# Select and Reorder: A Novel Approach for Neural Sign Language Production

**Harry Walsh, Ben Saunders, Richard Bowden**
CVSSP, University of Surrey
Guildford, United Kingdom
{harry.walsh, b.saunders, r.bowden}@surrey.ac.uk

**Abstract**

Sign languages, often categorised as low-resource languages, face significant challenges in achieving accurate translation due to the scarcity of parallel annotated datasets. This paper introduces Select and Reorder (S&R), a novel approach that addresses data scarcity by breaking down the translation process into two distinct steps: Gloss Selection (GS) and Gloss Reordering (GR). Our method leverages large spoken language models and the substantial lexical overlap between source spoken languages and target sign languages to establish an initial alignment. Both steps make use of Non-AutoRegressive (NAR) decoding for reduced computation and faster inference speeds. Through this disentanglement of tasks, we achieve state-of-the-art BLEU and Rouge scores on the Meine DGS Annotated (mDGS) dataset, demonstrating a substantial BLUE-1 improvement of 37.88% in Text to Gloss (T2G) Translation. This innovative approach paves the way for more effective translation models for sign languages, even in resource-constrained settings.

**Keywords:** Sign Language Translation (SLT), Natural Language Processing (NLP), Non-AutoRegressive (NAR) Generation

## 1. Introduction

Sign languages are multi-channel visual languages with complex grammatical rules and structure (Stokoe, 1980). The World Health Organisation estimates that 430 million people worldwide are Deaf or Hard of Hearing (HOH) (WHO, 2021), hence the need for accessibility and inclusivity. Sign languages are visual forms of communication, expressed through the manual articulation of gestures and non-manual features. The grammar and lexicon of the world's 300 sign languages are country-dependent and variations can develop from region to region, often sharing a large lexical overlap with each country's respective spoken language (National Geographic Society, 2017). In the USA, where 90% of deaf children are born to hearing families (Schein and Delk, 1974) sign languages may be acquired at different ages (LeMaster and Monaghan, 2005), resulting in potential grammar variations (Cormier et al., 2012; Skotara et al., 2012).

Sign Language Production (SLP) aims to generate sign language sequences from spoken language sentences, it is often decomposed into two concurrent tasks: Text to Gloss (T2G), translating spoken language to gloss sequences, and Gloss to Sign (G2S), creating sign language videos from gloss intermediaries. The quality of SLP videos depends on the initial T2G translation. However, current research has predominantly focused on G2S production (Saunders et al., 2020a; Hwang et al., 2021; Huang et al., 2021; Rastgoo et al., 2021; San José-Robertson et al., 2004), leaving a crucial gap in the SLP pipeline. This paper addresses this gap with a novel Select and Reorder (S&R) approach to T2G translation. While it is possible to directly synthesise a sign language sequence from a spoken language sentence (Text to Pose (T2P)), a two-step approach has been shown to yield superior translations (Saunders et al., 2020a).

To achieve an effective T2G translation, it is essential to transform the source spoken language sentence into the target gloss representation while preserving the original meaning. This transformation must include a change in lexicon and in order (El-dali, 2011), as shown by Figure 1. Semantic notations of sign language, such as gloss, share a large proportion of vocabulary with their country of origin. This causes T2G translation to have a high lexical overlap between the source and target sequences, a unique property of sign language translation. By first formatting the gloss tokens with lemmatization we find that datasets such as Meine DGS Annotated (mDGS) Konrad et al. (2020) and RWTH-PHOENIX-Weather-2014**T** (PHOENIX14**T**) (Camgoz et al., 2018) have a lexical overlap of 35% and 33%, respectively.

Neural Machine Translation (NMT) typically requires around 15 million sequences of parallel data to outperform statistical approaches (Koehn and Knowles, 2017). By this definition, sign languages can be defined as low-resource languages, with the largest annotated datasets containing only 50k parallel examples Konrad et al.



*LREC-COLING 2024*, pages 14531–14542

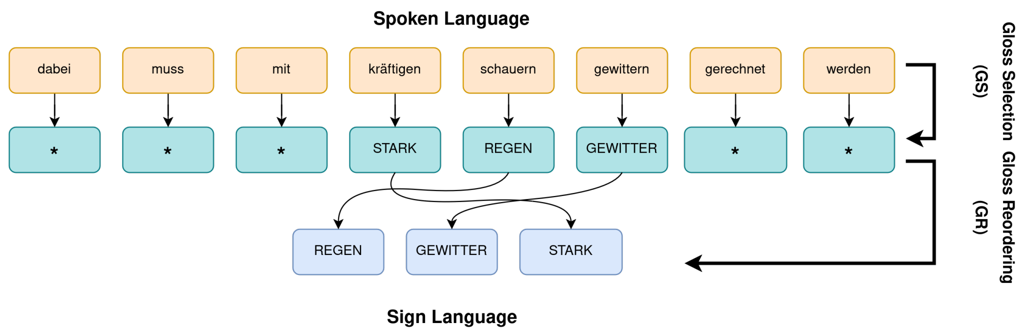


Figure 1: An example of Gloss Selection (GS) and Gloss Reordering (GR) being applied to a sentence from the RWTH-PHOENIX-Weather-2014**T** (PHOENIX14**T**) dataset.

(2020). In an attempt to circumvent this limitation and exploit the lexical overlap, we propose S&R, an approach that breaks down the translation task into two sub-tasks, Gloss Selection (GS) and Gloss Reordering (GR).
As the first step in the S&R pipeline, GS learns to predict the corresponding gloss for each word in the spoken language sentence, thus producing Spoken Language Order (SPO) gloss (Marshall and Hobsbaum, 2015). To create the ground truth SPO gloss for training, we obtain a one-to-one alignment between the text and gloss, exploiting the lexical overlap found using large spoken language models such as BERT and Word2Vec.
For the next step, GR changes the gloss sequence from SPO to Sign Language Order (SIO). We explore two approaches, a statistical based pre-reordering method (Nakagawa, 2015) and a deep learning approach. The statistical approach uses a Top-Down Bracketing Transduction Grammar (BTG) based pre-ordering model, that learns a number of reordering rules using our alignment, Part Of Speech (POS) tags and word classes. The corresponding deep learning approach uses a transformer with a reordering mask at inference time. The mask constrains the model to only predict tokens that are present at the input, meaning the model can only reorder.
Both GS and GR are Non-AutoRegressive (NAR) models, executing decoding in a single pass. This characteristic leads to decreased computational requirements and accelerates the inference process, which is a valuable asset for real-time translation.
The key contributions of this work can be summarized as the following:

- S&R, a novel two step approach to T2G translation.
- An approach to building a pseudo alignment between two paired sequences.
- State-of-the-art BLEU and Rouge scores on mDGS and PHOENIX14**T**.

The rest of this paper is organised as follows: In section 2 we provide an overview of the literature, then in section 3 we explain our S&R approach to T2G NMT. Section 4 explains the setup for the proceeding experiments in section 5 where we present quantitative and qualitative results. Finally, in section 6 we draw conclusions from the experiments and suggest possible future work.

## 2. Related Work

**Sign Language Recognition & Translation:** For the last 30 years computational sign language Translation has been an active area of research (Tamura and Kawasaki, 1988). Initial neural research focused on isolated Sign Language Recognition (SLR), where Convolutional Neural Network (CNN) were used to classify isolated instance of a sign (Lecun et al., 1998). Advancements in the field led to Continuous Sign Language Recognition (CSLR), where a video must first be segmented into constituent signs before being classified (Koller et al., 2015). The task of Sign to spoken language translation aims to convert continuous sign language to spoken language text, directly (Sign to Text (S2T)) or via gloss (Sign to Gloss to Text (S2G2T)) (Camgoz et al., 2018).
**Sign Language Production (SLP):** SLP is the reverse task of SLT, which aims to produce a continuous sequence of sign language given a spoken language sentence. As above, this can be performed either using gloss as an intermediate representation (Text to Gloss to Pose (T2G2P)) (Stoll et al., 2018) or directly from the spoken language (T2P) (Saunders et al., 2020a). State-of-the-art approaches use a transformer with Multi-Headed Attention (MHA) (Saunders et al., 2020c; Stoll et al., 2022). The output pose of these systems can be mapped to a photo-realistic signer

(Saunders et al., 2020b) or 3D mesh (Stoll et al., 2022). Older approaches used a parameterized gloss that is converted to a pose and mapped to a graphical avatar (Bangham et al., 2000; Cox et al., 2002; Zwitserlood et al., 2004; Efthimiou et al., 2012; Van Wyk, 2008), but this suffers from lack of non-manuals, under-articulation and robotic movement. Recently, alternate representations to gloss have been explored (Jiang et al., 2022; Walsh et al., 2022), namely SignWriting (Kato, 2008) and the Hamburg Notation System (HamNoSys) (Hanke, 2004). However, previous work has failed to achieve high T2G results, due to the limited dataset size. In this paper, we attempt to overcome the data deficiency by using a S&R approach.

**Machine Translation (MT):** MT is an NLP task that deals with the automatic translation from a source to a target language. Prior to the introduction of deep learning approaches to the field (Singh et al., 2017), statistical based methods were state-of-the-art (Della Pietra, 1994; Och and Ney, 2002; Koehn et al., 2003). However, these models struggled when the source and target languages had large changes in word order (Genzel, 2010). To overcome the issues with long-distance word dependencies, pre-reordering was used, where the source language is reordered into the target language order. This was shown to improve the performance of phase based statistical machine translation systems (Neubig et al., 2012; Hitschler et al., 2016; Nakagawa, 2015). To train these statistical models an alignment between the source and target words is found (Della Pietra, 1994; Vogel et al., 1996). Since then pre-reordering has been applied to NMT with limited success (Zhao et al., 2018; Du and Way, 2017; Sabet et al., 2020). Recently word alignment has been used to train multilingual models and has shown good performance when applied to low resource languages (Lin et al., 2020).

**Low resource NMT:** NMT has shown significant performance in large data scenarios but often struggles on low-resource languages (Stoll et al., 2018). For SLP, there is a lack of large annotated text to sign corpora. To overcome this, common NLP approaches are transfer learning (Zoph et al., 2016), use of large language models (Zhu et al., 2020) or data augmentation (Moryossef et al., 2021).

## 3. Methodology

Text to Gloss (T2G) translation aims to learn the mapping from a source spoken language sequence $X = (x_1, x_2, ..., x_W)$ with W words, to a sequence of glosses, $Y = (y_1, y_2, ..., y_G)$ with G glosses. Therefore, a T2G model learns the conditional probabilities $p(Y|X)$.

A model that learns $p(Y|X)$ jointly learns a change in lexicon and order, a challenging task. In this paper, we disentangle the two tasks into GS and GR, as shown in Figure 2, and define a new task of Text to Spoken Language Order Gloss (T2SPOG).

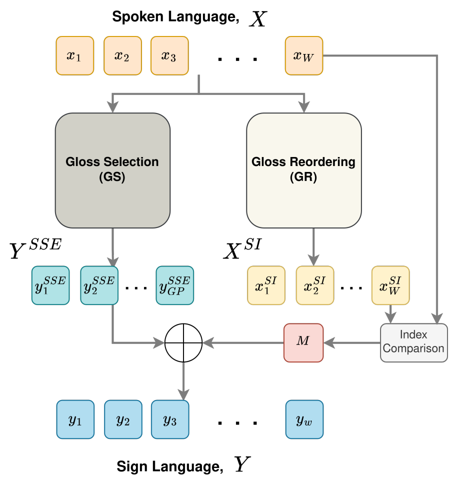


Figure 2: A overview of the Select and Reorder (S&R) approach

The GS model learns the mapping of a sequence of words $X = (x_1, x_2, ..., x_W)$, to a sequence of SPO glosses and pad tokens, $Y^{SPO} = (y_1^{SPO}, y_2^{SPO}, ..., y_W^{SPO})$. Our approach relies on creating a one-to-one alignment, $A$, of words to glosses, which limits us to sequences where ($W \geq G$), hence $X$ and $Y^{SPO}$ share the same sequence length, W. To create the gloss in SPO for the GS model the alignment, $A()$, is applied to the gloss;

$$Y^{SPO} = A(Y) \tag{1}$$

We define GR as a permutation task, where the model learns to reorder words in spoken language order, $X$, to words in sign order, $X^{SIO} = x_1^{SIO}, x_2^{SIO}, ..., x_W^{SIO})$. The source and target sequence share the same vocabulary and sequence length, W. Thus, GR learns $p(X^{SIO}|X)$. To create the text in sign order for the GR model the alignment, $A()$, is applied to the text;

$$X^{SIO} = A(X) \tag{2}$$

As shown by Figure 2, to obtain a full translation the outputs of the GS and the GR networks must be joined. We call this full method Select and Reorder (S&R). To correctly join the outputs the GR subtask creates a mapping $M()$. Applying the mapping to the SPO gloss gives a full translation (gloss in sign language order);

$$p(Y|X) = M(p(Y^{SPO}|X)) \tag{3}$$

Both input and target sequences are tokenized at the word level. The GS and GR networks are trained using cross-entropy loss, $L_{cross}$, calculated using the predicted target sequence, $\hat{x}$ and the ground truth sequence, $x^*$.

In the following sub-sections, we provide an overview of GS followed by GR. Firstly, we show how we create an alignment between the source and target languages, using two different word embeddings. Subsequently, we explain how this alignment is used in conjunction with the GS model to predict the intermediary SPO glosses. Finally, we outline two methods for GR, followed by an explanation of how the two sub-tasks are joined to obtain a full translation.

## 3.1. Select

As shown by Figure 1 (top to middle row), GS can be defined as the task of choosing the corresponding glosses for each word of a given spoken language sentence. To achieve this, an alignment must first be found to create a pseudo gloss sequence in SPO.

### 3.1.1. Alignment

Using the lexical overlap between the source and target language, a pseudo alignment can be found. For example given the sentence "*what is your name?*" it is clear to see which words correspond to which glosses in the translation *"YOU NAME WHAT?"*. Using two different word embedding techniques, Word2Vec (Mikolov et al., 2013) and BERT (Chan et al., 2020), we create a mapping between our spoken language words, X, and our glosses, Y. We can define a word gloss pair as a strong alignment if they share the same meaning. A strong connection can be established if the pair share a similar lexical form (e.g. word = run, gloss = RUN), for which we use Word2Vec. Where an accurate lexical mapping cannot be found, we use BERT to find connections based on meaning (e.g. word = weather, gloss = WEATHERFORECAST). When using German Sign Language - Deutsche Gebärdensprache (DGS) we first apply a compound word splitting algorithm (Tuggener, 2016) before creating the alignment.

For a sequence of words, $X$, and a sequence of gloss, $Y$ we apply Word2Vec as:

$$X_{Vec} = Word2Vec(X) \tag{4}$$

$$Y_{Vec} = Word2Vec(Y) \tag{5}$$

where $X_{Vec} \in \mathbb{R}^{W \times E}$ and $Y_{Vec} \in \mathbb{R}^{G \times E}$. We take the outer product between the resultant two embeddings to give us the Word2Vec alignment:

$$A_{Vec} = Y_{Vec} \otimes X_{Vec} \tag{6}$$

where $A_{Vec} \in \mathbb{R}^{G \times W}$. We filter the strongest connections, only keeping those that are above a constant, $\alpha$. Then we repeat the process this time using BERT:

$$X_{BERT} = BERT(x) \tag{7}$$

$$Y_{BERT} = BERT(y) \tag{8}$$

$$A_{BERT} = Y_{BERT} \otimes X_{BERT} \tag{9}$$

Where $X_{BERT} \in \mathbb{R}^{W \times E}$, $Y_{BERT} \in \mathbb{R}^{G \times E}$ and $A_{BERT} \in \mathbb{R}^{G \times W}$. When embedding with BERT, a wordpiece tokenizer is applied to the text. We average the sub-unit alignment in order to create an alignment at the word level. We find BERT embeddings capture the meaning of tokens, making this approach better for finding alignment between words and glosses that have different lexical forms. The BERT alignment is used to find any remaining connections not found by the Word2Vec alignment, where our final alignment is defined as;

$$A = A_{BERT} + (\alpha * A_{Vec}) \tag{10}$$

Note, as the alignment creates a one-to-one mapping, the proposed approach is limited to many-to-one sequences e.g. where the source sequence is longer than the target, ($W \geq G$). Furthermore, as the T2SPOG task is a many-to-one task, any words that are not aligned are mapped to a pad token, '*'. This ensures the sequence lengths of $Y^{SPO}$ and $X$ are the same.

Figure 3 and 4 shows a heat map of the alignment found between a German spoken language sentence and the corresponding gloss sequence from the PHOENIX14**T** and mDGS dataset, respectively. Figure 3 shows a clear alignment is found between the word and gloss "MORGEN", as they share the same lexical form. Additionally, an alignment is found between words with the same meaning e.g. "JETZT" and "nun".

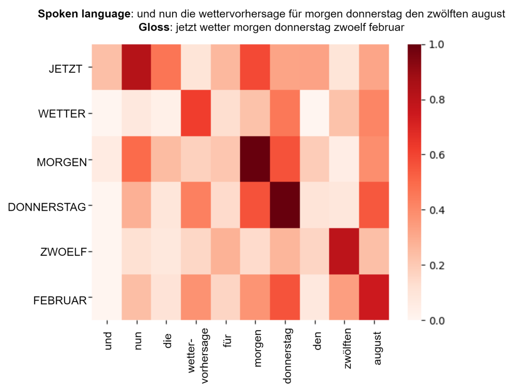


Figure 3: An example of the alignment found using BERT embeddings to connect the spoken language to the glosses on the PHOENIX14**T** dataset. (SRC: "and now the weather forecast for tomorrow thursday the twelfth of august", TRG: "now weather tomorrow thursday twelve february")

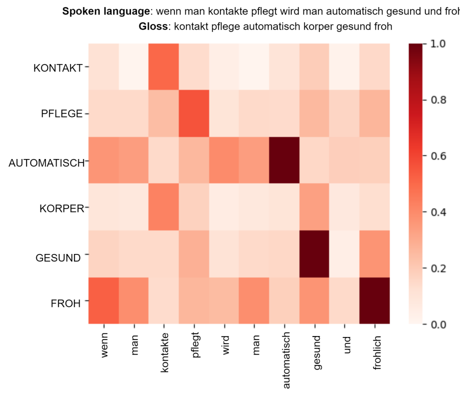


Figure 4: An example of the alignment found using BERT embeddings to connect the spoken language to the glosses on the mDGS dataset. (SCR: "when you keep in touch you automatically become healthy and happy", TRG: "contact care automatic body healthy glad")

#### 3.1.2. Architecture

We build our GS model as an encoder-decoder transformer (Vaswani et al., 2017). We pass the encoder and decoder the same spoken language sentence, whilst removing the auto-regressive feature of the decoder for reduced computational cost. This also makes each prediction independent of the previous, removing any possible negative feedback from incorrect guesses at inference time. Additionally, we alter the decoder's forward masking to allow the model to see all tokens in the sequence.

### 3.2. Reorder

The goal of the second sub-task, GR, is to create a mapping, $M()$, that reorders a sequence from SPO to SIO. By comparing the index movements of words between the input and output of the models we create a mapping. A visualization of applying $M()$ can be seen in Figure 1 (middle to bottom row).

To facilitate the creation of ground-truth data for this task, we can leverage the alignment discussed in section 3.1.1, referred to as $A$. This alignment enables us to generate SPO gloss and text in SIO. Consequently, we have the option to train our reordering model on either the gloss or the text. We opt to train on the text for two reasons. Firstly, gloss does not offer a perfect representation of sign language due to its inherent limitations. Secondly, we hypothesise that training on the higher-resourced language (text) will yield superior performance, as it contains richer structural information about the language.

In this section, we propose two approaches to tackle GR. We start by explaining the statistical approach from Nakagawa (2015), followed by our deep learning method.

#### 3.2.1. Statistical Approach

Our first approach uses the BTG method to learn a mapping from spoken to sign order (Nakagawa, 2015). The approach represents a source sentence as a binary tree, where each non terminal node can be one of three types: straight, inverted or terminal. The structure of the tree is dependent on the POS tags and word classes of the sentence. To create our word classes we use the Brown clustering method (Brown et al., 1992). Words are grouped into a single cluster if they are semantically related. Words are assumed to be semantically related if the distribution of surrounding words are similar. We use a pre-trained language model to tag the spoken language words with their POS. The model acquires a set of rules designed to restructure the tree in a manner that maximises reordering accuracy. To assess this accuracy, we rely on the alignment provided in section 3.1.1.

#### 3.2.2. Learned Approach

Our second approach uses deep learning to learn $p(X^{SIO}|X)$, using an encoder-decoder transformer. Once again we remove the auto-regressive feature from the decoder and change the mask to allow the model to see all tokens in the input sequence. At inference time we apply a mask to the output of the model, which ensures the model predicts all tokens that are present on the input, hence the model is limited to reordering. The mask is a binary vector with entries only in the index's of the tokens present in the input. At each decoding step, the predicted token is removed from the mask. If duplicates of the gloss are present then it is only removed once all copies have been predicted.

### 3.3. Select and Reorder

The GS model learns $p(Y^{SPO}|X)$ and the GR model learns $p(X^{SIO}|X)$. To obtain a full translation the output of the two models must be joined, as shown by Figure 2. As depicted, the predictions of the GR cannot be directly applied to the gloss in SPO. By analysing the index movement of words between the input and output a mapping, $M$, can be created that changes the order from spoken to sign. We train each task independently and join the outputs by applying the mapping, $M()$ from GR to the output of the GS model;

$$Y = M(Y_{SPO}) \tag{11}$$

This provides a full translation from a spoken language input sequence to a target gloss sequence.

## 4. Experimental Setup

In this section, we explain the experimental setup for the proceeding experiments.

| PHOENIX14**T** | DEV SET | | | | | TEST SET | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE |
| GS (Spoken order) | 62.69 | 41.22 | 29.04 | 21.31 | 58.32 | 60.12 | 39.22 | 27.40 | 20.19 | 57.10 |
| GS (Sign order) | 62.69 | 38.86 | 25.67 | 17.84 | 56.37 | 60.13 | 35.15 | 21.84 | 14.49 | 54.60 |

Table 1: A table showing the result of performing Gloss Selection (GS) on the RWTH-PHOENIX-Weather-2014**T** (PHOENIX14**T**) dataset.

| mDGS | DEV SET | | | | | TEST SET | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE |
| GS (Spoken order) | 42.91 | 23.21 | 12.47 | 6.89 | 42.63 | 43.06 | 23.23 | 12.71 | 7.02 | 42.65 |
| GS (Sign order) | 42.91 | 20.51 | 9.86 | 4.95 | 40.63 | 43.06 | 20.97 | 10.48 | 5.39 | 40.60 |

Table 2: A table showing the result of performing Gloss Selection (GS) on the Meine DGS Annotated (mDGS) dataset.

To initialize the encoder and decoder of the transformer we use xavier initializer (Glorot and Bengio, 2010) with zero bias and Adam optimization (Kingma and Ba, 2014). The initial learning rate is set to $10^{-4}$ with a decrease factor of 0.7 and patience of 5. During training we employ dropout connections, therefore we apply a dropout probability of 0.35 and 0.2 for the GS and GR models respectively (Srivastava et al., 2014). When decoding we apply a greedy algorithm on both models. We filter the confidence of the word2vec alignment, $A_{Vec}$, with a factor of 0.9. We train the BTG prereorder model for 30 iterations with a beam size of 20 on the training set only. We set the number of word classes to 50 when clustering with Brown et al. (1992) and we tag the spoken language with POS using the Spacy python implementation for German.

Our code base comes from the Kreutzer et al. (2019) NMT toolkit, JoeyNMT (Kreutzer et al., 2019) and is implemented using Pytorch (Paszke et al., 2019). When embedding with BERT, we use an open source pre-trained model from Deepset (Chan et al., 2020). Finally, we used fasttext's implementation of Word2Vec for word level embedding (Mikolov et al., 2013).

To evaluate our models, we use the Public Corpus of German Sign Language, 3rd release, the mDGS dataset (Konrad et al., 2020) and the PHOENIX14**T** dataset (Camgoz et al., 2018). mDGS contains aligned spoken German sentences and gloss sequences, from unconstrained dialogue between two native deaf signers (Konrad et al., 2020) and we use the translation protocol set in Saunders et al. (2022).

mDGS is 7.5 times larger compared to PHOENIX14**T** with 330 deaf participants performing free-form signing and a source vocabulary of 18,457. Note we remove the gloss variant numbers to reduce singletons. We use BLEU scores (BLEU-1,2,3 and 4) and Rouge score to evaluate all methods.

# 5. Experiments

## 5.1. Quantitative Experiments

In this section, we evaluate our proposed approaches on the mDGS and the PHOENIX14**T** dataset. We group our experiments in four sections:

1. Gloss Selection (GS).
2. Gloss Reordering (GR).
3. S&R (GS $+$ GR) and State-of-the-art Comparison.
4. Inference speed tests.

### 5.1.1. Gloss Selection

Firstly, we evaluate our GS approach. As discussed in section 3.1 we create an alignment for both datasets in order to perform GS. Table 1 and 2 show the results. In both cases, the GS output is the same but compared against the SPO (row 1) and the ground truth gloss (SIO) (row 2), hence the BLEU-1 score is the same. As the model was trained to predict SPO order, it is not surprising that the BLEU-4 score is higher. However, the performance drop when evaluated against sign order is small. The high BLUE-1 scores demonstrate the effectiveness of this method, achieving 42.91 on the challenging mDGS dataset.

### 5.1.2. Gloss Reordering

Next, we compare our different reordering approaches. When evaluating the GR model any words not present in the training set are replaced with unknown tokens. Thus, the BLEU score for the learnt method is not 100, even though the model has to predict all words that are present at the input. As can seen from table 3 and 4, the statistical method outperforms the learnt approach, with the statistical method achieving 26.51 and 28.64 BLEU-4 higher on the PHOENIX14**T** and mDGS dev sets respectively. Suggesting that POS tags and word classes are effective features for reordering. The learnt method is found to be

| PHOENIX14**T** | DEV SET | | | | | TEST SET | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Mapping: | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE |
| GT (Aligned Gloss) | 100.00 | 66.36 | 48.72 | 37.43 | 76.21 | 100.00 | 64.00 | 45.60 | 34.07 | 76.17 |
| Learnt | 99.14 | 60.14 | 39.50 | 26.93 | 59.17 | 99.20 | 59.43 | 38.58 | 25.74 | 58.28 |
| Statistical | 100.00 | 76.52 | 62.83 | 53.44 | 84.61 | 100.00 | 61.87 | 42.64 | 31.05 | 74.66 |

Table 3: A table showing the results of performing Gloss Reordering (GR) from Spoken Language Order (SPO) to Sign Language Order (SIO) on the RWTH-PHOENIX-Weather-2014**T** (PHOENIX14**T**) dataset.

| mDGS | DEV SET | | | | | TEST SET | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Mapping: | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE |
| GT (Aligned Gloss) | 100.00 | 65.20 | 43.72 | 29.89 | 80.20 | 100.00 | 64.69 | 42.98 | 29.35 | 80.37 |
| Learnt | 97.62 | 59.45 | 40.40 | 29.29 | 60.75 | 97.60 | 59.36 | 40.36 | 29.33 | 60.32 |
| Statistical | 100.00 | 82.12 | 68.67 | 57.93 | 91.24 | 100.00 | 60.06 | 36.87 | 22.91 | 77.47 |

Table 4: A table showing the results of performing Gloss Reordering (GR) from Spoken Language Order (SPO) to Sign Language Order (SIO) on the Meine DGS Annotated (mDGS) dataset.

detrimental to the ordering of the SPO gloss. We believe this result is due to the lack of large-scale training data. As suggested by Lin et al. (2020) 15 million parallel examples are needed for learnt methods to start outperforming statistical methods.

Row 1 of both tables shows the BLEU scores between the ground truth gloss and the SPO gloss, which gives an indication of the performance if the GS was 100% accurate. Table 3 shows a high BLEU-4 score of 37.43, which is the reordering score if we do not apply the GR mapping. Therefore, the GS output has the potential to generate a valid translation. Additionally, a proportion of the Deaf community are familiar with SPO (Lucas and Valli, 2014), whilst some may even prefer the SPO. However, further research is required to ascertain whether SPO translation is useful for the community.

### 5.2. State-Of-The-Art Comparison

The end-to-end S&R approach joins the output from the GS model and the mapping, $M()$, from GR to produce a full translation e.g. $p(Y|X) = M(p(Y^{SPO}|X))$. We used the mapping from the statistical approach as it was shown to give the best performance in section 5.1.2. In table 6 (PHOENIX14**T**) and 7 (mDGS) we compare our S&R approach to state-of-the-art work. Note we can only compare scores that are publicly available, therefore '-' denotes where the authors did not provide results.

For comparison on mDGS, we train a T2G transformer that achieves a competitive BLEU-4 score compared to (Saunders et al., 2022). The model is trained till convergence with a beam size of 5 and a word level tokenizer.

Our results show that reordering is beneficial to the GS model, increasing the BLEU-4 score by 1.23 and 1.11 on the PHOENIX14**T** Dev and Test sets respectively. On the mDGS dataset the reordering mapping was found to only benefit the dev set, increasing the BLEU-4 by 1.4, whilst being detrimental to the test set, decreasing the BLEU-4 by 1.25. The reordering performance is significantly reduced when applied to the output of the GS model, decreasing from the theoretical maximum of 53.44 to 19.07 on PHOENIX14**T**. We argue this is due to the number of false positives and false negatives in the output of the GS model.

As can be seen from table 6 and 7 our models outperformed all other methods on BLEU-1 score (Li et al., 2021; Saunders et al., 2020c, 2022; Stoll et al., 2018), setting a new state-of-the-art BLEU-1 on PHOENIX14**T** and mDGS, with a 12.65% and 37.88% improvement, respectively. We find our approach outperforms a neural editing program (Li et al., 2021), RNN (Stoll et al., 2018) and a basic transformer (Saunders et al., 2022) on BLEU-1 to 2 and Rouge scores on PHOENIX14**T**. While on mDGS our approach outperforms a traditional transformer on all metrics.

### 5.3. Model Latency

Table 5 demonstrates the significant advantages of our S&R model. It achieves an impressive 3.08 times speedup when compared to a traditional transformer architecture. Both GS and GR models utilize the same NAR decoder, but the incorporation of a reordering mask results in increased latency for the GR model. In contrast, our GS model, which has shown strong translation performance on its own, exhibits a large speed increase of 18.32 times. These findings highlight the practical utility of our approach, particularly in computationally constrained scenarios.

| Model | Latency | Speedup |
|---|---|---|
| T2G Transformer | 4380ms | 1.00x |
| GS | 239ms | 18.32x |
| GR | 1181ms | 3.71x |
| S&R | 1420ms | 3.08x |

Table 5: Inference latency comparison on mDGS.

| PHOENIX14**T** | DEV SET | | | | | TEST SET | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Approach: | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE |
| T2G Stoll et al. (2018) | 50.15 | 32.47 | 22.30 | 16.34 | 48.42 | 50.67 | 32.25 | 21.54 | 15.26 | 48.10 |
| T2G Saunders et al. (2020c) | 55.65 | 38.21 | 27.36 | 20.23 | 55.41 | 55.18 | 37.10 | 26.24 | 19.10 | 54.55 |
| T2G Li et al. (2021) | - | - | 25.51 | 18.89 | 49.91 | - | - | - | - | - |
| **GS** | 62.69 | 38.86 | 25.67 | 17.84 | 56.37 | 60.13 | 35.15 | 21.84 | 14.49 | 54.60 |
| **S&R** | 62.69 | 40.01 | 27.07 | 19.07 | 56.83 | 60.13 | 35.10 | 22.65 | 15.60 | 53.78 |

Table 6: Baseline comparison results for Text to Gloss (T2G) translation on the PHOENIX14**T** dataset.

| mDGS | DEV SET | | | | | TEST SET | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Approach: | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE |
| T2G Saunders et al. (2022) | - | - | - | 3.17 | 32.93 | - | - | - | 3.08 | 32.52 |
| T2G transformer | 31.12 | 14.32 | 6.49 | 3.04 | 34.71 | 31.25 | 15.08 | 7.26 | 3.38 | 34.98 |
| **GS** | 42.91 | 20.51 | 9.86 | 4.95 | 40.63 | 43.06 | 20.97 | 10.48 | 5.39 | 40.60 |
| **S&R** | 42.91 | 22.37 | 11.77 | 6.35 | 41.74 | 43.06 | 19.46 | 8.88 | 4.14 | 39.40 |

Table 7: Baseline comparison results for Text to Gloss (T2G) translation on the Meine DGS Annotated (mDGS) dataset.

## 5.4. Qualitative Experiments

Figure 5 shows example translations from the mDGS test set. We compare our approach to the baseline transformer that achieved 31.12 BLEU-1 and 3.04 BLEU-4. We show the output from the S&R approach as well as the intermediate output from GS.

| | |
|---|---|
| **Spoken Language:** | da hat man nur briefe geschrieben genau<br>(you only wrote letters there, exactly) |
| **GT:** | NUR SCHREIBEN SCHREIBEN SCHREIBEN ENDE<br>(WRITE ONLY WRITE END) |
| **GS:** | NUR BRIEF GENAU SCHREIBEN<br>(JUST WRITE LETTER ACCURATE) |
| **SNR:** | NUR BRIEF SCHREIBEN GENAU<br>(JUST WRITE ACCURATE LETTER) |
| **Baseline T2G:** | SCHREIBEN<br>(WRITE) |
| **Spoken Language:** | ich hatte es lieber knapp halten sollen<br>(I should have kept it brief) |
| **GT:** | ICH LIEBER KURZ<br>(I PREFER SHORT) |
| **GS:** | ICH KNAPP LIEBER<br>(I ALMOST PREFER) |
| **SNR:** | ICH LIEBER KNAPP<br>(I PREFER ALMOST) |
| **Baseline T2G:** | ICH IN-DER-KLEMME-STECKEN<br>(I'M-STUCK) |
| **Spoken Language:** | ach du hast oovoo so gebardet statt oovoo so zu gebarden<br>(oh, you behaved oovoo like that instead of acting oovoo like that) |
| **GT:** | MASS OOVOO<br>(MASS OOVOO) |
| **GS:** | DU OOVOO GEBARDEN OOVOO GEBARDEN<br>(YOU OOVOO ENTIRE OOVOO ENTIRE) |
| **SNR:** | OOVOO DU GEBARDEN OOVOO GEBARDEN<br>(OOVOO YOU ENTIRE OOVOO ENTIRE) |
| **Baseline T2G:** | OOVOO<br>(OOVOO) |

Figure 5: Example mDGS translations from a baseline transformer, the GS and S&R models )

These translations show that our approach is better at retaining the meaning of the spoken language sentence, likely due to the 37.88% improvement in BLUE-1 score. However, in some cases, GS can over predict the number of tokens, especially for long input sequences as shown by the third example.

## 6. Conclusion

In this paper we presented Select and Reorder (S&R), a novel two step approach to T2G translation, splitting the problem into two concurrent tasks of Gloss Selection (GS) and Gloss Reordering (GR). This approach disentangles the order from the vocabulary, allowing the GS model to focus on maximizing the correct vocabulary whilst leaving arguably the more difficult ordering task to a separate model. We showed our proposed GS model achieves a significant increase in BLEU-1 score of 11.79 on the mDGS dataset. In addition, we showed that reordering can be learnt by a GR model, but statistical based methods perform stronger with the current data limitations. Finally, we showed the result of combining the GS with the statistical reordering mapping, finding the S&R approach outperformed a neural editing program (Li et al., 2021), RNN (Stoll et al., 2018) and a basic transformer (Saunders et al., 2022).

It's clear that one major challenge to the field is the lack of quality gloss labelled data. Therefore, in the future it would be interesting to see if data augmentation could be used to pool sign language resources from different languages (e.g. DGS, BSL and ASL). A shared lexicon would need to be established across the datasets to combine all of the parallel bilingual data. Alternatively, using the proposed alignment a multilingual model could be trained using Randomly Aligned Substitutions (Lin et al., 2020).

## 7. Acknowledgements

We thank Adam Munder, Mariam Rahmani, and Abolfazl Ravanshad from OmniBridge, an Intel Venture. We also thank Thomas Hanke and the University of Hamburg for the use of the Meine DGS Annotated (mDGS) data. This work was supported by Intel, the SNSF project 'SMILE II' (CRSII5 193686), the European Union's Horizon2020 programme ('EASIER' grant agreement 101016982) and the Innosuisse IICT Flagship (PFFS-21-47). This work reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

## 8. Bibliographical References

Andrew Bangham, SJ Cox, Ralph Elliott, John RW Glauert, Ian Marshall, Sanja Rankov, and Mark Wells. 2000. Virtual signing: Capture, animation, storage and transmission-an overview of the visicast project. In *IEE Seminar on speech and language processing for disabled and elderly people (Ref. No. 2000/025)*, pages 6–1. IET.

Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kearsy Cormier, Adam Schembri, David Vinson, and Eleni Orfanidou. 2012. First language acquisition differs from second language acquisition in prelingually deaf signers: Evidence from sensitivity to grammaticality judgement in british sign language. *Cognition*, 124(1):50–65.

Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. 2002. Tessa, a system to aid communication with deaf people. In *Proceedings of the fifth international ACM conference on Assistive technologies*, pages 205–212.

Vincent J Della Pietra. 1994. The mathematics of statistical machine translation: Parameter estimation. *Using Large Corpora*, page 223.

Jinhua Du and Andy Way. 2017. Pre-reordering for neural machine translation: Helpful or harmful? *Prague Bulletin of Mathematical Linguistics*, 108:171–181.

Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. 2012. The dicta-sign wiki: Enabling web communication for the deaf. In *International Conference on Computers for Handicapped Persons*, pages 205–212. Springer.

Hosni Mostafa El-dali. 2011. Towards an understanding of the distinctive nature of translation studies. *Journal of King Saud University-Languages and Translation*, 23(1):29–45.

Dmitriy Genzel. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 376–384.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*.

Thomas Hanke. 2004. Hamnosys-representing sign language data in language resources and language processing contexts. In *LREC*, volume 4, pages 1–6.

Julian Hitschler, Laura Jehl, Sariya Karimova, Mayumi Ohta, Benjamin Körner, and Stefan Riezler. 2016. Otedama: Fast rule-based preordering for machine translation. *Prague Bull. Math. Linguistics*, 106:159–168.

Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. 2021. Towards fast and high-quality sign language production. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3172–3181.

Eui Jun Hwang, Jung-Ho Kim, and Jong C Park. 2021. Non-autoregressive sign language production with gaussian space. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2022. Machine translation between spoken languages and signed languages represented in signwriting. *arXiv preprint arXiv:2210.05404*.

Mihoko Kato. 2008. A study of notation and sign writing systems for the deaf. *Intercultural Communication Studies*, 17(4):97–114.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Philipp Koehn, Franz J Och, and Daniel Marcu. 2003. Statistical phrase-based translation. Technical report, University of Southern California Marina Del Rey Information Sciences Inst.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey nmt: A minimalist nmt toolkit for novices. *arXiv:1907.12484*.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.

Barbara LeMaster and Leila Monaghan. 2005. *Variation in Sign Languages*, chapter 7. John Wiley & Sons, Ltd.

Dongxu Li, Chenchen Xu, Liu Liu, Yiran Zhong, Rong Wang, Lars Petersson, and Hongdong Li. 2021. Transcribing natural languages for the deaf via neural editing programs. *arXiv preprint arXiv:2112.09600*.

Dongxu Li, Chenchen Xu, Liu Liu, Yiran Zhong, Rong Wang, Lars Petersson, and Hongdong Li. 2022. Transcribing natural languages for the deaf via neural editing programs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11991–11999.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. *arXiv preprint arXiv:2010.03142*.

Ceil Lucas and Clayton Valli. 2014. American deaf community. *The Sociolinguistics of the Deaf Community*, page 11.

Chloë R Marshall and Angela Hobsbaum. 2015. Sign-supported english: is it effective at teaching vocabulary to young children with english as an additional language? *International journal of language & communication disorders*, 50(5):616–628.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. *arXiv:2105.07476*.

Tetsuji Nakagawa. 2015. Efficient top-down BTG parsing for machine translation preordering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 208–218, Beijing, China. Association for Computational Linguistics.

National Geographic Society. 2017. Sign language. `https://education.nationalgeographic.org/resource/sign-language`, Retrieved September 7, 2022.

Graham Neubig, Taro Watanabe, and Shinsuke Mori. 2012. Inducing a discriminative parser to optimize machine translation reordering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 843–853.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics*, pages 295–302.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc.

Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2021. Sign language production: a review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3451–3461.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. *arXiv preprint arXiv:2004.08728*.

Lucila San José-Robertson, David P Corina, Debra Ackerman, Andre Guillemin, and Allen R

Braun. 2004. Neural systems for sign language production: mechanisms supporting lexical selection, phonological encoding, and articulation. *Human brain mapping*, 23(3):156–167.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020a. Adversarial training for multi-channel sign language production. *arXiv preprint arXiv:2008.12405*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020b. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020c. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jerome Daniel Schein and Marcus T. Delk. 1974. The deaf population of the united states.

Shashi Pal Singh, Ajai Kumar, Hemant Darbari, Lenali Singh, Anshika Rastogi, and Shikha Jain. 2017. Machine translation using deep learning: An overview. In *2017 international conference on computer, communications and electronics (comptelix)*, pages 162–167. IEEE.

Nils Skotara, Uta Salden, Monique Kügow, Barbara Hänel-Faulhaber, and Brigitte Röder. 2012. The influence of language deprivation in early childhood on l2 processing: An erp comparison of deaf native signers and deaf signers with a delayed language acquisition. *BMC neuroscience*, 13:1–14.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*.

William C Stokoe. 1980. Sign language structure. *Annual Review of Anthropology*.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British Machine Vision Conference*.

Stephanie Stoll, Armin Mustafa, and Jean-Yves Guillemaut. 2022. There and back again: 3d sign language generation from text using back-translation. Unpublished.

Shinichi Tamura and Shingo Kawasaki. 1988. Recognition of sign language motion images. *Pattern Recognition*.

Don Tuggener. 2016. *Incremental coreference resolution for German*. Ph.D. thesis, University of Zurich.

Desmond Eustin Van Wyk. 2008. *Virtual human modelling and animation for real-time sign language visualisation*. Ph.D. thesis, University of the Western Cape.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Harry Walsh, Ben Saunders, and Richard Bowden. 2022. Changing the representation: Examining language representation for neural sign language production. *arXiv preprint arXiv:2210.06312*.

WHO. 2021. Deafness and hearing loss.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, 23(3):377–403.

Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2018. Exploiting pre-ordering for neural machine translation. In *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018)*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

Inge Zwitserlood, Margriet Verlinden, Johan Ros, Sanny Van Der Schoot, and T Netherlands. 2004. Synthetic signing for the deaf: Esign. In *Proceedings of the conference and workshop on assistive technologies for vision and hearing impairment (CVHI)*.

## 9. Language Resource References


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

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.