Effect of Data Augmentation with Multi-View Perspectives of Signers on the DGS-Fabeln-1 Dataset

Fabian Renner, Daksitha Withanage Don, Elisabeth André, Cristina Luna-Jiménez
Chair for Human-Centered Artificial Intelligence,
University of Augsburg, Augsburg, Germany
{fabian.renner, daksitha.withanage.don, elisabeth.andre, cristina.luna.jimenez}@uni-a.de

Abstract

Sign languages constitute the principal form of communication for deaf communities across the globe. Nevertheless, the development of reliable Continuous Sign Language Translation (CSLT) systems is constrained by the lack of sufficient data and models able to handle spatio-temporal information. In this article, we explore the effect of adding multiview perspectives of the signer to the training set as data augmentation using the UniSign framework for the DGS-Fabeln-1 dataset. Our results reveal that increasing dataset size and using multiple camera perspectives significantly improve performance, with the best configurations achieving BLEU-4 scores of 4.20%. These results provide a competitive baseline for the DGS-Fabeln-1 dataset and guidance for further optimizations of CSLT systems.

Keywords: Sign Language Translation, Transformers, Transfer-Learning, Multi-view perspectives data augmentation

1. Introduction

By 2050, an estimated 2.5 billion people worldwide will experience some degree of hearing loss (Chadha et al., 2021). For many of these individuals, sign language serves as their primary mode of communication (Macht and Steinbach, 2019). However, due to fundamental differences in linguistic structure between sign languages and spoken languages, effective communication between user groups remains a persistent challenge (Macht and Steinbach, 2019), relying always on the availability of interpreters. Due to the existing research challenges and the motivation to develop technology that can foster independent communication, Sign Language Translation (SLT) is an active research topic.

Traditionally, SLT has followed the Isolated Sign Language Recognition (ISLR) approach, in which glosses are recognized from short fragments of videos. Although ISLR systems could be effective in reduced vocabulary scenarios (Luna-Jiménez et al., 2025) and useful for sign language spotting (Mercanoglu Sincan and Bowden, 2025); more recently, the SLT research community is driving their efforts to investigate Continuous Sign Language Translation (CSLT). This new paradigm is more natural to the translation task, receiving videos in Sign Language and translating them into Spoken Languages. Additionally, the absence of tedious gloss annotation saves human effort since only alignment between segments is required, making recording and data annotation more feasible for researching with large visual-language models (Li et al., 2024).

For this reason, the present article investigates the optimization of a Transformer-based model for SLT from DGS (German Sign Language) to German. Therefore, the contributions of this article are threefold:

- Considering the creative and varied nature of the narrative domain, we analyzed the impact of dataset size by comparing training configurations of six versus seven fairy tales. This assessment aimed to determine if incorporating additional, distinct story content improves the ability of the model to generalize and translate sentences across different narrative structures.
- Given the success of transfer learning in multiple tasks (Esteban-Romero et al., 2024; Luna-Jiménez et al., 2024), we investigated the transferability of knowledge within the UniSign transformer architecture to the DGS-Fabeln-1 dataset. This was conducted by comparing models initialized with random weights against models initialized with the weights learned from multilingual datasets.
- Additionally, since training large-scale models with more samples usually results in higher performance, we compared the effect of including only frontal camera perspectives against incorporating all eight synchronized perspectives available in the DGS-Fabeln-1 dataset.

To our knowledge, this is the first work proposing a competitive baseline for the DGS-Fabeln-1 for further investigations in SLT in a creative domain, as it is fairy tales.

The remaining part of the article is structured as follows: Section 2 provides foundational knowledge

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 429–437
16 May 2026.

on existing datasets annotated in DGS, and models employing Transformer-based architectures for performing SLT. Section 3 details the applied methodology, including the dataset preparation, feature extraction, and the experimental setup. Section 4 and 5 evaluate the experimental results by analyzing the influence of the distinct parameters. Finally, Section 6 outlines conclusions and future work.

2. Related Works

In this section, we introduce the datasets available for solving SLT in DGS. Next, we summarize existing models in the literature, focusing on transformers and Language Models-based architectures.

2.1. Datasets in German Sign Language

German Sign Language (Deutsche Gebärdensprache, DGS) is the main sign language employed by the Deaf community in Germany (Bross, 2020). DGS is expressed and perceived through the visual-gestural modality, encoding meaning in manual and non-manual articulators, including hand configurations and positions, facial expressions, head orientation, and body posture (Bross, 2020). Not only do manuals and non-manuals play a crucial role in Sign Language, but also the signing space.

Despite the relevance of the location of different events in the space, most of the available datasets for researching ISLR and CSLT only contain the frontal perspective of the signers. For example, RWTH-PHOENIX Weather 2014 T (Camgoz et al., 2018), one of the most used corpora by the machine learning community, only contains frontal perspectives of a signer interpreting broadcast news. Similarly, the AVASAG dataset (Bernhard et al., 2022) has only frontal-view videos of an interpreter signing sentences related to train station scenarios and wearing a motion tracking suit. From a linguistic point of view, one of the most popular datasets is the DGS-Korpus (Konrad et al., 2020), with more than 50 hours of videos of dialogues recorded with a frontal perspective and a profile perspective of the signers, although this profile perspective has the moderator plus the signers which would require further pre-processing to detect and extract the profile camera view of each individual. More recently, Avramidis et al. (2025) collected more than 900 hours of DGS in the multilingual TUB corpus covering up to 12 languages with 1.3 million subtitle segments containing 14 million tokens. The dataset covers topics related to politics, news, education, and social content. However, as in previous datasets, most videos are only available from the frontal perspective of the signer.

Finally, to support research in DGS and study the impact of having videos with different perspectives of the signer, the DGS-Fabeln-1 dataset (Nunnari et al., 2024) was introduced as a publicly available corpus for the linguistic and computer science community. It presents a native DGS signer retelling a set of seven German fairy tales, recorded from seven distinct camera angles as well as an additional backchannel perspective, to capture detailed spatial movements and facial expressions. The dataset has a total duration of 1 hour and 32 minutes, aligned with 1,428 written German sentences, making it a valuable resource for training and evaluating SLT systems.

2.2. Transformer Architectures for SLT

Since its introduction, Transformer architecture has been the leading model for sequence-to-sequence tasks, especially in natural language processing applications like machine translation. Its key innovation, the self-attention mechanism, allows the model to assess the relevance of input elements regardless of their position, making it highly effective at capturing long-term dependencies. Unlike recurrent neural networks, Transformers process all elements in parallel, which greatly improves training efficiency. The architecture typically follows an Encoder-Decoder structure: the Encoder transforms input sequences into contextual representations, which the Decoder uses to generate the output (Vaswani et al., 2017).

In SLT, this setup is particularly effective due to the sequential nature of both visual sign streams and textual output. The Encoder processes spatial and temporal patterns from sign language videos, such as hand positions and facial expressions, as well as their timing, learning rich, context-aware representations. The Decoder then uses these context-rich embeddings to produce fluent written or spoken translations in a differnt language (Vaswani et al., 2017; Duarte et al., 2021; Damdoo and Kumar, 2025; Cihan Camgöz et al., 2020).

From this family of models, UniSign (Li et al., 2024) is a Transformer-based architecture designed for sign language understanding that unifies several core tasks of this domain within a single framework: ISLR, which focuses on classifying individual signs; continuous sign language recognition (CSLR), which learns the alignment between sign sequences and their corresponding glosses; and SLT, which generates natural language descriptions from signed input. It leverages generative pretraining followed by task-specific fine-tuning, which enables effective knowledge transfer and optimization across tasks. UniSign is designed to be modular and extensible, allowing researchers to explore diverse input configurations and pretraining strategies across a range of sign language understanding scenarios.

Additionally, recent studies have further emphasized the potential of leveraging diverse camera angles; for instance, Ranum et al. (2024) demonstrated that incorporating multi-view data during training significantly enhances the model's ability to learn robust spatial representations, even when inference is restricted to a single frontal perspective, highlighting the advantages of this strategy as data augmentation for improving the performance of the model.

3. Methodology

This section describes the methodology used to adapt and evaluate the UniSign Transformer for DGS translation using the DGS-Fabeln-1 dataset. The process includes preparing the dataset by extracting pose landmarks and training multiple models under different configurations.

3.1. Dataset Preparation

The dataset employed in this article is the DGS-Fabeln-1 dataset. The corpus comprises 573 video segments of 7 fairy tales. Each sentence was recorded simultaneously from eight distinct camera perspectives. A central tablet provided a frontal view, while six additional tablets were positioned in a semi-circle at angles of 30°, 60°, and 90° on both sides of the signer. Finally, an eighth overhead camera captured the director's back-channeling, documenting the full interaction between the signer and the interlocutor.

For our experiments, we created three different configurations of the training and validation sets of the dataset according to the effect to explore in each of them:

1. The first dataset configuration contains six fairy tales and exclusively uses the front-view camera perspective. For the analysis, the samples are randomly split into an 80:10:10 train:validation:test ratios. This resulted in 338 videos for the training split, 42 videos for the validation split, and 43 videos for the test split. The last fairy tale ("Snow White") was left out of the test sets for evaluating data augmentation strategies and comparing the same 42 videos of the test set across experiments.

2. The second configuration also uses only the frontal-view videos, but includes the seventh fairy tales samples in the training and validation sets. In this case, the dataset resulted in 476 videos for the training set, 52 videos in the validation set and 43 in the test set.

3. Finally, the third configuration uses seven fairy tales and all perspectives, following the same allocation adopted in the second dataset configuration. In this way, if a sentence was in the train or validation split in the second configuration, it is also in the same split in this third configuration, with all available video perspectives for that sentence included. This ensures that all perspectives of a sentence always appear in the same subset. In these experiments, the train split contains 3,749, the validation split has 409 videos, and the test split has 43 videos.

As can be observed, the test split is identical across all the dataset configurations to ensure comparability of results in each experiment.

3.2. Landmark Extraction

For the feature extraction, we followed the UniSign pipeline (Li et al., 2024) to compute pose landmarks from the raw videos using OpenPose (Cao et al., 2021). OpenPose is an open-source tool that detects human keypoints from images and videos. This process produced frame-wise skeleton data capturing 134 keypoints of the hands, body and face.

3.3. UniSign Transformer

Figure 1 shows the architecture of the UniSign model. The UniSign model is capable of processing skeletal pose landmarks either independently or in combination with RGB video data, utilizing a multimodal fusion strategy to learn fine-grained spatio-temporal representations. The model contains several encoders that pass each of the features extracted from different body parts (i.e. hands, pose or face) to a projection layer and a three-layers Spatial Graph Convolutional Network (GCN) per modality. The outputs are then fed into the temporal encoders that are Spatio-Temporal Graph Convolutional Networks (ST-GCN) and the outputs are then combined with an average pooling to maintain the dimension of the embedding space of the multilingual language model mT5. Subsequently, the LM receives the resultant tokens to predict the most probable sentence represented in the video.

The vision encoder consists of an EfficientNet-B0 that receives the images of the hand for the experiments with RGB videos and fusion them through the Prior-Guided Fusion (PGF) module that combines the landmarks information from the hands with the visual features extracted from the hands of the original videos via an attention mechanism.

3.4. Evaluation Metrics

For comparing the model's performance across experiments, we used common metrics employed in

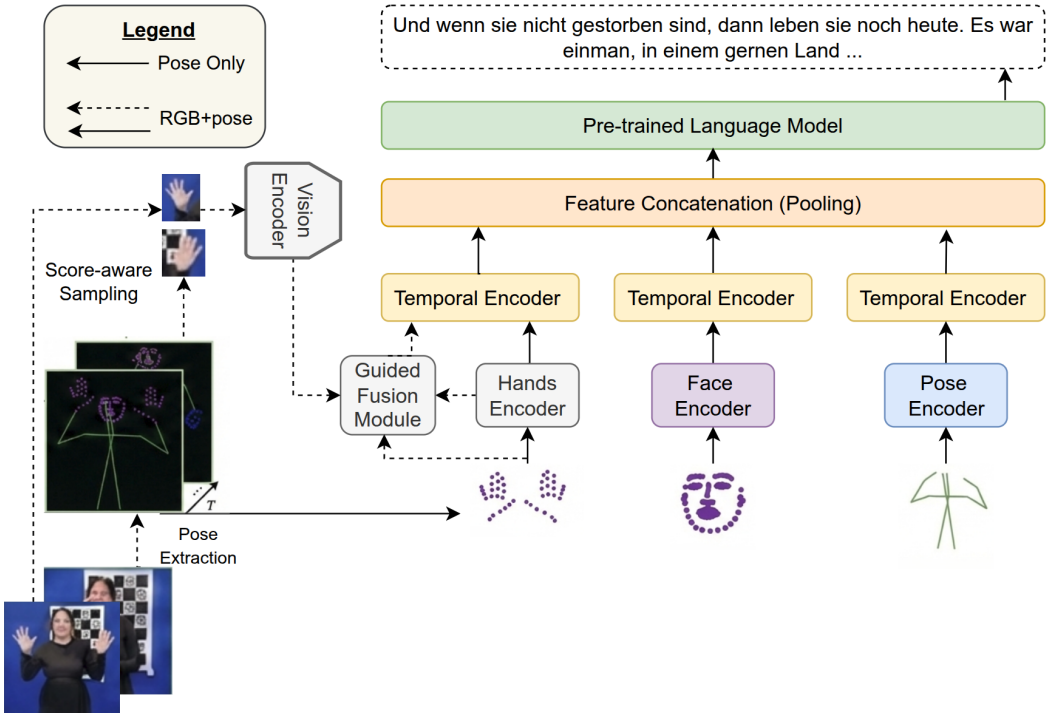

Figure 1: Overview of the UniSign architecture.

previous literature in SLT and MT (Li et al., 2024; Duarte et al., 2021; Camgoz et al., 2018; De Coster et al., 2023; Qin et al., 2025): BLEU-1 to BLEU-4 (Papineni et al., 2002), and ROUGE-L (ROUGE for short in the results tables) (Lin, 2004). These metrics are automatically computed within the UniSign framework by comparing the generated written German sentences against the corresponding ground-truth references. BLEU-1 assesses lexical accuracy through unigram precision, defined as the ratio of matching words to the total word count in the generated sentence, irrespective of their position. Conversely, BLEU-4 examines the alignment of four-word sequences (4-grams), providing an evaluation of the syntactic fluency. ROUGE-L quantifies structural overlap through the Longest Common Subsequence (LCS). This metric identifies the longest shared sequence of words that maintains the same relative order between the hypothesis and the reference. Unlike the fixed n-gram constraints in BLEU, the components of an LCS do not need to be contiguous. This characteristic allows the metric to accommodate variations in sentence structure while still rewarding correct word-order patterns. Additionally, to evaluate the efficiency in training times of each configuration, we included the training time. Finally, we added an auxiliary metric to measure the number of unique generated lemmas in inference time. Each lemma is the basic dictionary form of a word or sign, without grammatical endings or variations in form (e.g., “run” as the lemma for “running,” “ran,” and “runs”). Using lemmas in evaluation allows different word forms of the same meaning to be matched, making the metric less sensitive to inflectional variation. This metric also provides insight into the diversity of vocabulary. To operationalize this, we used the spaCy library (de_core_news_sm model) to process the predicted sentences and ground-truth texts. The process involved word-level tokenization, filtering for alphabetic characters and using the German language model to find the dictionary base form (lemma) of each word (e.g., converting "ging" to "gehen") in lowercase. The unique base forms were then aggregated to compare the overall vocabulary size of the translations and reference sentences. An illustrative application of these metrics on a sample sentence pair is provided in Table 1.

4. Experiments

In this section, we introduce the main experiments performed.

4.1. Experimental Set-Ups

To optimize our model, we conducted controlled experiments varying parameters while keeping other conditions constant. Specifically, we explored:

- *Effect of the Dataset size*: 6 vs. 7 fairy tales. As the domain is quite creative and fairy tales normally differ from each other except in certain sentences at the beginning and at the end,

Reference (German): "Der schlaue Fuchs bemerkte den Raben mit dem Käse auf dem Ast." **Reference (English Translation):** "The clever fox noticed the raven with the cheese on the branch."		
Prediction (German): "Fuchs sehen Raben mit Käse auf Ast." **Prediction (English Translation):** "Fox see ravens with cheese on a branch."		
Metric	**Value**	**Calculation Basis**
BLEU-1	85.71%	6 matches / 7 words in prediction. Matches: {Fuchs, Raben, mit, Käse, auf, Ast}.
BLEU-4	0.00%	0 matches. No sequence of four consecutive words matches the reference exactly.
ROUGE-L	63.16%	F1-score derived from LCS length 6. Precision: 6/7 (prediction), Recall: 6/12 (reference).
Lemmas (Ref/Pred)	9 / 7	Unique lemmas Ref: {der, schlau, fuchs, bemerken, rabe, mit, käse, auf, ast}. Unique lemmas Pred: {fuchs, sehen, rabe, mit, käse, auf, ast}.

Table 1: Illustrative metric calculation for a sample sentence pair.

we evaluated whether adding data from another fairy tale could improve the performance of the model for predicting sentences of the other fairy tales.

- *Effect of transfer learning*: Weights were either randomly initialized or initialized from the best pre-trained mT5 checkpoint provided by UniSign, which was derived from the multilingual mT5 Base model trained on large-scale multilingual text data.

- *Data Augmentation with additional Camera perspectives*: Front-only vs. All perspectives for training and validation. These experiments aim to evaluate whether adding data augmentation (i.e. training with more perspectives) could result in a final gain in the performance of the model.

For each experiment, we saved the best checkpoint based on the top BLEU-4 score in the test set, which was used for comparisons in the result section.

5. Results

In this section, we evaluate the results of the experiments described before. For this purpose, we first analyze the influence of each varied parameter individually, followed by a general overview. Specifically, we first conducted experiments using only the first configuration of the dataset with 6 fairy tales and the standard frame rate as a baseline. Subsequently, we kept the number of epochs constant and transferred the weights between models and then we evaluated the use of more camera perspectives as data augmentation to compare the performance of the model in each scenario.

5.1. Baseline on the Dataset size

In order to establish a first baseline, initial comparisons were conducted on the number of epochs from 50 to 100, which caused mostly no changes in the scores but increased the training times; for that reason, further experiments were performed only with 50 epochs. A complete set of all the performed experiments is added as supplementary material [1].

Regarding the experiments related to the dataset size, we employed a batch size of 4 with 50 epochs, random weights initialization and the regular number of frames of the original dataframe, corresponding with the first and second configurations of the datasets presented in Section 3.1. As can be observed in Table 2, the scores for experiments with 7 fairy tales were mostly significantly higher across all metrics than in the trials with only 6 fairy tales. For this specific configuration, an increment of 1.19 percentage points was obtained in BLEU-4 using the seven fairy tales in the training and validation sets, resulting in a generative model with a larger vocabulary, as can be observed in the lemmas. These results indicate that even in 'creative' topics in which sentences could differ from one fairy tale to another, adding more fairy tales to the training set results in higher performance of the models for predicting next sentences in other fairy tales. For this reason, in the next experiments, 7 fairy tales were kept for the training and validation.

Tales	BLEU-1	BLEU-4	ROUGE	Lemmas
6	13.66%	0.97%	13.73%	9
7	18.17%	2.16%	15.80%	16

Table 2: Results on the evaluation of the number of fairy tales used for training.

[1] https://cloud.hcai.eu/s/jJo2pdKEnZXqzLJ

5.2. Effect of Transfer Learning

As previous literature in deep learning has evidenced in different tasks, pre-training normally improves performance across similar tasks. Accordingly, we evaluated the previous top configuration trained from scratch with random weights versus applying Transfer Learning (TL) using pretrained weights, as shown in Table 3.

Overall, the training times were similar and pretrained weights delivered superior performance across the majority of experiments, with a particularly strong advantage for the Pose + RGB input type. This indicates that the pre-trained model, having learned foundational features from a large, diverse dataset, generalizes better and fine-tunes more effectively. Based on these findings, we concluded to use pre-trained weights for further model optimization.

Weights	BLEU-1	BLEU-4	ROUGE	Lemmas
Scratch	18.17%	2.16%	15.80%	16
TL	17.88%	4.15%	16.09%	28

Table 3: Results on Transfer Learning (TL).

5.3. Data Augmentation with additional Camera perspectives

Considering that different perspectives of the camera represent the same meaning when signing, another goal of the experiments was to explore whether adding other perspectives from the same sentences and person could help to improve the performance of the model when doing predictions on the frontal videos test set, following the third dataset configuration described in Section 3.1.

For the experiments with the Pose + RGB input type, no clear influence of the video perspective could be determined, especially when exploring the same model with different batch sizes. As shown in Table 4, results were generally similar but the training time increased significantly in trials using all perspectives by a factor of about three to four times. Given that in some cases, such as with the batch size of 8, the use of more perspectives increased BLEU-4 by +2.06 percentage points, for further optimizations, we continued using all the perspectives.

Batch	Persp.	BLEU-1	BLEU-4	ROUGE	Lemmas
2	Frontal	17.88%	4.15%	16.09%	-
2	All	17.88%	4.20%	**16.12%**	-
4	Frontal	17.88%	4.15%	16.09%	28
4	All	17.88%	4.20%	16.12%	43
8	Frontal	**18.32%**	2.15%	11.47%	12
8	All	15.21%	**4.21%**	11.43%	18

Table 4: Results on Camera Perspectives as data augmentation.

6. Conclusion

Sign language is the means of communication for deaf communities around the world. However, Sign Language Machine Translation solutions are still under research. In this article, we explored optimization strategies as well as data augmentation approaches in the UniSign architecture to perform SLT from DGS to spoken German. The DGS-Fabeln-1 dataset was employed, given its flexibility and the recorded multi-view perspectives of the same signer. From the different explorations, we observed that employing pre-trained weights of the multi-language model improved the final performance of the model. Results from the data augmentation techniques suggest that for some configurations, employing all the camera perspectives resulted in an increment in the BLEU-4, however these additional number of training samples resulted in longer training times that should be considered. Similarly, adding more fairy tales to the training helped to predict fragments of other fairy tales, being in line with the idea that more data results in better performance. In summary, in this article, we propose a competitive baseline for the DGS-Fabeln-1 dataset that achieved a BLEU-1 of 17.88%, BLEU-4 score of 4.20% and ROUGE of 16.12% in the test set of 6 of the fairy tales of DGS-Fabeln-1. This model was trained with all the camera perspectives of the videos and a batch size of 2, achieving an increment of +4.22 percentage points in BLEU-1, and +3.23 in BLEU-4 with respect to the initial model. This baseline is in line with other proposals in the field for SLT employing transformers, as the solution proposed by Tarrés et al. (2023) for the How2Sign dataset that achieved a BLEU-4 of 8.03% having a considerable larger dataset. This study establishes a baseline for further improvements and explorations in the field.

As future work, further optimization steps can be taken by introducing and varying other parameters, such as the learning rate or warmup epochs. Additionally, exploring other libraries and datasets could provide insights about the limitations of the pose extraction modules, as OpenPose in this case, given that they play a critical role in the pose detections and hence in the sign predictions. Once higher scores are achieved, future work should also involve human evaluation to test the real-world application of the translations. Furthermore, changes to the model architecture, including more advanced LLMs could provide a boost in the final performance that could be explored as future work.

7. Limitations

From the experiments, the test loss indicates possible overfitting, a similar effect has been observed in similar works of SLT under limited resource scenarios (Ko et al., 2019). A plausible explanation for this is the small size of the dataset, which contains only about 4,000 videos, even when using all videos and perspectives. To address this issue, further optimization would need to increase the size of the dataset.

8. Acknowledgments

This contribution is funded by the German Ministry for Education and Research (BMBF) through the BIGEKO project, grant number 16SV9094. It has also received funding from the project FORSocialRobots, financed by the Bavarian Research Foundation (AZ1594-23).

9. Bibliographical References

Eleftherios Avramidis, Vera Czehmann, Fabian Deckert, Lorenz Hufe, Aljoscha Lipski, Yuni Amaloa Quintero Villalobos, Tae Kwon Rhee, Mengqian Shi, Lennart Stölting, Fabrizio Nunnari, and Sebastian Möller. 2025. The TUB Sign Language Corpus Collection. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, pages 1–7, Berlin Germany. ACM. doi:10.1145/3742886.3756709.

Lucas Bernhard, Fabrizio Nunnari, Amelie Unger, Judith Bauerdiek, Christian Dold, Marcel Hauck, Alexander Stricker, Tobias Baur, Alexander Heimerl, Elisabeth André, Melissa Reinecker, Cristina España Bonet, Yasser Hamidullah, Stephan Busemann, Patrick Gebhard, Corinna Jäger, Sonja Wecker, Yvonne Kossel, Henrik Müller, Kristoffer Waldow, Arnulph Fuhrmann, Martin Misiak, and Dieter Wallach. 2022. Towards automated sign language production: A pipeline for creating inclusive virtual humans. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '22, page 260–268, Corfu, Greece. Association for Computing Machinery. URL: https://doi.org/10.1145/3529190.3529202, doi:10.1145/3529190.3529202.

Fabian Bross. 2020. Object marking in German Sign Language (Deutsche Gebärdensprache): Differential Object Marking and Object Shift in the Visual Modality. *Glossa: a journal of general linguistics*, 5(1). doi:10.5334/gjgl.992.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7793, Salt Lake City, UT, USA. IEEE Computer Society. URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Camgoz_Neural_Sign_Language_CVPR_2018_paper.html.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):172–186. doi:10.1109/TPAMI.2019.2929257.

Shelly Chadha, Kaloyan Kamenov, and Alarcos Cieza. 2021. The World Report on Hearing, 2021. Global Report. World Health Organization Press. URL: https://www.who.int/publications/i/item/9789240020481.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10020–10030, Online. IEEE Computer Society. doi:10.1109/CVPR42600.2020.01004.

Rina Damdoo and Praveen Kumar. 2025. SignEdgeLVM transformer model for enhanced Sign Language Translation on Edge Devices. *Discover Computing*, 28(1). doi:10.1007/s10791-025-09509-1.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine Translation from Signed to Spoken Languages: State of the art and Challenges. *Universal Access in the Information Society*, 23(3):1305–1331. doi:10.1007/s10209-023-00992-1.

Amanda Cardoso Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giró i Nieto. 2021. How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language. In *CVPR*, pages 2735–2744, Online. Computer Vision Foundation / IEEE. URL: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2021.html#DuartePVGDMTG21.

Sergio Esteban-Romero, Iván Martín-Fernández, Manuel Gil-Martín, David Griol-Barres, Zoraida Callejas-Carrión, and Fernando Fernández-Martínez. 2024. LLM-Driven Multimodal Fusion for Human Perception Analysis. In *Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor*, MuSe'24, page 45–51, Melbourne VIC, Australia. Association for Computing Machinery. doi:10.1145/3689062.3689084.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural Sign Language Translation Based on Human Keypoint Estimation. *Applied Sciences*, 9(13):2683. doi:10.3390/app9132683.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release. Universität Hamburg. Dataset. doi:10.25592/dgs.corpus-3.0.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2024. Uni-Sign: Toward Unified Sign Language Understanding at Scale. In *The Thirteenth International Conference on Learning Representations*, Singapore, Singapore. URL: https://iclr.cc/virtual/2025/poster/31250.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. URL: https://aclanthology.org/W04-1013/.

Cristina Luna-Jiménez, Lennart Eing, Annalena Bea Aicher, Fabrizio Nunnari, and Elisabeth André. 2025. Lightweight Transformers for Isolated Sign Language Recognition. In *Proceedings of the 27th International Conference on Multimodal Interaction*, ICMI '25, page 155–163, Canberra, Australia. Association for Computing Machinery. doi:10.1145/3716553.3750772.

Cristina Luna-Jiménez, David Griol, and Zoraida Callejas. 2024. Zero-shot ensemble of language models for fine-grain mental-health topic classification. In *Artificial Intelligence for Neuroscience and Emotional Systems: 10th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2024, Olhâo, Portugal, June 4–7, 2024, Proceedings, Part I*, page 88–97, Olhâo, Portugal. Springer-Verlag. doi:10.1007/978-3-031-61140-7_9.

Claudia Macht and Markus Steinbach. 2019. *33. Regionalsprachliche Merkmale in der Deutschen Gebärdensprache*, pages 914–935. De Gruyter Mouton, Berlin, Boston. URL: https://doi.org/10.1515/9783110261295-033 [cited 2026-01-26].

Ozge Mercanoglu Sincan and Richard Bowden. 2025. Spotter+GPT: Turning Sign Spottings into Sentences with LLMs. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, Berlin, Germany. Association for Computing Machinery. doi:10.1145/3742886.3756708.

Fabrizio Nunnari, Eleftherios Avramidis, Cristina España-Bonet, Marco González, Anna Hennes, and Patrick Gebhard. 2024. DGS-Fabeln-1: A Multi-Angle Parallel Corpus of Fairy Tales between German Sign Language and German Text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4847–4857, Torino, Italia. ELRA and ICCL. URL: https://aclanthology.org/2024.lrec-main.434/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. doi:10.3115/1073083.1073135.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A Survey of Multilingual Large Language Models. *Patterns*, 6(1):101118. doi:https://doi.org/10.1016/j.patter.2024.101118.

Oline Ranum, David Wessels, Gomèr Otterspeer, Erik J Bekkers, Floris Roelofsen, and Jari I. Andersen. 2024. The NGT200 dataset - geometric multi-view isolated sign recognition. In *ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling*. URL: https://openreview.net/forum?id=idkNzTC67X.

Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. 2023. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR): Workshops, pages 5625–5635, Vancouver, Canada. IEEE Computer Society. URL: https://doi.ieeecomputersociety.org/10.1109/CVPRW59228.2023.00596.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Long Beach, California, USA. Association for Computing Machinery. URL: https://dl.acm.org/doi/10.5555/3295222.3295349.