Dunder Mifflin at SPA-MSLG 2026: Sign Language Translation with No Language Left Behind (NLLB): A Mexican Sign Language Case Study

Ana Sofía Rentería-Palomares[1,†], Scarlette Berenice Alcaraz-Muñiz[1,†], Ever Essaú Rodríguez-Sandoval[1,*,†] and Jorge Omar Pérez-Villalvazo[1,†]

[1]*Tecnológico Nacional de Mexico / Instituto Tecnológico de Ciudad Guzmán, Jalisco, 49100*

Abstract

This paper details the participation of MSLG-SPA 2026: Bidirectional Translation between Mexican Sign Language Glosses and Spanish. The proposed approach adapts the distilled 600M-parameter NLLB-200 model to support Mexican Sign Language Glosses (MSLG) within a multilingual translation framework. Since MSLG is not originally included in NLLB, the tokenizer and vocabulary were extended by adding a new language token initialized using Spanish embeddings. Experiments were conducted using a low-resource parallel corpus composed of aligned Spanish sentences and MSLG glosses. Due to the unavailability of the official test sets during development, the original training subset was further divided using a 70/30 split for training and validation. The model was fine-tuned in both translation directions: Spanish-to-MSLG and MSLG-to-Spanish. Performance was evaluated using BLEU, METEOR, chrF, and COMET metrics. Results show moderate translation quality, mainly limited by the reduced dataset size. However, the model achieved stable and competitive performance while requiring fewer computational resources than larger multilingual transformer architectures. These findings demonstrate the potential of adapting multilingual neural machine translation models for low-resource sign language translation tasks involving MSLG.

Keywords

Transformer, Mexican Sign Language, NLLB-200, Natural Language Processing

1. Introduction

Worldwide, more than 1.5 billion people suffer from some level of hearing loss, of whom at least 430 million present moderate-to-profound loss [1, 2]. Furthermore, congenital hearing loss is the most common sensory deficit in newborns, with an incidence of 1.5 per 1,000 live births [3]. In Mexico, more than 7.3 million people face auditory or verbal communication barriers, of whom approximately 2.3 million have hearing disabilities [4, 5]. In this country, the Deaf community possesses its own linguistic and cultural identity; however, its minority status often leads to social exclusion within a predominantly hearing society [6, 7].

Mexican Sign Language (MSL) is legally recognized as part of the national linguistic heritage and consists of a set of gestural signs articulated with the hands, complemented by non-manual markers such as facial expressions and body posture. It also has its own syntax, grammar, and lexicon [5, 8]. This recognition is formally supported by the General Law for the Inclusion of Persons with Disabilities (2005), which also promotes bilingual education for the Deaf community [9].

Despite this legal framework, significant challenges remain regarding the dissemination of MSL and access to interpretation services. The country has only around 40 certified interpreters, with one quarter of them concentrated in Mexico City [7, 5]. When attempting to document or translate MSL, gloss notation is used — a written representation system that maps signs, movements, and expressions

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

anasofiarent@gmail.com (A. S. Rentería-Palomares); scarlettealcaraz15@gmail.com (S. B. Alcaraz-Muñiz); everrodriguez7@gmail.com (E. E. Rodríguez-Sandoval); omar.perez.vzo@gmail.com (J. O. Pérez-Villalvazo)

0009-0001-5902-4547 (A. S. Rentería-Palomares); 0009-0004-7409-0691 (S. B. Alcaraz-Muñiz); 0009-0001-9740-4479 (E. E. Rodríguez-Sandoval); 0009-0005-6540-211X (J. O. Pérez-Villalvazo)

onto words of the spoken language. However, the spatial and grammatical complexity of sign languages makes computational translation an enormous technological challenge [10].

MSLG-SPA 2026, organized as part of IberLEF 2026, is the first shared task dedicated to bidirectional translation between MSL glosses and Spanish. It addresses the critical lack of publicly available datasets, benchmarks, and computational resources for MSL by establishing a structured and reproducible evaluation framework for models capable of translating MSL gloss sequences into Spanish and Spanish text into MSL gloss sequences [11, 12].

This work presents an automatic translation system focused on converting Mexican Spanish text into Mexican Sign Language glosses (MSLG). To address this challenge, a deep learning approach is proposed through fine-tuning of the pre-trained NLLB-200 model. Specifically, the vocabulary of its tokenizer is expanded to treat MSL glosses as a new language, leveraging transfer learning from Spanish. This methodology aims not only to achieve competitive results on the proposed task, but also to establish a robust intermediate written representation that fosters digital inclusion and reduces the communication gap for the Deaf community in Mexico.

2. Related Work

The evolution of Sign Language Translation (SLT) systems—from rule-based pipelines to neural architectures—has been thoroughly documented in [10] and [13], where the canonical taxonomy of translation paradigms and benchmark architectural choices are established. At the rule-based end, Porta et al. [14] demonstrated both the linguistic precision and the brittleness of hand-crafted transfer rules for Spanish Sign Language (LSE), while Perea-Trigo et al. [15] showed that automatically generated synthetic corpora can effectively bootstrap deep learning models when parallel sign language data is scarce—a strategy directly applicable to (MSL).

The Transformer architecture introduced by Vaswani et al. [16] forms the backbone of most modern SLT systems, and its pretraining variants—BERT, proposed by Devlin et al. [17], and XLNet [18]—established the fine-tuning paradigm that enables strong performance under limited supervision. For low-resource translation specifically, the NLLB initiative [19] demonstrated that multilingual pretraining combined with data augmentation can substantially narrow the performance gap between high- and low-resource language pairs. De Genaro et al. [20] confirmed that fine-tuning NLLB-200 on small in-domain corpora outperforms training from scratch—a finding with direct implications for MSL gloss generation—while Zhang et al. [1] and Bhuvaneswari et al. [21] further support combining transfer learning with hybrid statistical-neural approaches in data-scarce settings.

On the visual side, Camgöz et al. [22, 23] established the neural SLT benchmark with an end-to-end CNN–attention framework and subsequently the Sign Language Transformers model, which jointly performs recognition and translation using CTC-guided glosses without treating them as a hard bottleneck. Chen et al. [24] extended this line by demonstrating that initializing sign language models with mBART multilingual weights yields state-of-the-art results across multiple benchmarks. For MSL specifically, González-Rodríguez et al. [25] and Caballero-Hernández et al. [4] addressed bidirectional MSL–Spanish translation using MediaPipe keypoint extraction with deep learning and IoT-integrated classifiers, respectively, while García-Gil et al. [26], Fernández-Rodríguez et al. [8], Fregoso et al. [27], and Ríos-Figueroa et al. [5] contributed recognition-focused work covering angular feature extraction, VGG16 transfer learning, PSO-optimized CNNs, and 3D invariant representations.

Two cross-cutting challenges remain open across the field. Müller et al. [28] demonstrated that BLEU correlates poorly with human judgments of SLT quality and that LLM-based evaluators provide more reliable assessments, motivating the use of complementary metrics in this work. In parallel, Gebru et al. [29] proposed the Datasheets for Datasets framework as a standard for corpus documentation, whose adoption in MSL data collection would facilitate reproducible comparisons and more generalizable models.

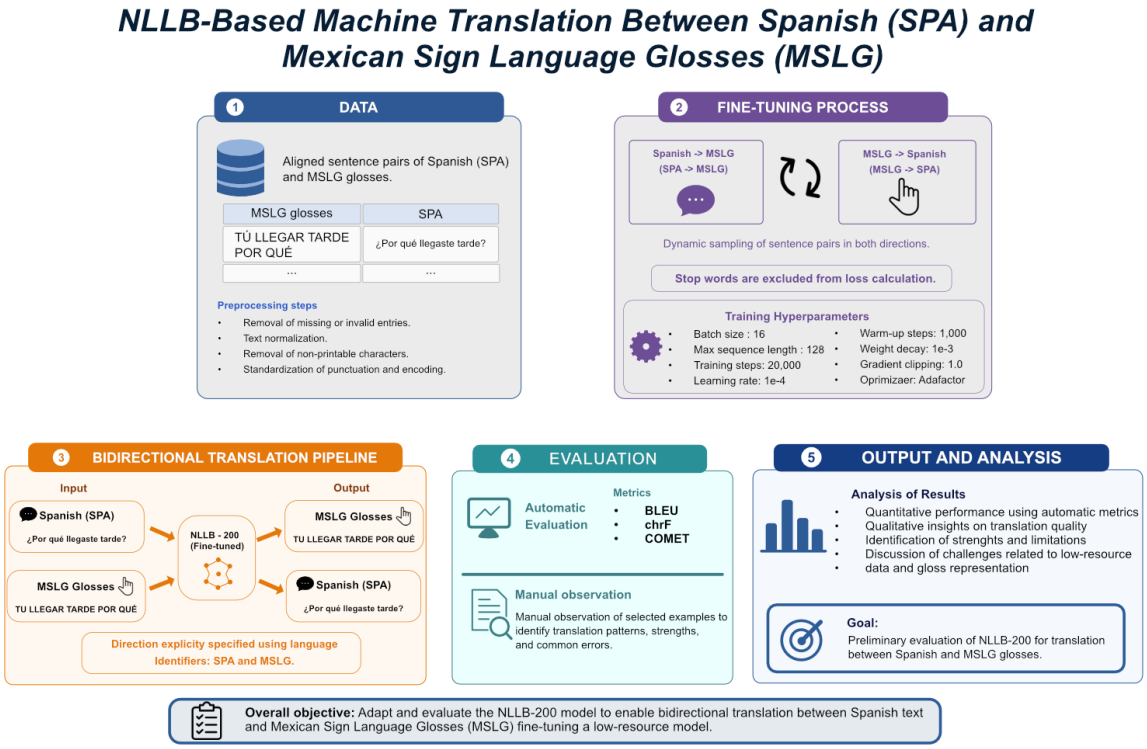

Figure 1: NLLB-Based Machine Translation between Spanish (SPA) and Mexican Sign Language Glosses (MSLG)

3. Methodology

This study follows an experimental and exploratory methodology aimed at evaluating the feasibility of adapting a neural machine translation model for translation between Spanish text and Mexican Sign Language Glosses (MSLG). For this purpose, the distilled version of the NLLB-200 model with 600 million parameters was selected as the base architecture. Rather than developing a production-ready translator, the objective of this work was to analyze the behavior of the model when fine-tuned on a specialized low-resource dataset and to evaluate its capability to generate approximate translations between both linguistic representations. Figure 1 summarize the overall process.

3.1. Data

The dataset provided for MSLG-SPA 2026 was designed to support reproducible and comparable research on bidirectional translation between MSLG and Spanish. The corpus consists of aligned sentence-level pairs of MSLG sequences and spanish text sentences.

Each pair represents semantically equivalent content in both modalities, enabling participation in the two complementary subtasks: Gloss-to-Spanish (MSLG2SPA) and Spanish-to-Gloss (SPA2MSLG).

The grammatical structure of MSL differs considerably from Spanish. Gloss sequences do not follow standard Spanish word order or grammatical inflection patterns, and grammatical markers are often omitted. Consequently, translation between MSL glosses and Spanish requires contextual inference and syntactic reordering rather than direct word-by-word mapping. In Table 1 shows the examples of these linguistic characteristics.

Some linguistic resources for MSLG include auxiliary annotations such as *pro-* markers for pronouns or grammatical information. However, for the MSLG-SPA 2026 shared task, these auxiliary annotations were removed, and only simplified gloss sequences were provided in order to ensure consistency across models and focus the evaluation on translation from incomplete gloss representations. These datasets

Table 1
Examples of translation between MSLG and SPA included in the dataset

MSLG	SPA
AMÉRICA-YO VIVIR	Vivo en América.
JUAN ASUNTO APARTE	Con Juan, es un asunto aparte.
TÚ LLEGAR TARDE POR QUÉ	¿Por qué llegaste tarde?

were converted into a CSV file format to facilitate its use, aditionally prior to training, a preprocessing stage was conducted to ensure data consistency and improve tokenization quality. This process included:

- Detection and removal of missing or invalid entries.
- Text normalization.
- Elimination of non-printable characters.
- Standardization of punctuation and encoding formats.

These preprocessing steps were particularly important because inconsistencies in gloss representations can negatively affect tokenization and model learning.

The dataset was originally divided into three disjoint subsets:

- **Training set** (50%, 489 pairs): Used for model development in both subtasks.
- **Test set MSLG2SPA** (25%, 244 pairs): Used exclusively for the final evaluation of the Gloss-to-Spanish subtask.
- **Test set SPA2MSLG** (25%, 244 pairs): Used exclusively for the final evaluation of the Spanish-to-Gloss subtask.

Since the official test sets were not available during the development stage, the original training subset containing 489 sentence pairs was further divided using a 70/30 split. Specifically, 70% of the pairs were used for training, while the remaining 30% were reserved for internal validation and testing purposes during experimentation and hyperparameter tuning. This strategy allowed model evaluation prior to the release of the official shared-task test sets while preventing data leakage between training and evaluation stages.

3.2. Fine-Tuning Process

Fine-tuning [30] in machine learning is the process of adapting a pre-trained model for specific tasks or use cases. It has become a fundamental deep learning technique, particularly in the training process of foundation models used for generative AI. In this particular task the model was fine-tuned in both translation directions:

- Spanish-to-MSLG.
- MSLG-to-Spanish.

During training, sentence pairs were dynamically sampled in both directions, allowing the model to learn bilingual correspondences simultaneously. Sentences were tokenized according to their respective source and target language identifiers.

Additionally, stop words were excluded from the loss calculation in order to reduce their influence during parameter optimization and prioritize semantically relevant tokens.

Table 2 shows the hyperparameters used on training.

The Adafactor optimizer was used during training due to its memory efficiency when working with transformer-based architectures. Table 3 summarizes the hardware and software configuration used to train the NLLB model.

During training, the process continuously monitored the training loss to evaluate convergence behavior. Additionally, the system periodically saved intermediate checkpoints of both the model and tokenizer to preserve training progress and enable future fine-tuning or recovery.

Table 2
Hyperparameters used in the training of the model

Hyperparameter	Value
Batch size	16
Maximum sequence length	128 tokens
Training steps	20,000
Learning rate	1×10^{-4}
Warm-up steps:	1,000
Weight decay	1×10^{-3}
Gradient clipping:	1.0
Optimizer	Adafactor

Table 3
Training environment configuration for the NLLB model

Component	**Configuration**
Processor	AMD Ryzen 7 5700X 8-Core Processor
GPU	NVIDIA GeForce RTX 4080
RAM	32 GB
Operating system	Ubuntu 24.04.4 LTS
Python Version	3.11.15
PyTorch	2.10.0
Transformers	4.32
Pandas	3.0.1

3.3. Bidirectional Translation Pipeline

A bidirectional translation pipeline was implemented to test the translation between Spanish and MSLG in both directions.

The translation process was based on the fine-tuned NLLB-200 model, where the translation direction was explicitly specified through language identifiers. The SPA token was used for Spanish text, while the newly introduced MSLG token represented Mexican Sign Language glosses within the multilingual architecture.

For the MSLG-to-SPA direction, gloss sequences were provided as input to the model, which generated corresponding Spanish sentences. This task required the system to reconstruct grammatical structures, word ordering, and linguistic information that are often omitted in gloss representations.

Conversely, in the SPA-to-MSLG direction, Spanish sentences were transformed into simplified gloss sequences. In this case, the model learned to reduce grammatical complexity and generate gloss representations aligned with the structure commonly used in Mexican Sign Language annotation.

During inference, the input sentences were tokenized according to the specified source language, and the model generated the corresponding translation conditioned on the target language token. Generated outputs from both directions were stored and later evaluated using automatic metrics.

This bidirectional framework enabled the model to learn semantic correspondences between Spanish and MSLG simultaneously, improving its adaptability across both translation tasks while maintaining a unified multilingual architecture. Additionally, a manual observation was conducted on selected examples to evaluate translation coherence, identify recurring patterns, and assess the strengths and limitations of the system. The observation revealed instances in which the generated translations closely matched the reference sentences, as well as cases where the outputs were incomplete, grammatically inconsistent, or semantically inaccurate.

3.4. Evaluation

The outputs from the proposed models were evaluated automatically using standard Machine Translation (MT) metrics selected according to the characteristics of each translation direction. Evaluation and ranking were performed independently for each subtask.

The MSLG2SPA subtask was evaluated using BLEU, METEOR, chrF, and COMET. Since this task generates well-formed Spanish sentences, both lexical-overlap and semantic-based metrics were employed to assess translation adequacy, fluency, and overall quality.

In contrast, the SPA2MSLG subtask was evaluated using BLEU, METEOR, and chrF only. COMET was not applied in this subtask because MSLG gloss sequences do not constitute fully natural language sentences and therefore are not suitable for fluency-oriented semantic evaluation.

BLEU (Bilingual Evaluation Understudy) [31] measures the similarity between machine-generated translations and one or more reference translations using modified n-gram precision combined with a brevity penalty. Higher BLEU scores indicate greater overlap with the reference text. The BLEU formulation is shown in Equation 1.

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{1}$$

where BP denotes the brevity penalty, p_n represents the modified n-gram precision, and w_n corresponds to the weight assigned to each n-gram order.

METEOR [32] evaluates translations by aligning generated and reference sentences using exact matches, stemming, synonymy, and paraphrase matching. Unlike BLEU, METEOR incorporates both precision and recall, making it more sensitive to semantic similarity and lexical variation. The METEOR score is defined in Equation 2.

$$F_{mean} = \frac{10PR}{R + 9P}, \tag{2}$$

where P represents unigram precision and R represents unigram recall.

chrF [33] is a character n-gram F-score metric designed to capture morphological variations and partial lexical matches more effectively than word-based metrics. This metric is especially useful for morphologically rich languages and noisy text generation tasks. The chrF formulation is presented in Equation 3.

$$chrF_\beta = (1 + \beta^2)\frac{chrP \cdot chrR}{\beta^2 chrP + chrR}, \tag{3}$$

where $chrP$ and $chrR$ correspond to character n-gram precision and recall, respectively. The parameter β controls the relative importance of recall over precision. When $\beta = 1$, both contribute equally to the final score.

COMET (Crosslingual Optimized Metric for Evaluation of Translation) [34] is a neural evaluation metric based on pretrained multilingual language models. Unlike traditional lexical-overlap metrics, it estimates translation quality by considering semantic relationships between the source sentence, the generated hypothesis, and the reference translation. The scoring function is defined in Equation 4.

$$\text{COMET}(x, y, r) = f_\theta(x, y, r), \tag{4}$$

where x represents the source sentence, y corresponds to the generated translation, r denotes the reference translation, and f_θ represents a neural regression model trained to predict translation quality.

Since this subtask produces fluent Spanish sentences, COMET complements traditional lexical-overlap metrics by capturing semantic adequacy and fluency more effectively.

Global Score Computation for system ranking, metric scores were standardized using z-score normalization across all submitted models. A subtask-specific Global Score was then computed as the arithmetic mean of the standardized metric values, as shown in Equation 5.

Table 4
Evaluation results for the MSLG2SPA subtask

Task Score	BLEU	BLEU z-score	METEOR	METEOR z-score	chrF	chrF z-score	COMET
-0.2294	21.6290	-0.3014	0.4110	-0.2678	50.5901	-0.2073	0.7250

$$z = \frac{x - \mu}{\sigma}, \tag{5}$$

where x represents the system score, μ is the mean score across all participating models, and σ is the corresponding standard deviation. Final rankings were determined in descending order according to the resulting Global Score.

As in the MSLG2SPA subtask, metric scores were standardized using z-score normalization, and a subtask-specific Global Score was computed as the arithmetic mean of the normalized BLEU, METEOR, and chrF scores. This methodology preserves meaningful performance differences between models while avoiding distortions introduced by naïve rank aggregation techniques.

Overall, the proposed evaluation methodology enabled a preliminary assessment of the NLLB-200 model for translation between Spanish and MSLG glosses while highlighting the challenges associated with low-resource model and the textual representation of sign language glosses.

4. Analysis Results

The proposed models were evaluated independently under the official evaluation protocol of the shared task. Rankings were reported at the team level, considering the best-performing run submitted for each subtask.

Evaluation was performed using BLEU, METEOR, chrF, and COMET (only for MSLG2SPA). Final rankings were computed using the official normalized Task Score provided by the shared task organizers.

Although the proposed NLLB-based approach achieved competitive results, performance was constrained by the limited size of the available dataset. Neural machine translation models such as NLLB generally benefit from large-scale parallel corpora, and the reduced number of training pairs limited the model's ability to learn robust semantic and syntactic patterns between Spanish and MSL glosses. This limitation particularly affected generalization capabilities in low-frequency expressions and complex grammatical structures.

Despite these limitations, the proposed approach presented an important practical advantage in terms of computational efficiency. Compared with larger transformer-based architectures and more resource-intensive multilingual models, the selected NLLB configuration required lower computational resources during both training and inference. This allowed faster experimentation, reduced memory consumption, and more accessible deployment conditions while still maintaining stable translation quality across both subtasks.

4.1. MSLG2SPA Results

Table 4 presents the results obtained for the Gloss-to-Spanish translation task.

The MSLG2SPA subtask obtained moderate performance across lexical and semantic evaluation metrics. The BLEU and METEOR scores indicate that the system was capable of generating partially accurate Spanish translations from gloss sequences, while the COMET score suggests acceptable semantic adequacy despite the structural differences between both languages.

However, negative z-score values across all metrics indicate that the system remained slightly below the average performance of participating models. This behavior can be attributed primarily to the limited amount of training data, which restricted the model's capacity to learn complex reordering patterns and contextual information required for fluent Spanish generation.

Table 5
Evaluation results for the SPA2MSLG subtask

Task Score	BLEU	BLEU z-score	METEOR	METEOR z-score	chrF	chrF z-score
0.0405	14.4749	0.3140	0.3904	-0.0898	50.7612	-0.1027

4.2. SPA2MSLG Results

Table 5 summarizes the results obtained for the Spanish-to-Gloss translation task.

For the SPA2MSLG subtask, the proposed model achieved a slightly positive Task Score, indicating competitive behavior relative to other participating models. The positive BLEU z-score suggests that the model performed adequately in generating gloss sequences with lexical overlap close to the reference glosses.

The translation from Spanish into gloss representations proved comparatively more stable than the inverse direction, likely because gloss sequences contain simplified grammatical structures and reduced linguistic variability. Nevertheless, performance remained limited by the reduced dataset size, which constrained the diversity of learned gloss patterns and affected the generation of more complex sign structures.

Overall, the experimental results demonstrate that lightweight NLLB-based approaches can provide a reasonable balance between translation quality and computational efficiency in low-resource sign language translation scenarios. While larger datasets would likely improve performance substantially, the proposed system shows that acceptable results can still be achieved without requiring highly demanding computational infrastructures.

Conclusions

This work presented an exploratory approach for bidirectional translation between Spanish and Mexican Sign Language Glosses using a fine-tuned NLLB-200 model. The results show that adapting a multilingual neural machine translation model to a low-resource sign language scenario is feasible, even when the target representation is not originally included in the model vocabulary.

The proposed system achieved moderate and competitive performance in both translation directions. In particular, the Spanish-to-MSLG direction showed slightly more stable behavior, which may be related to the simplified structure of gloss sequences compared with natural Spanish sentences. However, the Gloss-to-Spanish direction remained more challenging, as it requires the model to reconstruct grammatical structure, context, and fluency from incomplete gloss representations.

One of the main limitations of this work was the reduced size of the available dataset. Since neural translation models usually require large parallel corpora to learn robust linguistic patterns, the limited number of training pairs affected the model's ability to generalize to more complex expressions and less frequent structures. Despite this limitation, the model maintained acceptable translation quality while requiring fewer computational resources than larger transformer-based architectures.

Overall, the findings suggest that fine-tuning multilingual models such as NLLB-200 can be a promising strategy for low-resource sign language translation tasks. Future work should focus on expanding the MSLG–Spanish corpus, improving gloss normalization, incorporating more linguistic information, and exploring additional training strategies to enhance translation accuracy and semantic consistency.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) through the graduate scholarship program. This support has been fundamental for the continuation of their master's studies and for the development of the research activities presented in this work.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Z.-q. Zhang, J.-y. Li, S.-t. Ge, T.-y. Ma, F.-y. Li, J.-l. Lu, S.-r. Si, Z.-z. Cui, Y.-l. Jin, X.-h. Jin, Bidirectional associations between sensorineural hearing loss and depression and anxiety: a meta-analysis, Frontiers in Public Health 11 (2024) 1281689. URL: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1281689/full. doi:10.3389/fpubh.2023.1281689.

[2] World Health Organization, World Report on Hearing, Technical Report, World Health Organization, Geneva, 2021. URL: https://www.who.int/publications/i/item/world-report-on-hearing.

[3] G. Choe, S.-K. Park, B. J. Kim, Hearing loss in neonates and infants, Clinical and Experimental Pediatrics 66 (2023) 369. doi:10.3345/cep.2022.01011.

[4] H. Caballero-Hernandez, V. Muñoz-Jiménez, M. A. Ramos-Corchado, Translation of mexican sign language into spanish using machine learning and internet of things devices, Bulletin of Electrical Engineering and Informatics 13 (2024) 3369–3379. doi:10.11591/eei.v13i5.7853.

[5] H. V. Rios-Figueroa, A. J. Sánchez-García, C. O. Sosa-Jiménez, A. L. Solís-González-Cosío, Use of spherical and cartesian features for learning and recognition of the static mexican sign language alphabet, Mathematics 10 (2022) 2904. doi:10.3390/math10162904.

[6] W. M. Olvera, I. M. Vite, D. G. Miy, Inclusión de las personas sordas: El caso de veracruz, Inclusión y horizontalidad en la lectoescritura (2024) 307.

[7] B. Estrada Aranda, S. Romero Contreras, C. Delgado Álvarez, N. T. Kaufmann, Percepción de la violencia de género de las mujeres sordas mexicanas, Iztapalapa. Revista de ciencias sociales y humanidades 43 (2022) 189–220. doi:10.28928/ri/922022/aot2/estradaarandab/romerocontrerass.

[8] R. F. Rodríguez, F. J. P. Rosas, L. Á. Zuñiga-Madrid, P. Arguijo, Reconocimiento de las señas estáticas del lsm con características basadas en aprendizaje profundo., Research in Computing Science 150 (2021) 303–311.

[9] I. Moreno Vite, F. d. M. E. Gil-Bernal, A. M. Rivera Guerrero, C. E. Escobedo Delgado, The Status of Mexican Sign Language in Mexican Policy: A Case Report, Global Journal of Intellectual & Developmental Disabilities 13 (2024) 555875. doi:10.19080/GJIDD.2024.13.555875.

[10] A. Núñez-Marcos, O. Perez-de Viñaspre, G. Labaka, A survey on sign language machine translation, Expert Systems with Applications 213 (2023) 118993. doi:10.1016/j.eswa.2022.118993.

[11] A. Y. Rodríguez-González, D. Fajardo-Delgado, M. Á. Álvarez-Carmona, M. G. Sánchez-Cervantes, Á. Díaz-Pacheco, R. Aranda, S. Fajardo-Flores, C. H. Martínez-Rodríguez, Overview of MSLG-SPA at IberLEF 2026: Bidirectional Translation between Mexican Sign Language Glosses and Spanish, Procesamiento del Lenguaje Natural 77 (2026). URL: https://sites.google.com/cicese.edu.mx/mslg-spa2026/task-definition.

[12] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.

[13] N. Shahin, L. Ismail, From rule-based models to deep learning transformers architectures for natural language processing and sign language translation systems: Survey, taxonomy and performance evaluation, Applied Sciences 14 (2024) 320. doi:10.1007/s10462-024-10895-z.

[14] J. Porta, F. López-Colino, J. Tejedor, J. Colás, A rule-based translation from written Spanish to Spanish Sign Language glosses, Speech Communication 57 (2014) 312–324. doi:10.1016/j.csl.2013.10.003.

[15] M. Perea-Trigo, C. Botella-López, M. Á. Martínez-del Amor, J. A. Álvarez-García, L. M. Soria-Morillo, J. J. Vegas-Olmos, Synthetic corpus generation for deep learning-based translation of Spanish Sign Language, IEEE Access 12 (2024) 38112–38127. doi:10.3390/s24051472.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, volume 30, 2017, pp. 5998–6008. doi:10.48550/arXiv.1706.03762.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT 2019, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[18] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, XLNet: Generalized autoregressive pretraining for language understanding, in: Advances in Neural Information Processing Systems, volume 32, 2019, pp. 5753–5763. doi:10.48550/arXiv.1906.08237.

[19] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, L. Zettlemoyer, et al., No language left behind: Scaling human-centered machine translation, Technical Report, Meta AI, 2022. doi:10.48550/arXiv.2207.04672, arXiv:2207.04672.

[20] D. DeGenaro, T. Lupicki, Experiments in Mamba sequence modeling and NLLB-200 fine-tuning for low-resource machine translation, in: Proceedings of the AmericasNLP 2024 Workshop, 2024, pp. 92–100. doi:10.18653/v1/2024.americasnlp-1.22.

[21] K. Bhuvaneswari, M. Varalakshmi, Efficient incremental training using a novel NMT-SMT hybrid framework for low-resource neural machine translation, Expert Systems with Applications 238 (2024) 121868. doi:10.3389/frai.2024.1381290.

[22] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden, Neural sign language translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7784–7793. doi:10.1109/CVPR.2018.00812.

[23] N. C. Camgoz, O. Koller, S. Hadfield, R. Bowden, 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), 2020, pp. 10023–10033. doi:10.1109/CVPR42600.2020.01004.

[24] Y. Chen, F. Wei, X. Sun, Z. Wu, S. Lin, A simple multi-modality transfer learning baseline for sign language translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5120–5130. doi:10.1109/CVPR52688.2022.00506.

[25] J. R. González-Rodríguez, D. M. Córdova-Esparza, J. Terven, J. A. Romero-González, Towards a bidirectional Mexican Sign Language–Spanish translation system: A deep learning approach, Applied Sciences 14 (2024) 1350. doi:10.3390/technologies12010007.

[26] G. García-Gil, G. d. C. López-Armas, G. Ochoa-Ruiz, M. Gonzalez-Mendoza, H. J. Escalante, L. C. Solís-Reyes, Real-time machine learning for accurate Mexican Sign Language identification, Applied Sciences 14 (2024) 3432. doi:10.3390/technologies12090152.

[27] J. Fregoso, C. I. Gonzalez, G. E. Martinez, Optimization of convolutional neural networks architectures using PSO for sign language recognition, Axioms 10 (2021) 139. doi:10.3390/axioms10030139.

[28] M. Müller, Z. Jiang, C. Farinea, J. Ive, Considerations for meaningful sign language machine translation evaluation, in: Proceedings of the 1st International Workshop on Multimodal and Multilingual NLP (MM-NLP), 2023, pp. 1–12. doi:10.18653/v1/2023.acl-short.60.

[29] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, K. Crawford, Datasheets for datasets, Communications of the ACM 64 (2021) 86–92. doi:10.1145/3458723.

[30] D. Bergmann, What is Fine-Tuning? | IBM, 2024. URL: https://www.ibm.com/think/topics/fine-tuning.

[31] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.

[32] S. Banerjee, A. Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

[33] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, P. Pecina (Eds.), Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049/. doi:10.18653/v1/W15-3049.

[34] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A Neural Framework for MT Evaluation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2685–2702. URL: https://aclanthology.org/2020.emnlp-main.213/. doi:10.18653/v1/2020.emnlp-main.213.