Universität Zürich[UZH]

Bachelorarbeit

zur Erlangung des akademischen Grades

Bachelor of Arts

der Philosophischen Fakultät der Universität Zürich

Impact of Lingusitic Versus Data Properties for Swiss German ASR

Verfasserin: Iris Zoder

Matrikel-Nr: 20-722-302

Referentin: Prof. Dr. T. Samardžić

Institut für Computerlinguistik

Abgabedatum: 01.06.2023

Abstract

The topic of this thesis is automatic speech recognition (ASR) for Swiss German, an umbrella term for German dialects that are spoken in Switzerland. Although ASR models covering multiple of these dialects have been successfully built during the last few years, the performance of most systems is not stable over all of them. This fact was previously ascribed to either linguistic differences between the dialects or inequalities in the (evaluation) data but a conclusive answer has not yet been reached.

Therefore, this thesis analysed both these factors by using the distance of the dialects to a standardisation in terms of a dialectality score and measured data sparseness in type token and character ratio. From these analyses it was found that both the dialectality score and the sparseness of the evaluation data correlated with a drop in performance and therefore a mixture of both linguistic and data factors could be the cause for performance differences.

Additionally, it was analysed whether the system agreed on the difficulty of the dialects given by a performance ranking. This analysis revealed that when the distance between dialectality scores becomes smaller, the disagreement between the systems rises which highlights the need for additional research. This could possibly include calculating a more precise dialectality score to accurately capture smaller differences between the dialects or analyse further factors such as training data dialectality.

Some of these other measures were calculated for the papers covered in this thesis and are made available in the corresponding GitHub repository under https://github.com/siri-web/BA or in the appendix to constitute a starting point for future research.

Zusammenfassung

Diese Arbeit befasst sich mit der automatischen Spracherkennung (Automatic Speech Recognition [ASR]) für Schweizerdeutsch, einem Oberbegriff für die deutschen Dialekte, welche in der Schweiz gesprochen werden. In den letzten Jahren wurden erfolgreich ASR-Systeme für Schweizerdeutsch entwickelt, deren Input aus mehreren Dialekten bestehen kann, wobei jedoch die Qualität nicht auf allen davon gleich ist. Eine Vielzahl von Gründen hierfür wurde in der Forschung diskutiert, wie beispielsweise die sprachlichen Unterschiede zwischen den Dialekten oder ungleiche Testdaten.

Daher wurden in dieser Arbeit die sprachlichen Unterschiede zwischen den Dialekten, die in einem Dialektalitätswert gemessen wurden, mit Messungen der Verstreutheit der Daten wie Typ-Token- und Zeichenverhältnis der Auswertungsdaten verglichen. Dabei wurde festgestellt, dass sowohl der Dialektatilitätswert als auch die Verstreutheit der Testdaten eine Korrelation zu einem schlechteren Ergebnis aufwiesen und daher eine Mischung aus beidem ein Grund für diese Unterscheide sein könnte.

Zusätzlich wurde analysiert ob die Systeme sich bei der Schwierigkeit der Dialekte - gemessen in einer Rangfolge der Ergebnisse - einig waren. Diese Untersuchung ergab, dass bei kleineren Unterschieden im Dialektalitätswert die Übereinstimmung zwischen den Systemen abnimmt, was die Notwenidgkeit weiterer Forschung verdeutlicht. Diese könnte die Berechnung eines genaueren Dialektalitätswert beihhalten, um kelinere Unterschiede zwischen den Dialekten besser erfassen zu können oder zusätzliche Faktoren wie etwa den Dialektatlitätswert der Trainingsdaten analysieren.

Manche dieser anderen Faktoren wurden bereits in dieser Arbeit für die vorliegenden Studien berechnet und wurden auf GitHub unter https://github.com/siri-web/BA oder im Anhang zur Verfügung gestellt damit sie als ein Ansatz für zukünftige Forschung dienen können.

Acknowledgement

First and foremost I want to thank my assisting professor Mrs. Tanja Samardžić for giving me the opportunity to write a thesis about this topic. Without your continuous support and guiding help this would not have been possible.

Further, this work would not exist without access to the two main corpora it builds upon and therefore, my gratitude for the creation and access of the SwissDial data set goes to Dogan-Schönberger et al. and for the ArchiMob corpus I thank Scherrer et al..

Moreover, this thesis would not exist without the researchers who trained the analysed models and made their results publicly available and so I thank Khosravani et al. [2021a], Nigmatulina [2020], Scherrer et al. [2019], Schraner et al. [2022], and Sicard et al. [2023].

I also want to acknowledge my family and friends for their support. I am great-full to have you in my life and that I can always rely on your support.

Last but not least many thanks to my parents for proofreading the current text.

Contents

List of Figures

List of Tables

List of Acronyms

ASR	Automatic Speech Recognition
FHNW	Type of Transformer Network
NNET-DISC	Discriminative Hybrid Deep Neural Network
SMS	Short Message Service
STR	Speech to Text Recognition
STT	Speech To Text
TDNN	Time-Delay Neural Networks
TTS	Text To Speech
WER	Word Error Rate
XLS-R	Large-Scale Model for Cross-Lingual Speech Representation

1 Introduction

1.1 Motivation

The technology playing a main role in this thesis is called automatic speech to text (STT) also referred to as speech to text recognition (STR) or automatic speech recognition (ASR). Under all these names, it deals with receiving auditory input and automatically converting it into text and is commonly used for automatic transcriptions of audio data.

These transcriptions can be helpful in an educational context where they are used to help students understand a lecture, take notes and do their homework (Hwang et al. [2012]; Kuo et al. [2012]; Shadiev et al. [2013]). This is especially important for students with learning or physical disabilities, foreign students, and other at risk populations (Shadiev et al. [2013]; Wald and Bain [2008]).

For instance, the Speech Recognition in Schools Project (Nisbet and Wilson [2002b,a]; Nisbet et al. [2005]), concluded that it is possible to help students overcome difficulties in reading, writing, and spelling with the help of STR. These results were then confirmed in Wald and Bain [2008] where a better involvement of deaf students and non-native speakers in lectures was reached when using STR applications. Further, the students themselves perceived that STR-generated text was beneficial for their learning so long as its accuracy was fairly good (Colwell et al. [2005]; Wald and Bain [2008]).

Another context where the conversion from auditory input data to text plays an important role is a technological one. For instance, voice messages can be converted to written text as a feature of SMS providers by using a dictating function (Trivedi et al. [2018]). Moreover, it is used internally in many applications such as for instance in Apple's Siri which is an intelligent automated assistant meant to facilitate user interaction with a device (Shadiev et al. [2013]). A further case where ASR is used in the internal workflow of an application are especially systems designed for people with a disability.

These benefits are possible as systems trained on standardised high resource languages are able to perform with low word error rates (WERs) of less than 5 %. However, the performance on non-standardised varieties is often much lower (Nigmatulina [2020]) which can negatively impact the usefulness of the applications. Still, such systems are needed for automatic transcriptions in situations where people are less likely to want or be able to speak in a standard variety as when

analysing dialogues between doctors and patients to gain medical insights. Another example where ASR is needed for non-standardised varieties is for automatic subtitling of media that is only available in dialect, which includes regional news formats or media that uses dialect for artistic purposes.

These aspects are especially important for the collection of dialects spoken in the German part of Switzerland: The so-called Swiss German is held in a higher regard than most other dialects compared to their standard varieties and used frequently in a multitude of situations. It is even dominant over Standard German outside certain exceptions such as an educational context. Therefore, many children grow up with Swiss German as their native variety and are only able to communicate in Standard German after they learned it in school. Further, over time people become less accustomed to the standard variety once they are no longer using it for their education. Similarly, adult immigrants are often able to learn Swiss German faster compared to Standard German as they are more immersed in it. This leads to the problem that depending on factors such as age, job, and education having to deal with Standard German can be a substantial hindrance for certain groups.

However, building a system that is able to produce text from speech in non-standard varieties is difficult. Outside the lack of a standardised orthography, the application further needs to be able to handle multiple dialects which vary substantially from each other (Nigmatulina [2020]). How and to what extent this regional variation influences the performance is not entirely clear: While fluctuations between scores on different dialects have been reported (Nigmatulina [2020]; Scherrer et al. [2019]), the research has not yet been able to reach a consensus on the reasons for these differences.

One possible explanation would be the extent to which certain dialects are represented in the training data. However, this cannot be the only reason as for example Scherrer et al. [2019] observed higher error rates on dialects that had more training examples and vice versa. This has lead to a counter hypothesis focusing less on the aspects of the data composition and more on linguistic factors such as the distance of a dialedt to an underlying standard (Schraner et al. [2022]). This shift in focus further implies that different studies should agree on the difficulty of the dialects as the reason is linked ot the variety itself and not an aspect of the project. However, comparisons between papers are difficult as the data sets are considerably different from each other in a variety of factors. Therefore, it should first be established if the evaluation data itself does not currently have a higher impact on the performance of the different dialects than linguistic factors.

1.2 Research Question

This study will further examine the phenomenon outlined above that the performance of ASR systems varies between dialects and studies. Whether this fact is due to some dialects being

harder for applications to learn because of their linguistic properties is not trivial to answer as there are considerable differences in the evaluation data both between studies and linguistic varieties. It could be that variations in the data are more impact-full on the models' performance than linguistic dissimilarities to an underlying standard. Therefore, the thesis will answer the following question:

1. Do properties of the evaluation data have a greater impact on performance results of individual Swiss German dialects than the linguistic differences between them?

1.3 Thesis Structure

In this first chapter, Swiss German STT has been introduced and a motivation for continuing research in this field has been given. Further, it described the relevant phenomenon for which Chapter 2 examines the positions of previous literature and those were further analysed using the data and methods explained in Chapter 3. The findings that have arisen from the described procedures are presented in Chapter 4 and their discussion can be found in Chapter 5. Lastly, Chapter 6 presents a conclusion to this study.

2 Performance Differences

The last chapter introduced the fact that the performance of a model can vary on different Swiss German dialects which is further supported by Nigmatulina [2020], Khosravani et al. [2021a], Khosravani et al. [2021b], Scherrer et al. [2019] and Schraner et al. [2022]. While all of them report such findings, they do not agree as to their causes and mainly propose two different hypothesis:

The first one denies that varying results originate solely from linguistic differences between the dialects. This theory is claimed by researchers like Scherrer et al. and states that the results vary too much to be entirely attributed to known regional variation. However, Scherrer et al. themselves do not propose more probable causes for this fact nor did they include a more detailed explanation of the idea. This was instead done in Nigmatulina [2020] where - after a more in-depth analysis of the systems from Scherrer et al. [2019] -, it was concluded that data properties had an influence in this case. Specifically, Nigmatulina proposed high inter-annotator difference and variability in the training data due to the chosen transcription.

While sharing this opinion that the reason for performance differences can be found in the data, Khosravani et al. disagree on the relevant properties: In their study, annotator agreement and variability played less of a role than data set composition which showed imbalances in amount per dialect. The largest of which was between their Eastern dialect group - for which they had the most amount of data - and Grisons and Valais where the latter were evaluated separately and trained on smaller data sets. As they had more data for the group of Eastern dialects, they could cover more of the variation shown in them and therefore a better performance was achieved on those compared to the dialects from Grisons and Valais.

A direct contradiction to this importance of data amount is stated in Nigmatulina [2020], Khosravani et al. [2021a], and Schraner et al. [2022]. In the last, there was about ten to 45 times more training data for the Bern and Zurich dialect region then for Central Swiss dialects and still the best performances were reached on the latter group.

Similarly, in Nigmatulina [2020] better performances were achieved on Grisons dialect which was only represented by a single interview compared to Zurich dialect which contributed 14 interviews to the training data. Further, the systems performed well on the dialects of Basel-Landschaft and Schaffhausen which also only had one interview each.

These results gave rise to the second hypothesis currently discussed in research: The lower

performance does not originate from the data but from the dialects themselves, their linguistic characteristics, and differences between them. According to this thought, some dialects posses certain properties that make them harder for the models to learn as for example proposed in Nigmatulina [2020]. In this study, it was hypothesised that certain dialects vary more from others and therefore profit less from training on other varieties which makes them more difficult to learn. While this could not be analysed in depth in that thesis, certain tendencies were observable as for instance Valais dialect showed a worse performance relative to all others and Grisons with much less training data consistently performed well.

Following this thought from Nigmatulina [2020], the dialects with the lowest performance are the hardest because they exhibit certain linguistic traits more strongly. This links linguistic properties to performance, meaning the latter can be improved by addressing the former and dialects with the lowest performances should improve the most since they exhibit these traits more strongly.

An example of addressing such linguistic factors is Khosravani et al. [2021a] as they included a lexicon in their approach with the goal of reducing the influence of lexical variability on model performance. If this variability were greater in some dialects than in others and contributed to their lower performance, these dialects should now benefit more from the lexicon. As Valais dialect was identified as the variation with the most improvement, it should belong to the category of hard dialects and similarly, Grisons should be part of the group of easier dialects as its results increased the least.

These groups of harder and easier dialects should be the same across studies as it depends on the linguistic properties of the dialects and explicitly not on the data. However, this is not the case as can be seen by looking at the results of Grisons: It showed the smallest improvement in Khosravani et al. [2021a] and is the dialect with the highest performance in Nigmatulina [2020] - although being underrepresented in the training data. Therefore, it should belong to the category of easier dialects but contradictorily the worst performance in Khosravani et al. [2021a] was on Grisons, which would make it a harder dialect.

This disagreement on the difficulty of the dialects causes doubt on the view that the performance differences are solely due to linguistic factors. On the other hand, data properties are also not enough to explain the phenomenon in its entirety and so a further analysis of the impact of linguistic versus data properties on the performance differences between Swiss German dialects needs to be conducted.

3 Data and Methods

3.1 Included Papers and Data Sets

As the previous chapter showed, the reasons behind the performance fluctuations on different dialects are not yet certain. Therefore, the goal of this thesis is to further investigate this phenomenon through analysing the results found in previous studies and thereby functioning as a meta-study.

As such it builds on previous research and previously evaluated systems that were found through a Google Scholar search for the terms "Swiss German STT" and "Swiss German ASR". However, only studies showing a per dialect evaluation of their systems where included and additional papers were provided by my supervising professor. Therefore, the studies analysed in this thesis are Khosravani et al. [2021a], Nigmatulina [2020], Scherrer et al. [2019], Schraner et al. [2022], as well as Sicard et al. [2023] and the following paragraphs will introduce each of them in turn.

In Sicard et al. [2023], the goal was threefold: First, they proposed a new semantically informed metric to measure the performance of ASR systems, then used this metric to provide insight into state-of-the-art models on recently published datasets, and lastly further advanced on these results by fine-tuning OpenAI's whisper model. This model is especially interesting in the context of this thesis as it is the only part of the paper that included a step with a per dialect evaluation (on the varieties from the regions Aargau, Bern, Basel, Grisons, Lucerne, St. Gallen, Vallais, and Zurich). Unfortunately, this was not the final result but only a Zero-shot evaluation (i.e., without previous explicit training on Swiss German) that Sicard et al. did for computational reasons: The whisper model comes in five different sizes (large, medium, small, base, and tiny) and as it would be wasteful to train all of them, Sicard et al. decided to only use the most promising version for their further experiments.

A similar goal as in the previous paper was pursued in Schraner et al. [2022] as they wanted to test already existing commercial systems (anonymously named a to d) and compare their performance on datasets from different domains. In addition to these models, they trained two systems themselves - named FHNW XLS-R and FHNW Transformer - where the former was based on the XLS-R 1B model from Babu et al. [2021] which is pre-trained on 436,000 hours of unlabelled speech data from over 128 languages - explicitly excluding Swiss German. Therefore, Schraner et al. fine-tuned the model on a mixture of different popular Swiss German

ASR datasets such as SDS-200, the SwissDial dataset and the Swiss Parliament Corpus (SPC). Further, they used four-gram language models obtained from different corpora such as Europarl (Koehn [2005]) and news-crawl 2019 (Barrault et al. [2019]) during the decoding phase and then compared this model to an approach that was solely trained on supervised data (FHNW Transformer) which is more task specific but available in much smaller quantities. Then both models were evaluated on the STT4SG-350 test set and all the findings reported on the relevant dialects (Basel, Bern, Grisons, Central Swiss, East Swiss, Valais, Zurich) individually and therefore all of them could be analysed in this thesis.

Similarly, in Khosravani et al. [2021a] the aim was as well to improve on results from previous research. However, in this case the focus was on the challenge that Swiss German consists of non-standardised varieties and therefore lacks a defined orthography that is needed to establish lexical identity. This leads to the fact even experts transcribe spoken Swiss German differently as they orientate themselves mostly on the pronunciation which can very substantially between the different dialects. However, this is especially important in Khosravani et al. [2021a] as the goal was to improve a downstream task - namely the performance of the voice assistant for the Swisscom TV box which uses ASR internally and then needs to perform actions based on the generated transcriptions.

To establish this lexical identity, Khosravani et al. performed normalisation based on a lexicon that mapped different written variants of Swiss German words to a Standard German form and - where an appropriate word did not exist in the standard variety - to a Swiss German variant that is understood by the majority of dialect speakers. This lexicon was obtained by using the already existing Swisscom Dictionary (Schmidt et al. [2020a]) and generating additional written forms either manually through linguists that were native speakers of the respective dialect or automatically by a model trained on the pre-existing and manual candidates.

To evaluate the usefulness of this technique, Khosravani et al. tested four variants of their system: One using no lexicon, one using only the pre-existing dictionary (referred to as baseline), one using both the dictionary and the manually generated candidates for the normalisation, and lastly one using the dictionary and the automatically generated candidates. As the findings for all four systems were evaluated on the dialects from Bern, Central Swiss, Eastern Swiss, Grisons, Nidwalden, Valais, and Zurich individually, they could be included for the analyses in the current thesis.

In Scherrer et al. [2019] the aim was completely different as the main focus was the release of the ArchiMob corpus which is a collection of historical interviews consisting of personal testimonies about life in Switzerland between the years 1939 and 1945. While these interviews existed beforehand, Scherrer et al. organised manual transcriptions for the data and used different natural language processing tools to provide additional annotation layers such as a standardisation layer to facilitate search between the different dialects.

Moreover, they conducted different analysis on the first release of their corpus which included

the calculation of a dialectality score and the training of a first ASR baseline. For this baseline model, only interviews from the larger Zurich area were used and the system was then evaluated on the dialects from Basel, Bern, Grisons, Lucerne, Uris, and Valais individually which made it suitable to be analysed in this thesis.

However, this was only a first baseline as the overall focus of the paper was on the release of the corpus and so Nigmatulina [2020] aimed at analysing and improving these results using the second release of the ArchiMob corpus. This release had additional interviews available which were used to build different versions of models trained not only on data from the largr Zurich area but on all available dialects - namely Aargau, Basel-Landschaft, Basel-Stadt, Bern, Glarus, Grisons, Lucerne, Nidwalden, Schaffhausen, Schwyz, St. Gallen, Uri, Valais, and Zurich.

Further, the focus of the thesis was to mainly improve the acoustic modelling of multi-dialect systems in a limited data scenario. To this end, the systems were mainly compared over all dialects until two of the trained models were used for a per-dialect evaluation at the end of the thesis. These two models were both neural network based but used slightly different architectures to segment the audio signal: The first one required the separation into phonemes (referred to as discriminative and so the system was named NNET-DISC) and the other received the input time-delayed but required no previous splitting of the audio signal and additionally used speaker and dialect specific features (i-Vectors and therefore called TDNN-iVector). As only for these two models the results were reported for the relevant dialects individually, only those could be taken into account for the analyses of this current thesis.

Overall five papers, a total of 18 models were therefore included: Four from Khosravani et al. [2021a] (XLSR Lexicon-Free, XLSR Baseline, XLSR manual normalisation, and XLSR automatic normalisation), two from Nigmatulina [2020] (NNET-DISC and TDNN-iVector), one from Scherrer et al. [2019] (ArchiMob baseline), six from Schraner et al. [2022] (commercial systems numerated a to d, FHNW XLS-R, and FHNW Transformer) and five from Sicard et al. [2023] (whisper-large, whisper-medium, whisper-small, whisper-base, whisper-tiny).

However, these models were evaluated on different dialects as was mentioned when introducing the individual papers and can be seen in the overview table in the appendix. Moreover, the data these models were trained and evaluated upon originated from different data sets and could not be analysed for all models due to inaccessibility. For instance, the systems from Khosravani et al. [2021a] used a costume training set of voice commands for a Swisscom TV box and those from Schraner et al. [2022] were evaluated on the STT4SG-350 test set which is still in preparation as of the writing of this thesis. In these cases, the measures reported in the papers were used and if no information was available, it had to be replaced by a filler value for these models.

Still, two datasets were publicly available: Namely the ArchiMob corpus - which has been mentioned above - and the SwissDial dataset introduced in Dogan-Schönberger et al. [2021]. The latter is a parallel corpus of spoken Swiss German containing the dialects Aargau, Bern,

Basel, Grisons, Lucerne, St. Gallen, Valais, and Zurich as well as a Standard German reference. This reference is in fact the original sentence that was obtained by web-crawling which allowed the corpus to include multiple different genres and sources such as news stories, Wikipedia articles, weather reports, and short stories. The sentences obtained in this way were then manually translated into Swiss German, and recorded with one speaker per dialect.

3.2 Data Collection

System Name	Dialect	Performance	Performance Rank	Writing System	Corpus Name	Corpus Topic
ArchiMob baseline Scherrer et al.	Bern	29.83	5	Dieht	ArchiMob	historical interviews
ArchiMob baseline Scherrer et al.	Basel	45.4	2	Dieht	ArchiMob	historical interviews
ArchiMob baseline Scherrer et al.	Grisons	49.68	1	Dieht	ArchiMob	historical interviews
ArchiMob baseline Scherrer et al.	Lucerne	22.37	6	Dieht	ArchiMob	historical interviews
ArchiMob baseline Scherrer et al.	Uri	30.46	4	Dieht	ArchiMob	historical interviews
ArchiMob baseline Scherrer et al.	Valais	36.96	3	Dieht	ArchiMob	historical interviews
XLSR Lexicon-Free Khosravani et al.	Northwestern Swiss	17.2	3	in dialect	TV assistant	short commands, scripted or spontaneous
XLSR Lexicon-Free Khosravani et al.	Bern	18.2	5	in dialect	TV assistant	short commands, scripted or spontaneous
XLSR Lexicon-Free Khosravani et al.	Zurich	16.9	2	in dialect	TV assistant	short commands, scripted or spontaneous
XLSR Lexicon-Free Khosravani et al.	Grisons	20.8	7	in dialect	TV assistant	short commands, scripted or spontaneous
XLSR Lexicon-Free Khosravani et al.	Eastern Swiss	14.7	1	in dialect	TV assistant	short commands, scripted or spontaneous
XLSR Lexicon-Free Khosravani et al.	Valais	19.3	6	in dialect	TV assistant	short commands, scripted or spontaneous
XLSR Lexicon-Free Khosravani et al.	Central Swiss	17.4	4	in dialect	TV assistant	short commands, scripted or spontaneous
XLSR Baseline Khosravani et al.	Northwestern Swiss	15.8	3	in dialect	TV assistant	short commands, scripted or spontaneous

Figure 1: An exemplary excerpt from table A1 in the appendix which provides an overview of all systems and data sets included in this thesis.

While the last section described the included papers and models, this section presents the methods used to gain an overview of them. As a first step, the results and additional information about the data sets were summarised in a table for which an excerpt is shown in Figure 1. The full version can be found in the appendix of this thesis along with the code used to conduct the analyses described later in this chapter.[1]

This table is meant as an overview of what information is available about the systems and as a starting point for further analyses described in the following sections. To this end, a row constitutes the performance of a system on a dialect and the columns report information that

[1]Both are also available on GitHub under https://github.com/siri-web/BA.

could potentially influence this result which is rather wide so that the most promising could be chosen for further evaluation.

These factors can then be separated into two groups where the first reports general information about the system as well as its training corpus and consists of information reported directly in the literature. Among others, this includes factors such as the performance on different dialects, the training regime or the name and topic of the training corpus. The other group concerns itself with metrics that were calculated on the data sets for the purpose of this thesis such as measures of data sparseness or utterance lengths and both groups are described in detail in the following two subsections.

3.2.1 General System and Corpus Information

The first section of the table starts by giving the name of the system and its paper of origin which are repeated inside the bold borders. This was done so that the results can be given individually for each dialect the model was evaluated upon and therefore facilitate the comparison of different factors between the individual varieties.

After this column, the following two list the dialects and the performance on those varieties as reported in the literature. Because they were directly extracted from the papers, it must be noted that not all studies reported their results in the same detail - for instance in Nigmatulina [2020] and Sicard et al. [2023] the results were shown only in figures and these are less exact than when the findings were reported in numbers directly.

A further difference between the included studies was that even in cases where results were given as raw digits, not all papers used the same evaluation metric. For this reason, the performance is given in WER for all systems except for the ArchiMob baseline where it is reported using the F1 score. Therefore, an additional measure to make the results more comparable was needed and is shown in the column *Rank* where for each system, the performances on the different dialects were ranked in such a way that the first place was given to the dialect on which the best performance was achieved.

This column is followed by one named *Writing System* which summarises information on the output the model provides and consists of two options: The model could either produce writing in dialect (most often using Dieth transcription) or in Standard German. This is an important distinction to make as in Nigmatulina et al. [2020] it was reported that WER for systems using Standard German transcription was higher.

The next few columns form a subsection by themselves as they all contain information about the corpora used for training and evaluating a given system. In the first column of this section - called *Corpus Name* - the title of the corpora for training and evaluation are given. These corpora are further described in *Corpus Topic* where a brief overview over their included topics was given. This information was reported as certain measures described later in this chapter

vary considerably depending on the underlying corpus and its themes which bring some natural differences with them. For example, the ArchiMob corpus is build from interviews whereas the data used in Khosravani et al. [2021a] are voice commands aimed at digital agents. By virtue of their different topics and formats, one is a piece of free speech whereas the other consists of previously written commands that are read aloud. As read speech is produced from pieces of writing, it is naturally closer to written language than free speech and this could make it easier for ASR models to produce the transcription.

The subsection continues with *Automatic vs. Manual Transcription* which reports whether the papers stated that the transcription of their speech data was done manually by human annotators or automatically by using existing ASR systems. This could influence the system performance as automatically transcribed data is usually closer to what a machine would produce and might therefore be easier to replicate by a model.

The third last column in this subsection is called *Known Transcriber Differences* and reports whether the corpus is known to have considerable transcriber influences. It is followed by *Training Dialects* which states the dialects the system was trained on and was reported due to the fact that if the evaluated dialect was included in the training data it will have a higher score than if it was not. This information is made more explicit in the column *Overlap between Training and Test in Terms of Dialect*.

A new subsection concerns itself with information about the speakers that took part in the creation of the corpus. This information was either extracted from the metadata (in case of the ArchiMob corpus) or from the literature and filled with N/A where it was unavailable. It was reported as systems trained with multiple speakers are usually more robust and would therefore achieve higher scores.

Its first columns are *Number of Different Speakers (Training Overall)* and *Number of Different Speakers (Training Per Dialect)* that show the number of speakers in the training data as a whole and for each dialect specifically.

The same information was provided for the test data but this time only information about the relevant dialects was included and can be found in column *Different Speakers Test*. Additionally, it was extracted if test and training data overlapped in terms of speakers for a specific dialect which is stated in *Overlap Between Test and Training Speakers*. The reason behind including the last column was that if a dialect was tested on a speaker that was in the training data it could achieve a higher score than for an unknown speaker.

3.2.2 Data-Orientated Measures

This section of the table consists of measures for data related factors that could be the cause of the performance fluctuations on different dialects. These were extracted from the literature where possible and when no information was available, they were manually calculated. However,

this could not be done for all systems as for some of them the information was neither given in the papers nor could it be computed as the employed corpora where unavailable. This was the case for the costume dataset of Khosravani et al. [2021a] as well as the STT4SG-350 corpus which was used in Schraner et al. [2022].

In the latter, the models specifically build for this study were trained on a mix of two available corpora (the SDS-200 and the Swiss parliament corpus (SPC)) and the yet unpublished STT4SG-350. Therefore, calculating the measures for the first two corpora would have been possible but generate incomplete and possibly distorting results and for this reason, the values for the models trained in Schraner et al. [2022] were filled with N/A.

In the other cases, where the information was either available in the literature or could be successfully calculated, the table starts by reporting the training set size as it was among the most mentioned causes for performance differences between dialects discussed in chapter 2. This training set size is given in the columns *Training Set Size (Utterances Per Dialect)* and *Training Set Size (Words Per Dialect)* which show that it was measured in two different variants: Once by counting the number of utterances and once by counting the number of words.

While the notion of the latter is straightforward because it is the same as in written language (i.e., in the case of German white space separation in the transcription), the definition of the former is not as clear cut. Its underlying idea is that of an uninterrupted section of speech - usually congruent to the notion of sentences. This is due to the fact that speakers tend to produce entire sentences without longer breaks yet pause between them. However, utterances can also be much shorter when interjections, repetitions, or shorter pauses are counted as interruptions.

These discrepancies in the definitions were the reason for including multiple measure of the training set size. For instance, in a case where the utterance length is quite short, a system could be trained on a high amount of them while still having received less training data in terms of words than another model where the utterances were longer. Still, the size was given in terms of utterances as well for the sake of completeness and for the same reason, the test set size is reported in *Test Set Size (in Words)*.

However, both the size in terms of utterances and in terms of words could not be calculated the same way for both accessible corpora (ArchiMob and SwissDial) as their formats were quite different. The former was distributed in XML with the used tags including utterance and word segmentation so that the amount of utterances and words was the number of occurrences for the respective tag. The latter, was released in JSON format as a list of dictionaries where each dictionary constituted one utterance and contained a list of transcriptions in different varieties. Therefore, the number of utterances was the amount of dictionaries and the number tokens was the length of the list when the dialect transcription was separated at white space. For the remaining corpora, information was extracted from the literature or filled with N/A if not publicly available.

After the data set size, the table continues by reporting the mean length of utterances in words which was included to measure potentially varying complexity in syntax. The measure is given for each dialect individually (*Mean Utterance Length in Words (Training Per Dialect)/Mean Utterance Length in Words (Test)*) as well as the mean over all dialects (*Mean Utterance Length in Words (Training Overall)*). The latter was included only for the training set to analyse if distances between the training mean and the test data could have an effect on system performance. For both the overall and the dialect specific version on both train and test set, the calculation was the number of words over the amount of utterances, where both words and utterances were counted as explained in the previous paragraph.

The next section of the table consists of measures designed to capture data sparseness which aims at quantifying if the data consists of few elements that are repeated often or many elements that occur rarely. In the former case it is easier for the model to learn good representations as they can be refined with each of the many occurrences while in the latter this is not the case. Similarly, at test time it easier for the model to produce repetitions of frequent elements than to use rare or unseen tokens.

A first way to measure this is by using the absolute number of unique elements which is called the vocabulary size. To calculate this measurement, the amount of unique sequences between opening and closing tag was used for the ArchiMob corpus while for the SwissDial data it was the sequence of characters between white spaces. This number was again calculated for each dialect individually and for the training data overall due to the same reasons as above and can be found in the columns *Vocab Size (Training Overall)* and *Vocab Size (Training Per Dialect)/Vocab Size (Test).*

However, the vocab size does not take the length of a sequence into account and will be higher for longer and smaller for shorter texts. Therefore, two additional measures - the type token and character ratio - were used to stabilise the vocab size over different text lengths. In a first step, this meant that for the former the number of unique words (i.e., tokens) was calculated similarly to how it was done for computing the vocabulary size. For the latter, the number of unique symbols was used and then in both cases these unique amounts were divided by the total number of token types or characters respectively. The final result of this computation can be found in the columns *Token Ratio (Training Overall)*, *Token Ratio (Training Per Dialect)/Token Ratio (Test)* and *Character Ratio (Training Overall)*, *Character Ratio (Training Per Dialect)/Character Ratio (Test)* respectively.

A limitation especially of the variant using token types is that no real tokenisation was carried out. This leads to words with punctuation symbols being counted as different type tokens than without and therefore this measure should be seen as an approximation rather than an exact calculation.

3.3 Analyses

3.3.1 Standardisations

3.3.1.1 Performance Scores

The section above explained how a table summarising the findings from previous research was built in this thesis. However, as the table and the considerations mentioned above show, the performance of the included models differ considerably in multiple dimensions. One example are the evaluation metrics used as most papers report WER, but Scherrer et al. [2019] use precision, recall and F1 score. Further, as mentioned before, it was not possible to extract exact scores from all papers as some only showed figures.

An additional complication is that the training regime for the models was widely different in the included papers. For instance, in Nigmatulina [2020] all models were trained and evaluated on the same dialects, while in Scherrer et al. [2019] this was not the case. Instead, the ArchiMob ASR baseline was trained only on the dialect from the larger Zurich area and then evaluated on other varieties previously unseen by the model.

In both of the cases mentioned above, the systems were deliberately trained on Swiss German data. However, this was not done in Sicard et al. [2023] as these were zero-shot evaluations of whisper models not explicitly trained on Swiss dialects before. Therefore, the reported WER are relatively high as a comparison between whisper-tiny and the XLSR manual normalisation model from Khosravani et al. [2021a] shows: The former resulted in 121 % on Valais dialect while the latter achieved the best score with 11.6 % on the Eastern Swiss dialectal group.

For these reasons the performances for each system were ranked over all the dialects instead of working with the absolute scores. This was done by going over the results for each model and giving rank one to the dialect on which the best performance was achieved, rank two to the second best result and so on.

To illustrate this method, the following example considers the case of the ArchiMob baseline for which the results can be found in the table in the appendix. This model was evaluated using the F1 score where higher scores indicate a better performance and therefore, the ranks should be given in descending order of the scores. As the highest result was reported on Grisons dialect, this variety was given the first rank and the Basel variety with the second strongest performance received rank two. The remaining ranks in ascending order then are on Valais, Uri, Bern, and Lucerne dialect respectively.

A disadvantage of this method is that by transforming the metric scaled scores into ordinal scaled ranks, information about the distance between the performances is lost. However, this distance would have lead to a greater distortion of the results for the reasons mentioned, than it would have helped to gain a more detailed insight.

3.3.1.2 Dialects

Another dimension for standardisation is the number of the analysed dialects. As can be seen in the table in the GitHub repository and appendix, the dialects for which results were shown varied greatly between papers. Further, the dialects were grouped differently in the studies included in this thesis - mainly using two alternative: The first one emploid cantonal borders to discriminate between the different dialects, although the linguistic boundaries are of course not identical to the judicial ones. For example, all scores for the two systems from Nigmatulina [2020] were shown this way.

The other alternative was used in Khosravani et al. [2021a] and grouped multiple cantons together into a family of dialects - probably with the idea that these varieties are similar to each other in the opinion of the respective authors. For the latter it is hard to understand the exact motive as neither justifications were given nor the cantons contributing to a group were reported.

For this reason, it was necessary to leave out some dialects in order to gain a stable selection of varieties for the comparison between multiple systems. For this decision the aim was to follow guidelines from previous research such as adjustments in favour of densely populated areas and for a better correspondence to the perception of speakers (Schmidt et al. [2020b]). However, it was not possible to follow all suggestions from previous research as for instance in Scherrer and Stoeckle [2016] it is stated that the 'best cut' would be ten dialects. Yet in this case there were no ten dialects that were included in all or even most of the chosen papers and for the same reason it was only possible to choose a maximum of six different varieties.

These six dialects are Basel, Bern, Grisons, Lucerne, Valais, and Zurich and the locations of the respective cantons are shown in Figure 2. Basel includes both Basel-Landschaft and Basel-Stadt as the dialectality values reported for both in Scherrer et al. [2019] do not differ considerably. Further, only Nigmatulina [2020] explicitly differentiates between both versions and it was not always possible to extract information which variety or whether a mixture of both had been used (e.g., in the case of Khosravani et al. [2021a]).

However, even for these dialects results are not reported in all papers. In fact, only the Bern, Grisons, and Valais varieties were included in each studies while results on Zurich dialect were reported for every systems except the four models from Khosravani et al. [2021a] and the ArchiMob baseline.

Yet in the latter case the performance rank for the Zurich variety compared to the other dialects could be safely deduced. This is due to the training regime of the model as it was trained on data of the larger Zurich area and its performance was evaluated on different dialects it had not previously seen. Therefore, it is safe to assume that had it also been evaluated on data in Zurich Swiss German, this would have been the systems best result because of the familiarity with the data compared to the other dialects.

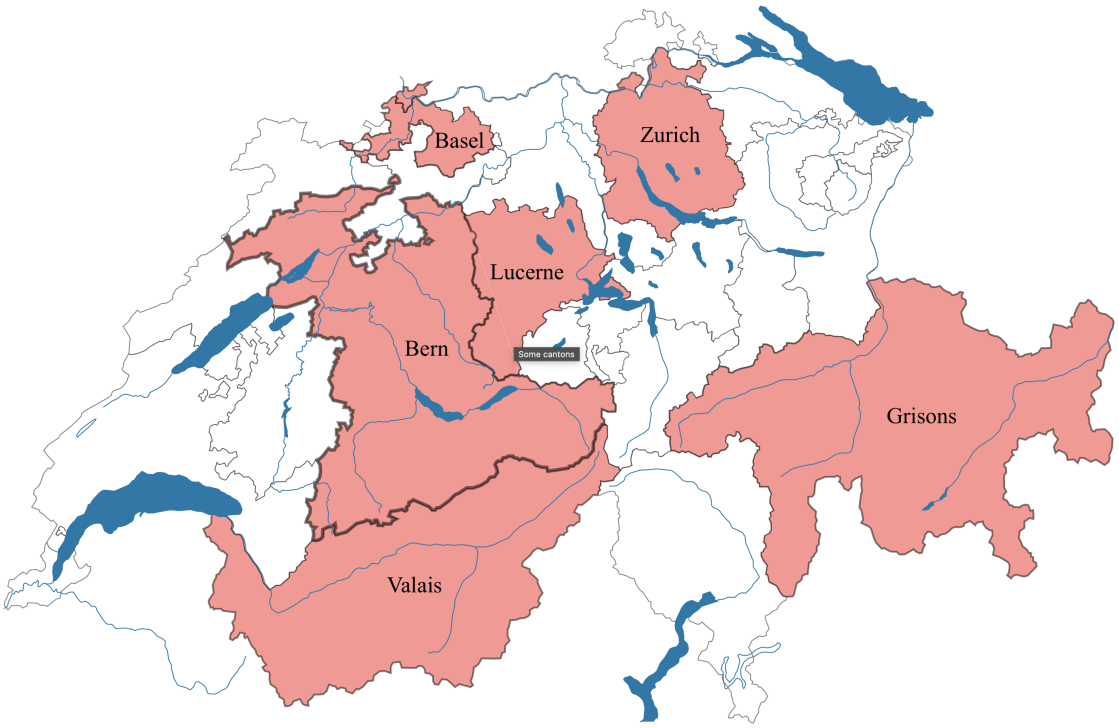

Figure 2: The regions for the six dialects which where included in the analyses of this thesis.

Unfortunately, such a deduction was not possible for the Basel and Lucerne varieties. The former is not reported in the four systems from Khosravani et al. [2021a] and values for the latter are missing for the same as well as for the six models in Schraner et al. [2022].

Table 1: An overview for how many of the 18 included systems, the results for the six dialects Basel, Bern, Grisons, Lucerne, Valais, and Zurich were reported, deduced, or no information was available.

Dialect(s)	**Results Reported**	**Results Deduced**	**No Information**
Bern, Grisons, Valais	18	0	0
Zurich	13	1	4
Basel	14	0	4
Lucerne	8	0	10

This information is summarised in table 1 which shows that an evaluation would be possible on all systems when using the three dialects Bern, Grisons, and Valais and on all six dialects but then only on eight models. To strike a balance between those two options, the analyses in this thesis were carried out in two different settings:

- One version included all systems but only on the four dialects Bern, Grisons, and Valais for which sufficient data was available.
- The other made use of all six dialects, but only on the systems for which results were

either directly available or could be deduced. This therefore constitutes the models from Nigmatulina [2020], Scherrer et al. [2019], and Sicard et al. [2023].

These two settings form the basis for all of the analyses in this thesis and the following two sections will discuss each of these analyses in turn.

3.3.2 Agreement on Performance Ranking of Dialects between Models

For each of the two settings described above, it was evaluated how well the models agreed upon the rankings of the dialects. As the first place of this ranking was given to the best performance, higher ranks indicate a worse performance for a given model. If the models then agree upon the order of the dialects in this ranking, this would be strong indication for that fact that there are factors in the dialects themselves which make them harder for ASR systems to learn.

To this end, the agreement of the rankings between the different models was calculated using the Kendall's Tau. This metric was chosen as it is appropriate for comparing agreement in terms of rank and often seen reported in literature. It indicates a perfect agreement with a score of 1 and a perfect negative agreement by -1, while 0 means no agreement. was calculated for the rankings of all system pairs through an own script using scipy.stats' kendallstau function.

The Kendall's Tau however only reveals how strongly the systems agree with each other over all the chosen dialects. It could however be that there are certain dialects for which the models strongly agree on their rank and others where there is less certainty. This could not be analysed in depth due to time constraints and instead the rankings of the dialects where evaluated manually to get a coarse overview if agreement was different on different dialects.

3.3.3 Correlation between Performance Rank, Dialectality, and Data-Orientated Measures

A second analysis aimed at finding correlations between either data orientated measures or linguistic factors and the performance rank of a dialect. The latter were given in ascending order starting with the dialect on which a model achieved the best performance. Therefore, dialects with a higher performance rank are those on which lower results were achieved and so should be the more difficult ones. The correlation then aims at investigating if this difficulty is due to properties of the evaluation data or linguistic factors of the dialects themselves.

A linguistic factor that has been discussed frequently in the past, is the variation between the individual varieties. For instance, in Nigmatulina [2020], the author argued that certain dialects might be less similar to others and therefore profit less from the training material of other varieties. Such dialects therefore have a larger distance to other varieties as well as to an underlying standard which can be quantified using a so-called dialectality score. This linguistic

measurement ascribes a number to each variety based on the similarity to a chosen standard which is calculated as the difference between the phonemes in the standard and non-standard variety.

The dialectality scores used in this thesis were extracted from Scherrer et al. [2019], which reported them in figure nine on the 25th page of their paper. This image consists of a map where each interview is represented by a circle and placed in the location the speaker originated from. The circles are differently coloured and each colour is associated with a score between a lower and upper bound via a legend on the side of the graphic. For this thesis the middle between the bounds was used and therefore, the score for each dialect was calculated by taking all points inside the cantonal borders, associating each of them to a number based on the colour code, and then calculating the average of all these circles.

On the other hand, the computation of the data orientated measures was done as the performance differences might not be caused by dissimilarities between the varieties but due the evaluation data not being equally difficult. One reason could be the varying sparseness as it is easier for models to produce a short text with repeating elements - especially if those were seen frequently during training. The latter is often the case when the sparseness of the evaluation data is low as it is usually a part of the training data held out for testing.

To investigate this hypothesis, this thesis used the two measurements of character and type token ratio. The former is the count of unique symbols over their total amount and similarly the latter is the count of unique tokens over their total sum. Therefore, both measures of sparseness are low when only few individual characters or tokens are used in a short text and higher when many of them occur uniquely in a long text.

While both scores are similar, they are still noticeable differences as the type token ratio can vary between speakers while the character ratio is less likely to do so. This is due to the fact that depending on aspects like education and age, people may have a wider range of vocabulary. Further, the notion of token is quite different between the data sets while the definition of individual symbols is more straightforward. For example in the ArchiMob corpus even interjections are marked as tokens while those are largely absent in the SwissDial data set. Therefore, the character ratio is more comparable between the two corpora used in this thesis.

Similarly utterance measures were excluded from the evaluation due to their variation between the two corpora: As the table in the GitHub repository and appendix shows, the utterances in the ArchiMob corpus are very short as many of them are only a few words long and interjections often constitute an entire utterance by themselves. For the SwissDial data set on the other hand, a mean utterance is longer and constitutes an entire sentence in most cases.

However, both character and type token ratio could not be calculated for all models as the data sets from Khosravani et al. [2021a] and Scherrer et al. [2019] were not publicly available. Still, an analysis between the rank of the dialect and its dialectality score was possible even for those systems and therefore they were included in the computation of the correlation in this case.

Due to this unavailability of some measurements for certain system and the reasons given in section 3.3.2, the same groupings as in that section were used. Therefore, the evaluation was carried out for six dialects (Basel, Bern, Grisons, Lucerne, Valais, and Zurich) when excluding the systems from Khosravani et al. [2021a] and Schraner et al. [2022] and on three dialects (Bern, Grisons, and Valais), where they were included.

In these setups the correlation measure used was the Spearman's rank correlation as it is most appropriate for measuring correlations for ordinal data. For this metric, 1 indicates a perfect positive, -1 a perfect negative and 0 no correlation. To do the calculations, the spearmanr function from scipy.stats was used.

4 Results

4.1 Kendall's Tau Agreement

4.1.1 18 Models with Three Dialects

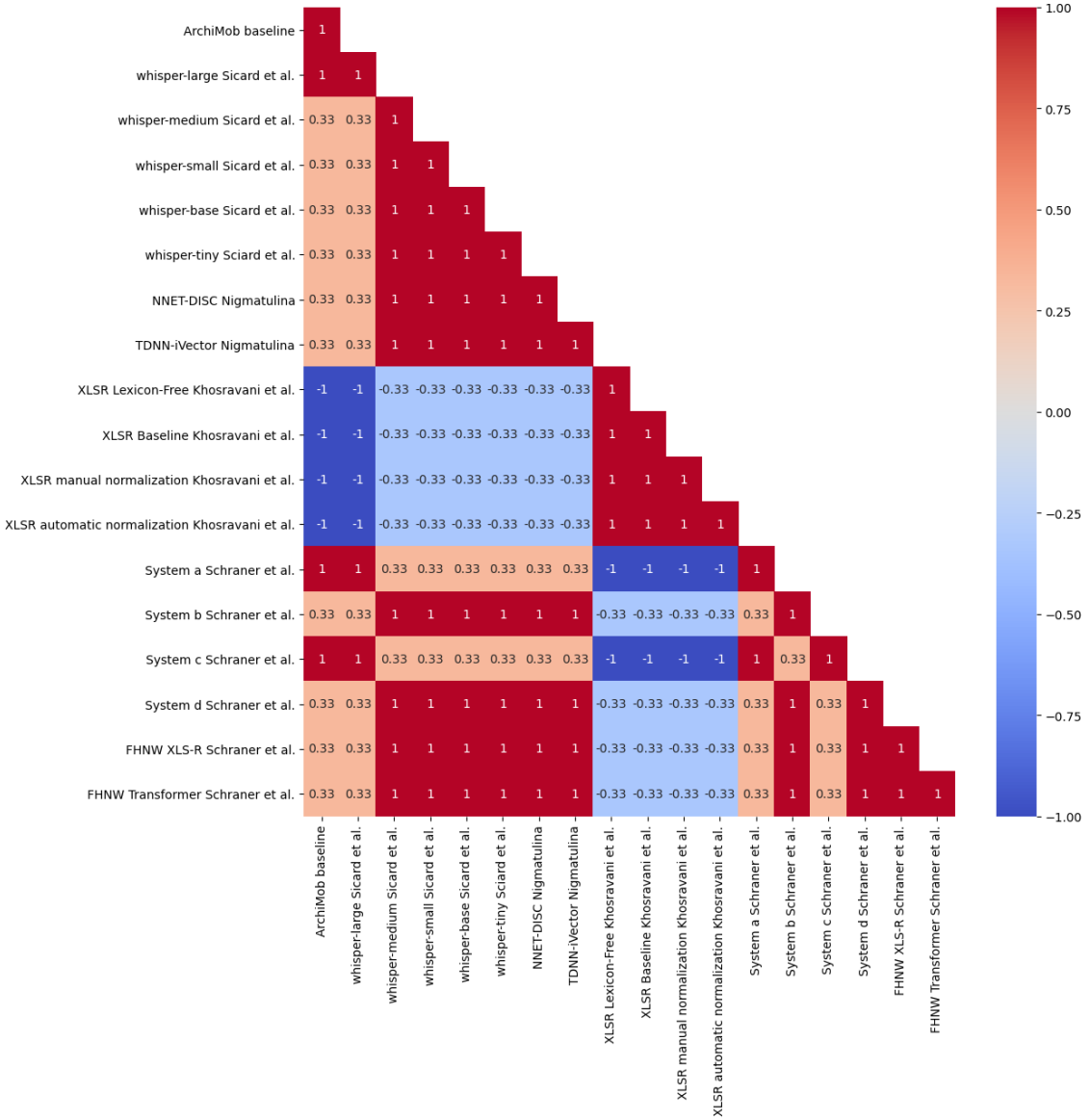

Figure 3: Pairwise Kendall's Tau on all 18 models for the three dialects Bern, Grisons, and Valais.

While the previous section described the settings used for the analyses in this thesis, this section will introduce the observed results. The first are given Figure 3 which shows in Kendall's Tau evaluation including all 18 systems on the three dialects Bern, Grisons, and Valais. Because it only includes three data points per pair, the results are coarse-grained: Only four states of agreement can be observed consisting of a perfect negative, a slightly negative, a slightly positive and a perfect positive one. This leads to an all-or-nothing type of concurrence when compared to Figure 4 where a more detailed picture can be seen. However certain tendencies can never the less be observed.

Overall the systems seem to agree strongly on how well they perform on the different dialects. For instance, the two systems from Nigmatulina [2020] align perfectly with the majority of the whisper models, the majority of the systems evaluated in Schraner et al. [2022], and both systems from Nigmatulina [2020]. Further, the ArchiMob baseline agrees perfectly with whisper-large and systems a and c from Schraner et al. [2022].

While the Kendall's Tau is high for most systems, the four from Khosravani et al. [2021a] make a noticeable exception. They agree perfectly between each other, but they are the only models that show negative scores when compared to systems from other papers. Their negative concurrence is only slight for most of the whisper models, both systems from Nigmatulina [2020] and four systems from Schraner et al. [2022]. However, they have a perfect negative agreement when compared to the ArchiMob baseline, whisper-large and systems a and c from Schraner et al. [2022].

These findings lead to the observation that the models can be separated into three clusters: The first consist of the models from Khosravani et al. [2021a] due to their negative concurrence with other applications. The second encompasses both models from Nigmatulina [2020], four whisper models excluding whisper-large together with four systems from Schraner et al. [2022] with the exceptions of systems a and c. The last one includes the exceptions from the previous one and the ArchiMob baseline. Further, these groups can be arranged into a continuum: On the one side is the ArchiMob baseline cluster, followed by the whisper group and on the far side the models from Khosravani et al. [2021a].

These groupings show that the models do not primarily form clusters based on the papers they originate from. For instance, whisper-large as well as system a and c from Schraner et al. [2022] fit better with the ArchiMob group instead of other models from the same paper. The only exception are the four systems in Khosravani et al. [2021a] which form a separate cluster by themselves.

The fact that the groupings do not strictly correspond to the papers implies that they are not primarily caused by the data as the latter is constant per study. Further evidence can be found when looking at the whisper models and the systems from Schraner et al. [2022]: They were evaluated on the same data and while they largely agree with each other, there are still the exceptions of systems a and c as well as the whisper-large model. Similarly, the ArchiMob

baseline and both systems from Nigmatulina [2020] were both trained and evaluated on the ArchiMob corpus, but belong to different clusters. For the last example it is important to note however, that the baseline was trained on the first release of the corpus and the models from Nigmatulina [2020] on the second one which had additional interviews available.

In summary, these findings showed that the systems agree well with each other in terms of a performance ranking between the dialects. Further, they can be clustered into three groups which neither reflect the papers the models originated from nor align with the usage of the same data. However, the systems from Khosravani et al. [2021a] are an exception as they agree negatively with other models and form a separate cluster by themselves. Still, the groupings are based on an all-or-nothing type of agreement as only three dialects were included. This causes the need for a more fine-grained evaluation on six dialects shown in the following section.

4.1.2 Eight Models with Six Dialects

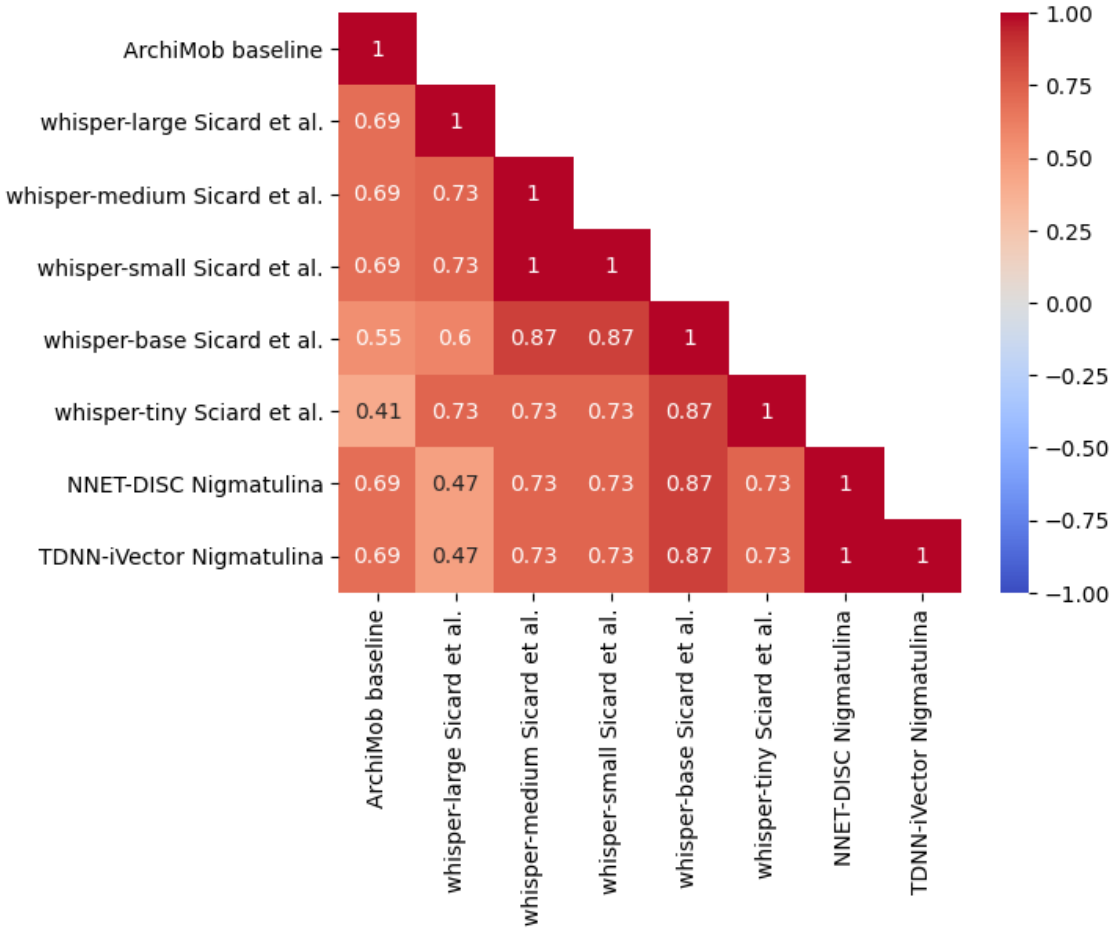

Figure 4: Pairwise Kendall's Tau on the eight models from Nigmatulina [2020], Scherrer et al. [2019], and Sicard et al. [2023] for the six dialects Basel, Bern, Grisons, Lucerne, Valais, and Zurich.

While the last section described the findings for the setting using three dialects and all systems,

Figure 4 shows the results of the pairwise calculated Kendall's Tau for only the eight systems from Nigmatulina [2020], Scherrer et al. [2019], and Sicard et al. [2023] but on all six chosen dialects (Basel, Bern, Grisons, Lucerne, Valais, and Zurich). Overall, the models agree strongly as all the Kendall's Tau scores are positive and lie between 0.41 and 1. Further, a high score of 0.73 is achieved in nearly half the cases (when excluding the values of each system with itself). However, the inclusion of more dialects gives more opportunities for the systems to disagree, which results in a smaller number of perfect concurrences.

Still, this constitutes an even better agreement than described during the previous section. One probable reason for this is the exclusion of the four models from Khosravani et al. [2021a] which were the only ones showing negative scores in the previous evaluation.

After this general analysis, each model will now be discussed in turn, starting with the ArchiMob baseline. This system shows a strong agreement for all comparisons except when contrasted with whisper-base and whisper-tiny. In both of those the score is still moderate with 0.55 and 0.41 respectively. However, the latter is the worst result seen in this figure.

Whisper-large very strongly agrees with whisper-medium, whisper-small, and whisper-tiny receiving a score of 0.73. The Kendall's Tau for the comparison to whisper-base is only slightly lower and achieves 0.6. The system however does not agree well with both models from Nigmatulina [2020] where it achieves a score of 0.47. While this is still a moderate agreement, it is the second worst seen in this setup.

In turn, whisper-medium shows strong concurrences for each comparison: Between this model and whisper-small one of the two perfect positive agreements can be seen (the other one being between both systems from Nigmatulina [2020]). Further, whisper-medium's ranking very strongly correlates with the one of whisper-base achieving a Kendall's Tau of 0.87. For the comparison to whisper-tiny and both models from Nigmatulina [2020] the score is 0.73 which is still high.

Further, whisper-small agrees most with whisper-base achieving a Kendall's Tau of 0.87. The comparison to whisper-tiny and both models from Nigmatulina [2020] results in a high score of 0.73.

The findings for whisper-base are similar: It achieves a score of 0.87 when compared to whisper-tiny and the models from Nigmatulina [2020]. To the latter also whisper-tiny agrees very strongly with a score of 0.73.

In summary, these analyses show that there are no outliers in this setup, but that between most system there were minor re-orderings in terms of the dialect ranking. Therefore, it is no longer possible to cluster the models into clearly defined groups. Still it needs to be noted that the agreement of systems originating from the same paper are higher than those of models across different ones. As the data is constant per paper, this further means that models with the same data show a higher agreement.

An exception are the whisper models which show some fluctuations although the same data was used in all of them. Further, the ArchiMob baseline and the systems from Nigmatulina [2020] show a lower score although both studies used the ArchiMob corpus. In the last case it needs to be said that in Scherrer and Stoeckle [2016] the first release was used, while Nigmatulina used the second one, containing additional interviews.

4.1.3 Agreement per Dialect

The previous sections inspected how strong the consent in the ranking of all dialects in a given setup was across different models. However, it could still be that the systems concur more on some varieties while achieving more mixed results on others. This was investigated by manually comparing the bare rankings which were given in ascending order according to the performance per model. Therefore, rank one was given to the dialect on which the best performance was achieved by a given system and the highest rank corresponds to the dialect on which the worst performance of the model was observed. The findings that originated from this analysis will be discussed in the following paragraphs.

For the setup using all systems and only the three dialects Bern, Grisons, and Valais, the agreement on Grisons appeared to be very strong as only the outlier systems from Khosravani et al. [2021a] do not give it the first rank. However, for the remaining two dialects the systems do not agree as strongly but still slightly more on the Valais than on the Bern variety. The former was ranked third by the eight systems of the whisper group while both other groups ranked it second. Yet for Bern dialect the agreement is slightly worse as the four models from Khosravani et al. [2021a] are the only ones giving it the first rank while the ten systems of the whisper group placed it in the middle and the remaining four at the end.

For the second setup using only eight models with the six dialects Basel, Bern, Grisons, Lucerne, Valais, and Zurich, Bern becomes the most agreed upon dialect: All systems rank it on position four except the ArchiMob baseline and whisper-large which give it rank five. Agreement on Grisons dialect remains strong as well as five models give it rank one and the remaining ones rank it second.

The same level of consistency in results does not apply to the other dialects as it is lesser on the remaining varieties. Still, some show a certain level of agreement as whisper-large and whisper-tiny give the Bern variety rank one, the remaining whisper systems rank it second and the ArchiMob baseline as well as both models from Nigmatulina [2020] obtain rank three on this dialect. Similarly, Valais dialect is ranked fourth by both the ArchiMob baseline and whisper-large, while whisper-medium and whisper-small achieve their second worst performance on it and whisper-base, whisper-tiny and the Nigmatulina [2020] models show their worst performance on this dialect.

However, even this lesser agreement is not present for the Zurich and Lucerne varieties as the

results on those two dialects are mixed: The former is ranked third on whisper-base only, while the models from Nigmatulina [2020] and whisper-tiny give it rank five and the ArchiMob baseline, whisper-large, whisper-medium, and whisper-small put it in the last position. Similarly, the Zurich variety is ranked first by the ArchiMob baseline, second by both models from Nigmatulina [2020], and third by all whisper systems, except whisper-base which ranked it fifth.

The previous analysis shows that the exact rank of a dialect is seldom exactly the same for all systems. However, they do have tendencies to be placed either towards the front or back as for example Grisons and Basel with 5:3:0:0:0:0 and 2:3:3:0:0:0 are both in the top half for all systems. Meanwhile Valais and Bern with 0:0:0:2:2:4 and 0:0:0:6:2:0 are in the back half for all models. Lucerne is in the middle but somewhat to the back with 0:0:1:0:3:4 and vice versa Zurich is in the middle but a bit towards the front with 1:2:4:0:1:0.

These tendencies suggest the building of two groups where one is placed the front and to the other receives higher ranks. The first consist of Grisons, Basel, and Zurich dialect while the latter encompasses the Lucerne, Bern, and Valais varieties.

In summary, this shows that there are indeed dialects - such as Grisons for instance - on which the models seem the agree strongly, while the rankings for varieties such as the Lucerne one are more mixed. However, even for dialects on which a stronger consensus was observed, the rank is rarely exactly the same while tendencies are nevertheless observable for all dialects: Some - as for instance the Basel variety - are placed more towards the front and others - such as Valais and Bern dialect - usually achieve higher ranks. This leads to two groups of dialects

4.2 Correlation between Rank and Dialectality, Type Token, and Character Ratio

Correlation Variables	6 Dialects on 8 Models	3 Dialects on 18 Models
Performance Rank & Dialectality	0.75	0.28
Performance Rank & Type Token Ratio	0.54	–
Performance Rank & Character Ratio	0.28	–

Table 2: Spearman's rank correlation

While the last analyses focused on the agreement between the system, this section looks into the correlation between the performance rank and the measures described in section 3.3.3 for which the findings are given in table 2. Specifically, it shows the Spearman's rank correlation between performance rank and either dialectality, type token ratio, or character ratio. For all these cases, the observed results are positive with the one between dialectality and performance rank achieving the highest score: On the setup with only eight models on all six dialects a correlation

of 0.75 was found - indicating a strong relation. However, when evaluated on less dialects and more models the number drops to 0.28.

Still, both results are positive which means that a higher dialectality correlates with a higher performance rank. The former rises with the distance of the dialect to the underlying standard and the latter with a decline in performance. In other words, a larger distance correlates to qualitatively poorer results on a dialect. However, it needs to be noted that the fluctuation spans the best and one of the two worst relations.

The other lowest score is the correlation between performance rank and character ratio - likewise with a value of 0.28. While this number might be low, it is still positive which implies that a higher character ratio co-occurs with a higher performance rank. As the character ratio rises when there is a greater number of individual symbols used over a shorter text, a higher outcome indicates sparser data. Therefore, a positive correlation between rank and character ratio signifies that sparser evaluation data indeed correlates with a lower performance on a dialect. The same argument applies to the correlation between performance rank and type token ratio which is even higher with a score of 0.54.

In summary, it can be said that dialectality as well as both sparseness measures show positive correlations to the performance rank. Therefore, both a further distance to an underlying standard as well as the sparseness of the evaluation data co-occur with poorer results on a dialect. However, both dialectality as a linguistic and sparseness as a data-dependant factor fluctuate considerably and achieve only a slight to moderate concurrence in most cases.

5 Discussion

5.1 Influence of Dialectality Versus Data Sparseness

In the previous section, both the two measures of data sparseness - namely type token and character ratio - as well as the dialectality score showed positive correlations to the performance rank on a dialect. Because the best performance received the first rank, this finding means that higher dialectality scores as well as a higher sparseness of the evaluation data co-occur with a worse performance of a model.

While the relation is only weak for the character ratio, it is moderate for the type token ratio and strongest for the dialectality - at least when evaluated on six dialects. Contrarily, on three dialects the relation is considerably weaker which could be due to the fact that in this setup the four models from Khosravani et al. [2021a] were included. Those were outliers in the analysis from section 4.1.1 as they were the only systems showing a negative agreement with other models.

However, the difference in dialectality needs to be sufficiently large in order to have an impact on the ranking of the dialects. For instance, the Grisons variety was ranked as the easiest dialect when compared to Valais and Bern with the exception of only a few systems. Yet, for both Valais and Bern dialects the agreement of the systems was not as strong which could be due to the fact that the distance between the Grisons and Valais varieties is about ten times larger then the distance between Valais and Bern dialect.

Further evidence for this hypothesis can be found in the results of the setup with six dialects where the dialectality scores suggest the formation of two groups: Grisons, Basel, and Zurich on the lower dialectality side and Valais, Bern, and Lucerne on the higher side. These clusters have been formed due the fact that between the groups the dialectality gap is larger (0.03 to 0.55) than inside the groups (0.005 to 0.02 and 0 to 0.005 respectively). These clusters co-inside with the ones discussed in section 4.1.3 as on the former group better performances were observed leading to lower ranks while the latter cluster had a tendency to receive higher ranks. Further, inside the groups where the dialectality differences are smaller, the systems agreed less on their exact rank.

Still, for the evaluation on six dialects the order of the dialect ranks does not exactly match the order of increasing dialectality for even a single system. This can be explained by different

reasons among them being that many factors can influence system performance as stated in section 3.2. However, multiple of them could unfortunately not be analysed in this thesis.

Further, the dialectality score used has two considerable limitations. For one, the evaluation could not use exact scores as they were extracted from a figure in Scherrer et al. [2019]. Secondly, the calculation of the score in Scherrer et al. [2019] is not exact either as it was extracted on the basis of the normalisation layer of the ArchiMob corpus. This is however a purely lexical normalisation and does not take other factors such as syntax into account which is normally done in the calculation of such scores. Moreover, the dialectality score was computed using the Levenshtein distance instead of pairwise comparisons of phonemes as would usually be the case. All these limitations lead to the fact that smaller differences between the dialects are likely not represented.

Another weakness of the score is that it was calculated based on only one corpus. However, there is a low number of data points per dialect in the ArchiMob corpus as a multitude of dialects is only represented by a single interview from a single speaker. This is unlikely to satisfyingly capture the variability in those dialects which has been stated to be a considerable factor (Khosravani et al. [2021a]). More variation might have lead to different dialectality scores and could lead to the the currently used one to be distorted. In order to gain more insight, future research should aim at calculating a more exact dialectality score on the evaluation data directly.

However, a noticeable exception to the influence of the dialectality score on the dialect ranking are the systems from Khosravani et al. [2021a]. One possible reason for this result could be that there are still factors for the difficulty of a dialect that could not be found in this thesis. However, their impact is on a smaller scale than big differences between dialectality scores for all other analysed models. Further, the systems from Khosravani et al. [2021a] are outliers in an otherwise strong agreement between the systems: They perform best on dialects with a high dialectality while for all other models the tendency is to obtain worse results on those varieties. Therefore it is unlikely that unobserved reasons for the difficulty of dialects are responsible for the result.

Another theory would be that this is explained by the approach that was taken in this paper: Khosravani et al. tried to improve the WER by including a lexicon during the decoding phase. This could help in reducing the dialectality of dialects that differ mostly in a lexical way from the standard. However, their lexicon free system shows the same disagreement with models from other papers as the ones using a lexicon.

A final factor that could cause this result is the training data which was not analysed due to unavailability. It is mentioned in Khosravani et al. [2021a] to be imbalanced in terms of dialect and it could therefore be that the dialectality during training was higher for these models than for others. This heightened familiarity in an environment with higher dialectality could then result in a performance drop on data with less dialectality.

This however would not fit the whisper models, the ArchiMob baseline and both systems from Nigmatulina [2020]. The dialectality of the training data for the whisper models should be the lowest as they were only trained on standard varieties and no Swiss German data. For the ArchiMob baseline it is slightly higher because it was trained on Swiss German but only from the larger Zurich area where the dialectality is rather low. The systems from Nigmatulina [2020] were trained on all varieties represented in the ArchiMob corpus and should therefore have the highest dialectality in their training data. However, the performance of the ArchiMob baseline follows the order of dialectality more closely than the whisper models, although the latter had a lower dialectality in their training data. Similarly, both systems from Nigmatulina [2020] saw the highest dialectality in their training data yet agree more with the whisper models trained on the lowest dialectality than with the ArchiMob baseline. Further research is therefore needed to verify if the dialectality of the training does have an influence on system performance per dialect.

In summary, for the analysed models and datasets both linguistic factors in terms of dialectality as well as measures of data sparseness on the evaluation data show correlations to the performance on a dialect. This relation is the strongest for the dialectality score but seems to only be reliable as long as the differences are sufficient in magnitude and is still influenced by the data sparseness and possibly other factors not analysed in this thesis. Another limitation of this thesis is that the used dialectality score has considerable weaknesses and therefore future research should aim at using a more exact score calculated directly on the evaluation data. Moreover, the training data could not be analysed due to its unavailability and warrants additional research as well.

5.2 Limitations

Some limitations such as the drawbacks of the dialectality score were already discussed in the previous section. However, those are not the only hindrances that this thesis faced as for one, the number of systems and datasets analysed was small: Although 48 data points were investigated, the measures were only calculated on three different datasets from which two are different versions of the same corpus. As each data set was repeated for each model that originated from the same paper, the SwissDial corpus had a great influence in contributing five of eight systems and therefore 30 of 48 data points. While this could distort the findings presented in this thesis, it was not possible to include more systems as not many exist at this point in time.

However, this is expected to change over the following years because one crucial bottleneck is beginning to improve: Before 2021 there were virtually no public data sets available for the task, but through the publications of new data sets training Swiss German ASR models is now feasible (Schraner et al. [2022]). With more systems trained and evaluated on different data in

combination with using different architectures and techniques, it is likely that a more detailed picture can be painted in the future.

While the richness in approach between the papers lead to interesting results, they additionally make absolute scores hard to compare. For instance, the approaches of papers analysed during this thesis range from Zero-shot evaluations over using different Swiss German varieties for training and testing to training on the same data set later used for evaluation (see 3.3.2). This was the reason for mainly evaluating dialect rank in this thesis but it comes with the downside of losing information about the magnitude of the performance differences. If future research wants to be able to work with metric variables in order to obtain a more detailed picture, measures to ensure the comparability must be taken first.

Comparability is further important for the dialects which are included in the different papers and how results are reported in previous research. For one, a wide range of dialects is covered between the studies analysed in this thesis but only a few of them appear in multiple or even all studies. Additionally, some papers group different dialects together and report their findings only for the entire group. However, it is often not clearly stated which dialects a cluster consists of - making it difficult to evaluate these findings across studies. Future research should be more transparent in these cases and provide reasons for their grouping.

Further, there is little data from Eastern dialects such like the Appenzell variety which means that little can be said about the performance of ASR models on them. Moreover, including them would be beneficial to the entire task as a broader range of dialectality scores could be covered. This is helpful in gaining a more detailed insight on how high the margin in dialectality needs to be in order to affect the performance. Future research should therefore cover more dialects, especially from the East of Switzerland.

6 Conclusion

In summary, this thesis investigated the influence both measures of data sparseness and linguistic factors might have on the performance differences between Swiss German dialects. To this end, 18 models from five different papers were analysed and compared in factors that might have an influence on this phenomenon and the results of this comparison are shown in a table in the appendix. This table could potentially provide a starting point for future research on this topic.

Moreover, it could be shown that both measures for data sparseness on the evaluation data as well as the dialectality score as a linguistic metric showed correlations to the system performance. While this is a first confirmation for two factors that have been hypothesised about in previous research, it is neither an exhaustive list of influences nor an analysis on the relation between those factors. Those need to be investigated by future research possibly by making use of the provided table.

A limitation of these findings is that they are based on little data both in terms of dialects and systems. While with more research in this field, more models should become available for comparison, the techniques in these systems will differ considerably from each other highlighting the need for better measures to ensure the comparability of the results. Similarly, the comparability between different dialect groupings must be improved by aiming at more clearly reporting the reasons behind the groupings and including more dialects, especially from the East of Switzerland.

Glossary

Dieth transcription As Swiss German is not standardised, it does not have a definitive orthography. Eugen Dieth was a Swiss linguist who proposed a spelling approach based on Standard German. The method can be used for all Swiss German dialects and does not introduce any new characters when compared to Standard German. It is widely used in an academic context today.

F1 score A widely used measure for system performance. It is the harmonic mean between precision and recall. Recall shows how many of the relevant instances have been retrieved, while precision is the fraction of relevant instance among the retrieved ones. For the context of ASR both definitions have been changed slightly as described in Scherrer et al. [2019].

Kendall's Tau A score used in many papers to compare agreement on ranks. It is calculated as follows:

$$Tau = (C - D)/(C + D)$$

where C is the amount of concordant and D the amount of discordant pairs between the systems. A score of 1 stands for a perfect agreement, -1 for a perfect negative agreement and 0 for no agreement.

Levenshtein distance A measure of distance between two sequences that is usually used for words. It is calculated by the minimum of deletions, insertions, and substitutions needed to transform one sequence into the other. As a substitution can be seen to correspond to a deletion followed by a insertion, it is usually weighted double.

Spearman's rank correlation A measure of rank correlation between two variables. It is an assessment of how well the relationship between two variables can be described using a monotonic function. A score of -1/1 denotes perfect negative/positive correlation, while 0 stands for no correlation. It is calculated by the following formula:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where n is the number of paired scores, $\bar{x}$ is the mean score of x, and $\bar{y}$ is the mean score of y.

Tokenisation Splitting a sentence or text into units called tokens. They do roughly correspond with the intuitive understanding of a word. However, also punctuation or other symbols are tokens.

WER A wildly used evaluation metric for ASR systems. It is calculated on word level and shows the minimum number of mistakes produced by a system relative to the number of words in a reference transcription, which makes it proportional to the correction cost. It can therefore reach over a 100 %. The following formula shows its calculation:

$$WER = (S + D + I)/N * 100\% = (S + D + I)/(S + D + C) * 100\%$$

where S, D and I is the minimum number of substitution, deletion and insertion operations required to transform the system prediction to the reference text, where N is the number of words in the reference text and C — the number of words correctly recognised by a system.

Zero-Shot A technique usually employed to set a baseline for future improvement. A system already trained on out-of-domain data is used on the evaluation set without having been fine-tuned to in-domain data.

References

A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli. XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, abs/2111.09296, 2021. URL `https://arxiv.org/abs/2111.09296`.

L. Barrault, O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, and M. Zampieri. Findings of the 2019 conference on machine translation (WMT19). In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, M. Turchi, and K. Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL `https://aclanthology.org/W19-5301`.

C. Colwell, A. Jelfs, and E. Mallett. Initial requirements of deaf students for video: lessons learned from an evaluation of a digital video application. *Learning, Media and Technology*, 30(2):201–217, 2005.

P. Dogan-Schönberger, J. Mäder, and T. Hofmann. SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German, Mar. 2021. URL `http://arxiv.org/abs/2103.11401`. arXiv:2103.11401 [cs].

W.-Y. Hwang, R. Shadiev, C.-T. Kuo, and N.-S. Chen. Effects of speech-to-text recognition application on learning performance in synchronous cyber classrooms. *Educational Technology & Society*, 15:367–380, 11 2012.

A. Khosravani, P. N. Garner, and A. Lazaridis. Modeling Dialectal Variation for Swiss German Automatic Speech Recognition. In *Interspeech 2021*, pages 2896–2900. ISCA, Aug. 2021a. doi: 10.21437/Interspeech.2021-1735. URL `https://www.isca-speech.org/archive/interspeech_2021/khosravani21_interspeech.html`.

A. Khosravani, P. N. Garner, and A. Lazaridis. 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, Cartagena, Colombia,

Dec. 2021b. IEEE. ISBN 978-1-66543-739-4. doi: 10.1109/ASRU51503.2021.9688249. URL https://ieeexplore.ieee.org/document/9688249/.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, Sept. 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.11.

T. C. Kuo, R. Shadiev, W.-Y. Hwang, and N.-S. Chen. Effects of applying str for group learning activities on learning performance in a synchronous cyber classroom. *Computers & Education*, 58(1):600–608, 2012. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2011.07.018. URL https://www.sciencedirect.com/science/article/pii/S036013151100176X.

I. Nigmatulina. Acoustic modelling for Swiss German ASR. 2020.

I. Nigmatulina, T. Kew, and T. Samardzic. 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, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics (ICCL). URL https://aclanthology.org/2020.vardial-1.2.

P. Nisbet and A. Wilson. *Introducing Speech Recognition in Schools: using Dragon Naturally Speaking*. CALL Centre, University of Edinburgh, 2002a. ISBN 1-898042-22-5.

P. Nisbet and A. Wilson. *Introducing Speech Recognition in Schools: using IBM ViaVoice*. CALL Centre, University of Edinburgh, 2002b. ISBN 1-898042-23-3.

P. Nisbet, A. Wilson, and S. Aitken. Speech recognition for students with disabilities. In *Proceedings of the Inclusive and Supportive Education Congress, ISEC 2005 Conference. Delph, UK: Inclusive Technology*, 2005.

Y. Scherrer and P. Stoeckle. A quantitative approach to Swiss German – Dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*, 24(1):92–125, Nov. 2016. ISSN 0942-4040, 1867-0903. doi: 10.1515/dialect-2016-0006. URL https://www.degruyter.com/document/doi/10.1515/dialect-2016-0006/html.

Y. Scherrer, T. Samardžić, and E. Glaser. Digitising swiss german: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53, 12 2019. doi: 10.1007/s10579-019-09457-5.

L. Schmidt, L. Linder, S. Djambazovska, A. Lazaridis, T. SamardÅ¾iÄ‡, and C. Musat. A swiss german dictionary: Variation in speech and writing. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2720–2725, Marseille, France, May 2020a. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.331.

L. Schmidt, L. Linder, S. Djambazovska, A. Lazaridis, T. Samardžić, and C. Musat. A Swiss German Dictionary: Variation in Speech and Writing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2720–2725, Marseille, France, May 2020b. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.331.

Y. Schraner, C. Scheller, M. Plüss, and M. Vogel. Swiss German Speech to Text system evaluation, Nov. 2022. URL http://arxiv.org/abs/2207.00412. arXiv:2207.00412 [cs].

R. Shadiev, W.-Y. Hwang, and Y.-M. Huang. Investigating learning strategies of using texts generated by speech to text recognition technology in traditional classroom. *Proceedings of the AECT International Conference on the Frontier in e-Learning Research*, pages 279–286, 01 2013.

C. Sicard, K. Pyszkowski, and V. Gillioz. Spaiche: Extending State-of-the-Art ASR Models to Swiss German Dialects, Apr. 2023. URL http://arxiv.org/abs/2304.11075. arXiv:2304.11075 [cs, eess].

A. Trivedi, N. Pant, P. Shah, S. Sonik, and S. Agrawal. Speech to text and text to speech recognition systems-areview. *IOSR J. Comput. Eng*, 20(2):36–43, 2018.

M. Wald and K. Bain. Universal access to communication and learning: the role of automatic speech recognition. *Universal Access in the Information Society*, 6:435–447, 2008.

Lebenslauf

Persönliche Angaben

Iris Zoder
Birkenmatt 4
6343 Rotkreuz
iris.zoder@uzh.ch

Schulbildung

bis 2020	Besuch der Kantsschule Zug
2020	Erwerb der Maturität
seit 2020	Bachelor-Studium Computerlinguistik und Sprachtechnologie an der Universtiät Zürich

Berufliche und nebenberufliche Tätigkeiten

2020-2023	Tierpflegerin für den Elfenhof
seit 2021	Webprogrammiererin für Zweieck Qt-Experts GmbH

A Tables

Please find the table for all included systems and corpora on the following pages.

Table A1: Overview of included systems and corpora

System Name	Dialect	Performance	Performance Rank	Writing System	Corpus Name	Corpus Topic	Automatic vs. Manual Transcription	Known Transcriber Differences	Training Dialects	Overlap in Training and Test in Terms of Dialect	Number of Different Speakers (Training Overall)	Number of Different Speakers (Training Per Dialect)	Different Speakers Test	Overlap between Test and Training Speakers	Training Set Size (Utterances Per Dialect)	Training Set Size (Words Per Dialect)	Mean Utterance Length in Words (Training Overall)	Mean Utterance Length in Words (Training per Dialect)	Vocab Size (Training Overall)	Vocab Size (Training Per Dialect)	Token Ratio (Training Overall)	Token Ratio (Training Per Dialect)	Charcter Ratio (Training Overall)	Character Ratio (Training Per Dialect)	Test Set Size (in Words)	Mean Utterance Length in Words (Test)	Vocab Size Test	Token Ratio (Test)	Character Ratio (Test)
ArchiMob baseline Scherrer et al.	Bern	29.83	5	Dieht	ArchiMob	historical interviews	manual	Yes	Zurich area	No	22	N/A	1	no	0	0	6.62	N/A	N/A	1757	0.096749	N/A	3.1E-05	0	7811	5.747608536	1757	1	0.0011018
ArchiMob baseline Scherrer et al.	Basel	45.4	2	Dieht	ArchiMob	historical interviews	manual	Yes	Zurich area	No	22	N/A	1	no	0	0	6.62	N/A	N/A	2697	0.096749	N/A	3.1E-05	0	12255	5.676239	2697	1	0.0005368
ArchiMob baseline Scherrer et al.	Grisons	49.68	1	Dieht	ArchiMob	historical interviews	manual	Yes	Zurich area	No	22	N/A	1	no	0	0	6.62	N/A	N/A	900	0.096749	N/A	3.1E-05	0	3037	6.886621315	900	1	0.0021887
ArchiMob baseline Scherrer et al.	Lucerne	22.37	6	Dieht	ArchiMob	historical interviews	manual	Yes	Zurich area	No	22	N/A	1	no	0	0	6.62	N/A	N/A	2515	0.096749	N/A	3.1E-05	0	13095	7.202970297	2515	1	0.0006819
ArchiMob baseline Scherrer et al.	Uri	30.46	4	Dieht	ArchiMob	historical interviews	manual	Yes	Zurich area	No	22	N/A	1	no	0	0	6.62	N/A	N/A	1969	0.096749	N/A	3.1E-05	0	10155	8.36490939	1969	1	0.0007124
ArchiMob baseline Scherrer et al.	Valais	36.96	3	Dieht	ArchiMob	historical interviews	manual	Yes	Zurich area	No	22	N/A	1	no	0	0	6.62	N/A	N/A	2236	0.096749	N/A	3.1E-05	0	11488	7.846994536	2236	1	0.0007826
XLSR Lexicon-Free Khosravani et al.	Northwestern Swiss	17.2	3	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR Lexicon-Free Khosravani et al.	Bern	18.2	5	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR Lexicon-Free Khosravani et al.	Zurich	16.9	2	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR Lexicon-Free Khosravani et al.	Grisons	20.8	7	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR Lexicon-Free Khosravani et al.	Eastern Swiss	14.7	1	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR Lexicon-Free Khosravani et al.	Valais	19.3	6	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

System Name	Dialect	Performance	Performance Rank	Writing System	Corpus Name	Corpus Topic	Automatic vs. Manual Transcription	Known Transcriber Differences	Training Dialects	Overlap in Training and Test in Terms of Dialect	Number of Different Speakers (Training Overall)	Number of Different Speakers (Training Per Dialect)	Different Speakers Test	Overlap between Test and Training Speakers	Training Set Size (Utterances Per Dialect)	Training Set Size (Words Per Dialect)	Mean Utterance Length in Words (Training Overall)	Mean Utterance Length in Words (Training per Dialect)	Vocab Size (Training Overall)	Vocab Size (Training Per Dialect)	Token Ratio (Training Overall)	Token Ratio (Training Per Dialect)	Charcter Ratio (Training Overall)	Character Ratio (Training Per Dialect)	Test Set Size (in Words)	Mean Utterance Length in Words (Test)	Vocab Size Test	Token Ratio (Test)	Character Ratio (Test)
XLSR Lexicon-Free Khosravani et al.	Central Swiss	17.4	4	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR Baseline Khosravani et al.	Northwestern Swiss	15.8	3	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR Baseline Khosravani et al.	Bern	17.1	5	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR Baseline Khosravani et al.	Zurich	15.7	2	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR Baseline Khosravani et al.	Grisons	20.4	7	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR Baseline Khosravani et al.	Eastern Swiss	13.6	1	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR Baseline Khosravani et al.	Valais	17.5	6	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR Baseline Khosravani et al.	Central Swiss	16.3	4	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

System Name	Dialect	Performance	Performance Rank	Writing System	Corpus Name	Corpus Topic	Automatic vs. Manual Transcription	Known Transcriber Differences	Training Dialects	Overlap in Training and Test in Terms of Dialect	Number of Different Speakers (Training Overall)	Number of Different Speakers (Training Per Dialect)	Different Speakers Test	Overlap between Test and Training Speakers	Training Set Size (Utterances Per Dialect)	Training Set Size (Words Per Dialect)	Mean Utterance Length in Words (Training Overall)	Mean Utterance Length in Words (Training per Dialect)	Vocab Size (Training Overall)	Vocab Size (Training Per Dialect)	Token Ratio (Training Overall)	Token Ratio (Training Per Dialect)	Charcter Ratio (Training Overall)	Character Ratio (Training Per Dialect)	Test Set Size (in Words)	Mean Utterance Length in Words (Test)	Vocab Size Test	Token Ratio (Test)	Character Ratio (Test)
XLSR manual normalization Khosravani et al.	Northwestern Swiss	13.7	2	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR manual normalization Khosravani et al.	Bern	15.9	5	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR manual normalization Khosravani et al.	Zurich	13.8	3	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR manual normalization Khosravani et al.	Grisons	19.6	7	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR manual normalization Khosravani et al.	Eastern Swiss	11.6	1	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR manual normalization Khosravani et al.	Valais	16.5	6	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR manual normalization Khosravani et al.	Central Swiss	14.8	4	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR automatic normalization Khosravani et al.	Northwestern Swiss	12.6	3	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

System Name	Dialect	Performance	Performance Rank	Writing System	Corpus Name	Corpus Topic	Automatic vs. Manual Transcription	Known Transcriber Differences	Training Dialects	Overlap in Training and Test in Terms of Dialect	Number of Different Speakers (Training Overall)	Number of Different Speakers (Training Per Dialect)	Different Speakers Test	Overlap between Test and Training Speakers	Training Set Size (Utterances Per Dialect)	Training Set Size (Words Per Dialect)	Mean Utterance Length in Words (Training Overall)	Mean Utterance Length in Words (Training per Dialect)	Vocab Size (Training Overall)	Vocab Size (Training Per Dialect)	Token Ratio (Training Overall)	Token Ratio (Training Per Dialect)	Charcter Ratio (Training Overall)	Character Ratio (Training Per Dialect)	Test Set Size (in Words)	Mean Utterance Length in Words (Test)	Vocab Size Test	Token Ratio (Test)	Character Ratio (Test)
XLSR automatic normalization Khosravani et al.	Bern	15.1	5	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR automatic normalization Khosravani et al.	Zurich	12.2	2	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR automatic normalization Khosravani et al.	Grisons	18.1	7	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR automatic normalization Khosravani et al.	Eastern Swiss	10.4	1	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR automatic normalization Khosravani et al.	Valais	15.7	6	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
XLSR automatic normalization Khosravani et al.	Central Swiss	13.7	4	in dialect	TV assistant	short commands, scripted or spontaneous	manual	No	Mixed: Northwestern Swiss, Bern, Zürich, Graubünden, Eastern Swiss, Valais, Central Swiss	Yes	3817	N/A	N/A	yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System a Schraner et al.	Basel	29.99	4	Standard German	STT4SG-350	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System a Schraner et al.	Bern	34.83	7	Standard German	STT4SG-351	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System a Schraner et al.	Grisons	28.72	3	Standard German	STT4SG-352	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System a Schraner et al.	Central	26.92	1	Standard German	STT4SG-353	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System a Schraner et al.	East	31.33	5	Standard German	STT4SG-354	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System a Schraner et al.	Valais	31.56	6	Standard German	STT4SG-355	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System a Schraner et al.	Zurich	27.7	2	Standard German	STT4SG-356	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System b Schraner et al.	Basel	29.47	6	Standard German	STT4SG-350	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

System Name	Dialect	Performance	Performance Rank	Writing System	Corpus Name	Corpus Topic	Automatic vs. Manual Transcription	Known Transcriber Differences	Training Dialects	Overlap in Training and Test in Terms of Dialect	Number of Different Speakers (Training Overall)	Number of Different Speakers (Training Per Dialect)	Different Speakers Test	Overlap between Test and Training Speakers	Training Set Size (Utterances Per Dialect)	Training Set Size (Words Per Dialect)	Mean Utterance Length in Words (Training Overall)	Mean Utterance Length in Words (Training per Dialect)	Vocab Size (Training Overall)	Vocab Size (Training Per Dialect)	Token Ratio (Training Overall)	Token Ratio (Training Per Dialect)	Charcter Ratio (Training Overall)	Character Ratio (Training Per Dialect)	Test Set Size (in Words)	Mean Utterance Length in Words (Test)	Vocab Size Test	Token Ratio (Test)	Character Ratio (Test)
System b Schraner et al.	Bern	28.75	5	Standard German	STT4SG-351	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System b Schraner et al.	Grisons	25.79	3	Standard German	STT4SG-352	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System b Schraner et al.	Central	24.52	2	Standard German	STT4SG-353	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System b Schraner et al.	East	28.1	4	Standard German	STT4SG-354	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System b Schraner et al.	Valais	34.38	7	Standard German	STT4SG-355	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System b Schraner et al.	Zurich	24.36	1	Standard German	STT4SG-356	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System c Schraner et al.	Basel	28.35	5	Standard German	STT4SG-350	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System c Schraner et al.	Bern	31.49	7	Standard German	STT4SG-351	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System c Schraner et al.	Grisons	24.71	2	Standard German	STT4SG-352	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System c Schraner et al.	Central	24.21	1	Standard German	STT4SG-353	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System c Schraner et al.	East	26.7	4	Standard German	STT4SG-354	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System c Schraner et al.	Valais	30.42	6	Standard German	STT4SG-355	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System c Schraner et al.	Zurich	24.94	3	Standard German	STT4SG-356	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System d Schraner et al.	Basel	28.58	5	Standard German	STT4SG-350	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System d Schraner et al.	Bern	28.72	6	Standard German	STT4SG-351	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System d Schraner et al.	Grisons	24.96	3	Standard German	STT4SG-352	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System d Schraner et al.	Central	23.63	1	Standard German	STT4SG-353	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System d Schraner et al.	East	26.76	4	Standard German	STT4SG-354	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System d Schraner et al.	Valais	32.76	7	Standard German	STT4SG-355	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
System d Schraner et al.	Zurich	24.85	2	Standard German	STT4SG-356	web crawl	N/A	No	N/A	N/A	N/A	N/A	10	No	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FHNW XLS-R Schraner et al.	Basel	16.3	5	Standard German	SwissDial, SDS-200, SPC, STT4SG-350	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FHNW XLS-R Schraner et al.	Bern	15.74	4	Standard German	SwissDial, SDS-200, SPC, STT4SG-351	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FHNW XLS-R Schraner et al.	Grisons	14.32	3	Standard German	SwissDial, SDS-200, SPC, STT4SG-352	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

System Name	Dialect	Performance	Performance Rank	Writing System	Corpus Name	Corpus Topic	Automatic vs. Manual Transcription	Known Transcriber Differences	Training Dialects	Overlap in Training and Test in Terms of Dialect	Number of Different Speakers (Training Overall)	Number of Different Speakers (Training Per Dialect)	Different Speakers Test	Overlap between Test and Training Speakers	Training Set Size (Utterances Per Dialect)	Training Set Size (Words Per Dialect)	Mean Utterance Length in Words (Training Overall)	Mean Utterance Length in Words (Training per Dialect)	Vocab Size (Training Overall)	Vocab Size (Training Per Dialect)	Token Ratio (Training Overall)	Token Ratio (Training Per Dialect)	Charcter Ratio (Training Overall)	Character Ratio (Training Per Dialect)	Test Set Size (in Words)	Mean Utterance Length in Words (Test)	Vocab Size Test	Token Ratio (Test)	Character Ratio (Test)
FHNW XLS-R Schraner et al.	Central	13.26	1	Standard German	SwissDial, SDS-200, SPC, STT4SG-353	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FHNW XLS-R Schraner et al.	East	16.45	6	Standard German	SwissDial, SDS-200, SPC, STT4SG-354	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FHNW XLS-R Schraner et al.	Valais	17.75	7	Standard German	SwissDial, SDS-200, SPC, STT4SG-355	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FHNW XLS-R Schraner et al.	Zurich	13.41	2	Standard German	SwissDial, SDS-200, SPC, STT4SG-356	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FHNW Transformer Schraner et al.	Basel	21.24	6	Standard German	SwissDial, SDS-200, SPC, STT4SG-350	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FHNW Transformer Schraner et al.	Bern	20.96	5	Standard German	SwissDial, SDS-200, SPC, STT4SG-351	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FHNW Transformer Schraner et al.	Grisons	17.29	2	Standard German	SwissDial, SDS-200, SPC, STT4SG-352	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FHNW Transformer Schraner et al.	Central	16.37	1	Standard German	SwissDial, SDS-200, SPC, STT4SG-353	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FHNW Transformer Schraner et al.	East	18.58	4	Standard German	SwissDial, SDS-200, SPC, STT4SG-354	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

System Name	Dialect	Performance	Performance Rank	Writing System	Corpus Name	Corpus Topic	Automatic vs. Manual Transcription	Known Transcriber Differences	Training Dialects	Overlap in Training and Test in Terms of Dialect	Number of Different Speakers (Training Overall)	Number of Different Speakers (Training Per Dialect)	Different Speakers Test	Overlap between Test and Training Speakers	Training Set Size (Utterances Per Dialect)	Training Set Size (Words Per Dialect)	Mean Utterance Length in Words (Training Overall)	Mean Utterance Length in Words (Training per Dialect)	Vocab Size (Training Overall)	Vocab Size (Training Per Dialect)	Token Ratio (Training Overall)	Token Ratio (Training Per Dialect)	Charcter Ratio (Training Overall)	Character Ratio (Training Per Dialect)	Test Set Size (in Words)	Mean Utterance Length in Words (Test)	Vocab Size Test	Token Ratio (Test)	Character Ratio (Test)
FHNW Transformer Schraner et al.	Valais	22.64	7	Standard German	SwissDial, SDS-200, SPC, STT4SG-355	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FHNW Transformer Schraner et al.	Zurich	17.3	3	Standard German	SwissDial, SDS-200, SPC, STT4SG-356	news, Wikipedia, wide topic range, translations, parlamentary debates	manual	No	All	Yes	N/A	N/A	10	No	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
whisper-large Sicard et al.	Aargau	75	6	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	11121	0	0	0	0	28132	10.23726346	11121	1	0.0004097
whisper-large Sicard et al.	Bern	81	5	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10614	0	0	0	0	28052	10.38962963	10614	1	0.0004395
whisper-large Sicard et al.	Basel-Stadt	90	1	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10316	0	0	0	0	30233	11.1437523	10316	1	0.0004177
whisper-large Sicard et al.	Grisons	75	2	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	33503	0	0	0	0	132717	12.66988067	33503	1	0.0001011
whisper-large Sicard et al.	Lucerne	90	7	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10372	0	0	0	0	29223	10.76353591	10372	1	0.0004285
whisper-large Sicard et al.	St. Gallen	93	8	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	11010	0	0	0	0	29176	10.60174419	11010	1	0.0003749
whisper-large Sicard et al.	Valais	82	4	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10988	0	0	0	0	28469	10.34108246	10988	1	0.0004106
whisper-large Sicard et al.	Zurich	76	3	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	14122	0	0	0	0	43429	10.68364084	14122	1	0.0002896
whisper-medium Sicard et al.	Aargau	88	6	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	11121	0	0	0	0	28132	10.23726346	11121	1	0.0004097
whisper-medium Sicard et al.	Bern	83	4	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10614	0	0	0	0	28052	10.38962963	10614	1	0.0004395

System Name	Dialect	Performance	Performance Rank	Writing System	Corpus Name	Corpus Topic	Automatic vs. Manual Transcription	Known Transcriber Differences	Training Dialects	Overlap in Training and Test in Terms of Dialect	Number of Different Speakers (Training Overall)	Number of Different Speakers (Training Per Dialect)	Different Speakers Test	Overlap between Test and Training Speakers	Training Set Size (Utterances Per Dialect)	Training Set Size (Words Per Dialect)	Mean Utterance Length in Words (Training Overall)	Mean Utterance Length in Words (Training per Dialect)	Vocab Size (Training Overall)	Vocab Size (Training Per Dialect)	Token Ratio (Training Overall)	Token Ratio (Training Per Dialect)	Charcter Ratio (Training Overall)	Character Ratio (Training Per Dialect)	Test Set Size (in Words)	Mean Utterance Length in Words (Test)	Vocab Size Test	Token Ratio (Test)	Character Ratio (Test)
whisper-medium Sicard et al.	Basel-Stadt	66	2	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10316	0	0	0	0	30233	11.1437523	10316	1	0.0004177
whisper-medium Sicard et al.	Grisons	65	1	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	33503	0	0	0	0	132717	12.66988067	33503	1	0.0001011
whisper-medium Sicard et al.	Lucerne	86	7	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10372	0	0	0	0	29223	10.76353591	10372	1	0.0004285
whisper-medium Sicard et al.	St. Gallen	93	8	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	11010	0	0	0	0	29176	10.60174419	11010	1	0.0003749
whisper-medium Sicard et al.	Valais	83	5	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10988	0	0	0	0	28469	10.34108246	10988	1	0.0004106
whisper-medium Sicard et al.	Zurich	76	3	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	14122	0	0	0	0	43429	10.68364084	14122	1	0.0002896
whisper-small Sicard et al.	Aargau	90	7	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	11121	0	0	0	0	28132	10.23726346	11121	1	0.0004097
whisper-small Sicard et al.	Bern	85	4	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10614	0	0	0	0	28052	10.38962963	10614	1	0.0004395
whisper-small Sicard et al.	Basel-Stadt	73	2	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10316	0	0	0	0	30233	11.1437523	10316	1	0.0004177
whisper-small Sicard et al.	Grisons	76	1	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	33503	0	0	0	0	132717	12.66988067	33503	1	0.0001011
whisper-small Sicard et al.	Lucerne	92	8	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10372	0	0	0	0	29223	10.76353591	10372	1	0.0004285
whisper-small Sicard et al.	St. Gallen	87	5	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	11010	0	0	0	0	29176	10.60174419	11010	1	0.0003749
whisper-small Sicard et al.	Valais	88	6	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10988	0	0	0	0	28469	10.34108246	10988	1	0.0004106

System Name	Dialect	Performance	Performance Rank	Writing System	Corpus Name	Corpus Topic	Automatic vs. Manual Transcription	Known Transcriber Differences	Training Dialects	Overlap in Training and Test in Terms of Dialect	Number of Different Speakers (Training Overall)	Number of Different Speakers (Training Per Dialect)	Different Speakers Test	Overlap between Test and Training Speakers	Training Set Size (Utterances Per Dialect)	Training Set Size (Words Per Dialect)	Mean Utterance Length in Words (Training Overall)	Mean Utterance Length in Words (Training per Dialect)	Vocab Size (Training Overall)	Vocab Size (Training Per Dialect)	Token Ratio (Training Overall)	Token Ratio (Training Per Dialect)	Charcter Ratio (Training Overall)	Character Ratio (Training Per Dialect)	Test Set Size (in Words)	Mean Utterance Length in Words (Test)	Vocab Size Test	Token Ratio (Test)	Character Ratio (Test)
whisper-small Sicard et al.	Zurich	76	3	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	14122	0	0	0	0	43429	10.68364084	14122	1	0.0002896
whisper-base Sicard et al.	Aargau	98	8	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	11121	0	0	0	0	28132	10.23726346	11121	1	0.0004097
whisper-base Sicard et al.	Bern	90	5	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10614	0	0	0	0	28052	10.38962963	10614	1	0.0004395
whisper-base Sicard et al.	Basel-Stadt	80	2	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10316	0	0	0	0	30233	11.1437523	10316	1	0.0004177
whisper-base Sicard et al.	Grisons	78	1	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	33503	0	0	0	0	132717	12.66988067	33503	1	0.0001011
whisper-base Sicard et al.	Lucerne	98	6	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10372	0	0	0	0	29223	10.76353591	10372	1	0.0004285
whisper-base Sicard et al.	St. Gallen	86	3	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	11010	0	0	0	0	29176	10.60174419	11010	1	0.0003749
whisper-base Sicard et al.	Valais	97	7	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10988	0	0	0	0	28469	10.34108246	10988	1	0.0004106
whisper-base Sicard et al.	Zurich	85	4	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	14122	0	0	0	0	43429	10.68364084	14122	1	0.0002896
whisper-tiny Sciard et al.	Aargau	115	6	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	11121	0	0	0	0	28132	10.23726346	11121	1	0.0004097
whisper-tiny Sciard et al.	Bern	100	4	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10614	0	0	0	0	28052	10.38962963	10614	1	0.0004395
whisper-tiny Sciard et al.	Basel-Stadt	90	1	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10316	0	0	0	0	30233	11.1437523	10316	1	0.0004177
whisper-tiny Sciard et al.	Grisons	90	2	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	33503	0	0	0	0	132717	12.66988067	33503	1	0.0001011

System Name	Dialect	Performance	Performance Rank	Writing System	Corpus Name	Corpus Topic	Automatic vs. Manual Transcription	Known Transcriber Differences	Training Dialects	Overlap in Training and Test in Terms of Dialect	Number of Different Speakers (Training Overall)	Number of Different Speakers (Training Per Dialect)	Different Speakers Test	Overlap between Test and Training Speakers	Training Set Size (Utterances Per Dialect)	Training Set Size (Words Per Dialect)	Mean Utterance Length in Words (Training Overall)	Mean Utterance Length in Words (Training per Dialect)	Vocab Size (Training Overall)	Vocab Size (Training Per Dialect)	Token Ratio (Training Overall)	Token Ratio (Training Per Dialect)	Charcter Ratio (Training Overall)	Character Ratio (Training Per Dialect)	Test Set Size (in Words)	Mean Utterance Length in Words (Test)	Vocab Size Test	Token Ratio (Test)	Character Ratio (Test)
whisper-tiny Sciard et al.	Lucerne	16	5	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10372	0	0	0	0	29223	10.76353591	10372	1	0.0004285
whisper-tiny Sciard et al.	St. Gallen	94	7	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	11010	0	0	0	0	29176	10.60174419	11010	1	0.0003749
whisper-tiny Sciard et al.	Valais	121	8	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	10988	0	0	0	0	28469	10.34108246	10988	1	0.0004106
whisper-tiny Sciard et al.	Zurich	98	3	Dialect	Swiss Dial	news, Wikipedia, wide topic range, translations	manual	No	Zero-Shot	No	0	0	N/A	No	0	Zero-Shot	0	0	0	14122	0	0	0	0	43429	10.68364084	14122	1	0.0002896
NNET-DISC Nigmatulina	Valais	71	14	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	1464	1464	6.7841	7.84699454	4452	4452	0.566125	0.000591	1.9E-05	0.00078265	1464	7.846994536	4452	0.127	0.0007826
NNET-DISC Nigmatulina	St. Gallen	69	13	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	3089	3089	6.7841	6.23599871	3606	3606	0.566125	0.000352	1.9E-05	0.000430426	3089	6.235998705	3606	0.16	0.0004304
NNET-DISC Nigmatulina	Luzern	62	12	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	5	5	Yes	6959	6959	6.7841	6.94252048	7085	7085	0.566125	0.00014	1.9E-05	0.000192523	6959	6.942520477	7085	0.144	0.0001925
NNET-DISC Nigmatulina	Bern	61	11	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	6	6	Yes	11417	11417	6.7841	7.02119646	10923	10923	0.566125	8.46E-05	1.9E-05	0.000114499	11417	7.021196461	10923	0.142	0.0001145
NNET-DISC Nigmatulina	Glarus	59	10	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	2	2	Yes	4561	4561	6.7841	5.88379741	4456	4456	0.566125	0.000253	1.9E-05	0.000345435	4561	5.883797413	4456	0.17	0.0003454
NNET-DISC Nigmatulina	Schwyz	59	9	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	1061	1061	6.7841	9.16588124	2298	2298	0.566125	0.000698	1.9E-05	0.000814093	1061	9.165881244	2298	0.109	0.0008141
NNET-DISC Nigmatulina	Basel-Stadt	58	8	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	5	5	Yes	11583	11583	6.7841	6.44375378	10637	10637	0.566125	9.09E-05	1.9E-05	0.000112671	11583	6.443753777	10637	0.155	0.0001127
NNET-DISC Nigmatulina	Uri	58	7	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	1214	1214	6.7841	8.36490939	1969	1969	0.566125	0.000668	1.9E-05	0.000712443	1214	8.36490939	1969	0.12	0.0007124
NNET-DISC Nigmatulina	Basel-Landschaft	57	6	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	2047	2047	6.7841	7.61016121	2686	2686	0.566125	0.000435	1.9E-05	0.000567537	2047	7.610161212	2686	0.131	0.0005675
NNET-DISC Nigmatulina	Schaffhausen	57	5	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	1806	1806	6.7841	9.57530454	3020	3020	0.566125	0.000392	1.9E-05	0.000375084	1806	9.57530454	3020	0.104	0.0003751
NNET-DISC Nigmatulina	Aargau	57	4	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	6	6	Yes	15590	15590	6.7841	6.59108403	13221	13221	0.566125	6.6E-05	1.9E-05	8.66818E-05	15590	6.591084028	13221	0.152	8.668E-05
NNET-DISC Nigmatulina	Nidwalden	56	3	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	2	2	Yes	3529	3529	6.7841	7.56446585	4394	4394	0.566125	0.000254	1.9E-05	0.000362605	3529	7.564465854	4394	0.132	0.0003626
NNET-DISC Nigmatulina	Zurich	55	2	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	14	14	Yes	29415	29967	6.7841	6.48104708	20320	20320	0.566125	3.56E-05	1.9E-05	3.11456E-05	29967	6.481047085	20320	0.157	3.115E-05
NNET-DISC Nigmatulina	Grisons	31	1	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	441	48313	6.7841	6.88662132	900	900	0.566125	0.002234	1.9E-05	0.002188679	48313	6.886621315	900	15.91	0.0021887
TDNN-iVector Nigmatulina	Valais	64	14	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	1464	1464	6.7841	7.84699454	4452	4452	0.566125	0.000591	1.9E-05	0.00078265	1464	7.846994536	4452	0.127	0.0007826
TDNN-iVector Nigmatulina	St. Gallen	60	7	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	3089	3089	6.7841	6.23599871	3606	3606	0.566125	0.000352	1.9E-05	0.000430426	3089	6.235998705	3606	0.16	0.0004304
TDNN-iVector Nigmatulina	Luzern	55	11	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	5	5	Yes	6959	6959	6.7841	6.94252048	7085	7085	0.566125	0.00014	1.9E-05	0.000192523	6959	6.942520477	7085	0.144	0.0001925
TDNN-iVector Nigmatulina	Bern	51	9	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	6	6	Yes	11417	11417	6.7841	7.02119646	10923	10923	0.566125	8.46E-05	1.9E-05	0.000114499	11417	7.021196461	10923	0.142	0.0001145
TDNN-iVector Nigmatulina	Glarus	51	4	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	2	2	Yes	4561	4561	6.7841	5.88379741	4456	4456	0.566125	0.000253	1.9E-05	0.000345435	4561	5.883797413	4456	0.17	0.0003454
TDNN-iVector Nigmatulina	Schwyz	49	12	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	1061	1061	6.7841	9.16588124	2298	2298	0.566125	0.000698	1.9E-05	0.000814093	1061	9.165881244	2298	0.109	0.0008141
TDNN-iVector Nigmatulina	Basel-Stadt	49	8	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	5	5	Yes	11583	11583	6.7841	6.44375378	10637	10637	0.566125	9.09E-05	1.9E-05	0.000112671	11583	6.443753777	10637	0.155	0.0001127
TDNN-iVector Nigmatulina	Uri	49	13	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	1214	1214	6.7841	8.36490939	1969	1969	0.566125	0.000668	1.9E-05	0.000712443	1214	8.36490939	1969	0.12	0.0007124

System Name	Dialect	Performance	Performance Rank	Writing System	Corpus Name	Corpus Topic	Automatic vs. Manual Transcription	Known Transcriber Differences	Training Dialects	Overlap in Training and Test in Terms of Dialect	Number of Different Speakers (Training Overall)	Number of Different Speakers (Training Per Dialect)	Different Speakers Test	Overlap between Test and Training Speakers	Training Set Size (Utterances Per Dialect)	Training Set Size (Words Per Dialect)	Mean Utterance Length in Words (Training Overall)	Mean Utterance Length in Words (Training per Dialect)	Vocab Size (Training Overall)	Vocab Size (Training Per Dialect)	Token Ratio (Training Overall)	Token Ratio (Training Per Dialect)	Charcter Ratio (Training Overall)	Character Ratio (Training Per Dialect)	Test Set Size (in Words)	Mean Utterance Length in Words (Test)	Vocab Size Test	Token Ratio (Test)	Character Ratio (Test)
TDNN-iVector Nigmatulina	Basel-Landschaft	43	5	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	2047	2047	6.7841	7.61016121	2686	2686	0.566125	0.000435	1.9E-05	0.000567537	2047	7.610161212	2686	0.131	0.0005675
TDNN-iVector Nigmatulina	Schaffhausen	42	3	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	1806	1806	6.7841	9.57530454	3020	3020	0.566125	0.000392	1.9E-05	0.000375084	1806	9.57530454	3020	0.104	0.0003751
TDNN-iVector Nigmatulina	Aargau	42	6	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	6	6	Yes	15590	15590	6.7841	6.59108403	13221	13221	0.566125	6.6E-05	1.9E-05	8.66818E-05	15590	6.591084028	13221	0.152	8.668E-05
TDNN-iVector Nigmatulina	Nidwalden	41	10	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	2	2	Yes	3529	3529	6.7841	7.56446585	4394	4394	0.566125	0.000254	1.9E-05	0.000362605	3529	7.564465854	4394	0.132	0.0003626
TDNN-iVector Nigmatulina	Zurich	41	2	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	14	14	Yes	29415	29967	6.7841	6.48104708	20320	20320	0.566125	3.56E-05	1.9E-05	3.11456E-05	29967	6.481047085	20320	0.157	3.115E-05
TDNN-iVector Nigmatulina	Grisons	22	1	Dieht	ArchiMob	historical interviews	manual	Yes	All	Yes	43	1	1	Yes	441	48313	6.7841	6.88662132	900	900	0.566125	0.002234	1.9E-05	0.002188679	48313	6.886621315	900	15.91	0.0021887