Evaluating a Multilingual Pre-trained Model for the Automatic Standard German Captioning of Swiss German TV

Johanna Gerlach, Pierrette Bouillon, Silvia Rodríguez Vázquez,
Jonathan Mutal and **Marianne Starlander**
Faculty of translation and interpreting, University of Geneva,
40, bd du Pont d'Arve, 1211 Geneva, Switzerland
johanna.gerlach, pierrette.bouillon, silvia.rodriguez,
jonathan.mutal, marianne.starlander@unige.ch

Abstract

In Switzerland, two thirds of the population speak Swiss German, a primarily spoken language with no standardised written form. It is widely used on Swiss TV, for example in news reports, interviews and talk shows, and captions are required for people who do not understand this spoken language. This paper focuses on the second part of a cascade approach for the automatic Standard German captioning of spoken Swiss German. We apply a multilingual pre-trained model to translate automatic speech recognition of Swiss German into Standard German suitable for captioning. Results of several evaluations, both human and automatic, show that the system succeeds in improving the content, but is currently not capable of producing entirely correct Standard German.

1 Introduction

In Switzerland, two thirds of the population speak Swiss German, a language that is therefore widely used on Swiss TV, for example in news reports, interviews and talk shows. Swiss German is primarily a spoken language, with many regional dialects and no standardised written form (Honnet et al., 2018). To make Swiss German content accessible to people who cannot understand spoken Swiss German, these TV programs need to be captioned in Standard German. The PASSAGE project (Bouillon et al., 2022), a collaboration between Geneva University, SRF (Schweizer Radio und Fernsehen) and recapp IT AG, financed by IMI ("Initiative for Media Innovation")[1] aims at generating captions for Swiss German TV shows in (Swiss) Standard German using a cascade approach. In the first step of this approach, illustrated in Figure 1, the automatic speech recognition (ASR) of Swiss German produces a normalised transcription that maintains the original Swiss German syntax and expressions, but uses German words. This is followed by a second step that involves machine translation (MT) into Standard German. Our contribution to this pipeline concerns the second step.

As far as we know, only a few studies have focused on this specific MT task. Arabskyy et al. (Arabskyy et al., 2021) developed a cascade approach using speech recognition for Swiss German followed by a lexical translation to standard German, but without syntactic restructuring. More recently, Plüss et al. (Plüss et al., 2022) developed an end-to-end system to transcribe spoken Swiss German and generate Standard German using a multilingual pre-trained model. In the present study, we also propose using a multilingual model, but only for the MT task. Although these models are known to be the best alternative in low-resource settings (Zanon Boito et al., 2022), models trained with Swiss German are currently unavailable (Plüss et al., 2022). In our case, the source is not Swiss German, but an automatic normalised transcription that combines language specific issues (Swiss German syntax) with spoken language phenomena (dysfluencies, informal language, ill-formed utterances) and ASR errors. Our objective is to see whether a multilingual model, fine-tuned on a small set of task-specific data, can solve these issues, while keeping as close as possible to what was actually said in order to produce coherent captions that will be useful to end-users.

In the course of the project, we investigated different models for this task and applied automatic metrics to determine which parameters led to the best performance. To follow up on these evaluations, in this study, we reuse the best performing system, but focus on human perception of the changes made to the ASR output. Our first hypothesis is that a multilingual pre-trained model that is fine-tuned on only a small amount of data can improve the quality of automatic transcriptions, in

[1] https://www.media-initiative.ch/project/subtitling-of-swiss-german-into-standard-german-automatic-post-editing/

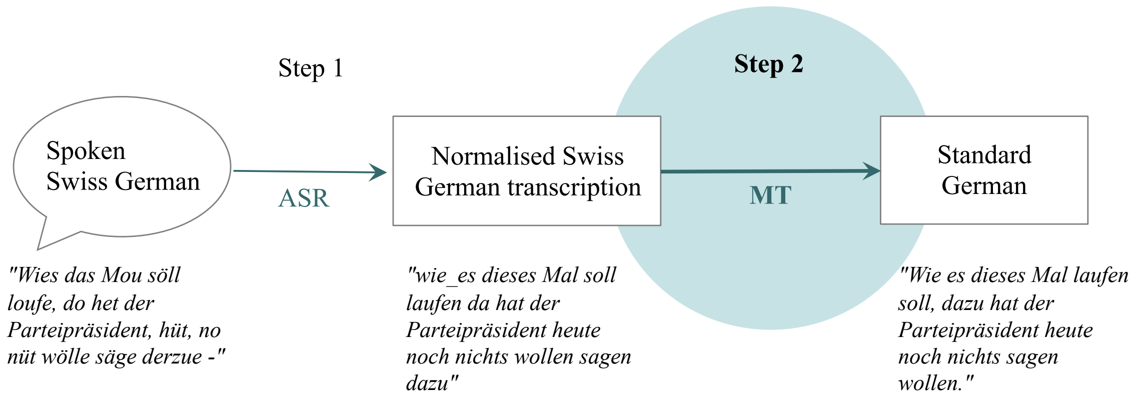
Step 1
Step 2
Spoken Swiss German
ASR
Normalised Swiss German transcription
MT
Standard German
"Wies das Mou söll loufe, do het der Parteipräsident, hüt, no nüt wölle säge derzue -"
"wie_es dieses Mal soll laufen da hat der Parteipräsident heute noch nichts wollen sagen dazu"
"Wie es dieses Mal laufen soll, dazu hat der Parteipräsident heute noch nichts sagen wollen."

Figure 1: Overview of the subtitling pipeline

terms of language and meaning. In addition, since ChatGPT was launched at the end of the project and has been found to outperform other models for many natural language processing tasks, including understanding low-resourced languages (Bang et al., 2023), we also provide a comparison with this large language model (LLM) based system for our task. Our second hypothesis is that a considerably larger model like ChatGPT will produce output with a better fluency, but sometimes to the detriment of meaning, with unnecessary transformations in this specific context.

This article is structured as follows: Section 2 introduces the system and data used in this study. We then present the human evaluations and their results in Section 3. Section 4 concludes and outlines future work.

2 System and data

For our particular task, i.e. transforming normalised Swiss German ASR output into correct Standard German, no parallel data was available. We therefore opted for a pre-trained multilingual model fine-tuned with data created specifically for the task.

We used mBART50 (Tang et al., 2020) – a multilingual sequence-to-sequence model pre-trained on large-scale monolingual corpora for 50 languages, including German. This model was fine-tuned on the task-specific aligned data sets described below, using the Transformers library (Wolf et al., 2020). Since mBART50 does not have a token for Swiss German, we reused the token for German in both encoder and decoder. We fine-tuned using 50,000 steps and two types of data: data derived from the Swiss German TV shows provided by our project partner SRF and the Swiss Parliaments Corpus, an automatically aligned Swiss German speech to Standard German text corpus (Plüss et al., 2021).

To produce aligned corpora suitable for fine-tuning, we applied different processes to these data, which we will now describe in further detail.

Swiss German TV shows This included data from a same set of talk shows and regional news, but in different unaligned forms, using different segmentation:

- **Human normalised transcriptions** (using Standard German words). These data were originally created to train the recapp Swiss German recogniser and were the only data available at the beginning of the project. While they include Swiss German syntax, as well as spoken language features, these data are devoid of ASR errors and therefore different from the input the MT system will have to handle;

- **Original subtitles**. These Standard German data follow subtitling standards;

- **ASR**. Automatic normalised transcriptions produced by recapp ASR;

We combined the above to produce the following aligned data sets:

- **Human normalised transcriptions to subtitles**. We used an algorithm proposed by (Plüss et al., 2021) to automatically align the transcriptions and original subtitles. We then reduced the noise by removing blank lines, joining chunks of words to create sentences and filtering sentences longer than 200 tokens. This filtered out 10% of the segments. In this corpus, the target segments have a much lower word count than the source segments;

- **Human normalised transcriptions to post-edited Standard German**. The target side of this corpus was produced by minimally post-editing the human transcriptions, i.e. post-editors were asked to correct syntax if it was not correct standard German, but not improve style or fluency;

- **ASR to subtitles** and **ASR to post-edited Standard German**. We manually aligned the automatic transcriptions to the original subtitles and the post-edited texts;

Swiss Parliaments Corpus By processing the speech part of this corpus with recapp ASR, we produced a large aligned ASR output to Standard German corpus.

Table 1 provides an overview of all the aligned corpora[2]. Combined, the corpora cover a vocabulary of 149,353 (source) and 147,507 (target) words.

3 Evaluation

Our system aims at converting the ASR output into correct Standard German while making as few changes as possible and preserving the exact meaning. To assess the system's performance, we have carried out several evaluations: an initial human evaluation that compares mBART output with the raw ASR output aims at quantifying the system's ability to improve the ASR output; a second comparative evaluation, which uses the same approach, aims at comparing mBART with ChatGPT for this task; a third human evaluation, which rates fluency and accuracy of system outputs, is intended to provide a more fine-grained image of the difficulties of the task; finally, an automatic evaluation will quantify the system's ability to produce output close to what can be achieved by a human post-editor.

3.1 Evaluation data

The evaluation data were taken from four different Swiss German TV shows: two debate shows (*Der Club* and *Ecotalk*), one magazine show that includes interviews in many dialects (*Gesichter und Geschichten*) and one daily news show (*Schweiz aktuell*). Based on the segmentation performed by the ASR, these data include 1,542 sentences. We processed them with mBART and randomly selected 100 segments per show in which the system output was different from the raw ASR. These 400 segments were then processed with ChatGPT. All evaluations were carried out in spreadsheets that included all the segments of the shows in the original order to provide context, but judges were only required to evaluate the selected segments.

3.2 Comparative Evaluations

In both comparative evaluations, we collected segment-level judgements on two aspects: language and meaning. We begin this section by describing the evaluation design, followed by the results of the two evaluations.

3.2.1 Comparative evaluations – design

For the first evaluation, we processed the ASR output with mBART. For the second, we processed it with ChatGPT, which was primed with the following prompt: "*Hi GPT, can you help me with the following task: I want to correct transcripts of swiss german audio files. These transcripts are in an incorrect german grammar, the goal is thus to correct the transcripts to create proper german texts.*". Each segment was then introduced by permutations of "*Korrigiere das*" (*correct this*), "*Verbessere den Satz*" (*improve the sentence*) etc. to help mitigate context diffusion between different queries.

The outputs of the systems (mBART and ASR for the first evaluation, mBART and ChatGPT for the second) were presented side by side, with the differences highlighted. No information regarding the systems was given to the evaluators and to prevent bias, the position of the systems (left or right) was randomised. An additional column provided a human Standard German transcription of the segments to serve as a reference for the meaning evaluation. Each segment pair was evaluated on two aspects:

- **Language**: participants were asked to consider whether the syntax, lexical choices and punctuation of the two outputs were correct in Swiss Standard German and provide a comparative judgement on a five-point scale: *A clearly better than B, A slightly better than B, A and B about the same, B slightly better than A, B clearly better than A*.

- **Meaning**: using the human transcription as a reference, participants were asked to compare

[2] Mutal et al. 2023 describes the impact of the individual corpora on system performance in detail; in this study, all corpora were used.

Domain	Aligned data		#Segments	#Words	
	source	target		source	target
TV shows	Hum. norm. transcr.	Original subtitles	59,932	910,597	649,039
TV shows	Hum. norm. transcr.	Post-edited hum. transcr.	76,684	1,009,749	968,360
TV shows	ASR	Original subtitles	12,393	197,689	161,047
TV shows	ASR	Post-edited hum. transcr.	9,223	213,185	201,328
Parl.	ASR	Stand. Ger. hum. transcr.	89,343	1,486,134	1,425,873

Table 1: Domain, composition, and number of segments and words in each of the aligned data sets

the two segments in terms of correctness and completeness of meaning and provide an assessment on a four point scale: *A better than B, both ok, both bad, B better than A*.

Additionally, participants were given the opportunity to append comments to individual segments. The spreadsheets were submitted to four native German speakers from Switzerland. Participants were compensated for the task.

3.2.2 Comparative evaluations – results

mBART vs raw ASR output Table 2 shows the results of the first comparative evaluation. In terms of language, with the original 5-point scale 39% of segments did not receive a majority judgement (3 or 4 judges agree). We have therefore condensed the scale by combining the “slightly better" and “clearly better" assessments. On the resulting 3-point scale, agreement between judges is fair (Light’s Kappa = 0.386) and 84% of segments received a majority judgement. For 71% of the segments, mBART’s output was preferred to the raw ASR, which shows that the system succeeds at improving the language of the speech recognition output.

In terms of meaning, mBART improves on ASR for 32% of the segments, degrading 1%. Inter-annotator agreement on this task is moderate (Light’s Kappa = 0.535), with 19% of segments left without majority judgement. For the remaining segments, which represent about half of the included data, the two system outputs were judged to be equivalent. For a high proportion of segments (29%), neither version was found to accurately convey the full meaning of the human transcription.

mBART vs ChatGPT Results of the second comparative evaluation are shown in Table 3. Overall, inter-annotator agreement was lower for this evaluation, with a higher proportion of segments without majority judgements for both language (29% on the 3-point scale) and meaning (35%), suggesting that differences between the two systems were less clear-cut than in the comparison with raw ASR. Overall, the ChatGPT output was largely preferred to mBART output in the language evaluation (45% ChatGPT better against 11% mBART better). A closer inspection of the outputs revealed that ChatGPT makes a higher number of edits, including stylistic improvements that bring the output closer to written language, additions that turn segment fragments into complete, coherent sentences, and context-dependent substitutions like replacing pronouns by their referent. While these transformations improve readability, they also increase the distance between the transcription and what was actually said, which may not be desirable for all applications, in particular when it comes to captioning. In the following example, ASR output consists of an ill-formed sentence that is typical of spoken language. ChatGPT has rephrased and lengthened it, but still fails to produce a grammatically correct sentence:

(1) HUMAN TRANSCR.: Kann nicht viel Originalität beisteuern, ein diversifiziertes Portfolio, gemanagt von der Bank Bär. (*Can’t contribute much originality, a diversified portfolio managed by Bank Bär.*)
ASR: Kann ja nicht viel Originalität beisteuern ein diversifiziertes Portfolio gemanagt von der Bank Bär (*Can’t contribute much originality a diversified portfolio managed by Bank Bär*)
CHATGPT: Kann ich nicht viel Originalität beisteuern, wenn es um die Verwaltung eines diversifizierten Portfolios geht, das von der Bank Bär gemanagt wird. (*Can’t I contribute much originality, when it comes to the administration of a diversified portfolio, which is managed by Bank Bär.*)

Language (5-point scale)		Language (3-point scale)		Meaning	
ASR clearly better	1 (0%)				
ASR slightly better	3 (1%)	ASR better	7 (2%)	ASR better	5 (1%)
Equivalent	45 (11%)	Equivalent	45 (11%)	Both ok	79 (20%)
mBART slightly better	97 (24%)	mBART better	284 (71%)	mBART better	127 (32%)
mBART clearly better	99 (25%)			Both bad	114 (29%)
No majority	155 (39%)	No majority	64 (16%)	No majority	75 (19%)
Light's Kappa	0.306	Light's Kappa	0.386	Light's Kappa	0.535

Table 2: mBART vs raw ASR output, majority comparative judgements for the 400 evaluated segments.

mBART: idem ASR except for punctuation

The preference for one system over the other is less clear in the meaning evaluation, with Chat GPT being judged as better in 26% of segments, as opposed to 15% for mBART. This evaluation also has a high proportion of segments in which no majority could be reached and there is merely fair inter-annotator agreement, according to Light's Kappa (K=0.404). A combination of poor segmentation, ill formed utterances and dysfluencies, common in spoken language, also make the human transcription that served as reference difficult to understand or ambiguous, compounding the difficulty of the evaluation. Finally, for 10% of segments, both system outputs were found to be unsatisfactory in terms of meaning. In these segments, neither of the systems could correct major ASR errors.

3.3 Fluency and accuracy rating

To refine the results of the comparative evaluation of mBART vs ChatGPT, we have extracted all the segments where the four judges unanimously agreed on the 3-point scale that one of the systems' outputs was better in terms of language (N= 106). The objective of this additional evaluation was to determine whether preference for one or the other output was caused by actual errors, or was more subjective, e.g. related to style.

3.3.1 Fluency and accuracy rating – design

To assess fluency, participants were asked to rate the correctness of the outputs on a 4-point scale (*no errors, minor errors, multiple or major errors, critical errors*). Since the outputs are direct transcriptions of mostly spontaneous speech, we asked participants to focus on correctness rather than style. In a second step, participants were asked to use the human transcription as a reference to rate how well the output conveyed the meaning of the utterance on a 4-point scale (*all meaning, most meaning, little meaning, none*). An additional option, "I don't understand the source", allowed participants to tag items where the human transcription was unclear. For this evaluation, we collected one judgement for each combination of segment, system and scale. The data were distributed to four judges in a crossover design.

3.3.2 Fluency and accuracy rating – results

Regarding fluency (see Table 4), for 70% of segments the mBART output was found to contain no (35%) or only minor (35%) language errors. This suggests that although the ChatGPT output was preferred for most of these segments, the mBART output is appropriate in terms of language. The remaining 30% include many cases where ASR errors such as misrecognised terms were not corrected. Another common issue relates to the difficulty of segmenting spontaneous speech, often resulting in long or ill-formed utterances that cannot be split into individual grammatical sentences.

The rating of meaning transfer, reported in Table 5, shows similar results for mBART, with the meaning fully (34%) or mostly (34%) preserved in 68% of segments. The distribution of common judgements suggests that the same sentences present difficulties for both systems.

Closer inspection of the segments where mBART failed to produce the same meaning as the human transcription shows that this is often due to ASR errors that were not corrected. In some cases this is due to misrecognised homophones, e.g. "geleert" (*emptied*), which was misrecognised as "gelehrt" (*taught*). ChatGPT was able to infer the correct verb, whereas mBART was not, resulting in incorrect meaning in the entire segment. As illustrated by example (2) below, in some cases, ChatGPT transforms poor ASR output with mul-

Language (5-point scale)		Language (3-point scale)		Meaning	
ChatGPT clearly better	51 (13%)				
ChatGPT slightly better	42 (11%)	ChatGPT better	181 (45%)	ChatGPT better	102 (26%)
Equivalent	60 (15%)	Equivalent	60 (15%)	Both ok	59 (15%)
mBART slightly better	16 (4%)	mBART better	44 (11%)	mBART better	41 (15%)
mBART clearly better	5 (1%)			Both bad	60 (10%)
No majority	226 (57%)	No majority	115 (29%)	No majority	138 (35%)
Light's Kappa	0.227	Light's Kappa	0.344	Light's Kappa	0.404

Table 3: mBART vs ChatGPT, majority comparative judgements for the 400 evaluated segments.

		mBART				
		4	3	2	1	
ChatGPT	4	29	27	24	2	82
	3	8	8	3	1	20
	2	0	0	1	1	2
	1	0	2	0	0	2
		37	37	28	4	106

Table 4: Contingency table for the fluency evaluation: distribution of segment scores for the two systems (4=no errors, 3=minor errors, 2=multiple or major errors, 1=critical errors)

		mBART					
		4	3	2	1	0	
ChatGPT	4	19	16	11	2	0	48
	3	13	17	9	0	0	39
	2	2	3	7	2	0	14
	1	1	0	1	1	0	3
	0	1	0	1	0	0	2
		36	36	29	5	0	106

Table 5: Contingency table for the accuracy evaluation: distribution of segment scores for the two systems (4=all meaning, 3=most meaning, 2=little meaning, 1=none, 0=I don't understand the source)

tiple misrecognised words into a correct sentence that is plausible in the given context, but distorts the actual meaning:

(2) HUMAN TRANSCR.: Ähm, du stehst am Start von einer Abfahrt in Kitzbühel und dein Inneres schreit und will nicht. (*Um, you are standing at the start of a downhill run in Kitzbühel and your guts are screaming and don't want to go.*)
ASR: du stehst am Start einer Abfahrt in Kitz Biel und Designer Schreit und will nicht (*you are standing at the start of a downhill run in Kitz Biel and designer Screams and doesn't want to go*)
CHATGPT: Du stehst am Start einer Abfahrt in Kitzbühel und dein Herz schlägt schneller, aber du willst nicht. (*You are standing at the start of a downhill run in Kitzbühel and your heart beats faster, but you don't want to.*)

In this example, we can also see that ChatGPT is able to correct the misrecognised proper noun "Kitzbühel", which mBART fails to do.

3.4 Judges feedback – Qualitative analysis

In the context of the two comparative evaluations, judges did not receive any specific instructions on what to elaborate on after making their judgements. As they were free to decide whether they wanted to include written feedback or not, they used the comment box as they saw fit. Hence, the comments that were collected varied in nature and number across judges and tasks. Nevertheless, we considered it relevant to look deeper into the additional qualitative data collected, as we believe it could provide insight into the challenges of the task, the quality of the output of the systems being assessed and the rationale behind the judges' preferences.

A total of 469 comments were retrieved after the study: N=126 from the first comparative evaluation, N=343 from the second one. This difference supports the challenges we highlighted in previous sections concerning the comparison of mBART and ChatGPT outputs. We first conducted a data cleaning stage, where empty-meaning comments such as "Option A is good" were not considered as usable data for the purposes of this paper. Based on the traditional thematic analysis approach, we then used a recursive process to explore and annotate the remaining data, during which one of the researchers generated initial codes, then searched for themes related to the topics mentioned above in subsequent rounds. For research reliability and reproducibility purposes, a second coder was asked to review a sample of the annotated comments against the coding scheme defined by the first coder. Despite the fact that the comments left by the judges were quite heterogeneous, a considerable high level of agreement was reached between the two coders.

Generally speaking, the judges' feedback indicates that not all preferred choices were errorless; i.e. most of the comments pointed at issues found in the system's output they had selected and not necessarily the one they had discarded. Since the main themes that emerged in the feedback for both comparative evaluations were very similar, we describe them jointly below, pointing to one or the other comparative evaluation task when needed.

ASR output errors A high number of comments — especially those attached to judgements where the two outputs proposed were rated as bad – referred to words that were not properly captured by the ASR system and therefore led to incorrect sentences. These were mostly related to wrong lexical choices (e.g. "'Booster Impfung' not recognized"). In particular, judges seemed to agree that the ASR system repeatedly failed to properly recognise numbers, technical terms and proper nouns (e.g. "B and A lack the technical term 'Omikron' and the name 'Huldrych Günthard'"). Certain judges also mentioned issues related to the identification of the type of sentence. For instance, questions were sometimes rendered as statements and vice versa in all the systems' output.

Importance of fidelity to the source When assessing whether the meaning was captured in the systems' output, judges showed considerable concern over the need to produce accurate sentences relative to the original message. Hence, they mostly pointed out issues related to content omission (e.g. "B is linguistically better, but in 'übriggeblieben' (left over) the 'anderes' (other) was omitted, which I still find important") and additions (e.g. " 'alle' (all) is superfluous, even distorts the meaning").

Interestingly enough, when mBART was compared against ChatGPT, in some of the comments received for segments where the latter was the preferred choice, negative remarks were made about the output's fidelity to the source. They mostly concerned reformulations – for instance, one judge said "A [ChatGPT] is bad because the sentence has been completely rearranged, which is not necessary." Paradoxically, we also identified a number of cases in which judges were positively influenced by the apparent fluency of ChatGPT's output, even when our model appeared to generate a more accurate sentence (e.g. "A [ChatGPT] is altogether more idiomatic, but the ending of B is closer to the source text.").

Appropriateness of the output for the expected purpose Some comments made direct reference to the context of use of the output being evaluated. In certain cases, judges assumed that captions had to be short, so lengthy outputs were penalised (e.g. "B [mBART] is linguistically much better, but probably much too long for a subtitle"; "A [ChatGPT] would be purely linguistically OK, but in my eyes it does not fulfil the requirements for subtitles").

Challenges of evaluating spoken language Finally, comments also indicate that judges struggled when having to make a choice between the two systems' output, due to the very nature of the source data (spontaneous spoken language). Concretely, judges experienced challenges related to the literal rendering of the text (e.g. "Too many errors by the speaker, the system has caught everything, making it too complicated to read."). In certain cases, this penalised our model, particularly in the second comparative evaluation. For instance, one judge said "A [ChatGPT] is bad because the text is changed too much and content is conveyed incorrectly, but B [mBART] is too literal and becomes incomprehensible due to the many filler words and incorrect grammar." Another issue that was put forward concerned excerpts where different people spoke at the same time or one person was interrupted by another, which sometimes led to syntactically incorrect sentences.

3.5 Automatic Evaluation

Our automatic evaluation aims at quantifying how close the system output is to a fully correct Standard German transcription produced by a human, with minimal edits. For the same set of 400 segments that were used for the human evaluations, we used the post-edited human transcription as a reference to assess the performance of the raw ASR and the two systems with BLEU (Papineni et al., 2001) and chrF (character n-gram F-score) (Popović, 2015). While BLEU allows us to quantify the performance on the word level, chrF allows us to quantify the performance on the character level – important for German, where small changes such as word endings indicating case are important. To calculate these metrics, we used the open-source library SacreBLEU (Post, 2018). Results are shown in Table 6.

	BLEU	chrF	Levenshtein to ASR
ASR	46.32	75.27	-
mBART	54.68	79.82	9.63
ChatGPT	39.12	68.71	32.59

Table 6: Automatic scores for raw ASR, mBART and ChatGPT outputs

In terms of automatic scores, mBART clearly outperforms ChatGPT, according to chrF and BLEU. These results confirm that the model fine-tuned with our data generates Swiss Standard German closer to the reference and suitable for the task.

We have also computed the Levenshtein distance between each of the outputs and the raw ASR. The average across the 400 segments is reported in Table 6. The distance for ChatGPT is more than three times that for mBART, confirming our observation that ChatGPT makes more changes.

4 Discussion and Conclusion

In this study we have performed several human evaluations of a system designed to produce Standard German suitable for captioning by correcting automatic normalised transcriptions of Swiss German TV content. The task, as well its evaluation, presents multiple difficulties.

The automatic transcriptions are very noisy: they combine Swiss German specific issues with spoken language phenomena and ASR errors. Additionally, the TV context continuously introduces new topics and people, along with their specific terminology and proper nouns. Building a system to address these issues requires task-specific data. In the absence of large parallel data sets for this task, we fine-tuned a multilingual model with several small data sets. The choices made for the creation of these data, e.g. minimal post-editing instead of full rephrasing, have influenced the system output. Aiming for minimal edits to stay close to the actual utterance means that many of the features of spontaneous spoken language remain, which is perceived as less fluent. This aspect could possibly be influenced by adjusting the instructions provided to the post-editors for training data creation, although beyond correcting grammar and obvious errors, the amount of required edits remains subjective, making it difficult to create consistent training data.

The evaluation itself was difficult and highly subjective. Agreement between judges was low, and a high number of items did not receive any majority judgement. The evaluation of accuracy is particularly complex. As in the case of spontaneous speech, the human transcription that serves as reference may also be unclear. We suspect that the fluidity of the system outputs may influence understandability and give a false impression of correct meaning.

To conclude, despite being fine-tuned on only a small set of data, mBART was able to improve a large proportion of the ASR output, in terms of language, and to some extent, in terms of meaning. Our comparison with ChatGPT on this task has shown that the fine-tuned mBART is less good at producing fluent standard German, because due to the training data, it only makes few changes. ChatGPT, on the other hand, produces mostly fluent output, but often transforms the meaning or generates paraphrases, which might not be desirable for captioning. Automatic scores confirm that mBART is better suited at producing the minimal edits required for this task. In future work, it would be interesting to investigate whether different prompt formulations could induce ChatGPT to generate output that remains closer to the original utterance. Overall, results of this study suggest that in this context where few task-specific resources are available, a fine-tuned multilingual pre-trained model is a promising approach.

Acknowledgements

This project has received funding from the Initiative for Media Innovation based at the EPFL's Media Center in Lausanne, Switzerland. We would like to thank our project partners David Imseng (recapp IT AG) and Florian Notter (SRF) for their contribution to data generation and recruiting participants for the human evaluation.

References

Yuriy Arabskyy, Aashish Agarwal, Subhadeep Dey, and Oscar Koller. 2021. Dialectal speech recognition and translation of swiss german speech to standard german text: Microsoft's submission to swisstext 2021. In *Proceedings of the Swiss Text Analytics Conference 2021*, Winterthur, Switzerland (online). CEUR-WS.org.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv:2302.04023*.

Pierrette Bouillon, Johanna Gerlach, Jonathan Mutal, and Marianne Starlander. 2022. The PASSAGE project : Standard German subtitling of Swiss German TV content. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 301–302, Ghent, Belgium. European Association for Machine Translation.

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Jonathan Mutal, Pierrette Bouillon, Johanna Gerlach, and Marianne Starlander. 2023. Improving standard german captioning of spoken swiss german: Evaluating multilingual pre-trained models. *To appear in Proceedings of MT Summit XIX – 2023, Macau, China*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, page 311. ACL.

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. *arXiv:2010.02810*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, page 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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv:2008.00401*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, page 38–45, Online. Association for Computational Linguistics.

Marcely Zanon Boito, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani, Ha Nguyen, Florentin Barbier, Souhir Gahbiche, and Yannick Estève. 2022. ON-TRAC consortium systems for the IWSLT 2022 dialect and low-resource speech translation tasks. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 308–318, Dublin, Ireland (in-person and online). Association for Computational Linguistics.