Improving Standard German Captioning of Spoken Swiss German: Evaluating Multilingual Pre-trained Models

Jonathan Mutal Jonathan.Mutal@unige.ch
Pierrette Bouillon Pierrette.Bouillon@unige.ch
Johanna Gerlach Johanna.Gerlach@unige.ch
Marianne Starlander Marianne.Starlander@unige.ch

TIM, Faculty of translation and interpreting, University of Geneva, Geneva, Switzerland

Abstract

Multilingual pre-trained language models are often the best alternative in low-resource settings. In the context of a cascade architecture for automatic Standard German captioning of spoken Swiss German, we evaluate different models on the task of transforming normalised Swiss German ASR output into Standard German. Instead of training a large model from scratch, we fine-tuned publicly available pre-trained models, which reduces the cost of training high-quality neural machine translation models. Results show that pre-trained multilingual models achieve the highest scores, and that a higher number of languages included in pre-training improves the performance. We also observed that the type of source and target included in fine-tuning data impacts the results.

1 Introduction

In Switzerland, over 60% of the population speaks Swiss German, which is a collection of spoken dialects with many regional variations. Swiss German is widely used in daily life and in the media, both on the radio and on Swiss TV. As these dialects lack a standardised written form, Standard German is often used for written communication, captions and subtitles. Standard German is also used to make Swiss German content accessible to people who cannot understand the dialects.

The PASSAGE project (Bouillon et al., 2022), which is the product of a collaboration between SRF and recapp IT, aims at making Swiss TV shows more accessible by automatically generating Standard German captions for Spoken Swiss German using a cascade approach. Figure 1 illustrates the two main steps. In a first step, our project partner's ASR transcribes spoken Swiss German into *Normalised Swiss German*, maintaining the original syntax and expressions of Swiss German, but using German words (Arabskyy et al., 2021). A second step, using machine translation (MT) approaches, aims at transforming these normalised transcriptions into fully correct *Standard German*.

Our contribution to this pipeline focuses on the MT step. In this context, MT could be used to different ends (Buet and Yvon, 2021). In our case, the objective is a minimal transformation to produce a correct Standard German transcription. This mainly involves resolving divergences between Swiss and Standard German by performing syntactical and lexical transformations, correcting speech recognition issues and removing spoken language phenomena such as dysfluencies. It would also be possible to condense and further transform content to achieve compliance with subtitling or captioning standards, which is not the aim of our task, but could be added as a subsequent step in the cascade approach. Since our input is not an actual language, but rather an artificial intermediate state between Swiss German and Standard German, the task is comparable to translation from low-resourced languages. We therefore propose to use multilingual pre-trained models which are often the best alternative in low-resource settings (Zanon Boito et al., 2022). In the absence of models for normalised Swiss German, we propose to use models trained on high-resource languages including German, that could generalise to normalised Swiss German (Kocmi and Bojar, 2018).

Some researchers have used multilingual pre-trained language models to generalise unseen languages – i.e. languages that are not covered in the pre-trained model (Wang et al., 2020; Pfeiffer et al., 2020; Muller et al., 2021). For example, (Muller et al., 2021) fine-tuned multilingual pre-trained models on 15 unseen languages to perform downstream tasks – POS (Part-of-Speech Tagging), NER (Named Entity Recognition) and DEP (Dependency Parsing). Their study shows that using multilingual pre-trained models increases the performance on these tasks for languages that have the same writing systems as the pre-trained languages. However, most of the researchers have applied these models to natural language understanding tasks from an unseen source language (Rust et al., 2021), but not many to machine translation. Multilingual models have already been applied to Swiss German, for example (Plüss et al., 2022) fine-tuned a multilingual pre-trained model to transcribe Swiss German and generate Standard German using an end-to-end approach. The model outperformed the transformer baseline by at least 8 BLEU points.

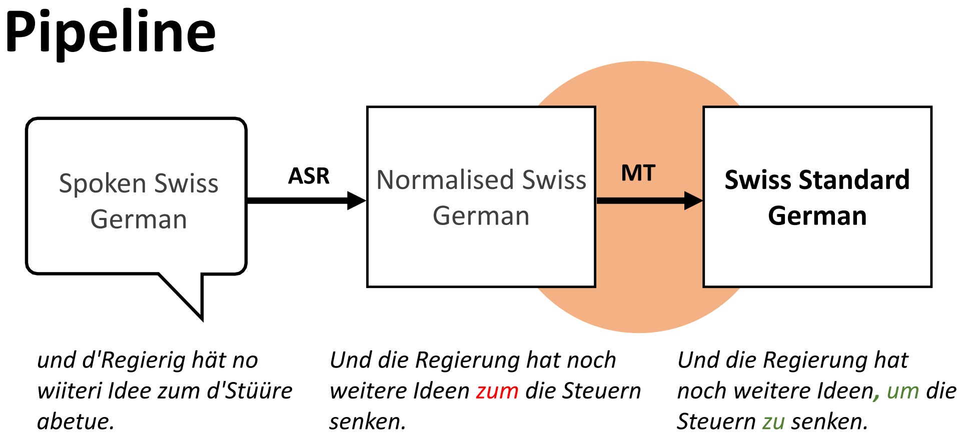
Pipeline
Spoken Swiss German
ASR
Normalised Swiss German
MT
Swiss Standard German
und d'Regierig hät no wiiteri Idee zum d'Stüüre abetue.
Und die Regierung hat noch weitere Ideen zum die Steuern senken.
Und die Regierung hat noch weitere Ideen, um die Steuern zu senken.

Figure 1: Overview of the pipeline from PASSAGE.

In this paper, we investigate different MT approaches for our task. Our first hypothesis is that in a low-resource setting such as this, pre-trained models will outperform a model trained from scratch with little data, as often shown in the literature. Our second hypothesis is that, in the absence of pre-trained models including our source language, models with a higher number of pre-trained languages will deliver better results. Finally, focusing only on fine-tuning of the best performing model, we will investigate the impact of different data characteristics, such as domain, provenance and quantity. The test data used in this study is available for research purposes.[1]

The remainder of this paper is organised as follows: Section 2 discusses the data, systems and evaluation methodology; Section 3 shows the results; and Section 4 presents the conclusion.

2 Methodology

In the following sections, we describe in more detail the data (Section 2.1), systems (Section 2.2) and evaluation methodology (Section 2.3).

2.1 Data

Due to the nature of the task, no large corpora were available. We therefore used data from two different sources: TV shows provided by our project partner SRF and the Swiss Parliaments Corpus, an automatically aligned Swiss German speech to Standard German text corpus, available for research purposes (Plüss et al., 2021). These data were processed in the following manner to produce aligned corpora:

Swiss German TV shows 1 This included data from a same set of talk shows and regional news, but in different unaligned forms, using different segmentation:

- GSW_NORM: normalised human transcriptions (using Standard German words). These data were originally created to train the Swiss German recogniser;
- DE_Subtitles: original Standard German subtitles. These data follow subtitling standards;
- ASR_NORM: automatic transcriptions produced by recapp IT ASR.

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

- **GSW_SubDE: Normalised Transcriptions to Subtitles** We used an algorithm proposed by (Plüss et al., 2021) to align GSW_NORM and DE_Subtitles. We then reduced the noise between the transcriptions and subtitles by removing blank lines, joining chunks of words together to create sentences, filtering out items based on length differences, and filtering sentences longer than 200 tokens. This filtered out 10% of the segments. Table 1 shows an extract of the automatic alignment.
- **GSW_PeDE: Normalised Transcriptions to Post-edited Standard German** We produced standard German by minimally post-editing the human transcriptions (GSW_NORM). The segments were provided to the post-editors in context. Table 2 shows examples of the transformations performed by the post-editors.
- **ASR_SubDE and ASR_PeDE: ASR output to Standard German** We manually aligned the automatic transcriptions (ASR_NORM) to the subtitles (ASR_SubDe) and the post-edited texts (ASR_PeDe).

Swiss Parliaments Corpus This corpus includes original Swiss German speech, automatically aligned with human transcription into Standard German. By processing the speech part of this corpus with recapp ASR, we produced a large aligned ASR output to Standard German corpus (**ASR_SwissPar**).

[1] https://doi.org/10/gr72xj

Transcription (GSW_NORM)	**Original Standard German Subtitle** (DE_Subtitles)
Weil	
dort steht eigentlich, was man mit dem, also was	Dort steht, was man tun muss.
man eigentlich muss machen. Also zum Beispiel, dass man ähm die neue Pensionskasse muss angeben.	Man muss z.B. seine neue Pensionskasse melden.
Oder	
wenn man jetzt zum Beispiel nicht gerade wieder geht gehen arbeiten, ähm in was für eine Freizügigkeitseinrichtung das Geld soll hin.	Wenn man nicht sofort wieder arbeiten geht, muss man angeben,an welche Freizügigkeitseinrichtung das Geld ausbezahlt werden soll.
Das ist ein so ein riesiges Volumen.	Das ist ein riesiges Volumen.

Table 1: Extract of the automatic alignment between the transcription (GSW_NORM) and the original subtitles (GSW_SubDE), (GSW_SubDE). The utterances with background colour were filtered out.

Transformation	**Normalised Human Transcriptions** (GSW_NORM)	**Post-edited Standard German** (GSW_PeDE)	**Literal Translation** (English)
Place modal after infinitive	Also, der einzige Ort, wo ich **würde gehen** ist Spanien.	Also, der einzige Ort, wo ich **hingehen würde**, ist Spanien.	The only place I would go to is Spain
Change subordinating conjunction	Wir haben es auch gesehen das letzte Jahr, **wo** ein Putschversuch [...]	Wir haben es auch im letzten Jahr gesehen, **als** ein Putschversuch [...]	We also observed it last year, when a coup attempt [...]
Disfluencies	die inländischen **produ- ähm** Produzenten geschützt sind	die inländischen Produzenten geschützt sind	the domestic producers are protected

Table 2: Extract of Normalised transcriptions and post-edited standard German (GSW_PeDe). Three examples of transformations performed by the post-editors on the transcriptions (GSW_NORM).

Swiss German TV shows 2 This second batch of more recent TV shows was also 1) transcribed automatically with the recapp ASR, 2) transcribed manually (normalised) and post-edited into Standard German. Part of the resulting aligned (ASR to Standard German) data was then put aside to be used as test data (cf. Section 2.3) and the remainder was used to constitute the **ASR_recent** data set.[2]

Table 3 summarises the data sets with the number of segments and words.

[2] The data set is available in `doi.org/10/gr72xj`

Data Set	#Segments	#Words		#Vocabulary	
		Source	Target	Source	Target
GSW_SubDE	59,932	910,597	649,039	47,846	65,918
GSW_PeDe	75,705	989,391	948,700	76,905	77,698
ASR_PeDe	9,223	213,185	201,328	24,899	27,216
ASR_SubDE	12,393	197,689	161,047	22,422	23,887
ASR_SwissPar	89,343	1,486,134	1,425,873	87,203	83,642
ASR_recent	979	20,358	19,660	4,639	5,024

Table 3: Number of segments, words and vocabulary (unique tokens) for each data set.

2.2 Systems and models

For our study, pre-trained models needed to fulfil several requirements 1) be available both in a monolingual German version and in a multilingual version including German in pre-training, 2) be adaptable to our particular task and 3) be computationally light-weight enough for use in production. Based on these criteria, we selected three different approaches: MT-based, which is trained to translate; Bert-based (Devlin et al., 2019), which is trained to predict words from a sentence; and Bart-based (Lewis et al., 2020), which is trained to reconstruct the original text and sentence order and exhibits increased robustness to language variation and noise.

We used a standard Transformer architecture for all the approaches (Vaswani et al., 2017). We carried out the training at FP16 precision. The models were trained and fine-tuned using the HuggingFace (Wolf et al., 2020) framework. We used default hyper-parameters for each approach.

MT-based We developed two neural machine translation models:

- **Monolingual** As no pre-trained model was available for normalised Swiss German, this model was trained using all our data (Normalised Swiss German to German).

- **Multilingual** We used a pre-trained machine translation model for Western Germanic languages, including Swiss German (Tiedemann, 2020).[3] We then fine-tuned it using all our data.

Bert-based We leveraged BERT for machine translation by initialising the Transformer architecture with BERT parameters as followed by (Rothe et al., 2020; Chen et al., 2022) and then fine-tuned with all our data:

- **Monolingual** The parameters of German BERT (Chan et al., 2020) were used to initialise the encoder and decoder of Transformer.

- **Multilingual** The parameters of Multilingual BERT – trained in 104 languages – were used to initialise the encoder and decoder of Transformer.[4]

Bart-based We used all our data to fine-tune two pre-trained models:

- **mBART25.** This model is pre-trained in 25 languages, including Standard German (Liu et al., 2020).

- **mBART50.** This model is an extension of mBART25 with 50 languages (Tang et al., 2020).

[3] https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-gmw-gmw

[4] https://huggingface.co/bert-base-multilingual-cased

2.3 Evaluation methodology

To address our first hypothesis, we compare the MT-system trained from scratch with pre-trained models for the three approaches. To address the second hypothesis (higher number of languages on pre-trained models deliver better results), we assessed each model with different numbers of seen languages. To address our third research question, we compare the performance of systems fine-tuned with the different types of data.

Models were compared based on automatic metrics, using a test data set consisting of 1,542 segments of recent TV shows provided by our partners, SRF and recapp. The data come from four TV shows: *Der Club*, which consists of debates led by journalists in local dialects; *Eco Talk*, which consists of debates on economic and business topics led by journalists; *Gesichter und Geschichten*, which reports people's stories and often involves interviews of the public; *Schweiz Aktuell*, which reports daily news and often involves interviews of the public. Table 4 details the number of segments for each TV show.

TV Show	**#Segments**
Der Club	780
Eco Talk	356
Gesichter und Geschichten	205
Schweiz Aktuell	201
Total	1,542

Table 4: Number of segments for each TV show.

We used the human post-edited version as a reference to compute the following metrics using the open-source library SacreBLEU (Post, 2018): chrF (Popović, 2015), which allows us to quantify performance on the character level – good for Germanic languages, where small changes such as word endings are important; BLEU, which allows us to quantify the performance on the word level. We also calculated the Levenshtein distance between the system outputs and the raw ASR to quantify the number of changes made by each system.

In addition, to corroborate and verify that the changes made by the systems are perceived as useful by end-users, we have carried out a human evaluation comparing the raw ASR with the output of the best performing system according to the automatic metrics. Specifically, a subset of 400 sentences was randomly selected from the test data for the evaluation. Participants were asked to provide segment-level judgements on two aspects: language (including syntax, lexical choices, and punctuation) and meaning. The objective was to compare the outputs and indicate whether users perceived the systems to be equivalent or if one system was preferred over the other. To assess language, participants were presented with a five-point scale consisting of the following options: "A clearly better than B", "A slightly better than B", "A and B about the same", "B slightly better than A" and "B clearly better than A". For meaning, a four-point scale was used, including the options: "A better than B", "both ok", "both bad" and "B better than A". The Standard German transcription was provided to serve as a reference for the meaning evaluation.

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. To prevent bias, the position of the ASR and MT output was randomised. The spreadsheets were submitted to four native German speakers from Switzerland. Participants were compensated for the task.

3 Results

In this section, we present the results by hypothesis.

3.1 Usefulness of pre-trained models

Results for the automatic metrics are shown in Table 5. Overall, the scores show that adding a MT step improves the ASR output by at least 1 point chrF and BLEU. Regarding our first hypothesis, we observe the pre-trained models outperform the MT-based monolingual trained from scratch. The two monolingual systems achieve the lowest scores, with BERT outperforming the MT-based, which is unsurprising since the latter is trained on very little data.

Approach	Systems	#Lang	#Params	chrF	BLEU	Levenshtein
raw ASR	none	-	-	75.27	44.72	23.81
MT-based	Monolingual	1	215M	76.27	47.53	12.30
	Multilingual	104		78.95	52.37	9.61
Bert-based	Monolingual	1	384M	77.23	48.52	11.26
	Multilingual	104		78.90	51.98	10.20
Bart-based	mBART25	25	610M	77.70	51.36	10.50
	mBART50	50		79.82	54.68	9.63

Table 5: The table presents the details of each system, categorised by approach and system. It includes the number of languages, parameters, as well as the chrF, BLEU, and Levenshtein distance to the raw ASR. The number of parameters was calculated using the Huggingface library.

If we compare the various multilingual models, the results show that mBART50 outperforms the other approaches. These results concur with findings by (Lewis et al., 2020; Anastasopoulos et al., 2022), who showed that fine-tuned BART models often outperform other approaches on machine translation in low-resource settings. However, our results also show that the multilingual MT-based model, which is three times smaller than BART, achieved a competitive score.

Looking at the Levenshtein distance between system output and raw ASR, we observe that the best performing systems are also those that make the least changes.

3.2 Impact of Number of Languages

To verify our second hypothesis, regarding the number of languages in pre-training, we comparatively assessed the same models with different numbers of languages. The results show that models with more pre-trained languages, although they do not include normalised Swiss German, outperform the others on chrF and BLEU. However, further work would be necessary to explain what influence each individual pre-training language, and its distance to Swiss German, has on the task.

3.3 Impact of Fine-tuning Data

Since our data come from different sources and are constituted of different types, we wanted to see which source and type was the most useful for fine-tuning for the task. To assess this, we used mBART50 – the highest performing system from our approach comparison – which we fine-tuned individually with each of the different data sets described in Table 3. We compare performance with mBART50 without any fine-tuning. Since in our previous evaluation we observed the same pattern for both BLEU and chrF, we decided to solely calculate chrF for this evaluation.

The first aspect where our aligned data sets differ is the provenance of the Standard German target: post-edited, original subtitles or human transcriptions (Swiss Parliament). The results, reported in Table 6, show that fine-tuning improves performance on the task, with the exception of the two cases where the data consist of original subtitles (GSW_SubDE and ASR_SubDE). This can be explained by the fact that, contrary to the task, subtitles are shortened and often simplified. Using the post-edited data (GSW_PeDe and ASR_PeDE) produced the highest chrF scores. The model that was trained only with GSW_PeDe achieves a comparable chrF (79.15) score to the model trained with all the available data (79.82). We also observed that absolute performance varied between the four TV-shows, with "Der Club" obtaining the worst results. However, the different models have the same ranking for each show.

Data	**Der Club**	**Ecotalk**	**Gesichter**	**Schweiz Aktuell**	**Total**
None	**66.75**	**76.59**	**70.65**	**80.04**	**71.12**
GSW_SubDE	64.97	74.09	72.61	75.78	69.30
GSW_PeDe	75.92	83.74	77.31	84.20	79.15
ASR_PeDe	75.17	83.73	77.33	84.11	78.65
ASR_SubDE	63.37	72.90	70.30	74.56	67.99
ASR_SwissPar	73.08	80.56	76.74	81.37	76.40
ASR_recent	73.68	81.91	76.30	83.51	77.25

Table 6: chrF for each TV show, by fine-tuning data set.

The second aspect differentiating the data sets is the domain (TV shows vs Swiss Parliament). Results suggest that using a larger out-of-domain data set (the Swiss Parliament corpus, ASR_SwissPar, 89,343 segments) has almost the same impact as using a small number of in-domain segments (TV_recent, 979 segments). To confirm these results and make a comparable evaluation, we sampled 1,000 segments from the out-of-domain data (ASR_SwissPar) to reduce the size difference. Table 7 shows that using in-domain data results in a better performance on the task.

Type	**Der Club**	**Ecotalk**	**Gesichter**	**Schweiz Aktuell**	**Total**
Out-of-domain	72.26	80.60	75.49	81.71	75.77
In-domain	73.68	81.91	76.30	83.51	77.25

Table 7: chrF for each TV show, by domain of fine-tuning data.

The third aspect of interest is the provenance of the source side of the aligned data sets. Fine-tuning with data using human transcriptions (GSW_PeDE) obtained almost the same score as using automatic transcriptions (ASR_PeDE). However, these results might have been influenced by the difference in size between the two data sets (75,705 and 9,223 segments). We therefore performed an additional experiment, fine-tuning mBART using only a subset of the human transcription segments (GSW_PeDE), namely those corresponding to the segments included in the automatic transcription data set (ASR_PeDE). The results reported in Table 8 confirm that there is almost no difference in performance when using human transcriptions (GSW_PeDE) compared to automatic transcriptions (ASR_PeDE).

The results of these fine-tuning experiments show that the choice of data, particularly in

Transcription	Der Club	Ecotalk	Gesichter	Schweiz Aktuell	Total
Automatic	75.17	83.73	77.33	84.17	78.68
Human	75.88	83.39	77.10	84.18	78.52

Table 8: chrF for each TV show, by source of transcription of the fine-tuning data.

terms of domain and target-side data-type (post-edited vs. subtitles data), has a significant impact on performance. We also observe that adding less specialised data, available in larger quantities, does not improve the system for our particular task (refer to Table 5).

3.4 Human Evaluation

Table 9 shows the results of the 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. Most of these cases can be attributed to incorrect lexical choices made by the ASR system, where the machine translation system was unable to generate the appropriate word. This finding highlights the need for improvement in accurately capturing the precise meaning of human transcriptions in these segments.

These evaluation results confirm that the improvements measured by the automatic metrics correspond to transformations perceived as useful by end users.

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

4 Conclusion

In this study we have applied different machine translation approaches to the task of improving ASR output for automatic Standard German captioning of Swiss German TV content in a cascade architecture. Overall, MT is able to improve the output, both according to automatic metrics and user perspective.

Our first hypothesis was confirmed by the better performance of all pre-trained models for this task. Among the tested approaches, Bart-based approach achieved the highest scores, possibly due to its ability to handle noisy text.

We also observed that a higher number of pre-trained languages improves performance on the task, meaning that more languages are useful to enable generalisation to an unseen language. However, we did not elucidate the influence of individual languages included in pre-training.

We also assessed the impact of fine-tuning data. In-domain data improved performance on the task, but there is no difference in performance when using as source side of the aligned data automatic or human transcriptions. Fine-tuning data for this task does not have to be clean human normalised transcriptions, which are more expensive to produce. Result shows that ASR output, despite being noisier, does not lead to significantly worse results. We also saw that target-side has a significant impact on performance. For the task as studied here, the subtitles, which contain many changes related to subtitling standards (shortening, simplification) are not ideal as training data.

The different systems described in this paper can be tested online at `https://passage-imi.unige.ch/demo/` and the test data is available at `https://doi.org/10/gr72xj`.

5 Acknowledgements

This project has received funding from the Initiative for Media Innovation, which is based at the EPFL's Media Center in Lausanne, Switzerland.

References

Anastasopoulos, A., Barrault, L., Bentivogli, L., Zanon Boito, M., Bojar, O., Cattoni, R., Currey, A., Dinu, G., Duh, K., Elbayad, M., Emmanuel, C., Estève, Y., Federico, M., Federmann, C., Gahbiche, S., Gong, H., Grundkiewicz, R., Haddow, B., Hsu, B., Javorský, D., Kloudová, V., Lakew, S., Ma, X., Mathur, P., McNamee, P., Murray, K., Nǎdejde, M., Nakamura, S., Negri, M., Niehues, J., Niu, X., Ortega, J., Pino, J., Salesky, E., Shi, J., Sperber, M., Stüker, S., Sudoh, K., Turchi, M., Virkar, Y., Waibel, A., Wang, C., and Watanabe, S. (2022). Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Arabskyy, Y., Agarwal, A., Dey, S., and Koller, O. (2021). Dialectal Speech Recognition and Translation of Swiss German Speech to Standard German Text: Microsoft's Submission to SwissText 2021. *arXiv:2106.08126 [cs, eess]*. arXiv: 2106.08126.

Bouillon, P., Gerlach, J., Mutal, J., and Starlander, M. (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.

Buet, F. and Yvon, F. (2021). Vers la production automatique de sous-titres adaptés à l'affichage. In Denis, P., Grabar, N., Fraisse, A., Cardon, R., Jacquemin, B., Kergosien, E., and Balvet, A., editors, *Traitement Automatique des Langues Naturelles*, pages 91–104, Lille, France. ATALA.

Chan, B., Schweter, S., and Möller, T. (2020). German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chen, C., Yin, Y., Shang, L., Jiang, X., Qin, Y., Wang, F., Wang, Z., Chen, X., Liu, Z., and Liu, Q. (2022). bert2bert: Towards reusable pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 2134–2148, Dublin, Ireland. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, page 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kocmi, T. and Bojar, O. (2018). Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7871–7880, Online. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *arXiv:2001.08210 [cs]*. arXiv: 2001.08210.

Muller, B., Anastasopoulos, A., Sagot, B., and Seddah, D. (2021). When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. (2020). MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Plüss, M., Hürlimann, M., Cuny, M., Stöckli, A., Kapotis, N., Hartmann, J., Ulasik, M. A., Scheller, C., Schraner, Y., Jain, A., Deriu, J., Cieliebak, M., and Vogel, M. (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.

Plüss, M., Neukom, L., Scheller, C., and Vogel, M. (2021). Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus. *arXiv:2010.02810 [cs]*. arXiv: 2010.02810.

Popović, M. (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.

Post, M. (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.

Rothe, S., Narayan, S., and Severyn, A. (2020). Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., and Gurevych, I. (2021). How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning.

Tiedemann, J. (2020). The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, Z., K, K., Mayhew, S., and Roth, D. (2020). Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (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.

Zanon Boito, M., Ortega, J., Riguidel, H., Laurent, A., Barrault, L., Bougares, F., Chaabani, F., Nguyen, H., Barbier, F., Gahbiche, S., and Estève, Y. (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.