# Leveraging Text-side Augmentation For Sign Language Translation

**Diandra Fabre[1], Julie Lascar[2], Julie Halbout[2], Markarit Vartampetian[1]**
[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
[2]Univ. Paris-Saclay, CNRS, LISN
{diandra.fabre, markarit.vartampetian}@univ-grenoble-alpes.fr
{julie.lascar, julie.halbout}@universite-paris-saclay.fr

**Abstract**

Sign language translation faces significant challenges due to the scarcity of annotated data and the inherent complexity of sign languages. This paper presents a method to improve sign-to-text translation models by augmenting data on the text side. We conduct experiments using two state-of-the-art models on two publicly available datasets: PHOENIX-2014T for German Sign Language and Mediapi-RGB for French Sign Language. Our main contributions are : (1) augmenting the training sets of both datasets on the text side using a generative model, (2) evaluating the impact of paraphrasing on BLEU and BLEURT scores, and (3) analyzing the impact of paraphrasing on translation outputs. We observed a significant improvement in translation for both languages. This suggests that adding variability to the training dataset through paraphrasing can lead to better generalization of the models. These results are comparable to state-of-the-art methods that use more complex approaches, such as Visual-Language fine-tuning, to improve translation.

**Keywords:** low-resources, sign language, data augmentation, translation

## 1. Introduction

Sign languages (SL) are visual languages used within Deaf communities. While spoken and written languages encode information sequentially, sign languages simultaneously use multiple articulators (hands, face, and body) to convey meaning (Sandler and Lillo-Martin, 2006). A single signed utterance can express information that would require multiple sequential units in a spoken language. Sign languages also make grammatical use of the signing space, encoding syntactic and referential relations spatially rather than through word order. As a result, sign and written languages differ fundamentally in their grammatical organization, and direct translation is not always possible. Speech-to-text or text-to-text translation likely requires less data than sign-to-text translation to reach the same performance. In the machine translation (MT) domain, a dataset of 6000 standard-length sentences is already considered tiny (De Coster et al., 2021; Gu et al., 2018). Given the additional complexity of sign languages described above, we hypothesize that reaching equivalent performance would require substantially more data than for sequential language pairs.

SL datasets are hard to collect and are several scales smaller than those used for speech or text translation [1]. However, text data are relatively easy to acquire as shown by the development of Large Language Models (LLM). These LLM are trained on an extensive amount of data, aggregated from open resources on the Internet, such as Wikipedia data or publicly available books. Millions of sentences are fed into these models, which learn language representations and perform on various tasks, such as summarizing, translation or generation. Using LLM knowledge to guide our translation system on the text side should improve translation quality and generate syntactically correct sentences.

In this paper, we relied on the paraphrasing capability of generative models. Today, human interpreters remain the most accurate translators between sign and spoken languages. A SL interpreter is a certified professional who facilitates communication between Deaf and hearing individuals, translating bidirectionally between spoken and signed language, as well as between written text and sign language. They act as cultural mediators, promoting accessibility in professional, social, and civic life. Their role requires bilingualism and deep cultural awareness, ensuring effective communication across both languages and communities.

Interpretation refers to real-time translation to or from signing, while translation is done offline with the possibility of correcting the proposed results. We will use the term "translation" in the following sections, and differentiate between the two terms if necessary. As with spoken language translation, variations in translation between different humans might be observed. Liu and Dou (2023) experienced the lexical diversity on simultaneous interpretation between different languages, and shows that interpreters used simplification techniques such as paraphrasing or summarizing, but are also required to add some explaining parts depending

[1]for more details see the "List of datasets": https://research.sign.mt/



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 129–139

on the specificities of the languages. Translating between any sign language in a 3D configuration, and a sequential language should require the same techniques.

When translating from spoken to signed language, interpreters must hear the speaker, meaning they usually work into a non-native language unless they are hCODAs (hearing Children of Deaf Adults). Emerging practices introduce Deaf interpreters as intermediaries between hearing interpreters and the audience, resulting in more natural signed expression. In different contexts, in a similar situation, hearing and deaf interpreters produce different signed language (Stone and Russell, 2011). Interference from the source language can cause the signed output to mirror spoken grammar rather than follow natural signed language conventions (Dayter, 2019). Halley (2020) also showed the contextual and personal influence of the interpreter on the translation. The work of Archambeaud (2022) also provides several potential influences of the interpretation. The media used for conveying the interpretation influence the sign space and thus, the way the interpreter signs. The physical space and work conditions can be factors of variability. The origin, political beliefs, or discrimination that could exist against the interpreter can also influence the way some specific subjects would be signed, even if the interpreter mission implies total neutrality.

We aim to reproduce this translation variability using generative models to artificially increase available datasets. These models, built on the Transformer architecture, are pretrained on large corpora, and have proven effective in diverse text generation tasks, including paraphrasing. By doing that, we hope to reduce model overfitting and increase its understanding of sign language by providing a bigger dataset. To our knowledge, this is the first systematic study of text-side paraphrasing across two SL datasets. While prior work explored paraphrasing in a gloss-to-text setting and reported no significant improvements (Angelova et al., 2022), we address the video-to-text translation task. Our main contributions are the following: (1) Augmenting the training set of two datasets (French and German SL) using a generative model on the text side; (2) Comparing two state-of-the-art architectures for small datasets with two different SL, and observing the impact of language-dependent pretrained models on the results ; (3) Evaluating the impact of paraphrasing on the scores and the translation output.

## 2. Related Works

Most research on Sign Language Translation (SLT) follows methodologies from text-to-text or speech-to-text translation. Recent works have focused on Transformer-based encoder-decoder architectures, originally introduced by Vaswani (2017).

The most straightforward application of Transformer-based encoder-decoder architecture for SLT was first proposed by Camgoz et al. (2020) and later optimized by Sincan et al. (2023), both on PHOENIX-2014T dataset (Forster et al., 2012). The encoder output was modified to simultaneously perform classification via a Connectionist Temporal Classification (CTC) decoder in parallel with the decoder used for next token prediction. This architecture enables both direct sign-to-text translation and translation guided by sign classification. These models show that gloss annotations can be useful, but glosses remain costly to annotate and are not always available. We use glosses in our study when provided, but our augmentation only increases the text side.

Recent research has explored integrating Large Language Models (LLMs) into Sign Language Translation (SLT). LLMs, based on the Transformer architecture, include encoder-only models for language representation, decoder-only models for language generation, and encoder–decoder models. In De Coster et al. (2021), a frozen BERT model was used to enhance SLT performance. Their architecture is an extension of the model introduced by Camgoz et al. (2020). Weights of either the decoder or both the encoder and the decoder were replaced by those of a BERT model: only the cross-attention and the fully connected layer on the decoder side were updated during training. They used initial BERT model trained on English. In our study, we evaluated the impact of a language-specialized BERT depending on our dataset.

Other works have investigated fine-tuning LLMs for SLT. In the works of Uthus et al. (2024) and Tarrés et al. (2023), a T5 encoder-decoder model was fine-tuned on an American Sign Language dataset using Mediapipe features (Lugaresi et al., 2019) extracted from YouTube videos, improving translation accuracy. Meanwhile, Sincan et al. (2024) trained a sign-spotting model on a large dataset, passing the sequence of detected signs to ChatGPT for sentence generation. Similarly, Wong et al. (2024) proposed Sign2GPT, a method using a frozen GPT decoder to enhance text generation. The authors reported a slight improvement on the BLEU score compared to previous methods, as we can observe from Table 1 for PHOENIX-2014T dataset if compared to the baseline Sign2gls2txt. Moreover, fine-tuning such large models requires a huge computational capability and also represents an ecological cost to take into account.

Recent studies have leveraged architectures that integrate both visual and language modalities to address Sign Language Translation (SLT). For example, Hwang et al. (2025) introduced a gloss-free

| Model | Dataset | Glosses | BLEU-4 | BLEURT |
|---|---|---|---|---|
| SLRT (Camgoz et al., 2020) | PHOENIX-2014T | yes | 20.85 | 49.02 |
| BERT2RND (De Coster et al., 2021) | PHOENIX-2014T | yes | 22.47 | x |
| BERT2BERT (De Coster et al., 2021) | PHOENIX-2014T | yes | 21.29 | 49.42 |
| Sign2GPT (Wong et al., 2024) | PHOENIX-2014T | no | 22.52 | x |
| DVE-SLT (Sincan and Bowden, 2025) | PHOENIX-2014T | no | 23.81 | 53.59 |
| SpaMo (Hwang et al., 2025) | PHOENIX-2014T | no | 24.32 | x |
| Youtube-ASL (Uthus et al., 2024) | How2sign, Youtube-ASL | no | 12.39 | 46.63 |
| SpaMo (Hwang et al., 2025) | How2Sign | no | 10.11 | 42.23 |

Table 1: Selected state-of-the-art models for SLT and their reported BLEU scores and BLEURT scores when available.

SLT framework that achieves state-of-the-art performance on Phoenix14T without relying on gloss annotations. Their approach employed off-the-shelf visual and video encoders (e.g., ViT and VideoMAE) combined with fine-tuned LLMs (e.g., mBART-L, mT0-XL, Flan-T5-XL, and Llama-2). Similarly, Jang et al. (2025) used BLIP-2 (Language–Image Pre-training with Frozen Image Encoders and LLMs) to extract visual cues from the video background, alongside a Video Swin Transformer pre-trained on a sign recognition task to encode the signs. Translation was then performed using a LLaMA model, resulting in improved SLT performance. Finally, the work of Sincan and Bowden (2025) presented DVE-SLT, a two-phase, dual visual encoder framework for gloss-free SLT, leveraging contrastive visual–language pretraining to align complementary visual features with sentence-level text embeddings. Most of these experiments results are reported in Table 1, on different datasets: PHOENIX-2014T (Forster et al., 2012), How2Sign (Duarte et al., 2021), Youtube-ASL (Uthus et al., 2024). When available, we reported both BLEU-4 and BLEURT scores (defined in Section 3.3). These approaches demonstrate the potential of LLMs in SLT, but they require large datasets for fine-tuning, and thus heavy computational resources and ecological impact.

In this paper, we focus on lightweight architectures, as opposed to the finetuning of LLMs. We conduct experiments on two datasets: German Sign Language dataset PHOENIX-2014T (Forster et al., 2012), and French Sign Language dataset Mediapi-RGB (Bull et al., 2024). We explore the data augmentation on the text side to improve translation performances, by taking advantage of generative models knowledge on languages in their text format. Data augmentation for low-resource machine translation has been widely explored for spoken languages (Fadaee et al., 2017), and more specifically applied to speech-to-text on low-resourced languages (Mi et al., 2022). For sign language, Jang et al. (2022) applied text-side augmentation with back-translation and paraphrasing on Korean Sign Language gloss data with modest improvements, and Angelova et al. (2022) experimented with paraphrasing on PHOENIX 2014T in a gloss-to-text setting. In contrast, our work applies LLM-based paraphrasing to the direct video-to-text translation task and obtains interesting improvements across two SL datasets.

## 3. Methodology

### 3.1. Datasets

**PHOENIX-2014T** is a dataset of interpreted German Sign Language (Forster et al., 2012). It consists of daily weather forecast translated in sign language. It is composed of 9 signers, with 1081 gloss vocabulary, and 7096 sentences for the training dataset. The video embedding is based on 2D CNN model fine-tuned on the same sign language dataset (Momeni et al., 2022), with a sliding window of 16 frames resulting in 1024-dimensional feature vectors (embeddings).

**Mediapi-RGB** is a dataset of news and reports in LSF, produced by deaf journalists between 2018 and 2022 (Bull et al., 2024). These videos were subtitled in French after recording. It is composed of 10 main signers, with a 445 sparse gloss vocabulary. These gloss annotations have been shown to improve translation performance (Fabre et al., 2025). Mediapi-RGB is composed 27 840 sentences after deduplication for the training dataset. The video embedding is based on SWIN model architecture from Liu et al. (2022) fine-tuned on a large-scale dataset of isolated signs in British Sign Language (Momeni et al., 2022), with a sliding window of 16 frames resulting in 768-dimensional embedding features. The authors of the Mediapi-RGB dataset provided SWIN features along with the dataset, we decided to redo the fine-tuning to have full control of our experiment, starting from the videos. Thus, baseline results may differ from other contributions on the Mediapi-RGB dataset. The main difference from the German Sign Language dataset is that the task is translation from sign to text, rather than interpretation from speech

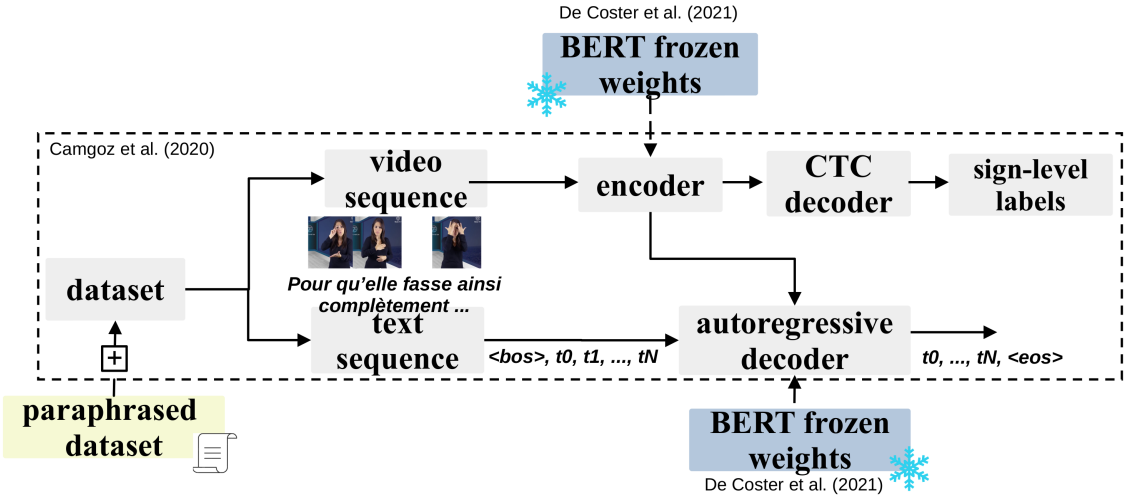


Figure 1: Simplified illustration of the architecture proposed by Camgoz et al. (2020) with our different proposed modifications: a language-specialized BERT model using De Coster et al. (2021) frozen-weights approach, and a data augmentation on the text side to increase dataset.

to sign. While the German SL dataset covers only one topic, the French SL dataset covers multiple topics accross multiple years.

## 3.2. Model Architectures

In this work, two architectures were used. The first architecture tested is the architecture proposed in Camgoz et al. (2020) and will be called *Baseline*. We used the configuration proposed by the authors, which consists of an encoder-decoder Transformer architecture with a CTC layer at the end of the encoder for classification purposes. This architecture provides both Word Error Rate (WER) for sign recognition, and BLEU score (defined in Section 3.3) for sign translation. The loss is computed as the weighted sum of CTC loss and cross-entropy loss. The second architecture, proposed by De Coster et al. (2021), consists of a Frozen Pre-trained Transformer. We will use the acronym *FPT* throughout this paper. It consists of using BERT pre-trained encoder-only model as initialization for some of the weights of both the encoder and the decoder. Only the cross-attention and the fully connected layer were updated during training. Figure 1 schematizes the architecture of the model and the two main modifications: the FPT approach using frozen weights, and the data augmentation on the text side. The encoder (resp. decoder) of the model is made of 4 (resp. 3) transformer blocks with 8 heads used in each self-attention layer. As Camgoz et al. (2020), we applied a weight of 5 for the CTC (classification) loss and 1 for the translation loss. For the remaining parameters, we kept the configuration proposed by Camgoz et al. (2020).

### 3.2.1. BERT Models

BERT-based models can also provide a structured representation space that captures linguistic proximities, and be used to tokenize the text before translation for the Baseline architecture. For *FPT* architecture, a BERT model is required to initialize some weights of the models.

The model initially used for FPT initialization was the initial `bert-base-uncased`. As we wanted to see the influence of language-specific model on performances, we changed this depending on the target language. For the German SL dataset, we chose the German BERT model `dbmdz/bert-base-german-uncased` . This is a German-only encoder, and it is uncased, as is the PHOENIX-2014T dataset. As for the model proposed by Chan et al. (2020), this model is trained on German Wikipedia dump, OpenLegalData dump and news articles. For the French SL dataset, we chose the French model CamemBERT [2] (Martin et al., 2020), trained on 138GB of French text.

## 3.3. Metrics

The BLEU metric is widely used in SL translation papers. It relies on the match of a chain of n-words between reference and estimated sequence Papineni et al. (2002). This metric suffers from its lack of adaptability regarding the possible variability when one sentence can have multiple translations. The BERT space can be used to evaluate the translation performances of a model. While BLEU score relies on a sequence of exactly matching words,

[2] https://huggingface.co/docs/transformers/en/model_doc/camembert

BLEURT score will rely on BERT space to determine the proximity between the predicted sequence and the reference sentence (Sellam et al., 2020). BLEURT score is based on a multilingual BERT space tested on 13 languages (including German, English and French for instance). It has been shown to be closer to human judgment than BLEU score for translation evaluation (Sellam et al., 2020). The BLEURT score range is between 0 and 100 (or 0 and 1) but may output values outside this range.

### 3.4. Text-Side Data Augmentation

Data augmentation is a way of artificially increasing the size of a given dataset to improve performance and reduce the risk of overfitting. In this work, we propose a dataset augmentation on the text side by doubling the training dataset. For this paraphrasing task, we used the GPT API (Paredes et al., 2023). We chose GPT-4o-mini for its strong instruction-following capabilities and paraphrase quality. While we acknowledge that using a closed-source model is not ideal for reproducibility, recent benchmarking from Beauchemin et al. (2025) shows that the best open-source alternatives (Qwen2.5-32b, Llama-3.1-8B-Instruct) perform approximately 20 points below GPT-4o-mini on French paraphrase quality metrics. Larger open-source models like Pixtral-Large (123B) approach comparable performance but require substantially more resources.

Using GPT-4o-mini allowed us to focus on data augmentation rather than model-specific tuning, as our pipeline is model-agnostic with minimal prompting (limited inference steps, low engineering overhead). Future work will systematically compare open-source alternatives to assess whether the performance gap justifies the trade-offs in cost and openness.

**German data augmentation** was performed with the following prompt: "Paraphrasiere einfach den folgenden Satz auf Deutsch:" (en: *Simply paraphrase the following sentence in German*).We calculated BLEURT and sacreBLEU scores as similarity metrics between the original training dataset and the paraphrased training dataset. We obtained a BLEURT score of `71.97` and a BLEU score of `17.82`. The training dataset was doubled using the proposed paraphrasing approach. All texts were provided in lowercase, as in the original version of dataset. This resulted in an increase in the size of the dataset vocabulary from 3000 to 4623 words. Table 2 shows some examples of paraphrasing.

**French data augmentation** Since the dataset is composed of parts of sentences, we used the contextual parts to paraphrase simultaneously all parts of a sentence. The following prompt was used : messages=[ "role": "system", "content": "Tu es un assistant linguistique expert en français." (en: *You are a language assistant specialising in French.*), "role": "user", "content": "Ta tâche est de paraphraser la phrase suivante en gardant la segmentation et le format. Portion à paraphraser :" (en: *Your task is to paraphrase the following sentence whilst retaining the line breaks and formatting. Portion to be paraphrased:*) ]

We obtained a BLEURT score of `72.78` and a BLEU score of `23.89`. This resulted in an increase in the size of the dataset vocabulary from 36160 to 37169 words. Table 3 shows some examples of paraphrases.

## 4. Experiments and Results

Each run of a model was estimated to produce about 262.88 gCO2e on an NVIDIA GTX 1080 Ti [3]. Each request on GPT API is estimated to produce 0.38 gCO2e / query [4], so a total of 2696.48 gCO2e for paraphrasing the German SL training dataset, and a total of 10579.58 gCO2e for the French SL training dataset. This last part could be reduced by running generative models locally.

### 4.1. German Dataset

As shown in Table 4, doubling the training dataset with paraphrases increases the overall BLEURT score on the test set, but results in a decrease of the BLEU score. This aligns with the observations made in Section 3.3. BLEU score is sensitive to word substitutions and perform poorly when synonyms are used, as it evaluates precise n-grams sequences. We can make the hypothesis that the BLEURT metric recognize paraphrasing as a different version of the same sentence, and can give a high score to a sequence that differs from the reference, whereas the BLEU score will assign a low score. Table 5 shows examples of the results on the test set after data augmentation.

We observed an improvement of more than 20 points of BLEURT for 113 sentences, and an improvement of more than 30 points for 48 of the test sentences (out of 642). Augmenting the dataset on the text side by paraphrasing the train dataset definitively improves the performances of the model. Table 5 shows examples of sentences where the BLEURT score increases when using data augmentation.

All pairwise differences between experiments were evaluated for statistical significance using a t-test ($p < 0.001$). We observed no significant difference between the Frozen Pre-trained Transformer and the random weight initialization, either with or without data augmentation. These results partially replicate those of De Coster et al. (2021) on

[3] https://calculator.green-algorithms.org/

[4] https://www.climateqa.com/docs/carbon-footprint/

| | |
|---|---|
| Sentence | **heute nacht kühlt die luft ab auf vierzehn bis acht grad.** *(tonight the air will cool down to fourteen to eight degrees.)* |
| Paraphrase | in der nacht wird es auf 14 bis 8 grad abkühlen. *(at night it will cool down to 14 to 8 degrees.)* |
| Sentence | **im süden kaum schauer.** *(hardly any showers in the south.)* |
| Paraphrase | es gibt im süden fast keine regenfälle. *(there is almost no rainfall in the south.)* |

Table 2: Examples of sentences and their paraphrases in German as generated by GPT API, along their translation in English.

| | |
|---|---|
| Sentence | **mais son apport à l'équipe n'avait pas été notable.** *(but his input to the team had not been noteworthy.)* |
| Paraphrase | mais sa contribution à l'équipe n'avait pas été significative. *(but his contribution to the team had not been significant.)* |
| Sentence | **handicap : mobilisation prévue le 5 mars.** *(handicap : mobilization scheduled for march 5.)* |
| Paraphrase | handicap : rassemblement programmé le 5 mars. *(handicap: rally scheduled for march 5.)* |

Table 3: Examples of sentences and their paraphrases in French as generated by GPT API, along their translation in English.

| German Sign Language (PHOENIX-2014T) | | | French Sign Language (Mediapi-RGB) | | |
|---|---|---|---|---|---|
| **Configuration** | **BLEURT** | **BLEU** | **Configuration** | **BLEURT** | **BLEU** |
| Baseline | 49.02 | 20.85 | Baseline | 10.55 | 4.14 |
| Bas. + Data aug. | **55.01** | 17.35 | Bas. + Data aug. | **11.56** | **4.71** |
| Bas. + Data aug. + BERT tok. | 51.70 | 20.71 | Bas. + Data duplication | 10.72 | 4.18 |
| FPT | 49.42 | **21.29** | FPT | 9.68 | 3.70 |
| FPT + Data aug. | 54.71 | 19.19 | FPT + Data aug. | 10.17 | 3.91 |

Table 4: BLEURT and BLEU scores for different configurations: baseline from Camgoz et al. (2020), BERT tokenizer and FPT (De Coster et al., 2021), adding paraphrasing.

PHOENIX-2014T. Data augmentation, in both configurations, surpasses the baseline dataset with significant difference. We conducted two supplementary experiments exploiting BERT space. We applied tokenization to PHOENIX-2014T text data. Adding a BERT tokenizer to the text increases drastically the size of the vocabulary (from 4623 words to 31106 tokens) and results in a slight drop of BLEURT score for our model.

## 4.2. French Dataset

Based on the results obtained on the German SL dataset, we reduced the number of experiments in French SL to four, as shown in Table 4: *Baseline* with and without data augmentation, and *FPT* with and without data augmentation. As for German SL dataset, paraphrase augmentation significantly improves the results. All experiments show significant differences (t-test with $p < 0.001$). However, unlike on PHOENIX-2014T, we observe a loss in performance when using the *FPT* architecture on Mediapi-RGB (scores of 10.55 to 9.68 on BLEURT). We attribute this to the dataset's larger size, multi-topic diversity, and vocabulary exceeding CamemBERT's token capacity, which frozen weights cannot accommodate. We observed an improvement of more than 20 points in BLEURT for 289 sentences, and an improvement of more than 30 points for 125 test sentences (out of 8060). We also compared the results between doubling the dataset by repeating each sentence twice, instead of using paraphrases. Dataset duplication resulted in a score close to the baseline (10.72 in BLEURT against 10.55), with no statistical difference between the two experiments.

From the examples displayed in Table 6, we observe the relevance of augmented datasets for the model. For the first example "mais le ministre français de l'économie bruno le maire", the sequence of sign produced in the video is as follows: " mais | ministre | économie | France | nom | POINTAGE | dire ", the last element "dire" ('to say') does not appear in the reference [5] but it ap-

[5]This last word belongs to the next subtitle, but since videos are overlapping for practical reasons, some signs appears in one video and are not translated in the subtitle.

| Model | Sentence | BLEURT |
|---|---|---|
| Ref. | **später ist es meist trocken.** *(later it is mostly dry.)* | |
| Baseline | ansonsten aber trocken ist es trocken. *(but otherwise dry it is dry.)* | 20.62 |
| Augm. | später wird es dort trocken sein. *(later it will be dry there.)* | 72.61 |
| Ref. | **morgen schwacher bis mäßiger ostwind.** *(tomorrow there will be a light to moderate easterly wind.)* | |
| Baseline | dort weht morgen meist schwach bis mäßig. *(tomorrow it will be mostly light to moderate there.)* | 40.07 |
| Augm. | morgen weht ein schwacher bis mäßiger wind aus östlichen richtungen. *(tomorrow there will be a light to moderate wind from the east.)* | 78.01 |
| Ref. | **in den alpen wird es dann sogar schneien.** *(it will even snow in the Alps.)* | |
| Baseline | am alpenrand da regen. *(on the edge of the Alps there will be rain.)* | 22.31 |
| Augm. | an den alpen fällt noch schnee. *(on the Alps there will still be snow.)* | 59.67 |
| Ref. | **an den temperaturen ändert sich wenig.** *(there is little change in the temperatures.)* | |
| Baseline | das ändert sich wenig. *(that changes little.)* | 39.58 |
| Augm. | die temperaturen ändert sich wenig. *(the temperatures changes little.)* | 79.10 |

Table 5: Examples of German Sign Language to text translation with an improvement of more than 20 points in BLEURT score between baseline and augmented dataset. *Ref.* is the reference, *Baseline* refers to Baseline model without augmentation, and *Augm.* refers to baseline + data augmentation.

| Model | Sentence | BLEURT |
|---|---|---|
| Ref. | **mais le ministre français de l'économie bruno le maire,** *(but french economy minister bruno le maire,)* | |
| Baseline | le ministre de la santé olivier véran a indiqué que le ministre de la france. *(health minister olivier véran said the french minister.)* | 11.10 |
| Augm. | le ministre de l'économie bruno le maire, a déclaré que *(economy minister bruno le maire said that)* | 43.81 |
| Ref. | **les six accusés encourent trois ans d'emprisonnement.** *(the six defendants face up to three years in prison)* | |
| Baseline | ces deux ans de prison. *(these two years in prison.)* | 23.22 |
| Augm. | ils encourent une peine de trois ans d'emprisonnement. *(they incur a three-year prison sentence.)* | 68.05 |
| Ref. | **23 titres du grand chelem à son actif,** *(23 grand slam titles to his credit,)* | |
| Baseline | il a été condamné à 23 ans de prison. *(he was sentenced to 23 years in prison.)* | 4.88 |
| Augm. | et a remporté 23 titres au grand chelem, *(and won 23 grand slam titles, )* | 55.57 |
| Ref. | **pour quelle raison ?** *(for what reason?)* | |
| Baseline | en allemagne, *(in germany,)* | 0.00 |
| Augm. | quelles sont les raisons ? *(what are the reasons?)* | 76.26 |

Table 6: Examples of French Sign Language to text translation with an improvement of more than 20 points in BLEURT score between baseline and augmented dataset. *Ref.* is the reference, *Baseline* refers to Baseline model without augmentation, and *Augm.* refers to baseline + data augmentation.

pears rightly in the prediction of the augmented model. For the second sentence, "les six accusés encourent trois ans d'emprisonnement.", the corresponding sequence of signs is: "eux | risquer | 3 ans | prison" (en: *"them | risk | 3 years | prison"*) the notion of "six defendants" is implied by the space position of the group "them", meaning they were previously mentioned. The sentence proposed by the augmented model is a good translation of this out-of-context isolated signed sentence.

As stated earlier, the Mediapi-RGB dataset is complex and includes multiple news topics that evolve over time. Let us take the example of three persons: Edouard Philippe (16 occurrences in training set) / Jean Castex (36) / Boris Johnson (10). All three of them are Prime ministers, and two of them were consecutive Prime ministers of France. We observed a confusion on model predictions between all of them. In the last example in Table 7, the signed sequence is: " venir | premier ministre |

| Model | Sentence | BLEURT |
|---|---|---|
| Ref. | **à l'époque des dinosaures,** *(during the time of the dinosaurs,)* | |
| Baseline | jean-luc mélenchon *(jean-luc mélenchon)* | -1.26 |
| Augm. | | |
| Ref. | **l'isnar-img reconnait qu'il est difficile de résister à la puissance de l'industrie pharmaceutique** *(the isnar-img acknowledges that it is difficult to resist the power of the pharmaceutical industry.)* | |
| Baseline | chine : un changement de jeu vidéo.*(china: a change in video games.)* | 1.21 |
| Augm. | ukraine : le grand public ? *(ukraine: the general public?)* | 1.39 |
| Ref. | **notamment édouard philippe et plusieurs de ses ministres,** *(including édouard philippe and several of his ministers)* | |
| Baseline | le premier ministre britannique boris johnson a donc été un premier ministre. *(british prime minister boris johnson was therefore a prime minister.)* | 5.58 |
| Augm. | le premier ministre, boris johnson, *(prime minister boris johnson,)* | 4.61 |

Table 7: Examples of French Sign Language to text translation where less than 1 point difference in BLEURT score was observed between baseline and augmented dataset. *Ref.* is the reference, *Baseline* refers to Baseline model without augmentation, and *Augm.* refers to baseline + data augmentation.

le-grand | plusieurs | ministre | aussi " (en: *" come | prime | minister | several | minister | too "*). Prime minister is followed by a sign supposed to represent the name of Édouard Philippe. However, at that time (before July 2020), there was no consensus on the signed name for him, and multiple propositions coexisted in the dataset. Both our models predicted "Boris Johnson" instead of "Édouard Philippe", ignoring this sign. Following the COVID lockdown and its numerous communications interpreted into French Sign Language, we observed a progressive convergence and standardization of many signed names. Rare vocabulary tends to translate randomly as opposed to frequent themes in the news report such as covid (793 out of 27841 sentences when counting only original training data) or deaf (949).

1390 sentences got the same score ( $\pm 1$ point) before and after augmentation. From this extract, we analyzed some sentences with lowest BLEURT scores. We picked three examples displayed in Table 7. The word "dinosaure" only appears twice in training dataset."Jean-Luc Mélenchon" sign is a two-hands signs where hand configuration is similar to the one chosen for "dinosaur", however placed differently (above the head versus shoulder height). The name "isnar-img" does not have a specific sign and is spelled using dactylology, currently lacking in translation system. As a consequence, output of baseline and augmented models seems random.

## 5. Discussion and Conclusion

### 5.1. Random Initialization Versus FPT

Our results partially replicate those of De Coster et al. (2021) on German SL. However, they reveal a critical limitation on Mediapi-RGB, where FPT seems to degrade performance on larger, multi-domain datasets. Our results suggest that the FPT approach is more suitable for small, narrow-domain datasets (7K sentences, weather-only, 3K vocab), where frozen weights help prevent overfitting. On larger, multi-domain datasets however, frozen weights appear to limit the model's capacity to adapt, leading to underfitting. The tokenizer might not match the requirements: 36K words vocabulary exceed CamemBERT's 32K token capacity. FPT seems more suitable to small and narrow-domain datasets. This highlights the importance of matching architectural choices to dataset characteristics, independently of data augmentation.

### 5.2. Paraphrasing Contribution

These experiments on two sign languages show the efficiency of data augmentation on the text side to improve SL translation models. In the French dataset, paraphrasing helps understanding of the main theme of a sentence. The examples in Table 6 confirm this tendency, showing that paraphrasing improves predictions on a specific theme, sometimes with better precision, such as "economy minister" versus "health minister."

Paraphrasing methods show promise for augmenting text data, but they also present limitations. The German weather forecast dataset is well-defined and scientifically constrained, reducing the likelihood of bias. In other domains, however, LLM-generated paraphrases may introduce distortions, altering the original meaning of sentences due to societal or stereotypical influences, as shown in prior work such as Ducel et al. (2024). Next steps would include a bias analysis of the French SL dataset paraphrasing. We evaluated paraphrasing using the open-source model Qwen2.5-32b mentioned previously, obtaining a BLEURT score of

75.23 and a sacreBLEU of 31.17, both higher than those achieved with GPT-4o-mini. However, a comparative qualitative analysis between Qwen-32b and GPT-4o-mini demonstrated that Qwen-32b revealed more gaps in its knowledge of Deaf culture. For example, it confuses key terms such as “Deaf” and “hCODA” and incorrectly uses *entendeuse* instead of the correct term *entendant* for "hearing person". Precise terminology is particularly critical when translating or paraphrasing Deaf-specific content. A second consideration arises in the German SL dataset, which involves interpreted data, moving from spoken language to sign language. We kept the sign as ground truth and paraphrased the speech. We have to verify if paraphrases would be interpreted in the same way as the original signed sequences.

Doubling the dataset through paraphrasing helps improve the translation task. We questioned the relevance of variability in translation between oral language and SL, or between SL and text. As most of the literature focuses on oral-to-sign language interpretation, the observations largely apply to that context, but they are also valid for sign-to-text translation. This proves the importance of adding this variability to our translation models, as there is not one single way to translate from French Sign Language to French.

For completing this evaluation, we will include in the next steps a human-in-the-loop. We would provide human translators, interpreters and also hCODA and Deaf participants with our out-of-context signed sentences and ask them to produce precise translations. In parallel, we would ask them to validate proposed automatic paraphrases and, when possible, generate new ones that accurately reflect the signed sequence. By doing so, we should reduce noise in our data and obtain more reliable sign-to-text sequences with multiple paraphrases.

### 5.3. Conclusion

In this study, we explored how generative models can help improve sign language translation through lightweight text-side data augmentation. We showed that paraphrasing the target text improves translation performance on two datasets (German and French SL), even with limited training resources. Next steps will include a human-in-the-loop for expert paraphrasing validation and generation, and a thorough analysis of the effect of associating two text translations to a single signed production.

## 6. Ethical Considerations

None of the authors is Deaf. As previously mentioned there might be biases in LLM-generated paraphrases alongside a risk of hallucinations. The dataset used for French Sign Language is sparse and covers multiple subjects with variable occurrences.

## 7. Bibliographical References


Galina Angelova, Eleftherios Avramidis, and Sebastian Möller. 2022. Using neural machine translation methods for sign language translation. In *Proceedings of the 60th annual meeting of the association for computational linguistics: Student research workshop*, pages 273–284.

Florine Archambeaud. 2022. *Vers un modèle des espaces en interprétation du Français vers la Langue des Signes Française*. Theses, Université de la Sorbonne nouvelle - Paris III.

David Beauchemin, Yan Tremblay, Mohamed Amine Youssef, and Richard Khoury. 2025. COLE: a comprehensive benchmark for French language understanding evaluation. *arXiv preprint arXiv:2510.05046*.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proc. of CVPR*, pages 10023–10033.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German’s next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Daria Dayter. 2019. *Collocations in non-interpreted and simultaneously interpreted English: a corpus study*. Routledge.

Mathieu De Coster, Karel D’Oosterlinck, Marija Pizurica, Paloma Rabaey, et al. 2021. Frozen pretrained transformers for neural sign language translation. In *18th Biennial Machine Translation Summit (MT Summit 2021)*, pages 88–97. Association for Machine Translation in the Americas.

Fanny Ducel, Aurélie Névéol, and Karën Fort. 2024. “You’ll be a nurse, my son!” Automatically assessing gender biases in autoregressive language models in French and Italian. *Language Resources and Evaluation*, pages 1–29.

Diandra Fabre, Julie Lascar, Julie Halbout, Yanis Ouakrim, Annelies Braffort, Thomas Hueber, et al. 2025. Exploring sign-level strategies to enhance automatic translation of French Sign Language. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, pages 1–7.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. LLMs are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18362–18372.

Jiatao Gu, Hany Hassan Awadalla, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354.

Mark Halley. 2020. Rendering depiction: a case study of an American Sign Language/English interpreter. *Journal of Interpretation*, 28(2):3.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. 2025. An efficient gloss-free sign language translation using spatial configurations and motion dynamics with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3901–3920, Albuquerque, New Mexico. Association for Computational Linguistics.

Jin Yea Jang, Han-Mu Park, Saim Shin, Suna Shin, Byungcheon Yoon, and Gahgene Gweon. 2022. Automatic gloss-level data augmentation for sign language translation. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6808–6813.

Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. 2025. Lost in translation, found in context: Sign language translation with contextual cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8742–8752.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211.

Zhibo Liu and Juhua Dou. 2023. Lexical density, lexical diversity, and lexical sophistication in simultaneously interpreted texts: a cognitive perspective. *Frontiers in psychology*, 14:1276705.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

Chenggang Mi, Lei Xie, and Yanning Zhang. 2022. Improving data augmentation for low resource speech-to-text translation with diverse paraphrasing. *Neural Networks*, 148:194–205.

Liliane Momeni, Hannah Bull, KR Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Automatic dense annotation of large-vocabulary sign language videos. In *European Conference on Computer Vision*, pages 671–690. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Cristian Mauricio Gallardo Paredes, Cristian Machuca, and Yadira Maricela Semblantes Claudio. 2023. ChatGPT API: Brief overview and integration in software development. *International Journal of Engineering Insights*, 1(1):25–29.

Wendy Sandler and Diane Carolyn Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge University Press.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Ozge Mercanoglu Sincan and Richard Bowden. 2025. Contrastive pretraining with dual visual encoders for gloss-free sign language translation. *arXiv preprint arXiv:2507.10306*.

Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. 2023. Is context all you need? Scaling neural sign language translation to large domains of discourse. In *Proc. of ICCV*, pages 1955–1965.

Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. 2024. Using an LLM to turn sign spottings into spoken language sentences. *arXiv preprint arXiv:2403.10434*.

Christopher Stone and Debra Russell. 2011. Interpreting in International Sign: Decisions of Deaf and non-deaf interpreters.

Laia Tarrés, Gerard I Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró-i Nieto. 2023. Sign language translation from instructional videos. In *Proc. of CVPR*, pages 5625–5635.

Inigo Jauregi Unanue, Jacob Parnell, and Massimo Piccardi. 2021. BERTTune: Fine-tuning neural machine translation with BERTScore. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 915–924.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2024. Youtube-ASL: a large-scale, open-domain American Sign Language-English parallel corpus. *Advances in Neural Information Processing Systems*, 36.

Ashish Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging large language models for gloss-free sign language translation. In *The Twelfth International Conference on Learning Representations*.

## 8. Language Resource References

Hannah Bull, Yanis Ouakrim, Julie Lascar, Annelies Braffort, and Michèle Gouiffès. 2024. Mediapi-RGB corpus.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, et al. 2021. How2sign: A large-scale multimodal dataset for continuous American Sign Language.

Jens Forster, Christoph Andreas Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H. Piater, and Hermann Ney. 2012. RWTH-PHOENIX-Weather: A large vocabulary Sign Language recognition and translation corpus.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2024. YouTube-ASL: A large-scale, open-domain American Sign Language-English parallel corpus. volume 36.