing the best systems from Experiments 1-5 in Figure 2, it becomes evident that Standard German data does not help improve the performance for any specific dialect but rather introduces more dialectal variability that negatively affects performance. The model *AllSwiss* performs best for the Berne dialect, possibly due to the additional Berne data from SPC. This demonstrates that more in-dialect data helps improve performance even if that data is from a completely different domain. However, the over-representation of Berne data resulted in performance drops for other dialects (e.g., Valais and Zurich) when comparing *AllSwiss* to our Baseline, which was trained on the STT dataset balanced by dialect. These drops are even more substantial for the model trained jointly on Standard and Swiss German data, resulting in a performance loss of 7.8 BLEU for Valais.

6 Dialect Transfer Experiments

In these experiments, we vary the number and diversity of dialects in the training data to study the effect of dialectal variability on performance and determine if there is an equivalent to the *Curse of Multilinguality* (Conneau et al., 2020) for dialects.

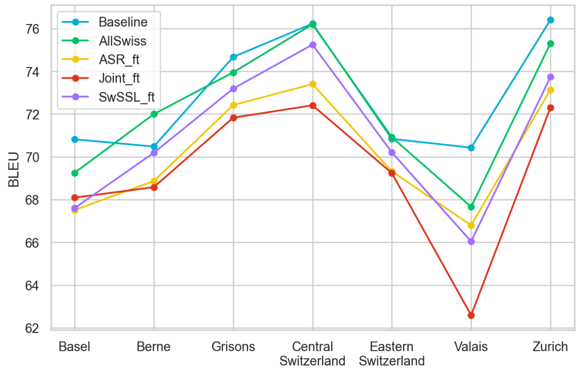


Figure 2: Per-dialect results of the fine-tuning experiments for the STT test set.

6.1 Overview and Setup

In the first set of experiments (DT1), we train a total of 7 models on 1, 2, 4, and 7 dialects. We use the Valais (VS) dialect region data as a starting point for one set of models, as this is the most distant dialect from all the others (Scherrer and Stoeckle, 2016; Paonessa et al., 2023). For a second set of models, we use the Zurich (ZH) region dialect because this was found to be the most similar to the other dialects. In the second set of experiments (DT2), we keep the dialect regions the same but add 10 minutes of speech data for every region that is not included. This allows us to investigate whether a small amount of data from different dialect regions can increase total performance. Table 5 contains an overview of these experiments.

Training Configuration We use XLS-R 300M for all the dialect transfer experiments (see Ap-

Name	Base	Full data	10 min of data	h (DT1)	h (DT2)
0	-	-	VS, ZH, CS, GR, BS, BE, ES	0	1.16
vs_1	VS	VS	ZH, CS, GR, BS, BE, ES	34	35
zh_2	ZH	ZH	VS, CS, GR, BS, BE, ES	33.46	34.45
vs_2	VS	VS, ZH	CS, GR, BS, BE, ES	67.46	68.29
zh_2	ZH	ZH, CS	VS, GR, BS, BE, ES	66.96	67.79
vs_4	VS	VS, ZH, CS, GR	BS, BE, ES	135.92	136.42
zh_4	ZH	ZH, CS, GR, BS	VS, BE, ES	136.17	136.67
all	-	VS, ZH, CS, GR, BS, BE, ES	-	238.71	238.71

Table 5: Overview for the dialect transfer experiments. The column **Base** shows the base dialect, **10 min of data** shows the regions added for DT2. **h (DT1)** and **h (DT2)** are the amounts of speech data used to train the first and second sets of experiments, respectively.

pendix B for more details on why this was chosen). We train each of our models on the balanced STT train set, filtered to only include the respective dialects. This amounts to 34 hours of speech data per dialect region. We use the same setup as described in Section 5 with the hyperparameters from Table 8 for the column *All others*.

Evaluation With the BLEU score, we compare the STT test set performance of the models. To determine whether there is a *Curse of Multilinguality* (Conneau et al., 2020) in Swiss German ST, we look at how the performance of the base dialect develops when adding more dialects in DT1. To investigate the influence of small amounts of added dialectal variability, the models from DT1 and DT2 are compared. Whether performance differences are significant is determined by BLEU's bootstrapping resampling as described in Section 5.

6.2 Results

The results of the DT1 and DT2 are displayed in Table 6 and 7, respectively.

Table 6 shows that for VS, performance is highest when the model is only trained on VS data and lowest when the training data only contains ZH data. Adding any non-VS data decreases BLEU scores, hinting at a *Curse of Multidialectality*. ZH exhibits the highest performance when the model is trained on the closely related dialects CS, GR, and BS in addition to ZH data. For most other regions and overall performance, models are best when using all the dialects for training. For BS and CS, models perform best when trained only on ZH, CS, GR, and BS, suggesting that VS, BE, and/or ES data have a negative impact on performance. This is another indicator of a *Curse of Multidialectality*.

Table 7 shows similar trends as the first set of experiments: VS performance is highest when using the highest percentage of VS data for training, while ZH peaks at 4 dialects that are closely related. We observe similar results for BS, GR, and CS. In Figure 3 we see that VS performance is significantly lower when adding 10 minutes of speech from all other dialect regions, indicating again that VS is strongly affected by other dialects. ZH, on the other hand, seems to benefit from the additional variety, exceeding the results from the DT1 Experiments. BE and the overall performance also benefit.

Contrary to DT1, GR now performs best when the training set contains only 4 dialects, suggesting that GR benefits from small amounts of variability from other dialects but is negatively affected if this variability is too high (i.e., when using all data for BE, VS, and ES). Another explanation could be that the very distant dialects of VS and/or BE significantly affect performance for GR when used entirely, but might enhance the model's generalizability by introducing a beneficial amount of variability when only small amounts of data are used. Further experiments are necessary to investigate how much variability is beneficial and when it negatively affects performance.

The Curse of Multidialectality Even though the model trained on all dialects performs well for both regions, there is a drop of 3.37 BLEU for VS compared to vs_1, the model trained on the VS data only. Paonessa et al. (2023) report similar findings. They trained 7 XLS-R models, one on each of the 7 regions from the STT dataset and found that the model trained on VS data is the only one that outperforms the model trained on the full dataset on its base dialect (in this case, VS). All the other models showed a performance drop of 1-5%, suggesting that they strongly benefit from cross-dialectal transfer. For ZH (and BS, CS, GR), however, our results indicate that this is only the case up to a certain number of (similar) dialects $3 \leq D_{max} \leq 6$ before performance drops slightly but significantly (0.97 BLEU in our case when comparing the performance for ZH of zh_4 and the model trained on all dialects). To determine the exact value of D_{max}, we would need to train models on every number of dialects between 1 and 7. Furthermore, we conjecture that D_{max} is higher when more similar dialects are included in the training set and lower otherwise. The fine-tuning experiments also suggest this: adding Standard German data in Experiments 3 and 4 can be considered as introducing another "dialectal" variety. After doing this, we saw a performance drop for almost all dialect regions

Name	**Regions**	VS	ZH	BE	BS	GR	CS	ES	Overall
vs_1	VS	**<u>67.8</u>**	43.2	36.8	35.6	40.0	46.1	25.0	42.4
vs_2	VS, ZH	67.1	65.1	49.4	53.4	57.2	64.0	51.9	58.4
vs_4	VS, ZH, CS, GR	64.7	65.8	54.0	56.2	65.6	66.1	58.5	61.5
all	all	64.4*	67.2	**<u>62.0</u>**	63.7	**<u>67.2</u>**	68.3	**<u>65.6</u>**	**<u>65.5</u>**
zh_1	ZH	40.7	64.4	44.4	51.0	56.9	61.7	55.7	53.6
zh_2	ZH, CS	48.7	66.5	53.1	54.9	59.6	67.6	57.8	58.4
zh_4	ZH, CS, GR, BS	52.5	**<u>68.2</u>**	57.1	**<u>64.3</u>**	66.7	**<u>68.3</u>**	63.8	63.0
all	all	64.4	67.2	**<u>62.0</u>**	63.7	**<u>67.2</u>**	68.3	**<u>65.6</u>**	**<u>65.5</u>**

Table 6: BLEU scores of the DT1 Experiments using around 34 hours of speech data for each dialect region specified in the **Regions** column. The best result per region is underlined and bold. Insignificant changes in BLEU as per bootstrap resampling for a system compared with the system in the row above are marked with *.

Name	**Regions**	VS	ZH	BE	BS	GR	CS	ES	Overall
0+10	-	0	0	0	0	0	0	0	0
vs_1+10	VS	**<u>65.8</u>**	50.4	41.9	43.5	48.4	51.9	39.0	48.8
vs_2+10	VS, ZH	65.7*	63.9	49.6	53.3	58.0	63.2	54.3	58.3
vs_4+10	VS, ZH, CS, GR	65.7*	67.4	56.9	58.7	66.8	68.0	61.3	63.6
all	all	64.4	67.2*	**<u>62.0</u>**	63.7	67.2*	68.3*	**<u>65.6</u>**	**<u>65.5</u>**
zh_1+10	ZH	43.8	64.5	47.1	52.8	59.0	62.5	57.6	55.4
zh_2+10	ZH, CS	50.5	67.3	54.7	56.7	60.7	67.9	60.2	59.8
zh_4+10	ZH, CS, GR, BS	53.7	**<u>69.1</u>**	58.2	**<u>64.5</u>**	**<u>67.8</u>**	**<u>68.8</u>**	64.5	63.9
all	all	64.4	67.2	**<u>62.0</u>**	63.7	67.2	68.3	**<u>65.6</u>**	**<u>65.5</u>**

Table 7: BLEU scores of the DT2 Experiments using 10 minutes of speech data for all the regions that are not included fully (specified in the *Regions* column).

(see Figure 2). These findings are reminiscent of the *Curse of Multilinguality* but require a more thorough investigation.

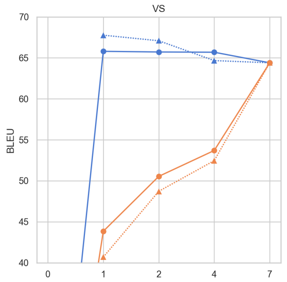
Introducing dialectal variability during training DT2 shows that the performance for almost all dialects improves when introducing dialectal variability through only 10 minutes of data per dialect. The improvements for the monodialectal VS model are the strongest: overall performance increases by 6.45 BLEU, ZH by 7.19 BLEU, and ES by 13.95 BLEU with only 60 minutes of additional but highly varied data. The models with ZH as the base dialect also benefit from this added data, increasing performance for all dialects when comparing zh_1, the model only trained on ZH data, and zh_1+10, which was trained on the complete ZH data and 10 minutes of all other dialects. This strongly suggests that even adding little dialectal variability is crucial to improve performance. This is an essential finding for dataset collection. When primarily data for a distant dialect is available (VS in our example), it is crucial to gather data from as many other regions as possible, even if that is only a small amount. In this way, overall model performance can be improved with little data, and underrepresented dialects can benefit.

7 Conclusion and Future Work

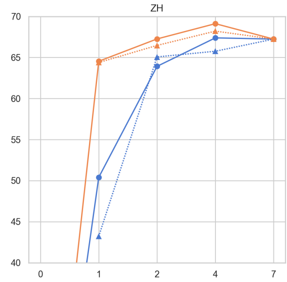
With respect to the research gaps identified in the introduction, the main findings of this paper are the following:

1. Standard German data is not beneficial for Swiss German ST performance when used in ASR pre-training or joint multilingual fine-tuning if a good amount ST data is available (> 500 h). Further pre-training XLS-R on noisy single-domain, single-dialect data does not improve performance.

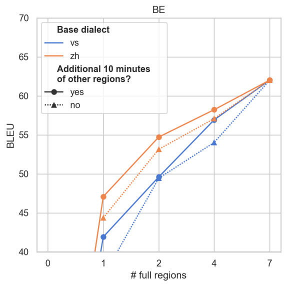
2. There are tendencies of a *Curse of Multidialectality* for Swiss German ST, especially when the dialects used for training are distant. Interestingly, Conneau et al. (2020) identified 7-15 languages as



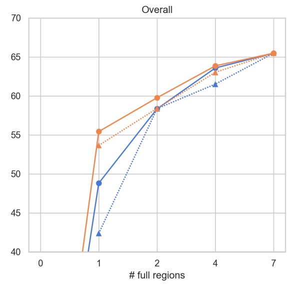
(a) Results for the VS region.



(b) Results for the ZH region.



(c) Results for the BE region.



(d) Results for all regions.

Figure 3: BLEU scores of the dialect transfer experiments with VS and ZH as base dialects. The models shown as dotted lines are from DT2, using 10 minutes of audio for all the dialect regions that were not included completely in the training set.

a breaking point. For ST, this number seems to be lower, and language similarity matters even more.

3. Using data containing rich dialectal variability is beneficial for the average performance of all dialects, even if the resulting training set is unbalanced and mainly contains distant dialects (VS in our case).

Future Work Imaizumi et al. (2022) introduced **dialect-aware modeling**, a promising and easy-to-implement approach that could help alleviate the *Curse of Multidialectality*. By performing dialect identification and ST simultaneously, the model might learn better to utilize dialect-specific acoustic/linguistic information for translation and more efficiently leverage cross-dialectal transfer. It is also worth investigating whether Standard German data proves beneficial for performance in this condition. A similar approach would be to introduce **dialect id tags** during training, as this has been shown to help with many-to-one translation performance in MT (Johnson et al., 2017; Fan et al., 2021). Furthermore, one could experiment with different approaches for **dataset balancing**, e.g., by considering the linguistic distances between the dialects as computed in Scherrer and Stoeckle (2016). Instead of employing ASR pre-training, an **existing ST model** (e.g., English →German) could be used to initialize the weights of the Swiss ST model. In contrast to an ASR model, an ST model has already learned non-monotonic mappings and vocabulary changes, which is crucial for Swiss German ST. Considering that there are no open-source ST systems for other German dialects, **benchmarking** our model on the performance of other, more **distant dialects** could be a fruitful experiment. This would be a step towards an ST system capable of translating all German dialects to Standard German, ultimately facilitating communication and cultural exchange between German-speaking countries immensely.

Limitations

Our work was constrained by computational resources, which prevented us from performing multiple training runs to draw statistically sound conclusions on whether performance differences between models were significant. Furthermore, we were unable to conduct the dialect transfer experiments for all dialect regions, which restricted the generalizability of our findings. As Swiss dialects vary significantly, dividing them into homogeneous regions remains a challenge. In our evaluations, we treat the dialect regions as homogeneous dialects even though they contain considerable variability. This might affect our results. Lastly, a thorough qualitative analysis of model outputs could have revealed region-specific error patterns and other limitations of our training and evaluation methods.

Acknowledgments

This work reports the findings of the first author's Master's thesis, which was carried out under a scholarship from the Erasmus Mundus European Master's Program in Language and Communication Technologies (EMLCT), EU grant no. 2019-1508. We want to express our gratitude to the HiTZ Center for Language Technology at the University of the Basque Country for allowing us to access their GPU clusters.

References

Ahmed Ali, Shammur Chowdhury, Mohamed Afify, Wassim El-Hajj, Hazem Hajj, Mourad Abbas, Amir Hussein, Nada Ghneim, Mohammad Abushariah, and Assal Alqudah. 2021. Connecting Arabs: Bridging the gap in dialectal speech recognition. *Communications of the ACM*, 64(4):124–129.

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.

Malajyan Arthur, Victoria Khurshudyan, Karen Avetisyan, Hossep Dolatian, and Damien Nouvel. 2024. Bi-dialectal ASR of Armenian from naturalistic and read speech. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 227–236, Torino, Italia. ELRA and ICCL.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Doris Baum, Daniel Schneider, Jochen Schwenninger, Barbara Samlowski, Thomas Winkler, and Joachim Köhler. 2010. DiSCo – a German evaluation corpus for challenging problems in the broadcast domain. *LREC 2010*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Verena Blaschke, Christoph Purschke, Hinrich Schütze, and Barbara Plank. 2024. What do dialect speakers want? a survey of attitudes towards language technology for German dialects. *Preprint*, arXiv:2402.11968.

Matthias Büchi, Malgorzata Anna Ulasik, Manuela Hürlimann, Fernando Benites de Azevedo e Souza, Pius von Däniken, and Mark Cieliebak. 2020. Zhaw-init at GermEval 2020 task 4: Low-resource speech-to-text. In *5th SwissText & 16th KONVENS Joint Conference, Zurich (online), 24-25 June 2020*. CEUR Workshop Proceedings.

Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards one model to rule all: Multilingual strategy for dialectal code-switching Arabic ASR. *Preprint*, arXiv:2105.14779.

Helen Christen. 2010. Vertikale und horizontale Variation: Beobachtungen zum Schweizerdeutschen. In *Variatio delectat. Empirische Evidenzen und theoretische Passungen sprachlicher Variation: für Klaus J. Mattheier zum 65. Geburtstag*, pages 145–159. Peter Lang.

Helen Christen. 2019. Alemannisch in der Schweiz. In *Deutsch*, volume 30/4, pages 246–279. De Gruyter, Inc, Germany.

Helen Christen, Andrea Ender, and Roland Kehrein. 2020. Sprachliche Variation in Deutschland, österreich, der Schweiz und Luxemburg. *Dialekt und Logopädie. Zürich/New York: Georg Olms Verlag*, pages 83–135.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 8440. Association for Computational Linguistics.

Zhengjia Dan, Yue Zhao, Xiaojun Bi, Licheng Wu, and Qiang Ji. 2022. Multi-task transformer with adaptive cross-entropy loss for multi-dialect speech recognition. *Entropy*, 24(10).

Amit Das, Kshitiz Kumar, and Jian Wu. 2021. Multi-dialect speech recognition in English using attention on ensemble of experts. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6244–6248.

Pan Deng, Shihao Chen, Weitai Zhang, Jie Zhang, and Lirong Dai. 2023. The USTC's dialect speech translation system for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 102–112, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timin Ding, Kai Sun, Xu Zhang, Jian Yu, and Degen Huang. 2024. Chinese multi-dialect speech recognition based on instruction tuning. In *Fourth Symposium on Pattern Recognition and Applications (SPRA 2023)*, volume 13162, page 131620A. International Society for Optics and Photonics, SPIE.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel multidialectal corpus of spoken Swiss German. *arXiv preprint arXiv:2103.11401*.

Andrea Ender and Irmtraud Kaiser. 2009. Zum Stellenwert von Dialekt und Standard im österreichischen und Schweizer Alltag – Ergebnisse einer Umfrage. *Zeitschrift für germanistische Linguistik*, 37(2):266–295.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Charles A Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

Philip N Garner, David Imseng, and Thomas Meyer. 2014. Automatic speech recognition and translation of a Swiss German dialect: Walliserdeutsch. In *Proceedings of Interspeech*, pages 2118–2122.

Andrea Gulli, Francesco Costantini, Diego Sidraschi, and Emanuela Li Destri. 2024. Fine-tuning a pre-trained Wav2Vec2 model for automatic speech recognition- experiments with de Zahrar Sproche. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7336–7342, Torino, Italia. ELRA and ICCL.

Robert Herms, Laura Seelig, Stefanie Münch, and Maximilian Eibl. 2016. A corpus of read and spontaneous Upper Saxon German speech for ASR evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4648–4651, Portorož, Slovenia. European Language Resources Association (ELRA).

Aaricia Herygers, Vass Verkhodanova, Matt Coler, Odette Scharenborg, and Munir Georges. 2023. Bias in Flemish automatic speech recognition. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pages 158–165.

Ranzo Huang and Brian Mak. 2023. wav2vec 2.0 ASR for Cantonese-Speaking Older Adults in a Clinical Setting. In *Proc. INTERSPEECH 2023*, pages 4958–4962.

Ryo Imaizumi, Ryo Masumura, Sayaka Shiota, and Hitoshi Kiya. 2020. Dialect-aware modeling for end-to-end Japanese dialect speech recognition. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 297–301.

Ryo Imaizumi, Ryo Masumura, Sayaka Shiota, Hitoshi Kiya, et al. 2022. End-to-end Japanese multi-dialect speech recognition and dialect identification with multi-task learning. *APSIPA Transactions on Signal and Information Processing*, 11(1).

Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 128–139, Dublin, Ireland. European Association for Machine Translation.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *Preprint*, arXiv:2401.05632.

Santosh Kesiraju, Marek Sarvaš, Tomáš Pavlíček, Cécile Macaire, and Alejandro Ciuba. 2023. Strategies for Improving Low Resource Speech to Text Translation Relying on Pre-trained ASR Models. In *Proc. INTERSPEECH 2023*, pages 2148–2152.

Abbas Khosravani, Philip N Garner, and Alexandros Lazaridis. 2021a. Learning to translate low-resourced Swiss German dialectal speech into Standard German text. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 817–823. IEEE.

Abbas Khosravani, Philip N Garner, and Alexandros Lazaridis. 2021b. Modeling dialectal variation for Swiss German automatic speech recognition. In *Interspeech*, pages 2896–2900.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Gongxu Luo, Yating Yang, Yang Yuan, Zhanheng Chen, and Aizimaiti Ainiwaer. 2019. Hierarchical transfer learning architecture for low-resource neural machine translation. *IEEE Access*, 7:154157–154166.

Jian Luo, Jianzong Wang, Ning Cheng, Edward Xiao, Jing Xiao, Georg Kucsko, Patrick O'Neill, Jagadeesh Balam, Slyne Deng, Adriana Flores, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, and Jason Li. 2021. Cross-language transfer learning and domain adaptation for end-to-end automatic speech recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Eduardo Medeiros, Leonel Corado, Luís Rato, Paulo Quaresma, and Pedro Salgueiro. 2023. Domain adaptation speech-to-text for low-resource European Portuguese using deep learning. *Future Internet*, 15(5):159.

Abir Messaoudi, Hatem Haddad, Chayma Fourati, Moez BenHaj Hmida, Aymen Ben Elhaj Mabrouk, and Mohamed Graiet. 2021. Tunisian dialectal end-to-end speech recognition based on DeepSpeech. *Procedia Computer Science*, 189:183–190. AI in Computational Linguistics.

Jonathan David Mutal, Pierrette Bouillon, Johanna Gerlach, and Marianne Starlander. 2023. Improving Standard German captioning of spoken Swiss German: Evaluating multilingual pre-trained models. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 65–76, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Seham Nasr, Rehab Duwairi, and Muhannad Quwaider. 2023. End-to-end speech recognition for Arabic dialects. *Arabian Journal for Science and Engineering*, pages 1–17.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. Asr for non-standardised languages with dialectal variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adapting multilingual speech representation model for a new, underresourced language through multilingual finetuning and continued pretraining. *Information Processing & Management*, 60(2):103148.

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.

Georgios Paraskevopoulos, Theodoros Kouzelis, Georgios Rouvalis, Athanasios Katsamanis, Vassilis Katsouros, and Alexandros Potamianos. 2024. Sample-efficient unsupervised domain adaptation of speech recognition systems: A case study for Modern Greek. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:286–299.

Phoebe Parsons, Knut Kvale, Torbjørn Svendsen, and Giampiero Salvi. 2023. A character-based analysis of impacts of dialects on end-to-end Norwegian ASR. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 467–476, Tórshavn, Faroe Islands. University of Tartu Library.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, et al. 2023a. STT4SG-350: A speech corpus for all Swiss German dialect regions. *arXiv preprint arXiv:2305.18855*.

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, Christian Scheller, and Manfred Vogel. 2020. Swiss parliaments corpus, an automatically aligned Swiss German speech to Standard German text corpus. *arXiv preprint arXiv:2010.02810*.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2021. Swisstext 2021 task 3: Swiss German speech to Standard German text. In *Proceedings of the Swiss Text Analytics Conference*, volume 2021.

Michel Plüss, Yanick Schraner, Christian Scheller, and Manfred Vogel. 2023b. 2nd Swiss German speech to Standard German text shared task at SwissText 2022. *arXiv preprint arXiv:2301.06790*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

Christian Rauh and Jan Schwalbach. 2020. The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. *Harvard Dataverse*.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. 2023. AudioPaLM: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

Charles Russ. 1990. *The dialects of modern German: A linguistic survey*. Routledge.

Tanja Samardzic, Yves Scherrer, and Elvira Glaser. 2016. Archimob-a corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4061–4066.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. The Edinburgh international accents of English corpus: Towards the democratization of English ASR. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. Digitising Swiss German: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53(4):735–769.

Yves Scherrer and Philipp Stoeckle. 2016. A quantitative approach to Swiss German – dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*, 24(1):92–125.

Katja Schlatter. 2024. «Schweizerdeutsch redet man paar Sachen komisch.» DaZ lernen und unterrichten in der diglossischen Deutschschweiz. In Stefan Hauser and Alexandra Schiesser, editors, *Standarddeutsch und Dialekt in der Schule*, pages 23–46. hep, Bern.

Barbara Schuppler, Martin Hagmueller, Juan A. 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'14)*, pages 1465–1470, Reykjavik, Iceland. European Language Resources Association (ELRA).

Clément Sicard, Kajetan Pyszkowski, and Victor Gillioz. 2023. Spaiche: Extending state-of-the-art ASR models to Swiss German dialects. *arXiv preprint arXiv:2304.11075*.

Mihaela C. Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing ASR pretraining for low-resource speech-to-text translation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7909–7913.

Artit Suwanbandit, Burin Naowarat, Orathai Sangpetch, and Ekapol Chuangsuwanich. 2023. Thai dialect corpus and transfer-based curriculum learning investigation for dialect automatic speech recognition. In *INTERSPEECH 2023*, pages 4069–4073.

Marc Tanti, Lonneke van der Plas, Claudia Borg, and Albert Gatt. 2021. On the language-specificity of multilingual BERT and the impact of fine-tuning. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 214–227, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021b. CoVoST 2 and massively multilingual speech translation. *Interspeech 2021*.

Saskia Wepner. 2021. Adaptation of automatic speech recognition systems to the needs of Austrian German. In *Phonetikworkshop 46. Österreichische Linguistiktagung 2020*.

Johannes Wirth and Rene Peinl. 2022. Automatic speech recognition in German: A detailed error analysis. In *2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–8.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Aditya Yadavalli, Mirishkar Sai Ganesh, and Anil Kumar Vuppala. 2022. Multi-task end-to-end model for Telugu dialect and speech recognition. In *Interspeech*, pages 1387–1391.

Jinyi Yang, Amir Hussein, Matthew Wiesner, and Sanjeev Khudanpur. 2022. JHU IWSLT 2022 dialect speech translation system description. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 319–326, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Yue Zhao, Jianjian Yue, Wei Song, Xiaona Xu, Xiali Li, Licheng Wu, and Qiang Ji. 2019. Tibetan multi-dialect speech and dialect identity recognition. *Computers, Materials & Continua*, 60(3).

A Language Model for decoding

We enhance XLS-R decoding by using LM fusion. We trained several language models of different sizes using the kenlm toolkit[11] and determined the best-performing model by evaluating the performance of our baseline model on the Swiss German test sets.

The best-performing LM is a 5-gram language model trained on 100M Standard German sentences compiled by concatenating EuroParl (Koehn, 2005)[12], NewsCrawl (Kocmi et al., 2022)[13], Tudatext[14], Parlspeech Bundestag + Nationalrat (Rauh and Schwalbach, 2020)[15] and the transcriptions of the STT, SPC, SDS, SDial train splits.

We fine-tuned the hyperparameters used for LM fusion by observing the performance of our Baseline model on the Swiss German test sets and ended up with `lm-weight=0.9`, `sil-weight=-1`, `word-score=1`, `nbest=1`. This configuration was used to obtain all our results.

B Training Hyperparameters

Table 8 lists the hyperparameters used for all experiments. These are mostly adapted from (Plüss et al., 2023a), who base theirs on (Babu et al., 2021).

Early stopping (or a maximum number of update steps) was set in every experiment to avoid overfitting and wasting resources. Learning rates were scheduled with Fairseq's tri-state scheduler, which warms up linearly for the first 6.25% of total steps, then keeps the learning rate constant for 25% of the total steps, and decays it exponentially afterward.

For the fine-tuning and pre-training experiments, XLS-R 1B was used. For the second fine-tuning step in Experiment 4, we had to adjust the learning rate to 1e-6 because the model had already seen the Swiss German data and did not converge with a higher learning rate.

For continued pre-training, we use the same configurations as (Babu et al., 2021) with modifications inspired by (Bartelds et al., 2023). As pre-training is computationally expensive and we train on one GPU (instead of 200 as (Babu et al., 2021)), we lower the batch size and apply gradient accumulation. All hyperparameters are listed in Table 8. If any parameters are not given, they were kept the same as in the pre-training config of XLS-R (Babu et al., 2021).

Unlike the fine-tuning experiments, the 300M version of XLS-R (Babu et al., 2021) was used for the dialect transfer experiments. The main reason for this is that we train 14 models for our dialect transfer experiments, and this would consume too many computational resources[16]. Additionally, Paonessa et al. (2023) showed that the results of XLS-R 300M are transferable to XLS-R 1B because both models have the same performance curve with a gap of around 5 BLEU per dialect region. All model trainings are conducted using the hyperparameters from Table 8 (column **All others**). However, for the first set of dialect transfer experiments, we use the STT validation set only containing the base dialect to track the model performance during training.

[11] `https://kheafield.com/code/kenlm/`

[12] `https://www.statmt.org/europarl/v7/`

[13] `http://data.statmt.org/news-commentary/v14/`

[14] `http://ltdata1.informatik.uni-hamburg.de/kaldi_tuda_de/`

[15] `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/L4OAKN`

[16] For instance, training the 300M version for 80k steps on the STT balanced train set took 28 hours in Paonessa et al. (2023). However, using XLS-R 1B with the same setup took 48 hours

	Ex 1	Ex 4.2	Ex 5.1	All others
learning rate	3e-5	1e-6	5e-5	3e-5
gradient accumulation	10	10	10	10
batch size (samples)	640k	640k	320k	640k
effective batch size	400 sec	400 sec	200 sec	400 sec
validation set	STT	SwG-all	GRZH	SwG-all*
validation interval	1000	1000	400	1000
early stopping patience	-	5	3	5
max. updates	80k	80k	80k	250k

Table 8: Hyperparameters for the fine-tuning, pre-training and dialect transfer experiments. The experiment numbers refer to Table 3. *SwG-all* refers to the combined STT, SDS, and SPC validation sets.
*For Experiment 3.1, the CV validation set was used.

C Performance comparison to SoTA models

Figure 4 shows the results of our models from the fine-tuning experiments compared to SoTA models for Swiss German ST. We hypothesize that the performance difference between our baseline and the baseline published in Plüss et al. (2023a) has two main reasons: (1) we trained on one GPU only, resulting in a different batch size and overall training time, (2) we used less data for training the language model and a potentially different ngram order.

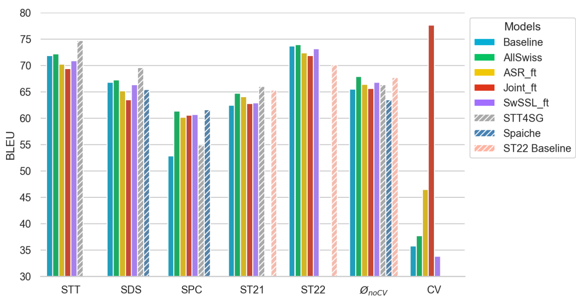


Figure 4: Results of fine-tuning Experiments 1-5, grouped by test set. STT4SG, Spaiche, and ST22 Baseline are the models published in (Plüss et al., 2023a), (Sicard et al., 2023), and (Plüss et al., 2023b) respectively. For these models, we only used the available performance metrics to compute the average ($\varnothing_{noCV}$).