Does Whisper Understand Swiss German? An Automatic, Qualitative and Human Evaluation

Eyal Liron Dolev
Linguistics Center Zurich
German Department
University of Zurich
eyalliron.dolev@uzh.ch

Clemens Fidel Lutz
Department of Computational Linguistics
Phonetics & Speech Sciences
University of Zurich
clemensfidel.lutz@uzh.ch

Noëmi Aepli
Department of Computational Linguistics
University of Zurich
noemi.aepli@uzh.ch

Abstract

Whisper is a state-of-the-art automatic speech recognition (ASR) model (Radford et al., 2022). Although Swiss German dialects are allegedly not part of Whisper's training data, preliminary experiments showed that Whisper can transcribe Swiss German quite well, with the output being a speech translation into Standard German. To gain a better understanding of Whisper's performance on Swiss German, we systematically evaluate it using automatic, qualitative, and human evaluation. We test its performance on three existing test sets: SwissDial (Dogan-Schönberger et al., 2021), STT4SG-350 (Plüss et al., 2023), and Swiss Parliaments Corpus (Plüss et al., 2021). In addition, we create a new test set for this work, based on short mock clinical interviews.

For automatic evaluation, we used word error rate (WER) and BLEU. In the qualitative analysis, we discuss Whisper's strengths and weaknesses and analyze some output examples. For the human evaluation, we conducted a survey with 28 participants who were asked to evaluate Whisper's performance.

All of our evaluations suggest that Whisper is a viable ASR system for Swiss German, so long as the Standard German output is desired.

1 Introduction

Swiss German is the name of a group of Alemannic (High German) dialects spoken in German-speaking Switzerland by around 5.5 million people.[1] German-speaking Switzerland displays a state of diglossia (Ferguson, 1959; Rash, 1998), more specifically a medial diglossia: spoken contexts evoke Swiss German, written contexts evoke Standard German (Kolde, 1983; Haas, 2004). According to this principle, Swiss German is used as a spoken language in almost all settings, with the exception of some restricted, specific formal settings in which Standard German is spoken, e.g., on the news or at school, as well as a *lingua franca* with non Swiss German speakers (Hogg et al., 1984).

Swiss German has no spoken standard variety and no written variety, and therefore no orthographic norms. In writing, Standard German is used. Thus, whenever spoken language (Swiss German) has to be written down, e.g., subtitles to a TV program or minutes of a meeting, Standard German is used. If Swiss German is written, it happens in situations that are conceptually spoken ("*konzeptionell mündlich;*" Koch and Oesterreicher, 1994) and which are situated on the immediacy-end of Koch and Oesterreichs's communication model (*Nähe-Distanz-Modell*, cf. Koch and Oesterreicher, 1985), e.g., ads or chat messages, cf. Ueberwasser and Stark (2017), which created a corpus of text messages written in Swiss German.

To summarize, in German-speaking Switzerland there is a state of medial diglossia: the spoken language is Swiss German (a group of dialects with no standard variety); the written language is a Swiss vareity of Standard German. Swiss German and Standard German are, although genetically and systematically very close, two different languages, whereas only Standard German has a codified written form (Berthele, 2004). The task of putting down Swiss German speech to written form is, therefore, not a transcription task, but rather a translation task, translating Swiss German to Standard German. This spoken–written juxtaposition of Swiss German and Standard German explains why almost all the automatic speech recognition

[1] Bundesamt für Statistik: Hauptsprachen seit 1910, accessed on 23.04.2024

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 28–40

efforts for Swiss German until now have dealt with Swiss German speech to Standard German text (see Section 2).

Whisper is a state-of-the-art multilingual model for automatic speech recognition (ASR) (Radford et al., 2022). Although Swiss German is not officially part of Whisper's training data[2], in preliminary trials, we observed that Whisper could recognize Swiss German quite well, with the output produced being Standard German. According to Ruder (2024) most large language models (LLMs) have likely encountered some data for most languages available on the web, which is probably the case here too.

We intentionally refrain from attempting to fine-tune Whisper. Not only did Sicard et al. (2023)'s fine-tuning attempts of Whisper on Swiss German data worsen the model's performance; we find Whisper's zero-shot performance on Swiss German, at this stage, already impressive and applicable. Before any costly GPU hours are spent in an attempt to improve Whisper, we think it should first be scrutinized and analyzed in its current state.

In this work, we evaluate Whisper's performance on Swiss German audio in different settings and modes. We automatically evaluated Whisper on three large corpora, namely SPC (Plüss et al., 2021), STT4SG-350 (Plüss et al., 2023), and SwissDial (Dogan-Schönberger et al., 2021), measuring word error rate (WER) and BLEU.

To test Whisper on real-life spoken language, we created a new test set for which we translated into Standard German mock clinical interviews held in Swiss German. The total length of the interviews is approx. 30 minutes. To test Whisper's performance on this test set, we offer a qualitative analysis of Whisper's output and a human evaluation based on a survey ($n = 28$).

2 Previous Work

ASR for Swiss German is an ambiguous term. While the audio input to the system is always Swiss German, the text output can be: (a) dialectal writing – loosely phonemic representation of Swiss German; (b) normalized writing – transcriptions resembling standard German that are relatively consistent but distant from the acoustic signal (Nigmatulina et al., 2020); (c) Standard German translation.

In recent years, Swiss German has enjoyed a proper boom in the field of speech corpora, ASR and speech generation. The first major corpus with Swiss German audio was ArchiMob, which includes dialectal as well as normalized writing (Samardžić et al., 2016; Scherrer et al., 2019). Nigmatulina et al. (2020) used the ArchiMob corpus to compare systems producing dialectal and normalized writing and concluded that performance is better with standardized writing. Dogan-Schönberger et al. (2021) created SwissDial, a large corpus containing Standard German as well as Swiss German transcriptions in eight dialects.

Some work concentrated on ASR with Standard German speech translation and leveraged existing Transformer and XLS-R ASR models, fine-tuning them with Swiss German data. Plüss et al. (2021) published the "Swiss Parliaments Corpus", and experimented further with ASR models for Swiss German with Standard German output. Plüss et al. (2022) presented SDS-200, a corpus of Swiss German dialectal speech with Standard German text translations containing 200 hours of speech. They also experimented with training Transformer models and fine-tuning Wav2Vec2 XLS-R models on their data. Their best model (XLS-R) reached a WER of 21.6 and a BLEU score of 64.0. Recently, Plüss et al. (2023) presented the as-of-today largest corpus of Swiss German dialectal speech with Standard German text, containing 343 hours of speech. They fine-tuned a Wav2Vec2 XLS-R model on the corpus and reached a WER of 14.0 and a BLEU score of 74.7 on their test set.

Most recently, Sicard et al. (2023) turned to Whisper and tested it in a zero-shot setting on select Swiss German/Standard German test sets (SwissDial, SDS-200, SPC). Reportedly, their fine-tuning experiments on Whisper (medium version) worsened performance, leading the model to suffer from catastrophic forgetting.

3 Test Sets

To evaluate Whisper's performance on Swiss German, we test it using WER and BLEU on three test sets: SwissDial (Dogan-Schönberger et al., 2021), Swiss Parliaments Corpus (Plüss et al., 2021), STT4SG-350 (Plüss et al., 2023). We additionally created a new test set based on short Swiss German mock clinical interviews, which we additionally evaluate using a qualitative analysis and a human survey.

[2] https://github.com/openai/whisper, accessed on 23.04.2024

3.1 Mock Clinical Interviews

This work serves as a preparation step towards a large longitudinal study in the field of suicide prevention.[3] During this study, patients from a Zurich-based psychiatric clinic will be interviewed several times. We test how reliable and viable Whisper is for transcribing/translating these interviews.

To this end, i.e., to test Whisper in a naturalistic and applied setting containing spontaneous speech, we used mock clinical interviews that were held in Swiss German and recorded for instructional and training purposes in a total length of approx. 30 minutes. The interviews were recorded with three women interviewees using a lapel microphone[4] and simple convertible laptops. We, the authors of this work, then translated these interviews into Standard German according to some basic translation guidelines we created to maintain consistency. We call this ad-hoc test set "Mock Clinical Interviews".

This test set will be automatically evaluated using WER and BLEU as well as using a qualitative analysis to discuss Whisper's strengths and weaknesses in Swiss German, and a human evaluation, for which we conducted a survey ($n = 28$).

3.2 SwissDial

For the creation of SwissDial, eight speakers, speaking eight different dialects[5] were asked to translate Standard German prompts to their own dialects and then record the translations. The prompts were made of sentences crawled from the internet, encompassing different text genres: news stories, Wikipedia articles, weather reports and short stories (Dogan-Schönberger et al., 2021). Because the prompts were translated into Swiss German by each of the speakers, sometimes greater departures from the Standard German source occur. See Figure 1, containing the first three dialect entries from the first entry in the corpus, for an example.

As can be seen in Figure 1, the German word *derzeit* "currently" was translated by the different dialect speakers as *zur Ziit*, *momentan* and *derziit*, respectively. One cannot, however, expect that Whisper translates all of these different Swiss German words back to the Standard German original, especially considering that the Swiss German words each have a closer Standard German equivalent (*zur Zeit*, *momentan* and *derzeit*, respectively).

To circumvent this problem and include prompts

```
{
  "id": 0,
  "de": "Derzeit␣ist␣er␣in␣
      ↪ \"Parasite\",␣dem␣
      ↪ Siegerfilm␣von␣Cannes,␣zu␣
      ↪ sehen.",
  "ch_sg": "Zur␣Ziit␣isch␣er␣in␣
      ↪ \"Parasite\",␣en␣
      ↪ Siegerfilm␣vo␣Cannes,␣
      ↪ zgseh.",
  "ch_be": "Momentan␣ischer␣in␣
      ↪ \"Parasite\"␣z␣gseh,␣em␣
      ↪ Siegerfium␣vo␣Cannes.",
  "ch_gr": "Derziit␣isch␣er␣in␣
      ↪ \"Parasite\",␣am␣
      ↪ Siegerfilm␣vu␣Cannes,␣z␣
      ↪ gseh.",
    ...
}
```

Figure 1: The first three dialectal translations of the first entry in the SwissDial corpus. The first word in the Standard German source ("de"), *derzeit*, is translated differently in each dialect: *zur ziit*, *momentan*, *derziit*.

that are less likely to contain major departures from the source, which might unfairly fail Whisper when the produced output is compared to the original prompt, we created an ad-hoc test set: We calculated for each Standard German prompt and its respective dialectal translations the chrF score (Popović, 2015) using SacreBLEU's implementation (Post, 2018). We then evaluated Whisper's performance on the 500 prompts with the best chrF scores for each dialect.

3.3 Swiss Parliaments Corpus

The Swiss Parliament Corpus (SPC) is a dataset containing sentences taken from speeches held at the Grand Council (Grosser Rat) of the Canton of Bern (Plüss et al., 2021).[6] Almost all speakers hold their speeches in Bernese German. For the creation of the corpus, Plüss et al. (2021) split the audio into segments, so-called sentences, whereas segments shorter than one second and longer than 15 seconds were discarded. The corpus creators also made sure that the speech segments were unique within the set. The speech segments were then force-aligned to the Standard German minutes (i.e., translations), which were created by the Canton of Bern. The result is parliament speeches split into segments (sentences) with their corresponding Standard German transla-

[3] MULTICAST

[4] RØDE smartLav+

[5] The dialects of Zurich, Bern, Basel, Aargau, Grisons, St. Gallen, Lucerne, and the Walser

[6] The name of the corpus is thus a misnomer – it is not a corpus representing the whole diversity of Swiss German.

tions from the minutes. We tested Whisper on the test set part of the corpus[7].

3.4 STT4SG-350

Like SPC, STT4SG-350[8] is a corpus containing single sentences of Swiss German speech with Standard German translations (Plüss et al., 2023). Unlike the former, STT4SG-350 includes an almost even split between seven different dialect regions.[9] The sentences produced by speakers were taken from Swiss newspapers and proceedings of two Swiss Parliaments. Participants, who were recruited either via a crowdsourcing platform or academic or personal channels as well as news ads, self-reported their dialect region, age group, gender, and where they grew up and/or went to school. The whole corpus consists of 343 hours of speech. We tested Whisper on the test set part which contains 34 hours of speech in approx. 25k sentences.

4 Evaluation

4.1 Automatic Evaluation

Usually, word error rate (WER) is used as a metric to automatically evaluate ASR systems. However, the type of ASR for Swiss German that we evaluate in this work is Swiss German audio input with Standard German text output – a speech translation task. This means, as is generally the case in translation, that it is not uncommon for a sentence to have several possible translations. Standard German translations of Swiss German are, in that sense, no different, although in many cases, there are clear one-to-one correspondences in vocabulary and grammatical structures between Standard and Swiss German. But when correspondences are ambiguous, the translator has to make a conscious decision on how to translate vocabulary or grammatical constructions. For example, Swiss German only has one tense referring to past events – the perfect. Standard German has, at least formally, two past tenses – the perfect and the preterite. The translator thus has to choose, according to context, how to translate the Swiss German perfect.

This ambiguity in translation, a typical problem in evaluating machine translation systems, makes the usual metric used for ASR systems – word error rate (WER) – not unproblematic. We thus additionally use BLEU (Papineni et al., 2002), a typical metric used to evaluate machine translation systems. This will also help compare the performance of Whisper to previous Swiss German ASR models, as previous work also reports WER and BLEU.

To compute WER, we used JiWER's[10] implementation. For BLEU we used SacreBLEU's implementation (Post, 2018).

Mode	WER	BLEU
Continuous recordings	**0.33**	**52.03**
Segmented clips	0.37	44.19

Table 1: Whisper's performance on our *Mock Clinical Interviews* test set, comparing continous recordings vs. segmented clips. Best results in bold.

4.2 Qualitative & Human Evaluation

In addition to testing Whisper's performance on several datasets and evaluating its performance automatically, we offer a qualitative and human evaluation of our Mock Clinical Interviews (see Section 3.1). In the qualitative evaluation, we will show examples of Whisper's output, analyze errors, and shed some light on the strengths and weaknesses of Whisper's performance.

Our human evaluation, in which we recruited 28 people – university students, colleagues, and acquaintances – via personal channels to evaluate Whisper's output, offers more informative feedback about how humans perceive Whisper's output.

5 Results: Automatic Evaluation

5.1 Mock Clinical Interviews

We tested Whisper's large-v3 model on our test set ("Mock Clinical Intverviews", see Section 3.1). We compared Whisper's performance on continuous recordings versus short clips containing single speech segments. Given segmented clips, WER and BLEU scores were 0.37 and 44.19, respectively. With the continuous recordings, WER and BLEU scores were 0.33 and 52.03, respectively, see Table 1. We conclude that Whisper performs better on longer, continuous recordings than on short clips.

This comes, however, at a slight risk of hallucinations: Four out of sixteen transcriptions/translations generated by Whisper included one sentence

[7] 6 hours, 3332 segments

[8] Standing for "Speech-to-text for Swiss German"

[9] These seven regions are Basel, Bern, Grisons, Central Switzerland, East Switzerland, Valais and Zurich.

[10] https://github.com/jitsi/jiwer, accessed on 23.04.2024

that was not uttered in the original audio, see Section 6.3 for more details.

5.2 SPC, STT4SG & SwissDial

We further tested Whisper's large-v3 model on the three other test sets: SPC, STT4SG-350, and SwissDial (see Section 3). The results, compared to results reported by other works, can be seen in Table 2. We always picked the best result reported in each of the other works.

Whisper's latest large model, version 3, outperforms Whisper's previous model, as reported by Sicard et al. (2023). However, fine-tuned Wav2Vec2 models on the SPC and the STT4SG-350 training sets outperform Whisper on the respective test sets, as reported by Plüss et al. (2023) and Schraner et al. (2022). Whisper does come close to the Conformer model pre-trained by Plüss et al. (2021) with a difference of only 1.7 *p.p.* and 1.6 *p.p.* in WER and BLEU, respectively.

For SPC, STT4SG-350, and SwissDial, we also computed WER and BLEU for each sentence separately and then computed the mean and standard deviation (so-called micro average). As can be seen in Table 3, the standard deviations for WER and BLEU are quite big, ranging at 0.24–0.25 for WER and 27.95–32.24 for BLEU. This shows that Whisper's performance measured in WER and BLEU fluctuates considerably. For some sentences in STT4SG-350 for example, BLEU scores went up to 100. See also Figures 2 and 3 in Appendix A.4.

It should be noted that the SPC corpus contains some considerable deviations between audio and reference translations, which were taken from the parliament's proceedings (see Section 3.3). For instance, in one clip[11], the heard audio is *und das isch schlächt*. The reference translation is "Das ist schlecht", excluding the coordinating conjunction *und* "and". Whisper perfectly transcribed this as "Und das ist schlecht", but this is penalized with a WER score of 0.33. It is not inconceivable, that the models trained by Plüss et al. (2021) learned these deviating translations, which might explain their better performance on the SPC test set. As the case may be, comparing WER and BLEU scores for SPC between Whisper's performance and Plüss et al. (2021) may raise concerns, and its meaningfulness can and should be questioned. For more examples of perfect output by Whisper penalized by diverging reference translations, see Table 10 in Appendix A.

For SwissDial, we also evaluated Whisper's performance on the different dialects. As can be seen in Table 4, the Grisons dialect has the best WER and BLEU scores; the Walser dialect has the worst scores.[12] Why Whisper performs differently on different dialects and which phonetic, phonological or grammatical traits affect Whisper's performance should be more closely examined in future work.

To conclude, we consider Whisper's results impressive, especially considering that it operates in a zero-shot setting. Its output is without doubt meaningful and useful.

6 Qulatitative Analysis

6.1 General Impression

In general, we were genuinely impressed with Whisper's performance. The Standard German output corresponds in meaning and style to the Swiss German audio to almost the full extent. Whisper generated entire error-free passages that are fluent, consistent in style, retain the original word order and correspond fully to the original (see Table 6 in Appendix A.1 for examples).

However, some things are not always consistent. For example, the Swiss German perfect tense is translated sometimes as the Standard German perfect tense and sometimes as the preterite. The output switches inconsistently between the two past forms within the same passage. See Table 7 in Appendix A.1 for examples.

We noticed that in certain cases, words are changed when translated to Standard German, even when the Swiss German word has an identical corresponding word in Standard German. One example of this is the Swiss German word *lässig* which is translated to Standard German *toll*, both meaning "cool, nice". In this case, this is desired since in Standard German, *lässig* means rather "casual, easy-going" – Swiss German *lässig* and Standard German *lässig* are false friends. Another example is the translation of Swiss German *Sache* to Standard German *Dinge*, both meaning "things", however, *Dinge* is used mostly for tangible things and in the given contexts *Sachen* would have been a better translation.

[11] `82495971-6523-4f96-be13-753b8bb564cf.flac`

[12] The Walser dialect is also considered in Switzerland the most difficult to understand.

Test Set	Model	Mode	WER	BLEU	
Mock Interviews	Whisper large-v3	zero-shot	0.372	44.3	This work
SPC	Conformer	pre-trained	0.278	58.6	Plüss et al. (2021)
	Wav2Vec2	**fine-tuned**	**0.237**	**60.7**	Schraner et al. (2022)
	Whisper large	zero-shot	0.332	55.6	Sicard et al. (2023)
	Whisper large-v3	zero-shot	0.295	57.0	This work
STT4SG-350	XLS-R	fine-tuned	0.153	72.2	Schraner et al. (2022)
	Wav2Vec2	**fine-tuned**	**0.140**	**74.7**	Plüss et al. (2023)
	Whisper large-v3	zero-shot	0.230	63.1	This work
SwissDial	Whisper large	zero-shot	0.294	56.2	Sicard et al. (2023)
	Whisper large-v3	**zero-shot**	**0.230**	**61.0**	This work

Table 2: WER (lower is better) and BLEU (higher is better) scores for our corpora, compared to results reported in previous works.

Test Set	WER	BLEU
SPC	0.30 (0.24)	54.01 (27.95)
STT4SG-350	0.24 (0.25)	60.61 (32.24)
SwissDial	0.25 (0.24)	57.23 (31.51)

Table 3: Mean and standard deviation WER and BLEU for the corpora when computed for each sentence separately.

Dialect	WER	BLEU
Aargau	0.272	55.40
Bern	0.210	64.95
Basel	0.209	63.24
Grisons	0.169	69.99
Lucerne	0.276	55.06
St. Gallen	0.209	64.03
Walser	0.297	53.46
Zurich	0.229	60.67

Table 4: WER and BLEU scores for each dialect in the SwissDial corpus.

6.2 Concise Style

We notice that Whisper's translations are of a style that is more concise than the original. This is especially noticeable in the removal of modal particles and conjunctions: Modal particles with little semantic content but with an information structural function like *halt* or *einfach* might disappear from the output. Conjunctions like *dann* "then" or *und* "and" are not always included. In one case, however, conjunctions and particles were hallucinated by Whisper. Whisper deals then inconsistently with particles and conjunctions, mostly removing them but rarely also adding them by hallucination.

It is a known phenomenon that during translation, the explicitness of cohesive markers, such as the particles and conjunctions mentioned above, can shift (Blum-Kulka, 1986). Leaving out such markers, as evidenced in Whisper's output, can be seen as a case of implicitation, cf. Lapshinova-Koltunski et al. (2022) (which refers to them as "discourse connectives"). If we assume that the target side of the training data was more concise and less explicit than the spoken Swiss German, then this would explain Whisper's behavior.

It should, however, be noted that such modal particles usually serve an information structural function (Musan, 2010). Thus, they do not necessarily affect the truth value of an utterance and, therefore, have little influence on the overall meaning (Krifka, 2007). For examples of removed particles, see Table 8 in Appendix A.1.

6.3 Hallucinations

Four out of sixteen transcriptions of whole conversations contained hallucinations – a sentence that was generated by Whisper without a corresponding utterance in the source audio.

In one conversation, in which the interviewee recounted the death of her mother, the following sentence was hallucinated:

> *Sie blieb nicht mehr in unserem pegen... Meine Frau, die ich so sehr liebte.* ("She didn't remain in our GIBBERISH... My wife, whom I loved so much.")

In another conversation, a sentence was continued by a hallucination (marked in bold):

> *Ähm ... Ja, jetzt bin ich immer noch etwas groggy, aber es geht etwas.* ***Ich bin ganz müde. Äh ...***

> ***Okay, ich kann ... Äh ... Zuerst schon.***, ("Ehm ... Yes, now I'm still somewhat groggy, but I'm managing. **I am really tired. Eh ... Okay, I can ... Eh ... Firstly.**")

In a different case, a sentence was preceded by a hallucination (in bold):

> ***Und ... äh ... Das hat mich sehr angestrengt. Äh ...*** *Das hat mich sehr viel aufgewühlt.* ("**And ... eh ... That really strained me. Eh ...** That really upset me.")

At the end of one conversation, *Untertitel von S G*[13] ("Subtitles by...") was added.[14]

We couldn't identify a pattern as to when and why hallucinations happen, but they seem to be a generally known problem with Whisper and are not specific to Swiss German audio.[15] Therefore, users should be aware that there is a possibility of hallucinations being added and in doubt re-check the audio.

7 Human Evaluation

7.1 Motivation

Performance of ASR systems is usually reported in WER , cf. Radford et al. (2022); Baevski et al. (2020). However, it is less meaningful for evaluating ASR for Swiss German with Standard German output since several outputs can be considered correct (see also Sections 1 and 4.1). Therefore, BLEU established itself as a second metric reported in works on ASR for Swiss German (Plüss et al., 2022, 2023; Schraner et al., 2022).

BLEU is meaningful mostly as a relative metric, comparing several systems; as an absolute score, it is less meaningful. It has been the object of criticism since Callison-Burch et al. (2006). Even its significance as a relative metric use has been harshly criticized, with Kocmi et al. (2021) complaining that "the sole use of BLEU impeded the development of improved models leading to bad deployment decisions." If we acknowledge that language technology is made for human beings, then its most important evaluation should be what humans think about it. We therefore conducted a short survey to evaluate how human beings perceive Whisper's performance.

[13]Whisper's output included a real person's name, which we anonymize here for privacy reasons.

[14]Obviously due to subtitles being part of the training data.

[15]A DuckDuckGo search for "openai whisper hallucination" returns many web pages discussing the issue.

7.2 Survey

For the survey, we randomly picked three of the conversations recorded as Mock Clinical Interviews (see Section 3.1) and extracted 119 sentence pairs consisting of our reference translation (sentence A) and Whisper's output (sentence B).

In the evaluation task, participants were asked, on a scale of 1 to 5, to rate:

1. To what extent is the meaning of sentence A retained in sentence B?
2. To what extent is sentence B fluent and natural?

with 1 being the worst and 5 being the best score. To assist the participants, each grade on the scale was given a verbal description (see Table 9 in Appendix A.2 for details). The participants were instructed to rate the fluency of sentence B (Whisper's output) independently from sentence A (reference) and to ignore punctuation marks.

Twenty-eight university students, colleagues, and acquaintances, who were recruited via personal channels, participated in the survey, all of them native speakers of German or Swiss German. The mean scores for meaning and fluency among all raters were 4.358 $\pm$ 0.046 (*SD* 0.239) and 4.39 $\pm$ 0.074 (*SD* 0.387), respectively, out of a maximum of 5 points. These scores suggest a very high human satisfaction with Whisper's performance.

7.3 Worst Rated Sentences

In an attempt to uncover some of Whisper's weaknesses, we picked the six sentences with the lowest mean score across all raters, see Table 5.

In sentence 1, the output includes the word *Riesiges* "huge" instead of the original *Kleines* "small", which is the exact opposite. In sentence 2, the subject of the sentence changes from the original *ich* "I" to *sie* "they", and the verb changes from *genommen* "took" to *liess* "let", causing the output to diverge greatly in meaning from the reference. Also, the preposition changes from *zur* in the reference to *an der* in Whisper's output. In sentence 3, the name of a train line in Zurich (*Forchbahn*) is "misheard" as *Furchtbahn* "fright train". Sentence 4 diverges greatly from the reference, with the use of the 3rd person accusative pronoun *ihn* without first introducing its referent, resulting in a genuine *non sequitur*. In sentence 5, the time mentioned in the original (*viertel nach sechs* "quarter past six") was changed to *4.15*. Finally, in sentence 6, the word *Schlummer-Taste* "snooze button" was

	Reference	Whisper	Mean
1	also meistens etwas **Kleines**, weil ich am Abend nicht so hunger habe	Also meistens etwas **Riesiges**, weil ich am Abend nicht so Hunger habe.	2.31
2	und habe meine Sachen genommen und dann **bin ich** auf den Bus gelaufen also **zur** Bushaltestelle	Ich nahm meine Sachen und **liess sie** auf den Bus. Also **an der** Bushaltestelle.	2.50
3	Auf das Tram gegangen, auf die **Forchbahn** und dann hierher gekommen.	Auf die **Furchtbahn** gingen. Dann kam ich hierher.	2.54
4	und dann bin ich bis zum **Stadelhofen** gefahren und habe ich noch die Tram genommen und dann bin ich hierher gekommen	Ich fuhr bis zum **Stadelhof** und nahm **ihn** aus der Rampe. Dann kam ich hierher.	2.58
5	Ja also ich bin um **viertel nach sechs** aufgestanden	Ich bin um **4.15** Uhr aufgestanden.	2.65
6	Da habe ich aber zuerst noch die **Schlummer-Taste** gedrückt zweimal.	Ich habe zuerst die **Schlamasseltasche** gedrückt, zweimal.	2.69

Table 5: The six sentences rated the worst in the human evaluation.

misheard as *Schlmasseltasche*, a gibberish word meaning "bad luck bag".

There is no recurring pattern in these sentences. It seems, however, that the transcription of named entities (*Forchbahn*, *Stadelhofen*, cf. sentences 3 and 4) and numbers (cf. sentence 5) might result in errors.

8 Conclusion

We evaluated Whisper's performance on Swiss German audio using automatic evaluation (WER and BLEU), a qualitative analysis and a human survey. All three evaluation types are evidential of very high performance: WER and BLEU are on par or slightly below other systems (cf. Table 2). The qualitative analysis revealed very high quality, retaining almost always the original meaning with only slight changes in style and some removal of cohesion markers such as particles and connectors. The human evaluation showed high human satisfaction (mean: 4.36/5.00, n=28).

We are therefore convinced that Whisper can be used, as is and out-of-the-box, without any further adaptations, for transcribing Swiss German, providing that the desired output is Standard German and that some loss of cohesion markers is acceptable.

However, as with any AI-based tool, Whisper should be used with caution. The qualitative analysis revealed some cases of changes in meaning, especially of numbers, as well as some hallucinations, though these were rare (one sentence in four out of sixteen 2-minute clips). In case of doubt, users should always refer to the original audio. Nevertheless, for the task of transcribing large portions of Swiss German audio or as a first step in a pipeline with other tasks in mind, such as keyword extraction or sentiment analysis, we think Whisper is a helpful, useful, and viable ASR tool.

Limitations

In this work, we evaluated Whisper's performance on Swiss German using automatic evaluation (WER and BLEU). We restricted ourselves to these metrics, since these are the metrics that are reported in previous works on ASR for Swiss German. Granted, other potentially better-suited metrics also come to mind, e.g., chrF (Popović, 2015) and BERTScore (Zhang et al., 2020). However, since models from previous works are not publicly available, we could not test them using different metrics besides WER and BLEU and had to rely on the scores reported in the respective works. Previous models not being publicly available also explains why we could not test the performance of previous models on our own test set (Mock Clinical Interviews), which would have been desirable.

Acknowledgements

Eyal Liron Dolev is a doctoral student at the MULTICAST project, Swiss National Science Foundation project no. 205913. He would like to express his gratitude to his doctoral supervisor,

Prof. Guido Seiler, for allowing him the freedom to conduct this work. This work was also supported by the Swiss National Science Foundation, project no. 191934. We thank the FHNW Institute for Data Science for making the SPC corpus available, as well as SwissNLP for making the STT4SG-350 dataset available. We also kindly thank the reviewers for their valuable feedback.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Raphael Berthele. 2004. Vor lauter Linguisten die Sprache nicht mehr sehen – Diglossie und Ideologie in der deutschsprachigen Schweiz. In Helen Christen, editor, *Dialekt, Regiolekt und Standardsprache im Sozialen und Zeitlichen Raum*, pages 111–136. De Gruyter, Vienna.

Shoshana Blum-Kulka. 1986. Shifts of cohesion and coherence in translation. *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies*, 272:17.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

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

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

Walter Haas. 2004. Die sprachsituation der deutschen Schweiz und das Konzept der Diglossie. In Helen Christen, editor, *Dialekt, Regiolekt und Standardsprache im Sozialen und Zeitlichen Raum*, pages 81–110. De Gruyter, Vienna.

Michael A. Hogg, Nicholas Joyce, and Dominic Abrams. 1984. Diglossia in Switzerland? A Social Identity Analysis of Speaker Evaluations. *Journal of Language and Social Psychology*, 3(3):185–196.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der Nähe — Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36(1):15–43.

Peter Koch and Wulf Oesterreicher. 1994. Schriftlichkeit und Sprache.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Gottfried Kolde. 1983. Sprachkontakte in gemischtsprachigen Städten: vergleichende Untersuchungen über Voraussetzungen und Formen sprachlicher Interaktion verschiedensprachiger Jugendlicher in den Schweizer Städten Biel/Bienne und Fribourg/Freiburg i. Ue. volume 37, Wiesbaden.

Manfred Krifka. 2007. Basic Notions of Information Structure. pages 13–56.

Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Heike Przybyl. 2022. Exploring Explicitation and Implicitation in Parallel Interpreting and Translation Corpora. *The Prague Bulletin of Mathematical Linguistics*, 119:5–22.

Renate Musan. 2010. *Informationsstruktur*, volume 9 of *Kurze Einführung in Die Germanistische Linguistik*. Universitätsverlag Winter, Heidelberg.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardžić. 2020. ASR for non-standardised languages with dialectal variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021. Swiss parliaments corpus, an

automatically aligned swiss german speech to standard german text corpus.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Felicity J. Rash. 1998. *The German Language in Switzerland: Multilingualism, Diglossia and Variation*. German linguistic and cultural studies. P. Lang.

Sebastian Ruder. 2024. True Zero-shot MT. `https://newsletter.ruder.io/p/true-zero-shot-mt`. Accessed: 2024-03-10.

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

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

Yanick Schraner, Christian Scheller, Michel Plüss, and Manfred Vogel. 2022. Swiss German speech to text system evaluation.

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

Simone Ueberwasser and Elisabeth Stark. 2017. What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik Online*, 84(5).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Examples of Whisper's Performance

Table 6 offers two excerpts from Whisper's output for our *Mock Clinical Interviews* test set (see Section 3.1). The excerpts exemplify Whisper's consistent high-quality performance over a longer passage of spontaneous and continuous speech.

Table 7 offers a speech excerpt from an interview in which the interviewee describes a past narrative (morning routine). It is an example of how Whisper inconsistently translates the Swiss German perfect sometimes as the Standard German preterite and sometimes as the Standard German perfect, cf. Section 6.

A.2 Human Evaluation Guidelines

Table 9 shows the rating guidelines for the raters in the human evaluation survey, cf. Section 7.

A.3 SPC Examples

Table 10 shows some examples of diverging reference translations that unjustly penalize Whisper's performance, cf. Section 5.2.

A.4 Plots

Figures 2 and 3 contain boxplots of the distribution of WER and BLEU scores of Whisper's performance on the three test sets: SPC, STT4SG and SwissDial. They show that Whisper's performance measured in WER and BLEU fluctuates considerably; for some sentences in STT4SG-350 for example, BLEU scores went up to 100.

Reference translation	Whisper
und es ist einfach so mega heimelig irgendwie also so	Es ist einfach so mega heimelig irgendwie.
es gibt mir so ein warmes Gefühl	Es gibt mir so ein warmes Gefühl.
und mir gefällt auch wie sie sich anziehen dort also,	Und mir gefällt auch, wie sie sich dort anziehen.
also ich finde es auch so schön es zu sehen,	Ich finde es auch so schön, es zu sehen.
nicht nur weil es lustig ist und beruhigend sondern auch, weil es schön ist zum anschauen	Nicht nur, weil es lustig und beruhigend ist, sondern auch, weil es schön ist, es anzuschauen.
also man hat Einzeltherapie zum Beispiel, oder Gruppentherapie mit anderen zusammen	also man hat Einzeltherapie zum Beispiel oder Gruppentherapie mit anderen zusammen
oder Musiktherapie, Maltherapie, oder auch so Entspannungsgruppen	oder Musiktherapie, Maltherapie oder auch so Entspannungsgruppen.
teilweise kann man auch selber Sport machen wenn man das will, also nicht in der Gruppe sondern alleine	Teilweise kann man auch selber Sport machen, wenn man das will, also nicht in der Gruppe, sondern alleine,
oder irgendwie so Walkinggruppen, oder so Achtsamkeitsgruppen, wo man in die Natur geht.	oder irgendwie so Walking-Gruppen oder so Achtsamkeitsgruppen, wo man in die Natur geht.

Table 6: Excerpts of Whisper's performance on continuous speech from our "Mock Clinical Interviews" test set (the segmentation into sentences is only for the sake of readability). These excerpts are evidential of Whisper high-quality performance.

Swiss German	Whisper
Denn **bin** i richtig **ufgstandfe**	Dann **bin** ich richtig **aufgestanden**.
Dänn **bin** i go dusche **ggange**	Ich **ging** duschen.
Dänn **han** i mi **aazoge**	Dann **zog** ich mich an.
Dänn mine chatz no fuetter **gää**, will si di ganzi ziit am maue **gsi isch** und unbedingt **het welle** esse	Ich **gab** meinen Katzen Futter, weil sie die ganze Zeit am Mauen **waren** und essen **wollten**.
Dänn **bin** i mit ire id stube abe ggange	Dann **ging** ich mit ihr in den Wohnzimmer.
Dänn **han** *ich* öppis **ggässe** dänn **het** *sii* öppis **ggässe**	Ich **habe** etwas **gegessen**, dann **hat** sie etwas **gegessen**.
Und dänn **bin** i wider uffe go zäh putze	Ich **ging** wieder hoch, um die Zähne zu putzen.

Table 7: Excerpt from a conversation in which the interviewee recounted what they did the same morning. It exemplifies how Whisper switches between perfect and preterite in Standard German. The input is always in the perfect tense. Perfect/preterites are marked in bold.

Reference	Hypothesis
“weil ich **dann halt** wieder auf mich gestellt bin.”	“weil ich wieder auf mich gestellt bin.”
“und darum ist es **ein bisschen** beides.”	“Darum ist es beides.”
“**ja und** ich find’s **einfach nur** spannend”"	“Ich finde es spannend”
“**Halt irgend so** eine Einschlafmeditation von einer Person...”	“Eine Einschlafmeditation von einer Person...”
“und er bekommt 50’000 Franken”	“und **dann** bekommt man **irgendwie noch** 50’000 Franken”"

Table 8: Examples for the removal of particles and conjunctions in Whisper’s output. Words in bold are particles/-conjunctions missing in the reference/hypothesis.

	Sinn – Ist der originale Sinn beibehalten? Entspricht Satz B Satz A?
5	Entspricht sinngemäss voll und ganz dem Original
4	Etwas ist verloren gegangen, die Bedeutung ist aber im grossen und ganzen gleich
3	Stimmt teilweise, aber nicht in allen Teilen
2	Entspricht kaum noch dem originalen Sinn
1	Gar nicht
	Flüssigkeit. Bezogen auf Satz B – ist das gutes Deutsch?
5	Ja, voll und ganz. Natürlich und einwandfrei.
4	Relativ flüssig
3	Nicht ganz flüssig, etwas merkwürdig
2	Kaum akzeptabel
1	Inakzeptabel

Table 9: Rating guidelines for the raters participating in the survey of human evaluation.

Swiss German Audio	SPC Reference	Whisper	WER
...nachhinei muss me döt de iibürgerigswillige sägge, er het scho...	So muss den Einbürgerungswilligen im Nachhinein gesagt werden:	Nachhinein muss man den Einbürgerungswilligen sagen, er hat schon	0.75
Dir wüssed scho vo de römerziite her	Aus Römerzeiten wissen Sie schon:	Ihr wisst schon von den Römerzeiten her,	1.4
...und das isch schlächt	Das ist schlecht.	Und das ist schlecht.	0.33
Während acht jahr isch s in betriib gsi	Während acht Jahren wurde es betrieben.	Während acht Jahren war es in Betrieb.	0.5
...u es het halt i Gotts name oo mitem finanzielle z tüe...	Und es hat halt auch wirklich mit dem finanziellen Aspekt zu tun.	Und es hat halt in Gottes Namen auch mit dem Finanziellen zu tun.	0.42
Töu vo euch erinnere sech müglicherwiis aa experiment ir physik oder chemie	Manche von Ihnen erinnern sich möglicherweise an missglückte Experimente in Physik oder Chemie.	Ein Teil von euch erinnert sich möglicherweise an Experimente in Physik oder Chemie.	0.38

Table 10: Examples for perfect performance of Whisper penalized by strongly divergent reference translations in the SPC corpus.

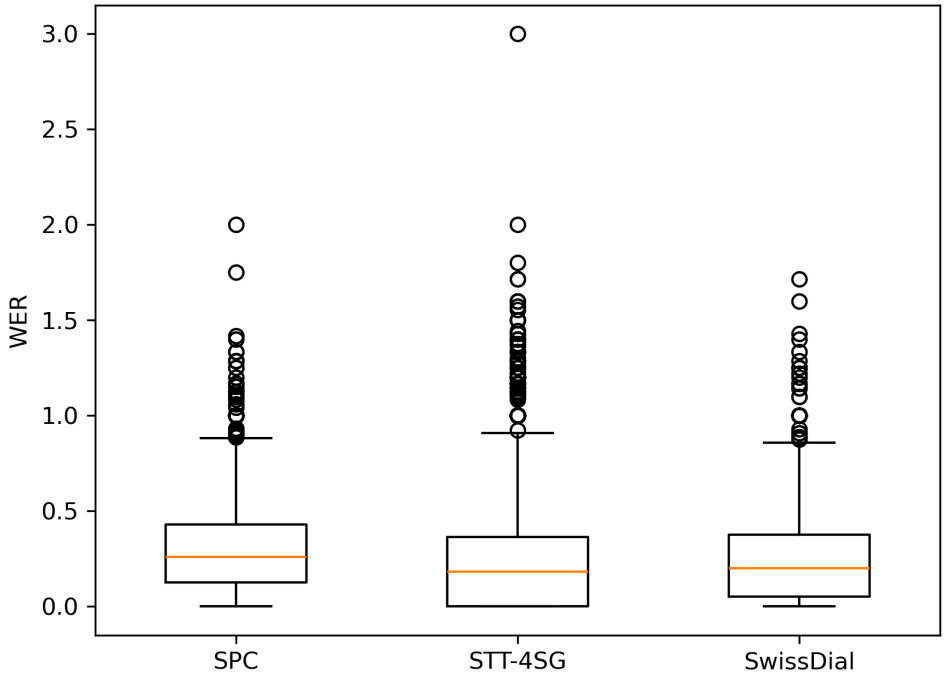

Figure 2: Distribution of WER scores for each corpus.

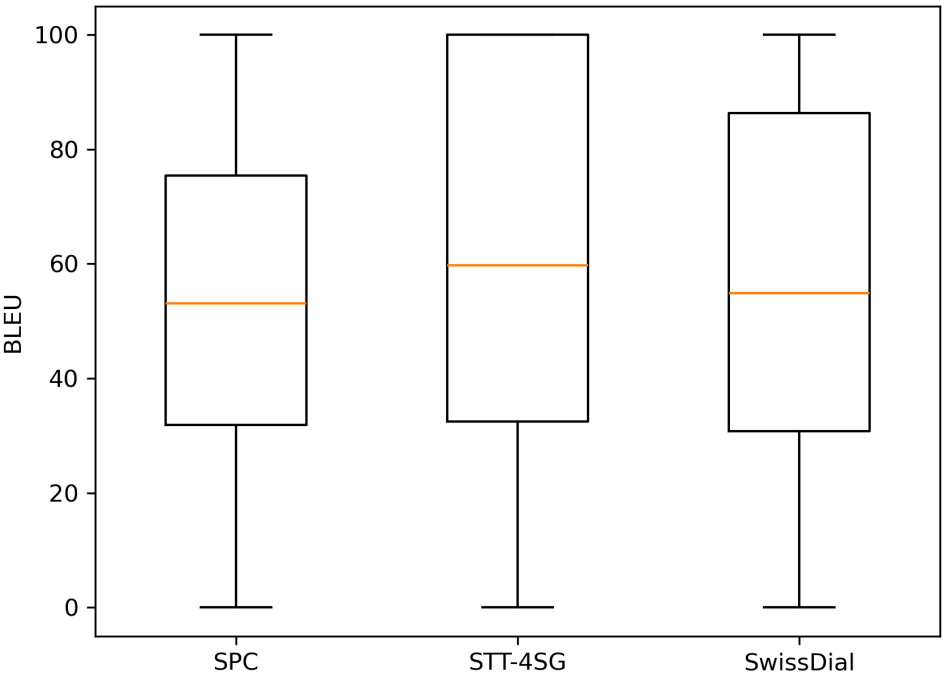

Figure 3: Distribution of BLEU scores for each corpus.