

# Neural Machine Translation Methods for Translating Text to Sign Language Glosses

Dele Zhu<sup>1</sup>, Vera Czehmann<sup>1,2</sup> and Eleftherios Avramidis<sup>2</sup>

<sup>1</sup>Technical University of Berlin, Berlin, Germany

<sup>2</sup>German Research Center for Artificial Intelligence (DFKI), Berlin, Germany  
dele.zhu@gmail.com, {vera.czehmann,eleftherios.avramidis}@dfki.de

## Abstract

State-of-the-art techniques common to low resource Machine Translation (MT) are applied to improve MT of spoken language text to Sign Language (SL) glosses. In our experiments, we improve the performance of the transformer-based models via (1) data augmentation, (2) semi-supervised Neural Machine Translation (NMT), (3) transfer learning and (4) multilingual NMT. The proposed methods are implemented progressively on two German SL corpora containing gloss annotations. Multilingual NMT combined with data augmentation appear to be the most successful setting, yielding statistically significant improvements as measured by three automatic metrics (up to over 6 points BLEU), and confirmed via human evaluation. Our best setting outperforms all previous work that report on the same test-set and is also confirmed on a corpus of the American Sign Language (ASL).

## 1 Introduction

Sign Language Translation (SLT) aims to break the language barrier between the deaf or hard-of-hearing communities and the hearing communities. One challenging aspect of SLT is the fact that Sign Languages (SLs) are multi-channeled and non-written languages (Langer et al., 2014). Therefore, Machine Translation (MT) for SLs cannot directly take advantage of the recent developments in text-based MT. For this purpose, previous work has used written representations of the SLs. One of these representations are *glosses*, where signs are labeled by words of the corresponding spoken language, often including affixes and markers.

It is known that glosses have strong limitations as a linguistic representation (Pizzuto et al., 2006). However, given the current status of SLT, we have indications that research on SL gloss translation can still be useful. For instance, translation from spoken language text to SL glosses can be useful for interpreters and educational uses (Collins



Figure 1: Text-to-video SLT using glosses as an intermediate step (Source of images: Müller et al., 2020).

et al., 2012). Secondly, SL glosses are the only SL representation having several parallel corpora big enough to train MT, and the results may provide indications for the future treatment of other more appropriate representations. Previous research on SLT has used glosses as an intermediate step to build MT systems for translating from SLs to spoken language text (Camgoz et al., 2017, 2018; Chen et al., 2023) or from spoken language text to SLs (Stoll et al., 2020; Saunders et al., 2020a,b, 2022). In the latter case, glosses allow building the system in two steps, *i.e.*, *text-to-gloss* translation and *gloss-to-video* production (Figure 1). The glosses can be given to a system for the generation of SL (avatar animations, autoencoders, GANs). Our work focuses on the first part of this pipeline, *text-to-gloss* translation, whose results are responsible for the generated sign animations. We find that prior research, despite its improvements, has still not made a big breakthrough in this direction (Rastgoo et al., 2021).

SLs are Low-Resource Languages (LRLs) with regards to MT, since there is little parallel data (Coster et al., 2022). Despite the recent progress of MT for LRLs (Sennrich et al., 2016a; Zoph et al., 2016; Sennrich and Zhang, 2019; Ranathunga et al., 2021), few of these methods have been used for MT of SLs, such as data augmentation (Moryossef et al., 2021; Zhang and Duh, 2021; Angelova et al., 2022)

and transfer learning (Egea Gómez et al., 2022). Other efficient techniques, *e.g.* semi-supervised NMT (Cheng et al., 2016) and multilingual NMT (Johnson et al., 2017) have not been explored. We are therefore inspired to extensively explore the effects of the relevant methods on *text-to-gloss* translation. To the best of our knowledge, this paper is the first work on *text-to-gloss*:

- to achieve significant improvements, as compared to the baseline methods, on the two known natural SL datasets annotated with glosses (namely for the German SL: Deutsche Gebärdensprache, further abbreviated as *DGS*),
- to perform extensive experimentation with most known LRL-related MT methods and their combinations and in particular:
- to apply semi-supervised NMT by copying the monolingual data to both the source and target side, for lack of monolingual corpora with glosses,
- to use transfer learning via the warm-start strategy, and
- to use a multilingual NMT setting with the focus on improving the *text-to-gloss* direction.

All code of this work has been open sourced.<sup>1</sup>

## 2 Related work

The early-stage of *text-to-gloss* translation systems were built using Statistical Machine Translation (SMT; San-Segundo et al., 2012; López-Ludeña et al., 2014), in an attempt to translate spoken language into a signing 3D avatar using SL glosses as intermediate. Although the system evaluations reported good results based on limited data and automatic metrics, deaf users assessed the system conversely. Recently, with the advance of NMT, more promising systems have emerged, based on RNNs (Stoll et al., 2020) or as parts of end-to-end transformer systems (Saunders et al., 2020b, 2022), which contrary to our work do not try particular LRL-related methods.

More related to our work, in terms of *text-to-gloss* translation using LRL-related techniques, Li et al. (2021) implement a transformer architecture equipped with an editing agent that learns to synthesize and execute editing actions on the source

<sup>1</sup><https://github.com/DFKI-SignLanguage/text-to-gloss-sign-language-translation>

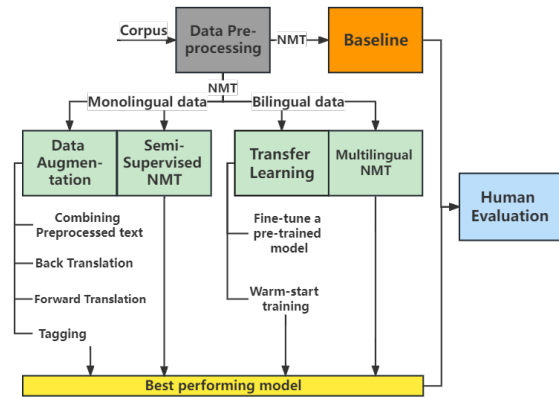


Figure 2: Scheme of experiments.

sentence. Walsh et al. (2022) examine the effect of different tokenization techniques and embedding approaches such as BERT and Word2Vec on the translation performance. Egea Gómez et al. (2021) propose a syntax-aware transformer injecting syntactic information into the word embeddings. In their follow-up work, Egea Gómez et al. (2022) achieve remarkable results with a transfer learning strategy that uses various ways of aggregating linguistic features and takes advantage of a pre-trained mBART model by filtering the original embedding and slicing model weights. In our work, we improve over these transfer learning methods using the *warm-start* strategy.

Data augmentation has been seen in *gloss-to-text* translation (Moryossef et al., 2021; Zhang and Duh, 2021; Angelova et al., 2022; Chiruzzo et al., 2022). Empirical comparison of our efforts with all state-of-the-art systems is presented in Section 5 (Table 4).

## 3 Methods

Our experiments (Figure 2) start from data preprocessing and setting the baseline. We then explore data augmentation, semi-supervised NMT, transfer learning and multilingual NMT as measured by automatic metrics. To confirm the consistency of system improvements between the best performing model and baseline, we conduct human evaluation.

### 3.1 Data augmentation

Data augmentation is a common technique used to face low resource conditions by adding synthetically generated data from various sources (Li et al., 2019). Here, we focus on the following methods:

**Combining preprocessing methods** is based on applying different preprocessing techniques on the source sentences and pairing them with copied target glosses. The differently pre-processed versions are concatenated into a new training dataset. This technique may be beneficial in that no changes are made to the target glosses and meanwhile the datasets are enlarged, being more robust to variable appearances of the spoken language sentences.

**Back-translation** is to obtain additional source-side data by translating a target-language monolingual dataset with target-to-source model (Senrich et al., 2016a). The generated source sentences are then paired with their target side into a synthetic parallel dataset. However, we lack a monolingual glosses dataset, so we use a *gloss-to-text* system to only translate the target-side glosses of the parallel corpus into spoken language text. This results in a synthetic version of the corpus, with the side of spoken language text modified. The synthetic corpus is then concatenated with the original one.

**Forward translation** or *self-learning* (Zhang and Zong, 2016) provides synthetic parallel pairs, in which the synthetic target data are obtained by translating an additional source-language monolingual dataset with the baseline system.

**Tagging** aims at informing the NMT model which sentences are original and which are synthetic, as the augmented data may be of lesser quality (Caswell et al., 2019). For this purpose, a special token is added in the beginning of each synthetic source sentence in the training data.

### 3.2 Semi-supervised NMT

To a certain extent, *text-to-gloss* translation can be regarded as a monolingual rephrasing task, as there is a large overlap in vocabulary of both sides. Thus, it triggers the assumption, that instead of generating synthetic data by models, we simply copy the monolingual data to both source and target side (Currey et al., 2017). This can be regarded as semi-supervised NMT, in which the model takes advantage of the concatenation of unlabeled monolingual data and labeled parallel data (Cheng et al., 2016). In this work, we do not delve into other potential effective factors of this method, *e.g.* size and domain of the monolingual data.

### 3.3 Transfer learning

Transfer learning uses learned knowledge to improve related tasks (Pan and Yang, 2010), *i.e.*, a parent model is pre-trained on a large corpus, used to initialize the parameters of the child model on a relatively small corpus. Zoph et al. (2016) first introduced the feasibility of transfer learning for NMT. We follow two approaches which differ in whether the child language pair (SL) is included during the parent model pre-training:

**Model fine-tuning** refers to fine-tuning a pre-trained model to train a child model. Although the pre-trained model usually contains a large vocabulary, it does not guarantee a full coverage of the child language pair. To alleviate this situation, the core operation of this approach is to modify the given vocabulary file manually. We tokenize the parallel SL dataset (*i.e.*, the child language pair) with the source-side tokenizer of the pre-trained model. Then, we append the vocabulary of the SL dataset into the pre-trained vocabulary. Since the vocabulary of the fine-tuned model has to be the same size as the original one, we replace the most frequent vocabulary occurrences of the pre-trained vocabulary with entries from the SL vocabulary.

Our method of fine-tuning by modifying the vocabulary is a simplification of the replacing algorithm used for Vocabulary Transformation (Kocmi and Bojar, 2020).

**Warm-start training** addresses the problem of vocabulary mismatch between parent and child models by introducing a joint vocabulary (Nguyen and Chiang, 2017). In this case, a parent model is pre-trained, but the training data of the child language pair is included during the pre-training of the parent model (Neubig and Hu, 2018). When the pre-training converges, this model is fine-tuned by training only on the child language pair. In order to select which language pair should be chosen as a parent one, Neubig and Hu (2018) suggest that using resources from related languages helps in improving the effectiveness of transfer learning, as it benefits from a high probability of words or characters overlapping within the related languages. For this reason we will be using a parallel dataset for paraphrasing of the spoken language.

### 3.4 Multilingual NMT

Multilingual NMT handles the simultaneous translation between more than one language through a

single model. We suggest multilingual NMT, considering that the amount of parallel data for our intended language direction is small but there is a larger parallel corpus for another related language direction (Johnson et al., 2017). Here we follow the case of *one-to-many translation*, i.e., one source language to multiple target languages. Parallel corpora from the two language pairs are concatenated and a target-language-indicator token is added at the beginning of each source language sentence. A joint vocabulary across all the training data is built.

In our case, the first target language refers to the SL glosses and the second target language is another spoken language. Contrary to other multilingual NMT experiments, we only focus on the performance of the *text-to-gloss* direction. An example of the combined parallel set follows:

- *German-to-English*: <2en> Wie heißt du? → What is your name?
- *German-to-DGSglosses*: <2gloss> im süden freundliches wetter → sued region besser

### 3.5 Evaluation

Following most of the MT tasks, we use three automatic evaluation metrics including BLEU-4 (Papineni et al., 2002), ChrF (Popović, 2015) and TER (Snover et al., 2006), with disabled internal tokenization, as suggested by Muller et al. (2022). Paired bootstrap resampling (Koehn, 2004) was performed to indicate the systems that are significantly better than the baseline, and the ones that are tied with the best-scoring system.

In order to confirm our conclusions and because the reliability of these metrics has not been confirmed for SL glosses, we conduct human evaluation. Since performing human evaluation for all system requires a lot of effort, we only collect human evaluation for the translation outputs from the best-scoring model and the baseline of every corpus, testing the hypothesis that the best-scoring system is significantly better than the baseline. Significance testing between pairs of systems is based on a one-tailed t-test, with a confidence threshold of  $\alpha = 0.05$ . As a means of quantitative human evaluation, we use Direct Assessment (Graham et al., 2013). Alternative translations of the same source by different systems are displayed shuffled at the same screen. A signer scores each output of shuffled systems from 0 to 6 (similar to Kocmi et al., 2022). Outputs marked with 0 fail to translate any of the contents of the original sentence, whereas

outputs marked with 6 show no significant mistakes in the translation.

## 4 Experiments

### 4.1 Datasets

We conduct our experiments on two parallel German SL (DGS) corpora containing gloss annotations.

**RWTH-PHOENIX-Weather 2014T** (Camgoz et al., 2018), abbreviated as *PHOENIX*, is a parallel corpus of SL containing weather forecasts. The original language was German, translated into DGS by professional interpreters and then annotated with DGS glosses. We use the provided split of parallel train-, dev- and test-set with respective sizes of 7,096, 519 and 642 sentences.

**The Public DGS Corpus** (Hanke et al., 2020; Konrad et al., 2020)<sup>2</sup>, further abbreviated as *DGS corpus*, contains conversations and narrations on topics culturally relevant to the deaf/Deaf community. The original language was DGS, which was then annotated with DGS glosses and German translation. We use the parallel corpus in plain text as extracted by Angelova et al. (2022)<sup>3</sup>, including the alignment of the DGS glosses to the German text by using the corresponding timestamps and prepending the gloss of the dominant hand to the non-dominant one, in case they co-occurred. We also follow the same data split into 54,325 training, 4,470 development and 5,113 test sentence pairs. Due to the big size of the test-set, for the human evaluation, we sample randomly 10% of the test sentences.

The DGS corpus gloss annotation (Konrad et al., 2022) includes suffixes to indicate different word variants, types, or groups. Muller et al. (2022) note that some annotation conventions may not be relevant to SLT and may make the problem unnecessarily harder. We confirmed this via our preliminary experiments (Appendix D), which yielded very low scores ( $\sim 1$  BLEU) when generating suffixes and we decided to strip all suffixes, for the following reasons. In order to be able to see the improvements of our methods we needed more generous references. Secondly, a criterion was to preserve basic lexical and syntactic information. A

<sup>2</sup>[https://www.sign-lang.uni-hamburg.de/meinedgs/ling/start\\_de.html](https://www.sign-lang.uni-hamburg.de/meinedgs/ling/start_de.html)

<sup>3</sup><https://github.com/dfki-signlanguage/gloss-to-text-sign-language-translation>

signer, part of our group, reviewed several gloss examples and noticed that while the suffixes might indicate lexical and phonological variants, so do the corresponding words in the German text, and with having written language as source there was in theory no way to determine which variant was necessary (except training a system to learn from context which seemed excessive at this point). Finally, PHOENIX glosses had no suffixes whatsoever, so by stripping the suffixes, the automatic metrics between the two corpora are comparable. Further work should focus on the importance of the suffixes, the right granularity for every purpose and how to optimize their generation. An example of suffix stripping follows:

- original: \$INDEX1\* SCHÖN1A ALLE2B ICH1 NICHT1\* \$GEST-OFF^
- stripped: \$INDEX SCHÖN ALLE ICH NICHT \$GEST-OFF

**NCSLGR** is a very small American Sign Language (ASL) parallel corpus (Vogler and Neidle, 2012), which we split ourselves to 1500, 177 and 178 sentences for train, development and test set.

**Other corpora** The German monolingual weather domain sentences (Angelova et al., 2022) and Europarl-v10 (Koehn, 2005) are used in data augmentation and semi-supervised NMT section (Sections 3.1 and 3.2) respectively. For training the parent model in transfer learning (Section 3.3), we use the parallel German paraphrasing corpus Tatoeba-Challenge (Tiedemann, 2020) for the main experiments in DGS and the synthetic text-to-gloss corpus ASLG-PC12 (Othman and Jemni, 2012) for the supplementary experiment in ASL. The German-English bilingual corpora News-commentary-v16 (Barrault et al., 2019) and Europarl-v10 are used in the section of Multilingual NMT (Section 3.4).

We report the statistics of vocabulary level and sentence lexical overlap of corpora with the custom split in Appendix F.

## 4.2 Data Preprocessing

For the data preprocessing, at source side, we perform lemmatization on both corpora and alphabet normalization specifically on the PHOENIX (the letters ü, ö, ä, and ß in the glosses are pre-normalized by dataset creators). We then apply Byte Pair Encoding (BPE; Sennrich et al., 2016b) to decompose the words and build vocabulary. In

the end, we set the lemmatized+normalized sentences with lowercased glosses of PHOENIX and lemmatized sentences with generalized glosses of the DGS corpus to train the models. We present the relevant statistics in Appendixes A and B.

## 4.3 Software

All software used is open source. MT models are trained with MarianNMT 1.11.0 (Junczys-Dowmunt et al., 2018). We also used Sentencepiece 0.1.97 (Kudo and Richardson, 2018), Moses-scripts (Koehn et al., 2007), Subword\_nmt 0.3.8 (Sennrich et al., 2016b), Hanover Tagger Lemmatization library 1.0 (Wartena, 2019), Scipy library 1.9.3 for t-test (Virtanen et al., 2020), SacreBLEU 2.2 (Post, 2018) for the automatic metrics and Streamlit 1.17 for the evaluation interface. To avoid model overfitting, we use several techniques such as early stopping (Zhang and Yu, 2005) during the model training.

## 4.4 Baselines

For the training hyperparameters, we start from the settings for a transformer (Vaswani et al., 2017) by the MarianNMT tutorial<sup>4</sup>. Specifically by baseline training, we take the advice of some paper that indicate in LRL MT scenarios with small data size, the model performance increases when the number of encoders/decoders are reduced compared to the original transformer architecture, e.g. one encoder and two decoders (Gu et al., 2018) and five encoders and five decoders (Chen et al., 2019; Araabi and Monz, 2020). After running extensive experiments with different combinations, which indicates we should reduce the encoder depth from 6 to 1 and the decoder depth from 6 to 2 to have the neural network fit better the small datasets. We present the baselines in Table 1. Our baseline models achieved a BLEU score of 22.78 on the PHOENIX dev set and 4.04 on the DGS dev set.

## 4.5 Effect of monolingual dataset

We first investigate the effect of using the additional monolingual dataset.

### 4.5.1 Data augmentation

**Combining preprocessed data** We collect different types of source text applied with different preprocessing methods of Section 4.2. For PHOENIX,

<sup>4</sup><https://github.com/marian-nmt/marian-examples/tree/master/transformer>

we combine the original, the normalized, the lemmatized, and the lemmatized+normalized text with the copied target glosses. For DGS, we mix the original and lemmatized text with the corresponding target glosses into a new training dataset.

**Back-translation** We first train simple *gloss-to-text* translation models for both corpora and then they generate sets of new source sentences from the target-side glosses. The synthetic texts are paired appropriately and then mixed with the original dataset.

**Forward translation** For PHOENIX, we use a German weather-domain monolingual dataset with the size of 1,203 to get a set of new glosses. Towards DGS, as it is a multiple-domain corpus, we obtain the new glosses by translating its source sentences with the baseline system.

We summarize the detailed statistics of the augmented datasets in Appendix C.

#### 4.5.2 Semi-supervised NMT

Here, we use the German monolingual dataset Europarl-v10 with a size of 2,107,971 sentences as auxiliary data. The monolingual data are copied to both the source and target side. To fit the neural network better with a larger training dataset, the encoder-depth is increased from 1 to 6, the decoder-depth from 2 to 6, the validation frequency from 500 to 5,000 and the max batch size from 64 to 1,000, as compared to the baseline. We build up a joint vocabulary of 32,000 entries after the corresponding BPE merge operations.

### 4.6 Effect of bilingual dataset

Then we start investigating the impact of the additional bilingual dataset on the model performance.

#### 4.6.1 Transfer learning

**Model fine-tuning** We take the German to English pre-trained model<sup>5</sup> from Opus (Tiedemann and Thottingal, 2020), whose vocabulary size is 65k. By applying the pre-trained tokenizer to both corpora, we get new vocabulary with size of 2,155 and 7,435 for PHOENIX and DGS corpus, respectively. We then crop the pre-trained vocabulary accordingly and merge the newly built vocabulary into it.

<sup>5</sup><https://opus.nlpl.eu/leaderboard/index.php?model=deu-eng%2Fopus-2021-02-22&pkg=Tatoeba-MT-models&scoreslang=deu-eng&test=all>

**Warm-start training** We select the Tatoeba challenge German paraphrasing dataset with a size of 4,574,760 as the parent language pair. In the first round of training, the training data contain German paraphrasing pairs and SL pairs. The parent model is trained using the Tatoeba challenge validation set. When it converges, we use this as pre-trained model to train with only the SL dataset. The child model is further trained using the SL development set for validation, until it converges too. We again build the joint vocabulary as in Section 4.5.2. During the two training phases, we reduce the validation frequency from 1,000 to 100 for a better observation.

#### 4.6.2 Multilingual NMT

We set up the identical source language in this part, *i.e.*, German. Only one additional language is selected to train the multilingual NMT, *i.e.*, English. We assume that a larger auxiliary dataset could be more helpful. Therefore, we set up two groups of sub-experiments with different sizes of auxiliary datasets in this section, *i.e.*, a relatively small dataset New-commentary-v16 with the size of 398,981 ("Multi") and a larger one Europarl-v10 with the size of 1,828,521 ("Multi-big"). Vocabulary and hyperparameters follow those of Section 4.5.2.

### 4.7 Effect of combining methods

We run the experiments independently and separately in Section 4.5 and Section 4.6. However, we cannot refuse the assumption that additional gain could be achieved by combining some or all of the best performing methods from above sections. Explicitly, we continue our experiments as following:

1. Combine all the data augmentation techniques of Section 4.5.1
2. Tag the monolingual data in the semi-supervised NMT setting of Section 4.5.2.
3. Combine multilingual NMT setting of Section 4.6.2 with combined preprocessed data and back-translation, respectively.

## 5 Results

In this part, we will present the performance of the various methods on both SL datasets and offer some further analysis.

Corpus	System	BPE Vocab	Dev			Test		
			BLEU	ChrF	TER	BLEU	ChrF	TER
PHOENIX	Baseline	2k	22.78	51.87	55.84	20.14	52.04	56.12
	Combine	2k	<u>24.01</u>	52.32	<u>53.20</u>	<u>21.88</u>	51.51	<u>54.53</u>
	Combine+Tag	2k	22.94	52.09	<u>52.88</u>	21.11	51.65	<u>54.81</u>
	Back	2k	23.63	52.03	<u>53.98</u>	21.04	51.59	<u>54.57</u>
	Back+Tag	2k	23.62	<u>52.85</u>	<u>52.88</u>	<u>21.57</u>	52.41	<u>53.94</u>
	Forward	2k	23.03	52.56	<u>53.71</u>	20.40	51.54	55.63
	Forward+Tag	2k	23.45	52.49	<u>54.16</u>	<u>21.64</u>	52.27	<u>54.57</u>
	All_combined	2k	23.63	52.32	<u>54.19</u>	21.04	51.97	<u>54.71</u>
	Semi	32k	<u>26.76</u>	<u>55.41</u>	<u>51.10</u>	<u>22.67</u>	<u>53.87</u>	<u>53.07</u>
	Semi+Tag	32k	<u>26.55</u>	<u>55.76</u>	<u>50.83</u>	<u>24.15</u>	<u>55.13</u>	<u>51.17</u>
	Fine-tune	65k	<u>26.39</u>	<b>56.84</b>	<u>50.88</u>	24.67	<b>55.97</b>	<u>52.86</u>
	Warm	32k	<b>27.62</b>	<b>56.92</b>	<b>49.25</b>	<u>24.89</u>	<u>55.46</u>	<b>50.40</b>
	Multi	32k	<b>28.34</b>	<b>57.29</b>	<b>48.48</b>	24.30	<u>55.71</u>	51.03
	Multi-big	32k	<b>27.45</b>	<u>56.52</u>	<b>48.77</b>	<b>24.97</b>	<u>55.75</u>	<b>49.89</b>
	Multi+combine	32k	26.61	<u>55.59</u>	<u>50.21</u>	23.22	<u>54.55</u>	<u>52.84</u>
	Multi-big+combine	32k	<b>28.02</b>	<b>57.07</b>	<b>49.31</b>	<u>24.94</u>	<u>55.89</u>	<u>51.01</u>
	Multi+back	32k	<b>28.41</b>	<b>57.54</b>	<u>49.39</u>	<b>26.32</b>	<b>56.70</b>	<u>51.15</u>
Multi-big+back	32k	<b>28.53</b>	<b>57.64</b>	<b>48.93</b>	<b>25.98</b>	<b>56.67</b>	<u>50.94</u>	
DGS	Baseline	5k	4.04	31.20	79.34	3.13	30.38	78.64
	Combine	5k	3.71	29.97	80.21	2.75	29.31	80.01
	Combine+Tag	5k	3.23	28.69	81.31	2.27	28.17	81.03
	Back	5k	3.83	30.08	82.75	3.06	29.30	80.94
	Back+Tag	5k	3.88	29.66	79.55	2.75	28.91	79.05
	Forward	5k	3.51	29.14	83.03	2.81	28.24	81.13
	Forward+Tag	5k	3.75	29.69	86.20	2.93	29.06	83.21
	All_combine	5k	3.14	28.37	81.61	2.43	27.87	81.83
	Semi	32k	<u>5.16</u>	<u>33.43</u>	<u>76.19</u>	<u>4.42</u>	<u>31.81</u>	<u>76.35</u>
	Semi+Tag	32k	<u>5.00</u>	<u>32.69</u>	<u>79.47</u>	<u>4.10</u>	<u>31.30</u>	<u>78.67</u>
	Fine-tune	65k	<u>5.82</u>	<u>35.05</u>	<u>79.92</u>	<u>4.53</u>	<b>34.14</b>	78.98
	Warm	32k	<u>5.87</u>	<u>33.42</u>	<b>74.07</b>	<u>4.55</u>	<u>31.90</u>	<u>74.54</u>
	Multi	32k	6.06	35.18	<u>74.51</u>	5.32	<u>33.55</u>	<u>74.71</u>
	Multi-big	32k	<b>6.60</b>	<b>35.26</b>	<b>73.25</b>	<b>5.46</b>	<u>33.49</u>	<b>73.53</b>
	Multi+combine	32k	<u>4.64</u>	<u>32.33</u>	<u>80.39</u>	<u>3.85</u>	<u>31.38</u>	<u>78.34</u>
	Multi-big+combine	32k	<b>6.79</b>	<b>35.50</b>	<b>73.98</b>	<b>5.61</b>	<b>33.88</b>	<b>73.94</b>
	Multi+back	32k	<u>5.35</u>	<u>33.43</u>	<u>78.30</u>	<u>4.85</u>	<u>32.16</u>	<u>76.76</u>
Multi-big+back	32k	<b>6.82</b>	<b>35.57</b>	<u>76.37</u>	<b>5.78</b>	<b>33.87</b>	<u>76.12</u>	

Table 1: Automatic metric scores of extensive experimentation search of the two DGS corpora. We **boldface** all the values that are not statistically significantly different from the best value of each evaluation metric and underline the results that are statistically significantly higher than baseline at the 95% confidence level.

## 5.1 Automatic evaluation

The performance of the various experiments, as measured with automatic metrics can be seen in Table 1. Looking at the scores on the test sets, we can observe that:

(1) Overall, the results on PHOENIX are better than on DGS corpus in all aspects. One of the reasons may be that DGS corpus is of broader domain and has a much bigger vocabulary. To support our assumption, we calculate the type-to-token ratio (Templin, 1957) for both corpora (PHOENIX: 2.2% and DGS corpus: 3.2%).

(2) For PHOENIX, data augmentation has shown a significant improvement in comparison with the

baseline, as measured by BLEU (+1.74) and TER (-1.59), although ChrF fails to measure a significant improvement. On the contrary, the performance on DGS corpus declines as compared to the baseline.

(3) Incorporating the large-scale monolingual dataset, (semi-supervised NMT), could further improve the scores of translation systems for both SL datasets. Tagging here seems to be of big importance for PHOENIX (+1.5 BLEU).

(4) Transfer learning incurs further improvement, with scores equal or better to the ones achieved with semi-supervised NMT. Here, each metric favors a different setting. ChrF indicates a significant improvement with fine tuning, TER prefers warm

System	BLEU	Test	
		ChrF	TER
Baseline	10.50	30.65	<b>78.95</b>
Back	9.37	28.25	<b>78.67</b>
Warm-start	<u>12.11</u>	<u>33.53</u>	83.44
Multi	<u>12.35</u>	<u>38.33</u>	<b>78.26</b>

Table 2: Automatic metric scores for NCSLGR corpus.

start, whereas BLEU indicates only a very small difference between the two.

(5) Multilingual NMT increases the automatic scores even further. The best scoring methods, favored by two automatic metrics each, and taking into consideration the significance tests, are (a) for PHOENIX the Multi with back-translation, Multi-big and Multi-big with back-translation, and (b) for DGS corpus the Multi-big, the Multi-big with back-translation, and the Multi-big with combined pre-processing.

In order to confirm the generalizability of our findings, we repeated the experiments with a very small corpus of the ASL, and the results are shown in Table 2. We observe that our best-scored method for the German SL (DGS) also gives the best performance for the ASL corpus, which is confirmed with two out of the automatic metrics.

In Table 4, we compare our best model to the approaches of recent work which have run experiments on PHOENIX *text-to-gloss* translation task. One can see that our best-scoring system performs 3.13 points BLEU higher than the closest result.

## 5.2 Quantitative Human Evaluation

As part of the human evaluation, an effort of approximately 40 hours for the PHOENIX test set and 20 hours for the DGS corpus was made. The results of the human evaluation (Table 3) confirm the basic hypothesis: that the best performing method of multi-NMT is statistically significantly better than the baseline. The density of the human evaluation scores of the two best scoring systems can be seen in Figure 3. One can see that more than half of the test-sentences of the best PHOENIX system are scored with a 5 or 6, whereas the corresponding percentage for the DGS corpus is only around 20%. Despite the extremely low automatic scores of the best model on the DGS corpus, it is promising that the human evaluator assigned the best score to 10% of the test sentences.

## 6 Conclusion

In this paper, we applied several techniques, commonly used in low resource MT scenarios, for MT from spoken language text to sign language glosses. We presented an extensive experimentation including data augmentation (combination of different pre-processing methods, back- and forward-translation), semi-supervised NMT, transfer learning with two different methods and multilingual NMT with different data sizes. The experiments were based on the two known natural datasets including gloss annotation, the RWTH-PHOENIX-Weather 2014T dataset and the Public DGS Corpus. Automatic metrics indicate significant improvement on the evaluation scores for both datasets when using most of the above methods, whereas the best results are achieved via a Multilingual NMT model (6.18 and 2.65 BLEU against the baseline respectively). Our best system outperforms all other state-of-the-art systems from previous work that report on the same test-set. Additionally, the best setting is confirmed with an experiment run on a corpus of the ASL. The conclusions are supported by human evaluation.

## Limitations

- These methods have been performed on three SL datasets (Section 4.1) as these were the only publicly available natural SL corpora found to contain gloss annotations. Therefore, the generalization of these conclusions to other SLs is limited and should be confirmed upon availability of suitable data.
- SL glosses are not an accurate representation of SLs and critical information can be missing, causing further limitations to the usability of the results (*e.g.* for SL video production) and the reliability of the automatic evaluation. However, as explained in the Introduction (Section 1), we think that given the current resource limitation, investigation of MT on glosses may be a research step to provide indications for other SL representations.
- As explained in Section 4.1, stripping the gloss suffixes from the DGS corpus was done in order to allow more clear comparisons with the automatic evaluation metrics, given the low scores incurred when the suffixes were there. It is clear that suffix stripping limits the



System	Size	automatic			human	
		BLEU $\uparrow$	ChrF $\uparrow$	TER $\uparrow$	Mean $\uparrow$	Std $\uparrow$
PHOENIX Egea Gómez et al. (2021)	642	13.13	46.86	73.33	2.74	1.64
PHOENIX baseline		20.14	52.04	56.12	3.85	1.58
PHOENIX Multi+back		<b>26.32</b>	<b>56.70</b>	<b>51.15</b>	<b>4.44</b>	<b>1.35</b>
DGS Baseline (sampled 10%)	511	3.44	29.56	78.55	2.49	1.81
DGS Multi-big (sampled 10%)		<b>6.97</b>	<b>33.16</b>	<b>73.45</b>	<b>3.28</b>	<b>1.60</b>

Table 3: System comparison based on the human evaluation. The **bold-faced** systems are significantly better than the respective baselines.

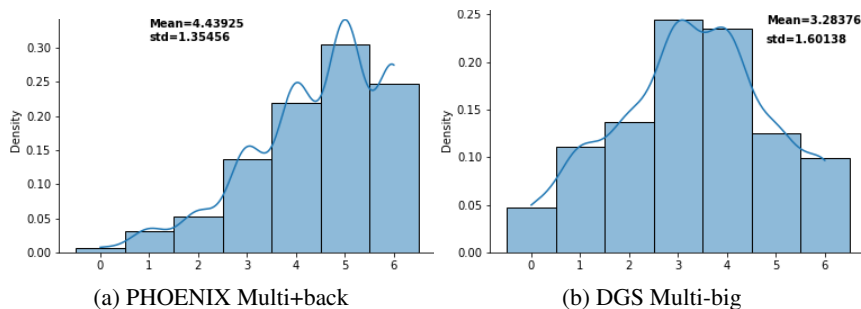


Figure 3: Density of human evaluation scores for the two best-scoring systems.

Approach	Dev BLEU $\uparrow$	Test BLEU $\uparrow$
Amin et al. (2021)	-	10.42
Egea Gómez et al. (2021) $\dagger$	-	13.13
Stoll et al. (2020)	16.34	15.26
Zhang and Duh (2021)	-	16.43
Li et al. (2021)	-	18.89
Saunders et al. (2020b)	20.23	19.10
Saunders et al. (2022)	21.93	20.08
Egea Gómez et al. (2022)	-	20.57
Walsh et al. (2022)	25.09	23.19
<b>Our PHOENIX Multi+back</b>	<b>28.41</b>	<b>26.32</b>

Table 4: Results comparison with recent work. ( $\dagger$ ) We compute the BLEU by ourselves, as the authors of paper only present the BLEU score in character level.

representational capacity of the glosses. As stated, further work should focus on the importance of the DGS gloss suffixes, the right granularity for every purpose and how to optimize their generation from MT.

- The original language direction of the DGS corpus was opposite to the one that we run our training and evaluation on. This is known to create translationese artifacts. Similar concerns have been expressed regarding the cleanliness of the PHOENIX corpus (Muller et al., 2022). Finally, whereas in MT of spoken language text, test-sets have been manually curated by professional translators for this purpose, in our experiments we use data splits,

whose test set quality may not have been confirmed.

- The human evaluation part (Section 3.5) was performed with one signer, but evaluation by more people and coverage of the Deaf community would be ideal. Additionally, due to the high effort required, we could only validate the hypothesis that the best system is significantly better than the baseline. Given more evaluation capacity one could verify whether there is a significantly perceived quality difference between methods that were scored closely by the automatic metrics (e.g. transfer learning and multilingual MT).
- The automatic metrics used have been designed for evaluating the textual output for MT of spoken languages. Whether they are applicable and reliable with regards to SLs and particularly to SL glosses has not been sufficiently analyzed and should be considered for further work. Any interpretation of the scores should consider this limitation.
- Despite the big progress regarding the model trained on the DGS-corpus (Section 5.1), the BLEU scores achieved indicate very low performance, if judged from the experience on the automatic scores for text translation for spoken languages. Whereas we tried to get

some information about this by looking at the distribution of scores, further investigation on whether such a system is usable with regards to particular use cases (interpretation, text-to-video) is needed.

## Ethical considerations

In our work, we present experiments on the German Sign Language (DGS) that are part of a broader research aiming to provide equal access to language technology for sign language users. Nevertheless, the fact that the majority of the researchers in NLP are hearing people entails the risk of developments that are not in accordance with the will of the respective communities, and therefore it is required that every research step takes them in constant consideration. In our broader research we have included members of the Deaf/deaf and hard of hearing communities as part of the research team, consultants and participants in user studies and workshops and we have been in co-operation with related unions and communication centers.

The fact that we are performing experiments on glosses, known to be inferior to the full linguistic capacity of the sign languages, should be seen as a methodological tool to aid further research.

The Public DGS corpus is provided under a limited license for linguistic research (Schulder and Hanke, 2022), prohibiting any further commercial usage. Any further usage of relevant artifacts from our work should respect the license of the original corpus. Removal of information that names or uniquely identifies individual people or offensive content was not deemed necessary. In the Public DGS corpus, participants provided consensus, whereas the content was carefully curated. The PHOENIX corpus does not pose any relevant risk because the content (weather forecasts) does not include any personal information. All other datasets used have been published with open or public domain licenses. Since our work does not use videos of SLs, there should be no ethical concerns regarding processing of human faces.

## Acknowledgements

The research reported in this paper was supported by BMBF (German Federal Ministry of Education and Research) via the project SocialWear (grant no. 01IW20002). We would like to thank Mathias Müller and Amit Moryossef for their advice with regards to DGS glosses.

## References

- Mohamed Amin, Hesahm Hefny, and Ammar Mohammed. 2021. [Sign language gloss translation using deep learning models](#). *International Journal of Advanced Computer Science and Applications*, 12(11).
- Galina Angelova, Eleftherios Avramidis, and Sebastian Möller. 2022. [Using neural machine translation methods for sign language translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 273–284, Dublin, Ireland. Association for Computational Linguistics.
- Ali Araabi and Christof Monz. 2020. [Optimizing transformer for low-resource neural machine translation](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. [Findings of the 2019 conference on machine translation \(WMT19\)](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.
- Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. [Subunets: End-to-end hand shape and continuous sign language recognition](#). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3075–3084.
- Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. [Neural sign language translation](#). In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.
- Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. [Tagged back-translation](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.
- Peng-Jen Chen, Jiajun Shen, Matthew Le, Vishrav Chaudhary, Ahmed El-Kishky, Guillaume Wenzek, Myle Ott, and Marc’Aurelio Ranzato. 2019. [Facebook AI’s WAT19 Myanmar-English translation task submission](#). In *Proceedings of the 6th Workshop on Asian Translation*, pages 112–122, Hong Kong, China. Association for Computational Linguistics.
- Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2023. [Two-stream network for sign language recognition and translation](#).
- Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. [Semi-supervised](#)

- learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany. Association for Computational Linguistics.
- Luis Chiruzzo, Euan McGill, Santiago Egea-Gómez, and Horacio Saggion. 2022. [Translating Spanish into Spanish Sign Language: Combining rules and data-driven approaches](#). In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 75–83, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Judith Collins, Granville Tate, and Paul Hann. 2012. [A translation studies approach to glossing using ELAN](#). *International Journal of Interpreter Education*, 4:83–91.
- Mathieu De Coster, Dimitar Shterionov, Mieke van Herreweghe, and Joni Dambre. 2022. [Machine translation from signed to spoken languages: State of the art and challenges](#). *ArXiv*, abs/2202.03086.
- Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. [Copied monolingual data improves low-resource neural machine translation](#). In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.
- Santiago Egea Gómez, Luis Chiruzzo, Euan McGill, and Horacio Saggion. 2022. [Linguistically enhanced text to sign gloss machine translation](#). In *Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings*, page 172–183, Berlin, Heidelberg. Springer-Verlag.
- Santiago Egea Gómez, Euan McGill, and Horacio Saggion. 2021. [Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation](#). In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 18–27, Online (Virtual Mode). INCOMA Ltd.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. [Continuous measurement scales in human evaluation of machine translation](#). In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.
- Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. [Universal neural machine translation for extremely low resource languages](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.
- Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. [Extending the Public DGS Corpus in size and depth](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. [Marian: Fast neural machine translation in C++](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.
- Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. [Findings of the 2022 conference on machine translation \(WMT22\)](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Tom Kocmi and Ondřej Bojar. 2020. [Efficiently reusing old models across languages via transfer learning](#). In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 19–28, Lisboa, Portugal. European Association for Machine Translation.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Philipp Koehn. 2005. [Europarl: A parallel corpus for statistical machine translation](#). In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. [Moses: Open source toolkit for statistical machine translation](#). In

- Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. [MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release.](#)
- Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. [Öffentliches DGS-Korpus: Annotationskonventionen / Public DGS Corpus: Annotation conventions.](#) Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.](#) In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Gabriele Langer, Susanne König, and Silke Matthes. 2014. [Compiling a Basic Vocabulary for German Sign Language \(DGS\) – lexicographic issues with a focus on word senses.](#) In *Proceedings of the 16th EURALEX International Congress*, pages 767–786, Bolzano, Italy. EURAC research.
- Dongxu Li, Chenchen Xu, Liu Liu, Yiran Zhong, Rongzhao Wang, Lars Petersson, and Hongdong Li. 2021. [Transcribing natural languages for the deaf via neural editing programs.](#) In *AAAI Conference on Artificial Intelligence*.
- Guanlin Li, Lemao Liu, Guoping Huang, Conghui Zhu, and Tiejun Zhao. 2019. [Understanding data augmentation in neural machine translation: Two perspectives towards generalization.](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5689–5695, Hong Kong, China. Association for Computational Linguistics.
- V. López-Ludeña, C. González-Morcillo, J.C. López, E. Ferreira, J. Ferreiros, and R. San-Segundo. 2014. [Methodology for developing an advanced communications system for the deaf in a new domain.](#) *Knowledge-Based Systems*, 56:240–252.
- Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. [Data augmentation for sign language gloss translation.](#) In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–11, Virtual. Association for Machine Translation in the Americas.
- Anke Müller, Thomas Hanke, Reiner Konrad, Gabriele Langer, and Sabrina Wähl. 2020. [From dictionary to corpus and back again – linking heterogeneous language resources for DGS.](#) In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 157–164, Marseille, France. European Language Resources Association (ELRA).
- Mathias Muller, Zifan Jiang, Amit Moryossef, Annette Rios Gonzales, and Sarah Ebling. 2022. [Considerations for meaningful sign language machine translation based on glosses.](#) *ArXiv*, abs/2211.15464.
- Graham Neubig and Junjie Hu. 2018. [Rapid adaptation of neural machine translation to new languages.](#) In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.
- Toan Q. Nguyen and David Chiang. 2017. [Transfer learning across low-resource, related languages for neural machine translation.](#) In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Achraf Othman and Mohamed Jemni. 2012. [English-asl gloss parallel corpus 2012: Aslg-pc12.](#) In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon LREC*.
- Sinno Jialin Pan and Qiang Yang. 2010. [A survey on transfer learning.](#) *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.
- 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.
- Elena Pizzuto, Paolo Rossini, and Tommaso Russo. 2006. [Representing signed languages in written form: questions that need to be posed.](#) In *Proceedings of the Workshop on the Representation and Processing of Sign Languages, 2006 Language Resources and Evaluation Conference*, pages 1–6. ELRA.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation.](#) In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. [Neural machine translation for low-resource languages: A survey](#). *ACM Computing Surveys*.
- Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and M. Sabokrou. 2021. [Sign language production: A review](#). *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3446–3456.
- Rubén San-Segundo, Juan Montero, Ricardo Córdoba, V. Sama, Fernando Fernández-Martínez, Luis D’Haro, Verónica López-Ludeña, D. Sánchez, and A. García. 2012. [Design, development and field evaluation of a spanish into sign language translation system](#). *Pattern Analysis and Applications*, 15.
- Ben Saunders, Necati Cihan Camgöz, and R. Bowden. 2020a. [Adversarial training for multi-channel sign language production](#). *ArXiv*, abs/2008.12405.
- Ben Saunders, Necati Cihan Camgöz, and R. Bowden. 2020b. [Progressive transformers for end-to-end sign language production](#). *ArXiv*, abs/2004.14874.
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. [Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5141–5151.
- Marc Schuler and Thomas Hanke. 2022. [How to be FAIR when you CARE: The DGS Corpus as a case study of open science resources for minority languages](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 164–173, Marseille, France. European Language Resources Association.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Rico Sennrich and Biao Zhang. 2019. [Revisiting low-resource neural machine translation: A case study](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Stephanie Stoll, Necati Camgoz, Simon Hadfield, and Richard Bowden. 2020. [Text2sign: Towards sign language production using neural machine translation and generative adversarial networks](#). *International Journal of Computer Vision*.
- MILDRED C. Templin. 1957. *Certain Language Skills in Children: Their Development and Interrelationships*, ned - new edition edition, volume 26. University of Minnesota Press.
- Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multi-lingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- Jörg Tiedemann and Santhosh Thottingal. 2020. [OPUS-MT – building open translation services for the world](#). In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.
- Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *ArXiv*, abs/1706.03762.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. [SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python](#). *Nature Methods*, 17:261–272.
- Christian Vogler and Carol Neidle. 2012. [A new web interface to facilitate access to corpora: development of the ASLLRP data access interface](#). In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*.

Harry Walsh, Ben Saunders, and Richard Bowden. 2022. [Changing the representation: Examining language representation for neural sign language production](#). In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 117–124, Marseille, France. European Language Resources Association.

Christian Wartena. 2019. [A probabilistic morphology model for german lemmatization](#). In *Conference on Natural Language Processing*, Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), pages 40 – 49.

Jiajun Zhang and Chengqing Zong. 2016. [Exploiting source-side monolingual data in neural machine translation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Tong Zhang and Bin Yu. 2005. [Boosting with early stopping: Convergence and consistency](#). *The Annals of Statistics*, 33(4).

Xuan Zhang and Kevin Duh. 2021. [Approaching sign language gloss translation as a low-resource machine translation task](#). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 60–70, Virtual. Association for Machine Translation in the Americas.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. [Transfer learning for low-resource neural machine translation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

## Appendix

### A Statistics of sign language datasets

We present the statistical analysis of all sign language corpora in Table 5 and Table 6. Out-of-Vocabulary (OOV) are the words that only appear in development or test set and singletons are the least frequent words appearing only once.

### B Effect of source-side preprocessing

Previous work on *gloss-to-text* translation (Moryossef et al., 2021) suggested the use of lemmatization of the spoken language words as part of their data augmentation pipelines. Lemmatization of spoken language words in the *text-to-gloss* is justified by the fact that they contain inflection (*e.g.* for nouns, or verb conjugation), something that does not exist in SL glosses (Moryossef et al., 2021). Therefore

we ran preliminary experiments on PHOENIX with lemmatization in order to determine the baseline settings. The results of these experiments on PHOENIX appear in Table 7. One can see that lemmatization incurs considerable automatic metric improvements, with an improvement of around 0.9 BLEU score on test set.

In Table 8, we demonstrate the statistics of the source side after the data preprocessing. We can see that the vocabulary size has dropped by around 23% and 25% after data preprocessing, respectively.

### C Statistics of augmented datasets

As an appendix for Section 4.5.1, we present here the statistics of the datasets through our data augmentation methods in Table 9.

### D Effect of DGS-Corpus gloss suffixes to the automatic evaluation

We train these two multilingual NMT systems on the DGS corpus under the same configurations but they are evaluated against two different types of reference translations: the original DGS glosses and the glosses with stripped suffixes. In Table 10 we can observe the results, indicating that generating glosses with correct suffixes is a much harder problem and that current automatic metrics are not optimized to measure that.

### E Statistics of additional datasets

The statistics of additional datasets used for data augmentation (Section 3.1), semi-supervised NMT (Section 3.2), warm-start of transfer learning (Section 3.3) and multilingual NMT (Section 3.4) are shown in Table 11.

### F Vocabulary-level and sentence-level overlap for custom split

We calculate the vocabulary overlap over the difference of the vocabulary counts and OOVs in Table 6 and Table 8 of the preprocessed datasets. The DGS vocabulary overlap is 79.61% between test and train set and 80.24% between dev and train set. The vocabulary overlap for NCSLGR is 69.50% and 70.60%, respectively. The official splits of the PHOENIX dataset (that has been used in most of the SoTA papers and related work) have a much higher vocabulary overlap, of 95.45% and 95.08% respectively.

	PHOENIX						Generalized DGS					
	Train	Text Dev	Test	Train	Glosses Dev	Test	Train	Text Dev	Test	Train	Glosses Dev	Test
<b>Sentences</b>	7,096	519	642	7,096	519	642	54,325	4,470	5,113	54,325	4,470	5,113
<b>Vocabulary</b>	2,887	951	1,001	1,085	393	411	20,868	4,617	4,992	19,521	4,894	5,688
<b>Tot. words</b>	99,081	6,820	7,816	55,247	3,748	4,264	472,609	36,629	44,452	301,772	21,715	28,405
<b>Tot. OOVs</b>	-	57	60	-	14	19	-	971	1,080	-	614	752
<b>Singletons</b>	1,077	-	-	355	-	-	9,946	-	-	6,286	-	-

Table 5: Statistics of both corpora.

	NCSLGR					
	Train	Text Dev	Test	Train	Glosses Dev	Test
<b>Sentences</b>	1,500	177	178	1,500	177	178
<b>Vocabulary</b>	2,796	745	754	2,287	662	639
<b>Tot. words</b>	13,904	1,860	1,832	11,064	1,471	1,449
<b>Tot. OOVs</b>	-	219	230	-	210	192
<b>Singletons</b>	1,665	-	-	1,209	-	-

Table 6: Statistics of NCSLGR.

Preprocessing	Dev			Test		
	BLEU	ChrF	TER	BLEU	ChrF	TER
<b>No lemmatization</b>	27.90	57.50	49.92	25.44	56.30	51.76
<b>With lemmatization</b>	28.41	57.54	49.39	26.32	56.70	51.15

Table 7: Effect of lemmatization on preliminary experiments of the PHOENIX corpus.

	PHOENIX						Generalized DGS					
	Train	Text Dev	Test	Preprocessed text			Train	Text Dev	Test	Preprocessed text		
<b>Sentences</b>	7,096	519	642	7,096	519	642	54,325	4,470	5,113	54,325	4,470	5,113
<b>Vocabulary</b>	2,887	951	1,001	2,216	793	836	20,868	4,617	4,992	15,170	3,497	3,791
<b>Tot. words</b>	99,081	6,820	7,816	99,081	6,820	7,816	472,609	36,629	44,452	472,609	36,629	44,452
<b>Tot. OOVs</b>	-	57	60	-	39	38	-	971	1,080	-	691	773
<b>Singletons</b>	1,077	-	-	765	-	-	9,946	-	-	6,929	-	-

Table 8: Statistics of preprocessed corpora.

	PHOENIX Text		PHOENIX Glosses		DGS Text		DGS Glosses	
	Authentic	Synthetic	Authentic	Synthetic	Authentic	Synthetic	Authentic	Synthetic
<b>Original</b>	7,096	-	7096	-	54,325	-	54,325	-
<b>Combining</b>	7,096	3 * 7096	4 * 7096	-	54,325	54,325	2 * 54,325	-
<b>Back-translation</b>	7,096	7,096	2 * 7,096	-	54,325	54,325	2 * 54,325	-
<b>Forward-translation</b>	7,096	1,023	7,096	1,023	2 * 54,325	-	54,325	54,325

Table 9: Statistics of augmented datasets

DGS gloss reference	Dev			Test		
	BLEU-4	ChrF	TER	BLEU-4	ChrF	TER
<b>Original_DGS</b>	1.21	32.67	92.24	0.81	31.34	91.78
<b>Generalized_DGS</b>	6.06	35.18	74.51	5.32	33.55	74.71

Table 10: Results comparison with different DGS gloss references.

	<b>Dataset</b>	<b>Language (pair)</b>	<b>#</b>
<b>Monolingual</b>	German weather domain sentences	de	1, 203
	Europarl-v10	de	2, 107, 971
<b>Bilingual</b>	Tatoeba-Challenge	de-de	4, 574, 760
	News-commentary-v16	de-en	398, 981
	Europarl-v10	de-en	1, 828, 521
	ASLG-PC12	en-ASL	87, 710

Table 11: Auxiliary language datasets overview.

For sentence lexical overlap within the DGS corpus there are no sentences with 100% lexical overlap, 6.45% of the test sentences had approximately 90% overlap with the train set and 1.51% of the test sentences had approximately 80% overlap with the train set, whereas the sentence-level overlap between the dev set and the train set is similar. For the NCSLGR test set, the overlaps are 0, 12.23% and 4.26% and for dev set 0, 14.44% and 5.88% respectively. The sentence-level lexical overlaps of our custom splits are lower or comparable to the ones of the official PHOENIX corpus. These are 0, 11.06% and 16.04% in the test set and 0, 7.32% and 15.42% in the dev set respectively.

## G Statistics on computational experiments

Experiments were run in a GPU computational cluster on an Nvidia RTX A-6000, using 1 GPU, 2 CPUs and 50 GB of RAM and summing approximately 100 hours of computational time.

## H Human Evaluation

The human evaluator and consultant is a user of the German Sign Language (DGS) and an employed member of our research team, having consented on the use of their evaluation effort for this research. The interface used for the human evaluation can be seen in Figure 4.

The evaluation rating was that outputs marked with 0 failed to translate any of the contents of the original sentence, whereas outputs marked with 6 show no significant mistakes in the translation. Insignificant mistakes or minor issues dropped the rating from 6 to a 5 or 4, some correctly translated words or phrases pushed it up from 0 to 1 or 2 or, if some information was conveyed but it was missing significant interrelations, to 3.



Click if it is a bad reference

PHOENIX GERMAN SENTENCE REFERENCE 0

regen und schnee lassen an den alpen in der nacht nach im norden und nordosten fallen hier und da schauer sonst ist das klar

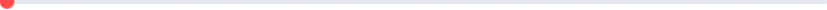
PHOENIX GLOSS REFERENCE 0

regen schnee region verschwinden nord regen koennen region stern koennen sehen

PREDICTION 1

nacht nordost schauer alpen ix region wolke klar


Score for prediction 1

0  6

PREDICTION 2

alpen regen schnee heute nacht region regen nord nordost klar stern koennen sehen

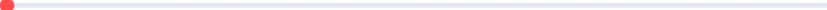
Score for prediction 2

0  6

PREDICTION 3

alpen regen schnee nacht nord nordost bisschen schauer sonst region klar

Score for prediction 3

0  6

Comment here if available

1 out of 642

NEXT

PREVIOUS

SAVE & PAUSE

Figure 4: The interface used for human evaluation.

## ACL 2023 Responsible NLP Checklist

---

### A For every submission:

- A1. Did you describe the limitations of your work?  
*Limitations have been described as separate section after the Conclusions, as required by the ACL instructions.*
- A2. Did you discuss any potential risks of your work?  
*Risks with regards to technological developments and their acceptance by communities using the Sign Languages are described in the Section of the "Ethical Considerations".*
- A3. Do the abstract and introduction summarize the paper's main claims?  
*Abstract and Section 1.*
- A4. Have you used AI writing assistants when working on this paper?  
*Left blank.*

### B Did you use or create scientific artifacts?

*Section 4*

- B1. Did you cite the creators of artifacts you used?  
*Section 4 (4.1, 4.3)*
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?  
*Section 4.3 (software), Ethical considerations section (datasets)*
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?  
*discussion in Ethical considerations section*
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?  
*discussion in Ethical considerations section*
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?  
*Section 4 (4.1)*
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.  
*Appendix*

### C Did you run computational experiments?

*Sections 4 and 5*

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?  
*Appendix*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?  
*Sections 4 and Appendix*
- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?  
*Section 5*
- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?  
*Sections 3.5, 4.2, 4.3*
- D**  **Did you use human annotators (e.g., crowdworkers) or research with human participants?**  
*Sections 3.5, 5.2*
- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?  
*Appendix*
- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?  
*Appendix*
- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?  
*We didn't create any data. Evaluator consented on the use of their evaluation effort.*
- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?  
*Not applicable. There was no data collection*
- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?  
*We only had one human evaluator. Further demographic and geographic characteristics are not relevant to the experiment, given the current state of research, and would unnecessarily reveal personal information of the evaluator.*