Clean Text and Full-Body Transformer:
Microsoft's Submission to the WMT22 Shared Task on
Sign Language Translation

*Subhadeep Dey, Abhilash Pal, Cyrine Chaabani, *Oscar Koller
Microsoft - Munich, Germany
`{subde,t-apal,t-cchaabani,oskoller}@microsoft.com`

Abstract

This paper describes Microsoft's submission to the first shared task on sign language translation at WMT 2022, a public competition tackling sign language to spoken language translation for Swiss German sign language. The task is very challenging due to data scarcity and an unprecedented vocabulary size of more than 20k words on the target side. Moreover, the data is taken from real broadcast news, includes native signing and covers scenarios of long videos. Motivated by recent advances in action recognition, we incorporate full body information by extracting features from a pre-trained I3D model and applying a standard transformer network. The accuracy of the system is further improved by applying careful data cleaning on the target text. We obtain BLEU scores of **0.6** and 0.78 on the test and dev set respectively, which is the best score among the participants of the shared task. Also in the human evaluation the submission reaches the first place. The BLEU score is further improved to **1.08** on the dev set by applying features extracted from a lip reading model.

1 Introduction

Sign languages are natural visual languages that are used by deaf and hard of hearing individuals to communicate in everyday life. Sign languages are actively being researched. However, there is a huge imbalance in the field of natural language and speech processing between oral and signed languages. Since recently, one observes the emergence of a transition shifting sign language processing to be part of the NLP mainstream (Yin et al., 2021). We embrace this development which manifests (among others) in the creation of the first shared task on sign language translation as part of WMT 2022 (Mathias et al., 2022). It is great to have real-world sign language data (Bragg et al., 2019; Yin et al., 2021) as the basis of this

---

*Equal Contribution

shared task, manifested in native signers content and an unprecedentedly large vocabulary. Nevertheless, this leads to a very challenging task with low performance numbers. When participating in the advances of sign language technologies it is worth recapping that deaf people have much at stake, both to gain and lose from applications that will be enabled here (Bragg et al., 2021). We aim to advance the field and the use-cases in a positive way and present our findings in this system paper.

In the remainder of this work we first present a brief view on the relevant literature in Section 2, then we present the employed data in Section 3. Subsequently, we describe our submission in Section 4, additional experiments in Section 5 and we end with a summary in Section 6.

2 Related Work

In this section, we present a limited overview of related work in sign language translation. We focus this review on the translation direction from sign language to spoken language and dismiss approaches that target the opposite direction, i.e. sign language production.

Sign language translation started targeting written sign language gloss to spoken language text translations, hence no videos were involved. Related works were mainly based on phrase-based systems employing different sets of features (Stein et al., 2007, 2010; Schmidt et al., 2013). Then, neural machine translation revolutionized the field. The first research publications on neural sign language translation were based on LSTMs either with full image input (Camgoz et al., 2018) or utilized human keypoint estimation (Ko et al., 2019). Transformer models then replaced the recurrent architectures (Camgoz et al., 2020; Yin and Read, 2020; Yin, 2020). These models perform a lot better, but suffer from a basic drawback that the input sequences must be limited to a maximum length. Previous work (Camgoz et al., 2018; Or-

bay and Akarun, 2020) has identified the need for strong tokenizers to produce compact representations of the incoming sign language video footage. Hence, a considerable body of publications target creating tokenizer models that are often trained on sign language recognition data sets (Koller et al., 2020, 2016; Zhou et al., 2022) or sign spotting data sets (Albanie et al., 2020; Varol et al., 2021; Belissen et al., 2019; Pfister et al., 2013).

There are several data sets relevant for sign language translation. Some of the most frequently encountered are RWTH-PHOENIX-Weather 2014T (Koller et al., 2015a; Camgoz et al., 2018) and the CSL (Huang et al., 2018) (which could be also considered a recognition data set). However, there are promising new data sets appearing: OpenASL (Shi et al., 2022a), SP-10 dataset (Yin et al., 2022) (covers mainly isolated translations) and How2Sign (Duarte et al., 2021).

3 Data

To train our system, we used the training data provided by the shared task organizers. The data can be considered real-life-authentic as it stems from broadcast news using two different sources: FocusNews and SRF. FocusNews, henceforth FN, is an online TV channel covering deaf signers with videos of 5 minutes having variable sampling rates of either 25, 30 or 50 fps. SRF represents public Swiss TV with contents from daily news and weather forecast which are being interpreted by hearing interpreters. The videos are recorded with a sampling rate of 25 fps. All data, therefore, covers Swiss German sign language (DSGS). Our feature extractors are pretrained on BSL-1k (Albanie et al., 2020) and AV-HuBERT (Shi et al., 2022b). Additionally, we evaluate the effect of introducing a public sign language lexicon that covers isolated signs [1], which we refer to as Lex. It provides main hand shape annotations, one or multiple (mostly one) examples of the sign and an example of how this sign is used in a continuous sentence. We choose a subset that overlaps in vocabulary with either FocusNews or SRF. As part of the competition, independent dev and test sets are provided, which consist of 420 and 488 utterances respectively.

Table 1 shows the statistics of the training data. We see, that there is about 35 hours of training data in total. In raw form without any preprocessing the data is case sensitive, contains punctuation and

---

[1]https://signsuisse.sgb-fss.ch/

	SRF	FN	Lex	Total
Videos	29	197	1201	1427
Hours	15.6	19.1	0.9	35.6
Raw: no preprocessing				
Vocabulary	18942	21490	–	34783
Singletons	12433	13624	–	22083
Clean: careful preprocessing				
Vocabulary	13029	14555	821	22840
Singletons	7483	7923	591	12290

Table 1: Data statistics on data used for training. SRF and FN refer to SRF broadcast and FocusNews data, while Lex stands for a public sign language lexicon. Singletons are words that only occur a single time during training.

digits. In this raw form the vocabulary amounts to close to 35k different words on the target side (which is written German). 22k words of these just occur a single time in the training data (singletons). Through careful preprocessing as described in Section 4.3 we can shrink the vocabulary to about 22k words and the singletons to about 12k.

4 Submitted System

Sign languages convey information through the use of manual parameters (hand shape, orientation, location and movement) and non-manual parameters (lips, eyes, head, upper body). To capture most information from the signs, we opt for an RGB-based approach, neglecting the tracked skeleton features by the shared task organizers. For the submitted system we rely on a pre-trained tokenizer for feature extraction and train a sequence-to-sequence model to produce sequences of whole words (no byte pair encoding). We further pre-process the sentences (ground truths of the videos) to clean it. This step is crucial to push the model to focus more on semantics of the data. Finally, in order to adhere to the expected output format for the submission, we convert the text back to display format using Microsoft's speech service. This applies inverse text normalization, capitalization and punctuation to the output text to make it more readable. The details of various components of the system are described in the next subsections.

4.1 Features

We use a pre-trained I3D (Carreira and Zisserman, 2017) model, based on inflated inceptions with 3D convolutional neural networks, to extract features

for our task. The model (Varol et al., 2021) was pre-trained to take consecutive video frames as input and predict over 1k sign classes. It was trained on BSL-1K (Albanie et al., 2020) consisting of about 700k spotted sign instances from the British broadcast news. The features are extracted with a context window of 64 frames and a temporal stride of 8. We use the model as a feature extractor, recovering embeddings before the final classification layer (layer: mixed_5c), yielding a sequence of 1024 dimensional vector, extracted for each video. Our input data is required to match the training conditions of the network, hence we apply gray background padding (adding 20% padding left and right, 7.5% up and down) and rescale the videos to 224×224 resolution. The front end features are subsequently fed to a sequence model.

4.2 Sequence model

A standard transformer network is trained to predict text sequences. We apply word-based units as the output instead of byte pair encoding. It seems that full words help reduce ambiguity in a data constrained scenario. The model is trained with the fairseq (Ott et al., 2019) toolkit. We apply 3 transformer layers at the encoder side and 2 layers at the decoder with 1024 hidden feed-forward dimension and 22k output units. The model is trained with Adam optimizer for 2k epochs with a learning rate of 1e-3. We found that a beam size of 1 works well on the dev set during decoding.

4.3 Data Cleaning

To reduce ambiguity and noise in the target text, we first applied manual cleaning by removing foreign (French and English) sentences. We then proceeded to removing sentences that start with a hashtag sign, as this seems to indicate inaccurate annotation. Further, we removed status messages that were added by the subtitling agency (such as "1:1-Untertitelung.", "Livepassagen können Fehler enthalten.", "Mit Live-Untertiteln von SWISS TXT") and patterns enclosed by an asterisk which indicate sounds occuring in the show (e.g. "* Beschwingte Blasmusik *"). As a next step, we expanded abbreviations like "Mrd." to "Milliarden" and applied text normalization to remove punctuation and special characters, lower case of the text and expand numbers and dates. As can be shown in Table 1, this plays a major role in reducing the total vocabulary.

Data cleaning	Dev (RedB)	Dev	Test
no	0.49	0.70	0.4
yes	**0.78**	**0.77**	**0.6**

Table 2: The Table shows the effect of data cleaning. Performance of translation systems trained on SRF and FocusNews evaluated on the WMT 2022 dev and test data is provided. We report reduced and standard BLEU score on the Dev set and only standard BLEU on the Test set. RedB stands for the reduced BLEU measure.

4.4 Evaluation metrics

A common evaluation metric for machine translation is BLEU (Papineni et al., 2002). However, the difficulty of the given task and the inherently low performance of the submitted systems cause a bias in the automated evaluation. Spoken languages like Swiss German, which is the target output space for the translation in this challenge, follow a statistical pattern where stop words or function words constitute the classes of words that occur most frequently. This explains one of the observations that we made in regard to the generated system output. In fact, the models which were producing more stop words achieved the highest BLEU scores. For example, the model we submitted achieved a BLEU score of 0.77 on the dev set, while an earlier, clearly worse, checkpoint achieved 0.91. Looking at the stop words, the submitted model output counts 2125 stop words, while the earlier checkpoint counts 2237 of such words. After our stop word removal the submitted reduced BLEU score is 0.78, while the earlier model achieves only 0.66. Hence, for model selection, we propose to use 'reduced BLEU' with an additional step: we first apply a blacklist to remove stop and function words. Using this approach, the model selection process based on the reduced BLEU metric turned to be much more reliable and more reflective of actual performance. In this work, we report both reduced BLEU and standard BLEU for all results on the dev set. Only standard BLEU is reported by the automatic evaluation through the shared task on the test set. The list of employed stop words can be found in the appendix.

4.5 Results

The results of the submitted systems are presented in Table 2, which allows to compare the effect of applying data cleaning. In terms of reduced BLEU, data cleaning improves the performance from 0.49

to 0.78 on the dev data. The test data shows similarly an improvement from 0.4 BLEU to 0.6. Based on the preliminary automatic BLEU scoring this result was the best among the participants of the shared task. However, it has to be noted that the final evaluation will be based on a human eval that has not yet been completed at the time of paper submission. It can be concluded that careful data preparation is fundamental for this data. Nevertheless, the shared task proves to be very challenging and overall, we observe rather low performance compared to published results on benchmark data sets like RWTH-PHOENIX 2014T (Camgoz et al., 2018). This further proves the large amount of variability in the task, which is amplified due to the presence of high numbers of singleton words.

5 Additional Experiments

We perform additional experiments to assess the impact of having dedicated mouth features as well as a lexicon data set on the model performance. The mouth carries important semantic information in sign languages. In the literature, exploiting lip information has shown to be fundamental for increased performance in sign language recognition and translation (Koller et al., 2015b; Shi et al., 2022a).

5.1 Mouth features

To extract features from the lip area, we employ a pre-trained AV-HuBERT model (Shi et al., 2022b) that has been trained on English data. AV-HuBERT is trained to learn a robust audio-visual representation in a self-supervised fashion. The success of the model is evident by the performance on an audio-visual speech recognition task. It has proven to be useful for sign-language task as well (Shi et al., 2022a). For us, the first step is to obtain mouth patches from the video frames. Hence, we rely on the dlib utility provided by the AV-HuBert authors for obtaining facial key point extraction. The face patch is cropped and re-scaled to match the input size (96 x 96) of the model. We use the AV-HuBERT model to extract 768 dimensional embedding (output of the ResNet layer) to obtain a feature vector per frame.

5.2 Comparison to RWTH-PHOENIX 2014 T

To underline the difficulty of the given shared task, we compare our employed pipeline on a standard benchmark data set for sign language transla-

	WMT 2022	PHOENIX 2014T	Factor
Hours	35.6	9.2	3.9
Vocabulary	22840	2887	7.9
Singletons	12290	1077	11.4

Table 3: Comparison between the training data for the WMT 2022 shared task and PHOENIX 2014T.

	Training data	Features	
		Full body	Mouth
Stand. RedB	SRF + FN	0.78	0.95
	SRF + FN + Lex	0.54	1.08
	SRF + FN	0.77	1.15
	SRF + FN + Lex	0.68	1.27

Table 4: The effect of adding lexicon data to systems trained with full body features (I3D) and mouth features (AV-HuBERT). Configurations are evaluated on the WMT 2022 Dev dataset using the reduced BLEU (RedB) and standard BLEU (Stand.) score as metric.

tion, namely Phoenix 2014T (Camgoz et al., 2018). We noticed a small difference between the original PHOENIX 2014T corpus as referenced and shared in of (Camgoz et al., 2018) and the publicly available embeddings and experiments of (Camgoz et al., 2020). In the latter full stops mark the end of each utterance, while in the original version this is not the case. The effect is small, but for the sake of completeness, we show it here. Furthermore, considering the statistics of PHOENIX 2014T and this WMT 2022 shared task, the difference becomes apparent. Table 3 shows the key statistics for the two tasks side by side. We can see that the WMT 2022 task has a nearly 8 times larger vocabulary, with 11 times more singletons that occur only once in training. However, it has not even 4 times more video material.

5.3 Results

Table 4 shows the results of applying AV-HuBERT features as input to the sequence model. It can be observed that the model trained with AV-HuBERT features performs better than the I3D model. In fact, it achieves a 0.95 reduced BLEU score, while the full body I3D features reach only 0.78. On visual inspection, we found that the model trained with AV-HuBERT is able to predict infrequent words but fails on simple words such as "auf wiedersehen". Therefore, we assess the effect of adding a lexical data set ('Lex') to boost representations

of those simple words. Table 4 shows that the addition of lexical data further improves the model performance which reaches 1.08 reduced BLEU. We believe that this is likely due to the matching lip movement patterns in the lexical training dataset and dev dataset. Unfortunately, due to time constraints, we were not able to submit this model.

Table 5 shows results on PHOENIX 2014T. We can see that our submitted pipeline matches the performance of (Camgoz et al., 2020) (19.80/20.24 BLEU on the dev/test sets compare to 20.69/20.17). However our employed embeddings, which were not trained on PHOENIX 2014T, do not generalize well to PHOENIX 2014T and are hence significantly outperformed by the ones employed in (Camgoz et al., 2020). The experiments show that the WMT 2022 shared task is significantly more challenging than PHOENIX 2014T. We also see that the addition of full stops at the end of each utterances in PHOENIX 2014T amounts to a difference in BLEU of about 1% relative (14.22/13.22 BLEU with full stops opposed to 14.06/13.13 without full stops).

6 Summary

In this paper, full body information is applied successfully for a challenging sign language task as part of the WMT 2022 competition. As such, we employed a pre-trained I3D model to extract an embedding for a sequence of frames of the video. The features are further fed as input to a standard transformer network. We obtain reasonable performance of 0.4 in terms of BLEU score on the test set. The model is further enhanced by applying careful cleaning to the text output. We obtain the result of 0.6 BLEU score on the official test data. Based on the automatic BLEU scoring this result was the best among the 7 participants of the shared task, but also in the human evaluation our submission reaches the first place. With additional experiments, we validate the usefulness of a pre-trained lip reading model for this task and the addition of a lexical data set. This improves the results to 1.08 reduced BLEU on the dev set.

Limitations

One major limitation to our work resides within the data set used for training our model. In fact, the signing interpreter is usually not a native signer and often seems to be heavily influenced by the source language, a.k.a the spoken language. As stated previously, we used a signing interpreter for SRF data. Another issue we have identified lies in the limited domain of the data, as it is constrained to Broadcast news. The trained model may therefore be too specialized to generalize well beyond this area. Furthermore, due to the small number of individuals present in the data set, it remains unclear if and how much ethnicity bias is introduced to the model. Our team did not proceed with any experiments to identify and measure this. However, we do believe that it is crucial to further analyze possible biases in the future. One evaluation metric that we did take into consideration for the model performance is the BLEU score. The experiments consistently returned extremely low values which reflects a poor accuracy. One thing that remains unclear to us is how significant small BLEU differences are for human perception and subjective evaluation.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision*.

Valentin Belissen, Michèle Gouiffès, and Annelies Braffort. 2019. Automatic recognition of Sign Language structures in RGB videos: The detection of pointing and lexical signs.

Danielle Bragg, Naomi Caselli, Julie A. Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E. Ladner. 2021. The FATE Landscape of Sign Language AI Datasets: An Interdisciplinary Perspective. *ACM Transactions on Accessible Computing*, 14(2):7:1–7:45.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *Proc. Int. ACM SIGACCESS Conf. on Computers and Accessibility (ASSETS)*, ASSETS '19, pages 16–31, New York, NY, USA.

Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7793, Salt Lake City, UT.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Trans-

Approach	Decoding Beam	Fullstops No	Fullstops Yes	PHOENIX 2014T Dev	PHOENIX 2014T Test
(Camgoz et al., 2020)	10		x	20.69	20.17
Our pipeline with embeddings from (Camgoz et al., 2020)	5		x	19.80	20.24
Our pipeline with WMT full body embeddings	5	x		14.06	13.13
Our pipeline with WMT full body embeddings	5		x	14.22	13.22

Table 5: Showing results on PHOENIX 2014T, a benchmark data set. The system which matches our submission to the shared task can be found on the last line.

formers: Joint End-to-End Sign Language Recognition and Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.

Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i-Nieto. 2021. How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2735–2744.

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *Proc. of the AAAI Conf. on Artificial Intelligence*, pages 2257–2264, New Orleans, Louisiana, USA.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural Sign Language Translation Based on Human Keypoint Estimation. *Applied Sciences*, 9(13):2683.

Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. 2020. Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(09):2306–2320.

Oscar Koller, Jens Forster, and Hermann Ney. 2015a. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding (CVIU)*, 141:108–125.

Oscar Koller, Hermann Ney, and Richard Bowden. 2015b. Deep Learning of Mouth Shapes for Sign Language. In *Proc. IEEE Int. Conf. on Computer Vision Workshops (ICCVW)*, pages 477–483, Santiago, Chile.

Oscar Koller, Hermann Ney, and Richard Bowden. 2016. Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3793–3802, Las Vegas, NV, USA.

Müller Mathias, Sarah Ebling, Avramidis Eleftherios, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2022. Findings of the WMT 2022 shared task on sign language translation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alptekin Orbay and Lale Akarun. 2020. Neural Sign Language Translation by Learning Tokenization. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 222–228.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. Fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Tomas Pfister, James Charles, and Andrew Zisserman. 2013. Large-scale learning of sign language by watching TV (using co-occurrences). In *Proc. British Machine Vision Conference (BMVC)*, pages 1–11, Bristol, UK.

Christoph Schmidt, Oscar Koller, Hermann Ney, Thomas Hoyoux, and Justus Piater. 2013. Using Viseme Recognition to Improve a Sign Language Translation System. In *International Workshop on Spoken Language Translation*, pages 197–203, Heidelberg, Germany.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022a. Open-Domain Sign Language Translation Learned from Online Video.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022b. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. In *International Conference on Learning Representations*.

Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, and Andy Way. 2007. Hand in Hand: Automatic Sign Language to Speech Translation. In *Proc. Conf. on Theoretical and Methodological Issues in Machine Translation*, pages 214–220, Sk"ovde, Sweden.

Daniel Stein, Christoph Schmidt, and Hermann Ney. 2010. Sign Language Machine Translation Overkill. In *International Workshop on Spoken Language Translation*, pages 337–344, Paris, France.

Gul Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and Attend: Temporal Localisation in Sign Language Videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16852–16861, Nashville, TN, USA. IEEE.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. MLSLT: Towards Multilingual Sign Language Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5109–5119.

Kayo Yin. 2020. Sign Language Translation with Transformers. *arXiv:2004.00588 [cs]*.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including Signed Languages in Natural Language Processing. 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 7347–7360, Online. Association for Computational Linguistics.

Kayo Yin and Jesse Read. 2020. Better Sign Language Translation with STMC-Transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2022. Spatial-Temporal Multi-Cue Network for Sign Language Recognition and Translation. *IEEE Transactions on Multimedia*, 24:768–779.

A List of Stop Words for Reduced BLEU Estimation

ab	diese	hatte	seit	wenn
als	diesen	hätte	sich	werde
als	dieser	hatten	sie	werden
also	dieses	hätten	sie	werdet
am	doch	her	sind	weshalb
am	dort	hin	so	wie
an	ein	ihm	solchen	will
an	eine	ihre	soll	wir
andere	einem	ihre	somit	wird
auf	einen	im	sowie	wirst
aus	einen	in	sowohl	wo
beim	einer	ins	statt	wohl
bin	eines	ist	über	wolle
bist	eines	könne	um	wollte
da	er	könnte	und	wollten
darauf	es	könnten	vom	worauf
das	es	man	von	wurde
dass	für	mehr	vor	würde
davon	gar	mit	war	würden
dazu	gegen	noch	war	zu
dem	geht's	nun	wäre	zudem
den	genau	ob	war's	zum
denen	gibt	oder	wars	zur
der	habe	quasi	warst	zur
des	haben	schon	wart	zur
des	habt	sehr	wegen	zwar
deshalb	hast	sei	weiteren	
dessen	hast	seid	weiterhin	
die	hat	seien	wem	
dies	hat	sein	wen	