

# Sign Language Translation

## A Review

Daniela Winter

December 22, 2022

### Abstract

Sign Language Translation is an important tool for deaf or hard-of-hearing people to communicate with people who do not know Sign Language or only speak a different Sign Language. This work focuses on Sign Language Translation (SLT) which describes the translation of sign videos into spoken language text in contrast to Sign Language Recognition (SLR) and Production. This review covers the common terms and main concepts related to the topic without requiring a detailed mathematical background. It introduces gloss transcription and different feature extraction methods and provides an overview of important datasets, current SLT architectures, and their performances. Currently, Transformers are outperforming other architectures and show that it can also be possible to perform SLT without the use of expensive glosses.

## 1 Introduction

The following work is a comprehensive overview of the key ideas, important background knowledge, and current approaches related to Machine Sign Language Translation, specifically continuous sign-video-to-spoken-language translation. This review starts with a general introduction to sign languages (SLs), transcription with glosses, and a distinction of Sign Language Recognition (SLR), Translation (SLT), and Production (SLP) in Section 1. In Section 2, we look at different feature extraction methods for SL videos followed by different evaluation metrics for continuous SLT in Section 3. After that, we introduce different datasets, that have been popular for SLT in the past and bigger, more recent datasets in Section 4. In Section 5, we provide an overview of architectures used for this task and their performances. At last, we finish with a conclusion in Section 6 and present an outlook on possible steps that can be taken next to improve the performance of current SLT models.

### 1.1 Sign Languages

Sign languages (SLs) are visual languages that are commonly used as the form of communication for deaf and many hard-of-hearing people. They are official languages with defined signs, grammar, and dialects. Different signs are expressed by different components of the articulators. The articulators include different body parts of the upper body with the hand being the most important articulator. They produce manual and non-manual components of the SL. Manual components include hand poses, the orientation of those, and the location of

the hands in space; while upper-body movements, lip movement, and other facial expressions constitute the non-manual components [Ananthanarayana et al., 2021].

## 1.2 Transcription with Glosses

Since SLs are visual languages, which are conceptually completely different from spoken languages, it is a challenge to convert them into written representations. Glosses are one way of representing SLs in written form. This form of transcription is a closer representation of SLs than spoken language text because the latter can not capture all the dimensions of the complexity of signs. Glosses on the other hand capture not only the meaning of the hand pose itself but also other SL specifics that carry additional information. Those specifics include among others the location of the sign in the signing space, if a sign is repeated, mouthings, facial expressions, and so-called classifiers [Crasborn, 2015]. The location of the sign and classifier signs are used as variables that can store information that can be referred to later. All this information is crucial for the transcription quality. Without it, transcriptions lose a big part of the information that is contained in the signs. Without this additional information, glosses can not function well as a basis of interpretation for spoken languages. For this reason, the gloss transcription uses so-called diacritics which are additional symbols, letters, or numbers as extra notes next to the translated word as can be seen in Figure 1. This enables the gloss transcription to carry more SL-specific information that is needed to interpret the meaning in the context. One disadvantage of gloss transcription is, that it still requires the sign to be translated into one single term which belongs to the spoken language. This can result in problems since signs can have a variety of meanings that are often not captured by one word of the respective spoken language and depend a lot on context [Hoiting and Slobin, 2002]. To tackle this problem, a lot of other transcription systems have been introduced which do not depend on a spoken language. This makes them less readable for readers that are unfamiliar with the transcription system [Hoiting and Slobin, 2002]. Moreover, they are not very substantial for Machine SLT since most of the relevant data sets for Machine SLT use gloss annotations.

## 1.3 Sign Language Translation as a Machine Learning Task

Sign Language Recognition (SLR) and Sign Language Translation (SLT) can be realized with Machine Learning methods. SLR describes the detection and identification of different signs in a video and the automatic transcription of an SL video into glosses or some other form of textual SL representation. SLR can be a categorization task where one video is only constituted of one sign. This kind of task is called Word-Level Sign Language Recognition (WSLR) or Isolated Sign Language Recognition (ISLR). The signing video is used as input and the model outputs a gloss for the sign. In contrast to WSLR, there is a continuous way of performing SLR and SLT. For this, a complete utterance is used as input. So the model needs to determine where a sign ends and a new one begins (which is not necessary for WSLR). After splitting the several signs, continuous SLR outputs an annotation for each of them separately in a sequence that is parallel to the input sequence. Since this continuous input forms a more realistic challenge and is more useful for real-world applications, this review is focusing mainly on continuous tasks.

Scheme	Example
gloss in capital letters	WIE-IMMER
finger spelling split by +	A+G+N+E+S
compound glosses split by +	V+LAND
numbers in written form	SIEBEN instead of 7
pointing gestures	IX
extended repetitions	SONNE++
pronunciation variants	TAG#1 TAG#2
classifier signs	cl-KOMMEN
lefthand only signs	lh-SONNE
signs negated by headshake	neg-WIND
signs negated by the alpha rule	negalp-MUESSEN
localization	loc-REGEN
additional mouthing	GLOSS-(mb:hill)
additional facial expression	GLOSS-(mk:strong)
additional localization	GLOSS-(loc:alps)
additional object of sign	GLOSS-(obj:cloud)

Figure 1: This is an example of the gloss annotations used in the RWTH-PHOENIX-Weather corpus [Forster et al., 2012]. The first section includes general rules of the gloss transcription. The reference spoken language word needs to be capitalized and if a word is spelled or a gloss is made up of two signs their transcriptions are split by a "+". "IX" stands for "index" and refers to locations in the signing space that can be assigned a meaning and referred to later. Also, repetitions and different pronunciations can be transcribed. The second section shows how the meaning of a gloss can be changed or adapted by negation, using it as a location or as a classifier for future reference. The last section shows additional information that was included by the usage of other articulators (mouth, face) or for example by additional localization.

SLR is a visual-spatial recognition task, that poses typical challenges of Computer Vision tasks: orientation and 3D understanding, occlusion difficulties, and differentiating multiple components (e.g. different signers in one video). SLR is sometimes termed as SLT which is misleading since the produced transcriptions like glosses do not provide a translation that is readable for most humans. SLT on the other hand also belongs to the field of Natural Language Processing since it is a Machine Translation task that aims to translate the signed video into a spoken language text form. This can either be done directly in an end-to-end manner by using video frame embeddings as input or with an intermediate step of gloss transcription. Some authors argue that the gloss representation as an intermediate step poses a disadvantage as it can act as a bottleneck and make the performance depend solely on the quality of the glosses obtained by the continuous SLR process [Camgöz et al., 2020b]. Others argue that this does not have to be the case and show results where the final performance using the predicted glosses exceeds the performance with ground truth glosses [Yin and Read, 2020].

Generally, it would be of great advantage if intermediate glosses are not required for good performances of SLT models because creating big data sets with glosses is very expensive.

SL datasets without gloss annotations and just spoken language ground truths on the other hand are very easy to acquire since a signer can translate spoken language to SL in real time. While it is desirable to skip the step of gloss prediction also for reasons of simplicity, the performance of such pipelines is often worse. This could be the case because the task of SLT is too complex to learn for the proposed models so some form of intermediate supervision helps guide towards the right direction [Camgöz et al., 2020b]. While machine translation works quite efficiently for spoken languages already, SLT is still a very challenging task. This may be due to the fact that sign languages are conceptually different from spoken languages. In comparison to spoken language translation, in SLT, the mappings of the SL and its respective spoken language are often one-to-many mappings since a lot of information can be contained in one sign. Moreover, the mappings are highly non-monotonic which means that the signed and the spoken sentence often have a completely different word order and grammar [Camgöz et al., 2020b].

The opposite direction of the SLR and SLT pipeline from spoken language text or speech to SL is called Sign Language Production (SLP). Since this task is very different from what is described above, this review will not focus on SLP. But since it is a translation process as well, the term “SLT” is sometimes used in a way that includes SLP. In the following, the term ”SLT” describes the translation process starting with a video as an input and outputting the spoken language text. In addition to that, it will be used to differentiate the text-to-text translation task from the SLR step in pipelines where both steps are separated. This usage conforms with the literature and the exact meaning can be determined by context.

## 2 Feature Extraction from Sign Language Videos

### 2.1 Features from CNNs

In order to perform SLR and SLT, the information of the SL video frames needs to be extracted to be fed to the different network models. This process is very crucial for the performance of the models since the quality of the information contained in the input feature acts as a bottleneck for the performance of the following steps. One possible way of feature extraction is via Convolutional Neural Networks (CNNs) that output multi-dimensional vectors in which the information is encoded. It can be useful to use CNNs that were pre-trained on big data sets. Those can either be used directly or be fine-tuned to encode SL frames specifically by some additional training with SL data. Ananthanarayana et al. [Ananthanarayana et al., 2021] compared the SLT performance of a transformer when using different CNNs for the input feature extraction. The ResNet50 model with 24048-dimensional feature vectors pre-trained on the ImageNet dataset proved to give the best performances, so the authors concluded that deeper networks like ResNet50 provide better features than shallow networks or even wider networks.

### 2.2 Features from OpenPose

OpenPose is a system to extract key points (2D-coordinates) of joints or facial landmarks that can be used to represent a certain pose or facial expression of a human in an image [Cao et al., 2019]. Those key points are useful for the interpretation of human actions in

a video and therefore fit well with the SL interpretation task. OpenPose input vectors are considerably smaller than the CNN feature vectors since they only include two values per key point and the number of key points is relatively small. In addition to that OpenPose features are less expensive to extract than CNN features and the visualized feature vectors are human interpretable (see Figure 2). While those advantages speak for the usage



Figure 2: This is an example of the sign "exaggerate". On the left, there is the visualization of the OpenPose key points, on the right, the original video frame with the key points overlapped [Amorim et al., 2019]. It can be seen that the number of key points needed to represent the sign is not very large.

of OpenPose features, there is also a certain loss of information that seems to be greater than when using CNN features. The effects of this information loss become clear when letting human SL translators interpret those OpenPose extracted signs. Ananthanarayana et al. [Ananthanarayana et al., 2021] show that the performance of the human translator suffers strongly from only being presented with OpenPose features. Compared to when using the original videos, the translators were only able to reach an average BLEU-1 Score (see Section 3) of 6.32 instead of 24.86 and an average BLEU-4 Score of 3.81 instead of 0.14 for 340 ASL videos. This problem also shows when Neural Networks are used for SLT. De Coster et al. [De Coster et al., 2020] compared two Transformer networks that use OpenPose and CNN input features respectively for sign classification. The Top-1 to Top-10 accuracy values (see Section 3) are around 10% better for the CNN input features than for the OpenPose features. This shows that it is important to choose the input features carefully since they might include weaknesses for the whole performance.

### 2.3 Multiple-Channel Inputs

In order to mitigate those weaknesses, it is possible to combine different inputs by using multiple input channels. Input features must not exclusively be one or the other, it is possible to use multiple inputs when training the model. The big advantage of OpenPose features for example shows when it is combined with the CNN input features. De Coster et al. were able to consistently achieve better Top-n accuracy values (see Section 3) when using both input features [De Coster et al., 2020].

Another way to use multiple input channels is to divide the different information channels of SL. Since multiple articulators such as hands, torso, lips, and facial expressions are used in SLs, another plausible approach is to use several separate input channels for the different articulators. Camgöz et al. were the first that used a multi-channel architecture for end-to-end SL video to spoken text SLT translation [Camgöz et al., 2020a]. They used three different channels which focus on the hands, mouth, and upper-body pose respectively. The input for the hands and mouth channels are 1024-dimensional feature vectors from CNNs and the upper-body pose is represented in a vector of OpenPose key points. While the performance of their model does not surpass previous performances, it is still worth mentioning because due to the usage of multiple input channels, their model functions stable enough to not be dependent on the usage of glosses. As mentioned previously, this is highly desirable since, without the need for glosses, much bigger unannotated data sets become useful for training SLT networks.

### 3 Evaluation Metrics

To evaluate and compare different models, a uniform and precise way of assessing their performances is necessary. The following evaluation metrics are used for calculating the output quality of architectures that perform continuous SLR and SLT.

*Bilingual Evaluation Understudy (BLEU)*: The BLEU-Score is the most used evaluation metric for SLT. It describes the n-gram precision of the predicted translation when compared to the ground truth. N-grams are tuples of consecutive words, a 4-gram, therefore, is a 4-tuple of four successive words in a sentence. If those four words in the predicted translation overlap with the ground truth translation, the 4-gram precision increases. The more words the n-gram includes the more the precision represents the fluency of the resulting phrase. A shorter n-gram overlap rather shows the adequacy of the translated words. The standard BLEU-Score includes uni-, bi-, tri-, and four-grams and is often called BLEU-4. In comparison to that, BLEU-1 only compares unigrams. The BLEU-Score ranges from 0 to 1 with 1 being a translation that is identical to the ground truth one. Figure 3 shows the interpretation of those values.

Since the precision divides the correct n-grams by the total predicted n-grams, it gets higher the lower the total predicted n-grams are. This means that short translations would get a better precision value. Therefore, the BLEU-Score incorporates the Brevity Penalty. It is used to punish the translation if it is much shorter than the ground truth [Google, 2022].

*Metric for Evaluation of Translation with Explicit Ordering (METEOR)*: The METEOR-Score is an evaluation metric that tries to tackle the problem of the BLEU-Score which does only reward exact overlaps. In the METEOR evaluation also words with the same stem, synonyms, and paraphrases are counted. For this reason, METEOR-Scores are closer to human evaluations. It is calculated by the mean of the unigram recalls and precisions combined with a penalty for unmatched words or matchings between prediction and ground truth with different word order. The disadvantage of METEOR is that the evaluation currently only exists for a limited number of languages [Denkowski and Lavie, 2014].

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

Figure 3: This is an interpretation of the BLEU-Score values in percent [Google, 2022]. The score ranges from 0 to 1 (0 to 100%) and good translations are achieved from a score of around 30-40%.

*Recall-Oriented Understudy for Gisting Evaluation (ROUGE)*: The ROUGE evaluation outputs multiple values which include the n-gram overlap precision, recall, and F1-Score separately. This makes it more interpretable than other scores. The precision shows the percentage of correctly predicted n-grams of all the predicted n-grams. The recall shows the percentage of correctly predicted n-grams of all n-grams in the ground truth phrase. And the F1-Score as a mean of precision and recall.

*Word Error Rate (WER)*: The Word Error Rate is a word-level or unigram evaluation metric where the overlap of the predicted and the ground truth sentences is tested. It is calculated as the fraction of word mistakes over the whole number of ground truth words. Word mistakes are counted after matching the words of the prediction to the words of the ground truth. Mistakes are if either a ground truth word or a predicted word does not have a matched word or if the matched word is not identical.

Even if the focus of this review lies on continuous SLT and not WSLR, for the sake of completeness, the following WSLR metric is included, too.

*Top-n accuracy*: Since WSLR is a classification task, those networks output probability values for different sign translations. The Top-n accuracy represents the frequency with which the ground truth translation can be found among the top-n predictions of the network.

## 4 Datasets

How well a model performs SLT tasks depends highly on the data that it is trained on. Therefore, it is crucial to choose well-balanced and large enough datasets for the chosen task. Since this report focuses on continuous SLT, the following section will only present datasets suitable for this task. A lot of available SL datasets only offer SLR training data, namely videos with gloss annotations, or gloss-to-text translation data without any signing videos. The first publicly available continuous SLT dataset with both gloss and spoken language an-



notations was the RWTH-PHOENIX Weather 2014T dataset. This dataset is an extension of the previously released RWTH-PHOENIX Weather 2014 dataset and was itself released in 2018 [Camgöz et al., 2018]. It has been widely used since then and is one of the most important datasets for continuous SLT. Boggaram et al. [Boggaram et al., 2022] analyzed a collection of datasets (continuous and isolated SLR and SLT). Of the 19 datasets they chose in their review, most of them were German, American, and Chinese. Since those are also the SLs whose data sets are often used in current research papers, the following sections will focus on data sets in those languages.

Generally, the problem with the datasets used for most of the past research is that they are very small with a maximum of about 10000 unique utterances in the Chinese SL dataset and that their environment is not controlled. Moreover, the resolution of those videos is really low and not matching current video quality standards [Ananthanarayana et al., 2021]. Therefore, in the past, researchers urged for the creation of bigger datasets and there have been more recent releases for Flemish SL, Swiss-German SL and German SL such as the Content4All datasets [Camgöz et al., 2021a] and the DGS Corpus [Schulder et al., 2020]. This section will also give an overview of those new datasets. The Content4All datasets and the DGS Corpus include 30h and 50h of annotated videos respectively. Since not that much research has been done on them yet, it is still useful to have a look at the less recent datasets to be able to compare new performance results to other models’ performances.

#### 4.1 Commonly Used Datasets in Previous Research

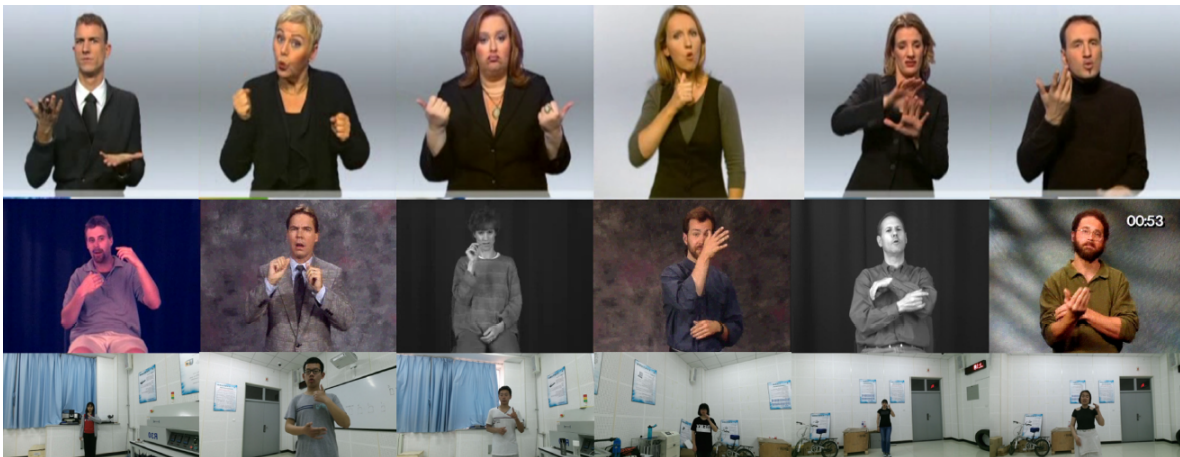


Figure 4: These are samples from the RWTH-PHOENIX Weather dataset (top), the ASLL-RPR dataset (middle), and the CSL dataset (bottom) [Ananthanarayana et al., 2021]. It can be seen that the upper two datasets use a less detailed background and have a more controlled video setting than the CSL dataset. The most consistent setting is used in the RWTH-PHOENIX Weather dataset, where the background is grey and the signers wear dark clothing to ensure contrast.



#### 4.1.1 RWTH-PHOENIX Weather Dataset

The data in the RWTH-PHOENIX Weather dataset [Forster et al., 2012] consists of videos from the weather forecast of the Phoenix TV channel in German Sign Language. Therefore, the topic and the SL vocabulary are quite restricted. The resolution of the videos is 210x260px with a frame rate of 25fps. Nine different people are used as signers and in total 1078 different words were used [Ananthanarayana et al., 2021]. The big advantage of this dataset is, that it has a controlled environment which makes it possible to easily detect the signer in front of a constant, homogeneous, undetailed background (see Figure 4). The annotations originally consisted only of glosses (only useful for SLR, not SLT, see Figure 1) but were extended by Camgöz et al. in 2018 with spoken language text annotations [Camgöz et al., 2018]. Since then the dataset has been used as a popular dataset for SLT.

## 4.2 American Sign Language Dataset

Compared to the RWTH-PHOENIX Weather dataset, the American Sign Language Linguistic Research Project (ASLLRPR) dataset [Neidle and Opoku, 2020] is not restricted to one topic but consists of videos of seven signers telling stories about different topics. The videos were recorded at 25fps but the resolution is not fixed and varies between 216x218px and 312x324px. There were about 1457 different words used in the videos and the background is always the same and without details [Ananthanarayana et al., 2021]. Some of the videos are gray-level videos and the color distribution in the colored videos varies (see Figure 4). An exemplary annotation can be seen in Figure 5. Another commonly used dataset for American Sign Language is the WLASL dataset [Li et al., 2020] with 2000 different words used.








<b>topic/focus:</b>	<u>topic</u>				
<b>dh:</b>	<u>POSS-3p:i</u>	<u>FATHER</u>	<u>GO-OUT</u>	<u>5 "perplexed"</u>	
<b>dh shape:</b>					
<b>ndh:</b>				<u>5 "perplexed"</u>	
<b>ndh shape:</b>					
<b>engl trans:</b>	His/Her father left.				

Figure 5: This is an example of the gloss annotations of the ASL Dataset. The glosses and shapes of the dominant (dh shape) and non-dominant hand (ndh shape) are annotated as well as the spoken English translation (engl trans) of the utterance [Neidle and Opoku, 2020].

### 4.3 Chinese Sign Language Dataset

The Chinese Sign Language (CSL) dataset [Yuan et al., 2019] includes 10000 different signed words of different topics in 50000 sign videos signed by 50 different signers. The videos were recorded with a high resolution of 1920x1080px at 30fps. While that resolution makes it preferable to the other two datasets shown in Figure 4, the background is uncontrolled and complex with the signers making up a smaller part of the frame. This makes it harder to extract the signs from the signers and that shows in the performance of the models when comparing the different datasets [Ananthanarayana et al., 2021]. The annotations do not include gloss transcriptions but only Chinese and English spoken language translations [Yuan et al., 2019].

### 4.4 More Recent, Bigger Datasets

#### 4.4.1 DGS Corpus

The publicly available, annotated part of the DGS Corpus <sup>1</sup> consists of a total of 50h of videos about different topics collected by researchers of the Institute for German Sign Language and Communication of the Deaf at Hamburg University. The videos include 330 different signers from different areas of Germany and 374800 different words were used in the publicly available dataset [Isard and Konrad, 2022]. This makes it considerably larger than the previously used datasets. The published videos were downsampled to a resolution of 640x360px at 50fps [Schulder et al., 2021] and consistently have a blue background while the signers wear dark clothing to ensure contrast to the background (see Figure 6 <sup>2</sup>). The project includes several online tools to explore the dataset and the annotations used. An exemplary annotation including gloss and spoken language transcription can be seen in Figure 7 <sup>3</sup>. In the online tool, the gloss annotations are linked to the respective SL dictionary entry of this sign to provide more information. For all the videos there is OpenPose keypoint information provided [Schulder and Hanke, 2019]. To our current best knowledge, no work has been published yet on this dataset regarding end-to-end SLT, but Angelova et al. successfully used the dataset for gloss-to-text translation and compared the results from the DGS Corpus to the results from the RWTH-PHOENIX Weather dataset [Angelova et al., 2022].

#### 4.4.2 Content4All Datasets

Similarly to the RWTH-PHOENIX WEATHER dataset, the Content4All datasets have been collected from TV channels and consist of weather and news broadcasts. The SL used in the datasets are the Flemish SL for the VRT-NEWS dataset and the Swiss-German SL for the SWISSTXT-NEWS and the SWISSTXT-WEATHER datasets. For each of the three datasets, 10 hours of video have been annotated, which results in 30 hours of data. Those hours have been chosen specifically to resemble the data distribution of the RWTH-PHOENIX WEATHER dataset. The SWISSTXT datasets have been recorded with a resolution of 1280x720px at 50fps and the VRT datasets with a resolution of 1280x720px at 25fps. The SWISSTXT-NEWS dataset contains 9623 different words in the training set while the VRT-NEWS dataset contains 6325 different words [Camgöz et al., 2021b]. The signers are in the

---

<sup>1</sup><https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html>

<sup>2</sup>retrieved from [https://www.sign-lang.uni-hamburg.de/meinedgs/ling/start\\_de.html](https://www.sign-lang.uni-hamburg.de/meinedgs/ling/start_de.html), 13.12.2022

<sup>3</sup>retrieved from [https://www.sign-lang.uni-hamburg.de/meinedgs/ling/start-name\\_de.html](https://www.sign-lang.uni-hamburg.de/meinedgs/ling/start-name_de.html), 13.12.2022



Figure 6: This figure shows frames from different videos in the DGS Corpus. The background is blue and the signers wear black/dark clothing to ensure sufficient contrast to the background.

Ach, und Fisch kann natürlich auch gegrillt werden.		
AUCH1A*		auch
FISCH1*		fisch
AUCH1A		auch
FISCH1*		fisch
SSEXTRA-LING-MAN*		[MG]

Figure 7: This is an example of the spoken language (left) and gloss (right) annotations of the DGS Corpus. The right side shows the gloss as it is written in the dictionary in the middle column in addition to the mouth gestures (MG) in the rightmost column. In the online tool, the gloss annotations are linked to the respective SL dictionary entry of this sign to provide more information.

same frame as the news/weather broadcast and therefore not the only person in the frame (see Figure 8). The videos have been published as original, anonymized and OpenPose data has been made available as well <sup>4</sup>. The annotations contain spoken language sentences that were manually aligned to the time frame their signs occur in the videos. Gloss transcription is not available. The SWISS-NEWS and the VRT-NEWS datasets have been used to train the end-to-end SLT model from Camgöz et al. [Camgöz et al., 2020b] and baseline performances have been reported [Camgöz et al., 2021b].

<sup>4</sup><https://www.cvssp.org/data/c4a-news-corpus/>



Figure 8: This figure shows two frames from videos of the SWISSTXT dataset (Swiss-German SL) on the top and from the VRT dataset (Flemish SL) on the bottom. The signers are in the same frame as the news/weather broadcast and therefore not the only person in the frame [Camgöz et al., 2021b].

## 5 Currently Used Architectures and Their Performances

In this section, we will describe different Neural Network architectures that have been used for SLT. Figure 9 gives an overview of the development of those architectures. The earlier implementations include RNNs that do not make use of attention mechanisms. More recent implementations like the Transformer architecture include different attention mechanisms.

### 5.1 Non-Attention-Based Recurrent Neural Networks

The earlier SLR and SLT implementations use Recurrent Neural Networks (RNNs). RNNs are a type of neural network that can handle sequential data such as natural language or speech. They get the input element-wise but maintain an internal state that functions as memory so that all earlier inputs can affect later predictions. Since the elements are input into the model element-wise, the total input length is not limited per se. For long input sequences, however, the problem of long-term dependencies occurs which refers to the difficulty of the network in learning patterns that span a long sequence of inputs. Moreover, vanishing gradients make it difficult for the model to learn. To address those problems, Long Short-Term Memory (LSTM) networks have been introduced. They include additional gates and a

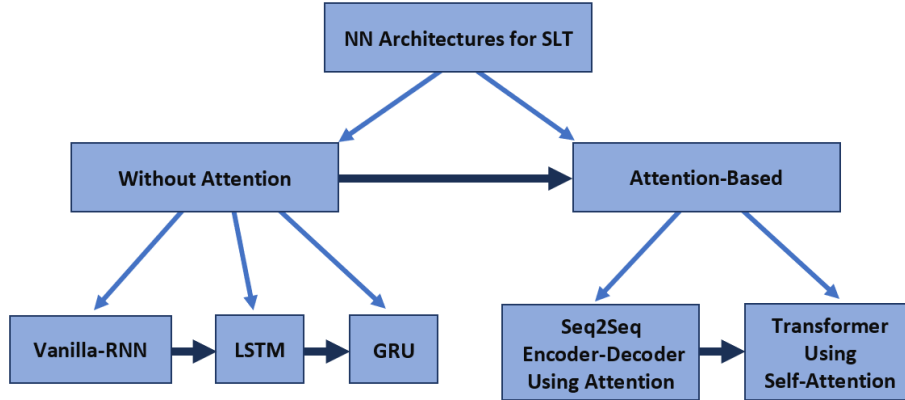


Figure 9: This figure gives an overview of the development of different Neural Network architectures that were used for SLT. The earlier implementations were using RNNs without any attention mechanism while more recent implementations use attention and self-attention to tackle the SLT task. The horizontal arrows indicate the temporal development.

memory cell with which the vanishing gradients effect can be dampened so that LSTMs can handle longer input sequences. Another variant of RNNs are Gated Recurrent Units (GRUs), which also help to mitigate the long-term dependency problem. In comparison to LSTMs, GRUs do not have that kind of memory cell and use fewer gates. Camgöz et al. compared the performance of LSTMs and GRUs in SLT tasks and showed that GRUs outperform LSTMs [Camgöz et al., 2018]. They assume this to be due to less overfitting as GRUs have fewer parameters than LSTMs since they use fewer gates. While LSTMs and GRUs mitigate the above-mentioned issues, for very long input sequences they are still showing those problems. Therefore, they can be used for SLR tasks where inter-sign dependencies are not relevant but not for SLT tasks that produce complete spoken language output. For SLR, LSTMs can be used even on videos with several signs when a reset gate is included in the LSTM that resets the memory as soon as a new word is signed in the video [Mittal et al., 2019].

## 5.2 Attention-Based Sequence-to-Sequence Models and Transformers

A significant step forward from earlier RNNs has been made with the introduction of attention mechanisms for Sequence-to-Sequence models. Sequence-to-Sequence models are a type of RNN with an encoder-decoder architecture. The encoder processes input sequences and generates a fixed-length context vector as a representation. The decoder uses this vector, along with the previous decoder output, to produce the output sequence, such as a spoken language sentence. This architecture can be improved by using attention. Attention helps the model to focus on the relevant parts of the input. The model generates hidden states for each part of the input. They can then be used to calculate a weighted sum that helps to determine which parts should be attended to when generating the output (see Figure 10).

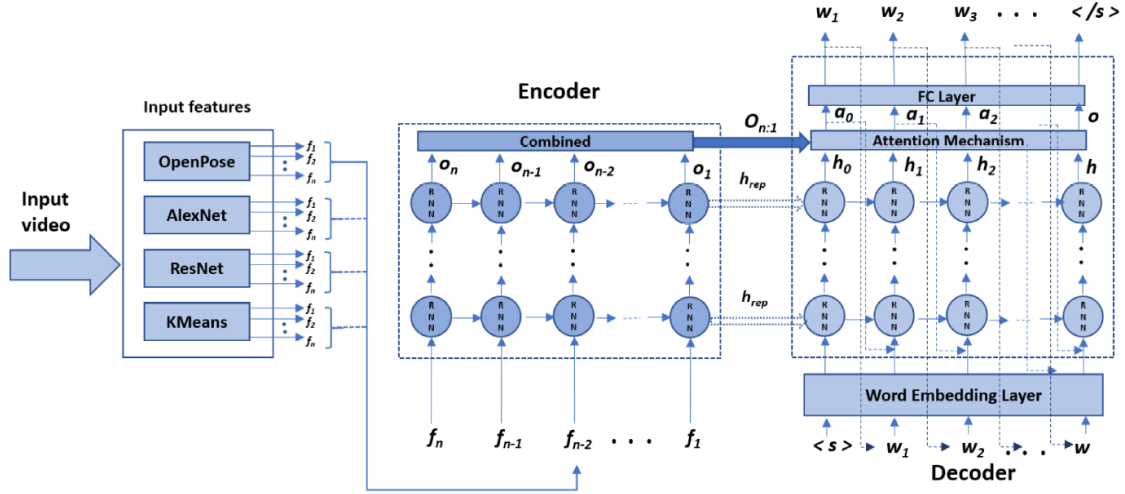


Figure 10: This is an example of an Encoder-Decoder Architecture with attention. The attention layer is included in the decoder combining the outputs of the encoder and the hidden states of the decoder [Ananthanarayana et al., 2021].

Legend:

$f_n$ : input feature at timestep  $n$ ;  $o_n$ : output of the encoder at timestep  $n$ ;  
 $h_{rep}, h_i$ : hidden representations;  $a_i$ : attention vector;  
 $< s >, < /s >$ : beginning and end of sentence token;  $w_i$ : input/output words

The weights to compute the weighted sum can be acquired in different ways. Luong attention uses dot product multiplication to combine the context of the input and the output while Bahdanau uses concatenation. Camgöz et al. compared those two attention mechanisms for SLT tasks and found that Luong attention performs better and overfits less to the training data, which then allows better generalization to the test set [Camgöz et al., 2018].

After the introduction of Transformers in 2017 [Vaswani et al., 2017], they have become most popular for solving SLT tasks and translation tasks in general. Transformers use self-attention which differs from attention as explained above in the way that it allows the model to focus on different parts of the input sequence at the same time. The entire input sequence can be processed at once and with the self-attention mechanism the model can find relationships between different input elements. This helps when generating spoken language text since focusing on relationships between different words and phrases is needed to produce grammatically correct sentences. Yin et al. use a Transformer in combination with a Spatial-Temporal Multi-Cue (STMC) network where they use different articulators and different feature extraction methods for the input [Yin and Read, 2020]. The input is then evaluated with regard to the spatial and temporal relationships in the data by the spatial and temporal modules. The STMC network outputs gloss transcriptions which are used as input for the Transformer network. The Transformer then performs a gloss-to-text translation task and outputs the spoken language sentence. There are also works that use a Transformer in an end-to-end manner for the whole SLT process. The first model of this kind was published



by Camgöz et al. in 2020 [Camgöz et al., 2020b]. They let the Transformer learn SLR and SLT jointly which means the output does not depend on gloss prediction as an intermediate step. While avoiding the potential bottleneck of the gloss prediction, they still rely on gloss annotations in the dataset to provide some gloss-level supervision. The architecture of this model is shown in Figure 11. As the first Transformer-based end-to-end SLT architecture it serves as a good basis for further adaptations and improvements and has been used as such in several other publications [Camgöz et al., 2020a] [Voskou et al., 2021]. As it is explained in

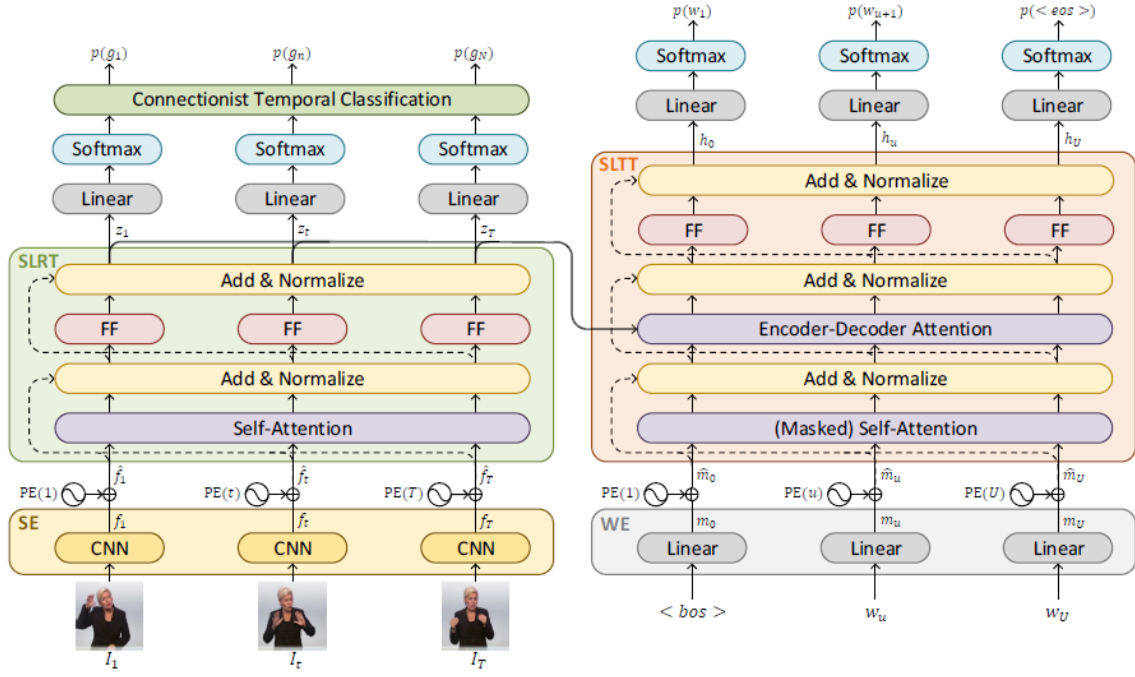


Figure 11: This is an example of a Transformer Architecture for end-to-end SLT. Self-attention is used in the encoder as well as the decoder and (encoder-decoder) attention in the decoder [Camgöz et al., 2020b].

Legend:

$I_t$ :	input frame at timestep $t$ ;	$f_t, \hat{f}_t$ :	CNN feature extracted at timestep $t$ (position-encoded);
$z_t$ :	SLRT output at timestep $t$ ;	$p(g_n), p(w_u)$ :	probability distribution of the $n/u^{th}$ predicted gloss/spoken language word;
$h_u$ :	$u^{th}$ SLTT output	$w_u$ :	input/output spoken language words
$< bos >$ :	beginning-of-sentence token;		
$< eos >$ :	end-of-sentence token;		
$m_u, \hat{m}_u$ :	$u^{th}$ word embedding (position-encoded)		
SE:	Spatial-Embedding;	WE:	Word-Embedding;
PE:	Positional-Encoding;	FF:	Feed-Forward
SLRT:	SLR Transformer;	SLTT:	SLT Transformer

Section 1, it is desirable to have architectures that can perform SLT completely without the use of glosses. There have been publications that implemented such architectures. By using multiple channels for the different articulators and channel-wise self-attention, Camgöz et al.

propose a method to guide the training with losses for the separate channels instead of using gloss supervision [Camgöz et al., 2020a]. Voskou et al. improve this approach by introducing a new type of layer that combines local winner-takes-all (LWTA) layers with stochastic winner sampling [Voskou et al., 2021]. While both approaches do not need gloss data for training, their performances are still below the methods that rely on gloss annotations.

Authors	Model	Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	Glosses needed
[Camgöz et al., 2018]	Seq2Seq with attention	Phoenix	32.24	19.03	12.83	9.58		31.80	no
[Camgöz et al., 2018]	Seq2Seq with attention	Phoenix	43.29	30.39	22.82	18.13		43.80	yes
[Ananthanarayana et al., 2021]	Seq2Seq with attention	Phoenix	21.71	13.89	10.07	7.87			yes
[Ananthanarayana et al., 2021]	Transformer	Phoenix	25.52	15.24	11.00	8.67			yes
[Yin and Read, 2020]	SMTc + Transformer decoder	Phoenix	<b>50.63</b>	<b>38.36</b>	<b>30.58</b>	<b>25.40</b>	<b>47.60</b>	<b>48.78</b>	yes
[Camgöz et al., 2020b]	Transformer (end-to-end SLT)	Phoenix	45.54	32.60	25.30	20.69			(yes)
[Camgöz et al., 2020a]	Transformer (multi-articulator)	Phoenix				18.31		<b>43.75</b>	no
[Voskou et al., 2021]	Stochastic Transformer	Phoenix	<b>48.61</b>	<b>35.97</b>	<b>28.37</b>	<b>23.65</b>			no

Table 1: Test performances of current architectures. To make comparisons more meaningful, all performances are based on the RWTH-PHOENIX Weather data set. The best performances for models that do not need gloss transcription and for those that need gloss transcription respectively are boldfaced.

The performances of the previously mentioned SLT approaches are summarized in Table 1. Among the most used datasets in recent years, the RWTH-PHOENIX Weather dataset was rated as the most scientifically suitable one [Ananthanarayana et al., 2021] and it is also one of the most used datasets for SLT. Therefore, the table shows the test performances of those models on the RWTH-PHOENIX Weather dataset. It can be seen that the combination of the SMTc modules and the Transformer by Yin et al. shows the best performance among the models that include gloss transcriptions in the training process. Among the models that do not depend on glosses, the approach of Voskou et al. with the LWTA layers in the Transformer outperforms the other approaches. Overall, the Transformer models show better performances than the Sequence-to-Sequence models.

## 6 Conclusion and Outlook

In this review, an overview of the important concepts of Sign Language Translation (SLT) has been presented. There is an important difference between Sign Language Recognition, the task of transcribing single signs, and SLT, a continuous translation task that results in full spoken language sentences. A realization of SLT without the need for datasets that include gloss transcription is desirable (transcription is expensive; glosses induce a bottleneck for the performance) and there are already promising Transformer architectures that achieve this. It is made clear, that the data plays a core role in the quality of the performance of the model: which dataset it is trained on, how the features are extracted and which articulators and number of channels are used as input. The potential of inputting the right data is a promising focus for future improvements in the area of SLT. Based on the first End-to-End SLT using a Transformer architecture [Camgöz et al., 2020b] several improvements have been made including using multiple input channels for different articulators. This idea could be extended by experimenting with other feature extractors and other articulators. Moreover, the recent releases of bigger datasets offer the possibility of better training for the models in future works and therefore, a chance of better performing models for real-world applications.

## References

- [Amorim et al., 2019] Amorim, C. C. d., Macêdo, D., and Zanchettin, C. (2019). Spatial-temporal graph convolutional networks for sign language recognition. In *International Conference on Artificial Neural Networks*, pages 646–657. Springer.
- [Ananthanarayana et al., 2021] Ananthanarayana, T., Srivastava, P., Chintla, A., Santha, A., Landy, B., Panaro, J., Webster, A., Kotecha, N., Sah, S., Sarchet, T., et al. (2021). Deep learning methods for sign language translation. *ACM Transactions on Accessible Computing (TACCESS)*, 14(4):1–30.
- [Angelova et al., 2022] Angelova, G., Avramidis, E., and Möller, S. (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.
- [Boggaram et al., 2022] Boggaram, A., Boggaram, A., Sharma, A., Ramanujan, A. S., and Bharathi, R. (2022). Sign language translation systems: A systematic literature review. *International Journal of Software Science and Computational Intelligence (IJSSCI)*, 14(1):1–33.
- [Camgöz et al., 2018] Camgöz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.
- [Camgöz et al., 2020a] Camgöz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020a). Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision*, pages 301–319. Springer.
- [Camgöz et al., 2020b] Camgöz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020b). Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Camgöz et al., 2021a] Camgöz, N. C., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., and Bowden, R. (2021a). Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5.
- [Camgöz et al., 2021b] Camgöz, N. C., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., and Bowden, R. (2021b). Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE.
- [Cao et al., 2019] Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [Crasborn, 2015] Crasborn, O. A. (2015). Transcription and notation methods. *Research methods in sign language studies: A practical guide*, pages 74–88.

- [De Coster et al., 2020] De Coster, M., Van Herreweghe, M., and Dambre, J. (2020). Sign language recognition with transformer networks. In *12th international conference on language resources and evaluation*, pages 6018–6024. European Language Resources Association (ELRA).
- [Denkowski and Lavie, 2014] Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.
- [Forster et al., 2012] Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J., and Ney, H. (2012). Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 3785–3789.
- [Google, 2022] Google (2022). Automl translation documentation: Evaluating models. <https://cloud.google.com/translate/automl/docs/evaluate#interpretation>. Accessed: 2022-12-06.
- [Hoiting and Slobin, 2002] Hoiting, N. and Slobin, D. I. (2002). Transcription as a tool for understanding. *Directions in sign language acquisition*, 2:55.
- [Isard and Konrad, 2022] Isard, A. and Konrad, R. (2022). MY DGS – ANNIS: ANNIS and the Public DGS Corpus. In Efthimiou, E., Fotinea, S.-E., Hanke, T., Hochgesang, J. A., Kristoffersen, J., Mesch, J., and Schulder, M., editors, *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 73–79, Marseille, France. European Language Resources Association (ELRA).
- [Li et al., 2020] Li, D., Rodriguez, C., Yu, X., and Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469.
- [Mittal et al., 2019] Mittal, A., Kumar, P., Roy, P. P., Balasubramanian, R., and Chaudhuri, B. B. (2019). A modified lstm model for continuous sign language recognition using leap motion. *IEEE Sensors Journal*, 19(16):7056–7063.
- [Neidle and Opoku, 2020] Neidle, C. and Opoku, A. (2020). A user’s guide to the american sign language linguistic research project (asllrp) data access interface (dai) 2—version 2. american sign language linguistic research project report no. 18, boston university.
- [Schulder et al., 2020] Schulder, M., Blanck, D., Hanke, T., Hofmann, I., Hong, S.-E., Jeziorski, O., König, L., König, S., Konrad, R., Langer, G., et al. (2020). Data statement for the public dgs corpus. Technical report, Project Note AP06-2020-01, DGS-Korpus project, IDGS, Hamburg University.
- [Schulder et al., 2021] Schulder, M., Blanck, D., Hanke, T., Hofmann, I., Hong, S.-E., Jeziorski, O., König, L., König, S., Konrad, R., Langer, G., Nishio, R., and Rathmann, C. (2021). Data statement for the public dgs corpus.

- [Schulder and Hanke, 2019] Schulder, M. and Hanke, T. (2019). Openpose in the public dgs corpus. *Hamburg: University of Hamburg*. doi, 10.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- [Voskou et al., 2021] Voskou, A., Panousis, K. P., Kosmopoulos, D., Metaxas, D. N., and Chatzis, S. (2021). Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11946–11955.
- [Yin and Read, 2020] Yin, K. and Read, J. (2020). Attention is all you sign: sign language translation with transformers. In *Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts*, volume 4.
- [Yuan et al., 2019] Yuan, T., Sah, S., Ananthanarayana, T., Zhang, C., Bhat, A., Gandhi, S., and Ptucha, R. (2019). Large scale sign language interpretation. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE.