Multilingual Transformer Models with LoRA and Reranking for Mexican Sign Language Gloss Translation

Jean Michel Arreola Trapala[1,*], Ester Aguayo Moreno[2] and Angelica Verenice Villavicencio Soriano[3]

[1]Center for Research in Mathematics, México

Abstract

This paper describes the SignNexus submission to the MSLG-SPA 2026 shared task on translation between Mexican Sign Language glosses (MSLG) and Spanish. We explore multiple multilingual sequence to sequence architectures including mBART, NLLB, MT5, and ByT5, combined with parameter efficient fine tuning techniques such as LoRA and custom gloss adapters. We further investigate reranking strategies, constrained decoding, ensemble selection, backtranslation, and paraphrase based data augmentation. Our best validation systems achieved 49.29 chrF for MSLG-to-Spanish translation and 40.64 chrF for Spanish-to-MSLG translation, showing that multilingual pretrained models with parameter-efficient adaptation are effective for this low-resource gloss translation task.

Keywords

Mexican Sign Language, Sign Language Translation, Neural Machine Translation, Multilingual Transformers, LoRA, Low-Resource NLP

1. Introduction

Sign languages constitute a fundamental communication medium for deaf communities around the world. In Mexico, Mexican Sign Language (Lengua de Señas Mexicana, LSM) is widely used by the deaf population and represents an essential component of cultural identity and accessibility. Despite recent advances in Natural Language Processing, automatic translation systems involving sign languages remain significantly underexplored, especially for low resource settings such as LSM.

One common strategy for computational sign language translation consists of representing signs through glosses, which are textual annotations that approximate the meaning or structure of signs using written tokens. Gloss based translation tasks simplify the problem by transforming sign language sequences into textual representations that can be processed using standard sequence to sequence neural architectures. However, glosses still present substantial challenges due to their non standardized syntax, sparse vocabularies, limited training data, and structural differences from spoken languages.

The MSLG-SPA 2026 shared task [1], organized as part of the IberLEF 2026 evaluation campaign [2], focuses on automatic translation between Mexican Sign Language glosses (MSLG) and Spanish in two directions: (1) gloss-to-Spanish translation, and (2) Spanish-to-gloss translation. These tasks are particularly challenging because the available parallel corpora are relatively small, noisy, and domain diverse, making the problem highly sensitive to data sparsity and generation errors.

To address these challenges, we developed a multilingual neural translation system based on transformer encoder-decoder architectures combined with parameter efficient fine tuning techniques. Our approach explores several multilingual pretrained models, including mBART, NLLB-200, MT5, ByT5 and FLAN T5, adapted to the MSLG-SPA task using Low Rank Adaptation (LoRA), custom gloss adapters, constrained decoding strategies, and heuristic reranking methods. In addition, we investigated data augmentation techniques such as backtranslation and paraphrase generation in order to mitigate the low resource nature of the dataset.

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

jean.arreola@cimat.mx (J. M. A. Trapala); ester.aguayo@cimat.mx (E. A. Moreno); angelica.villavicencio@cimat.mx (A. V. V. Soriano)

0000-0002-7626-7428 (J. M. A. Trapala); 0000-0002-2751-5595 (E. A. Moreno); 0009-0009-3684-230X (A. V. V. Soriano)

Our experiments show that multilingual pretrained translation models can effectively adapt to the MSLG-SPA task when combined with lightweight adaptation methods and decoding heuristics. In particular, NLLB achieved the strongest performance for gloss to Spanish translation, while mBART obtained the best results for Spanish to gloss generation. We also evaluated several ensemble and reranking approaches, including oracle analysis and heuristic candidate selection, to study complementarities between models.

The remainder of this paper is organized as follows. Section 2 describes the dataset and preprocessing pipeline. Section 3 presents the proposed modelling approaches, including model architectures, adaptation strategies, and decoding techniques. Section 4 discusses the experimental results and qualitative analysis. Finally, Section 5 summarizes the main conclusions and future directions for sign language translation research.

2. Dataset

The dataset used in this work was provided as part of the MSLG-SPA 2026 shared task and consists of parallel sentence pairs between Mexican Sign Language glosses (MSLG) and Spanish. The corpus is designed for bidirectional translation and supports two subtasks: gloss to Spanish translation and Spanish to gloss translation.

The training corpus contains aligned pairs of MSLG gloss sequences and their corresponding Spanish translations. Each example includes an identifier (`ID`), a gloss sentence (`MSLG`), and its Spanish counterpart (`SPA`). The gloss notation uses uppercase lexical tokens and special markers such as `dm-`, `#`, `+`, and hyphenated compounds to represent sign language structures and named entities. Examples of the dataset are shown in Table 1.

Table 1
Examples from the MSLG-SPA dataset.

MSLG Gloss	Spanish Translation
APAGAR LUZ POR-FAVOR	Por favor, apaga la luz.
YO GUSTAR PAPA	Me gustan las papas.
MI PAPÁ TRABAJAR CONSTRUCCIÓN	Mi papá trabaja en la construcción.
MÉXICO BIBLIOTECA GRANDE	La Biblioteca de México es muy grande.

The corpus contains conversational expressions, cultural references, named entities, geographical locations, colloquial language, and domain diverse vocabulary. Several examples include linguistic phenomena that are challenging for machine translation systems, such as omitted function words, compressed semantic structures, and gloss specific grammatical patterns. In addition, the glosses are not fully standardized, which introduces lexical and structural variability across examples.

The train dataset was divided into training and validation partitions. The training set was used for model fine tuning and adaptation, while the validation set was employed for hyperparameter optimization and model selection. The test sets provided for each task contain only source side inputs without references and were used exclusively for final prediction generation.

The MSLG notation preserves several important structural conventions:

- `dm-`: named entities or fingerspelled names.
- `#`: lexicalized borrowed forms.
- `+`: compound signs or linked concepts.
- Hyphenated expressions such as `CIUDAD-DE-MÉXICO`.

To preserve these structures, the preprocessing pipeline avoided aggressive normalization and maintained the original gloss formatting whenever possible. Minimal preprocessing operations included whitespace normalization and controlled handling of punctuation. Additional postprocessing rules were applied after decoding in order to restore special gloss markers consistently.

The dataset is relatively small compared to standard neural machine translation corpora, making the task particularly challenging for large transformer architectures. To mitigate data sparsity, several augmentation techniques were explored, including synthetic backtranslation pairs and rule based Spanish paraphrases. These synthetic examples were incorporated selectively during training after applying filtering heuristics designed to remove noisy or invalid gloss generations.

For evaluation purposes, the shared task employed standard machine translation metrics including BLEU, chrF, and METEOR. Since gloss generation does not always follow strict grammatical conventions, character based metrics such as chrF proved particularly informative for evaluating translation quality in the SPA to MSLG task.

3. Modelling Approach

Our approach is based on multilingual transformer encoder-decoder architectures [3] adapted to the MSLG-SPA translation task through parameter efficient fine tuning and decoding optimization techniques. Since the dataset is relatively small and highly specialized, we focused on methods capable of leveraging multilingual pretrained knowledge while minimizing the number of trainable parameters.

3.1. Multilingual Transformer Models

We experimented with several multilingual sequence to sequence architectures:

- **mBART**[4]: a multilingual denoising autoencoder designed for machine translation.
- **NLLB-200 Distilled 600M**[5]: Meta's multilingual translation model specialized for low resource languages.
- **MT5** [6]: a multilingual text to text transformer model.
- **ByT5** [7]: a byte level multilingual transformer robust to noisy and uncommon tokens.
- **FLAN-T5** [8]: an instruction tuned multilingual transformer model explored to evaluate whether instruction tuning could improve generalization in low resource gloss translation tasks.

All models were trained using a sequence to sequence formulation in which a source sentence is encoded and decoded autoregressively into the target language or gloss representation.

3.2. Directional Prefixing

To support bidirectional translation using the same architecture, we introduced task specific prefixes at the beginning of the input sequence:

- `<2spa>` for gloss to Spanish translation.
- `<2mslg>` for Spanish to gloss translation.

For example:

- Input: `<2spa> APAGAR LUZ POR-FAVOR`
- Output: *Por favor, apaga la luz.*

This strategy allows the model to learn translation direction jointly while preserving a unified architecture.

3.3. LoRA Fine Tuning

To reduce memory usage and computational cost, we employed Low Rank Adaptation (LoRA) [9] for parameter efficient fine tuning. Instead of updating all transformer parameters, LoRA injects trainable low rank matrices into attention projection layers while keeping the pretrained weights frozen.

The main LoRA configuration used in our experiments was:

- Rank (r) = 16
- Alpha = 32
- Dropout = 0.1

For most transformer architectures, LoRA was applied to the query and value projection layers of the attention mechanism.

3.4. Training Configuration

All models were fine-tuned using the official training split. The experiments employed the AdamW optimizer with a learning rate of 3×10^{-4} and early stopping with patience 2.

The main training configuration was:

- Learning rate: 3×10^{-4}
- Epochs: 10
- LoRA rank (r): 16
- LoRA alpha: 32
- LoRA dropout: 0.1
- Beam size: 8
- Number of generated candidates: 6
- Batch size: 1

All experiments were conducted using the same validation split for model selection and hyperparameter tuning.

Table 2
Main training hyperparameters.

Parameter	Value
Learning Rate	3×10^{-4}
Epochs	10
LoRA Rank	16
LoRA Alpha	32
LoRA Dropout	0.1
Beam Size	8
Candidates	6
Batch Size	1

3.5. Gloss Adapter

In addition to LoRA, we experimented with a custom gloss adapter designed specifically for MSLG representations. The adapter introduces an additional transformation layer over the token embeddings, allowing the model to specialize its internal representations for gloss processing.

The adapter consists of:

- A two layer feed forward projection.

- A gating mechanism controlling the contribution of the adapted representation.
- Selective activation depending on translation direction.

This mechanism was intended to improve the handling of gloss specific structures without fully modifying the pretrained multilingual encoder-decoder model.

3.6. Decoding Strategy

During inference, we employed beam search decoding with multiple candidate generation. Typical configurations used:

- Beam size = 8
- Number of returned hypotheses = 6
- Length penalty = 0.9
- No repeated bigrams

Sampling based generation methods were also explored during preliminary experiments, but deterministic beam search consistently produced more stable and higher quality outputs.

3.7. Candidate Reranking

To improve final prediction quality, multiple generated hypotheses were reranked using heuristic scoring functions. Different heuristics were designed for each translation direction.

For MSLG-SPA, the reranker considered:

- Output length consistency.
- Repetition penalties.
- Detection of malformed or nonsensical tokens.
- Presence of proper punctuation.

For SPA-MSLG, the reranker additionally penalized:

- Lowercase tokens.
- Spanish punctuation.
- Tokens outside the expected gloss vocabulary.
- Mixed language artifacts.

The final score combined normalized model confidence scores with heuristic penalties and bonuses.

3.8. Constrained Decoding

For the Spanish to gloss task, we also explored constrained decoding using a custom logits processor. A vocabulary extracted from the training glosses was used to softly penalize unlikely tokens during generation. This strategy attempted to reduce malformed outputs and encourage valid gloss structures without imposing hard decoding restrictions.

3.9. Data Augmentation

Because the dataset is relatively small, several augmentation strategies were explored:

- **Backtranslation**: synthetic glosses were generated from Spanish sentences using previously trained models.
- **Synthetic filtering**: noisy synthetic examples were removed using heuristic quality filters.
- **Rule based paraphrases**: additional Spanish paraphrases were generated through lightweight lexical transformations.

Although some augmentation strategies produced limited improvements, they provided useful insights into the challenges of low resource sign language translation.

3.10. Ensemble Experiments

We also explored several ensemble strategies combining outputs from multiple models, particularly NLLB and mBART. Oracle analysis showed strong complementarity between systems, motivating heuristic ensemble selection methods based on sentence level scoring and consensus reranking. While the oracle ensemble significantly improved upper bound performance, practical ensemble gains were more modest due to prediction instability and error propagation.

4. Results and Discussion

The proposed systems were evaluated on both translation directions of the shared task: Mexican Sign Language glosses to Spanish (MSLG-SPA) and Spanish to Mexican Sign Language glosses (SPA-MSLG). Evaluation was performed using the official validation split and the standard machine translation metrics BLEU [10], METEOR [11], and chrF [12].

4.1. MSLG to Spanish Translation

Table 3 presents the main results obtained for gloss to Spanish translation.

Table 3
Validation results for MSLG-to-Spanish translation.

Model	BLEU	METEOR	chrF
mBART + LoRA + Reranking	12.49	35.19	44.45
NLLB + LoRA + Reranking	13.75	32.71	47.59
Heuristic Ensemble	14.85	34.60	49.29
flan-t5-small	5.6	23	33.79
mt5-small	5.84	22.2	28

Among individual models, NLLB achieved the best overall performance in terms of chrF and BLEU, suggesting stronger lexical and structural alignment with Spanish outputs. In contrast, mBART obtained higher METEOR scores, indicating that it often generated semantically closer paraphrases even when lexical overlap was lower.

The ensemble approach combining NLLB and mBART produced the best validation performance overall, reaching approximately 49.3 chrF. Oracle experiments further showed that the two models were highly complementary, with an oracle upper bound exceeding 53 chrF. This indicates that different architectures captured distinct translation patterns and errors.

Qualitative inspection revealed that the models were generally capable of generating fluent Spanish sentences from compressed gloss structures. However, several common error types remained:

- Incorrect semantic substitutions (e.g., *busto - beso*).
- Hallucinated entities or locations.
- Misinterpretation of gloss compounds.
- Repetition and malformed syntax in long sequences.

Despite these issues, the multilingual pretrained models demonstrated strong adaptation capabilities for low resource gloss translation.

4.2. Spanish to MSLG Translation

Table 4 presents the results for the Spanish to gloss task.

The Spanish to gloss task proved considerably more challenging than the reverse direction. Unlike natural language generation, gloss generation requires producing compact and highly structured symbolic sequences that differ substantially from spoken Spanish syntax.

Table 4
Validation results for Spanish-to-MSLG gloss translation.

Model	BLEU	METEOR	chrF
mBART + LoRA + Reranking	11.06	33.07	40.64
NLLB + LoRA + Reranking	0.19	17.52	1.12
ByT5 + LoRA	28	12.70	17.33
MT5 + Adapter	0.09	3.53	0.30
mt5-small	11.84	0.29	38.7
flan-t5-small	9.8	0.284	40.61

Among all evaluated systems, mBART achieved the best performance by a large margin. The model was able to preserve many gloss specific structures and generated outputs that more closely resembled the training distribution. In contrast, NLLB struggled significantly in this direction, frequently producing malformed glosses or outputs that resembled natural Spanish rather than gloss notation.

ByT5 demonstrated moderate robustness to unusual gloss tokens due to its byte level representation, but overall translation quality remained substantially below mBART. Similarly, MT5 with the gloss adapter showed unstable behavior and often generated noisy or semantically inconsistent outputs.

4.3. Effect of Reranking

Reranking strategies played an important role in improving output consistency. Instead of selecting only the top beam search hypothesis, the system generated multiple candidates and applied heuristic scoring functions to select the most plausible output.

For MSLG-SPA, reranking improved fluency and reduced malformed outputs by favoring:

- reasonable sentence lengths,
- lower repetition rates,
- proper punctuation,
- and semantically coherent hypotheses.

For SPA-MSLG, reranking attempted to discourage outputs containing lowercase spanish words, punctuation, and out of vocabulary tokens. While reranking improved some cases, it occasionally introduced instability when model confidence scores favored malformed hypotheses.

4.4. Data Augmentation and Backtranslation

Several augmentation strategies were explored to compensate for the limited size of the parallel corpus. Synthetic glosses generated through backtranslation were filtered using vocabulary and formatting heuristics before being merged into the training data. In addition, lightweight Spanish paraphrases were generated using rule based transformations.

Although these methods provided moderate diversity improvements, their overall impact on final validation metrics was limited. In some cases, noisy synthetic examples degraded generation quality, particularly in the gloss generation task.

4.5. Discussion

Overall, the experiments demonstrate that multilingual pretrained translation models can be adapted effectively to low resource sign language gloss translation tasks using lightweight fine tuning techniques. The strongest systems combined multilingual pretrained knowledge, LoRA adaptation, beam search decoding, and heuristic reranking.

Several important observations emerged during experimentation:

- Deterministic beam search consistently outperformed sampling based generation.
- Few shot prompting strategies degraded translation quality for multilingual NMT models.
- LoRA provided strong performance improvements while keeping training efficient.
- Gloss generation remains substantially harder than natural language generation.
- Ensemble methods revealed strong complementarities between multilingual architectures.

The results also highlight the challenges of translating gloss based sign language representations, particularly under low resource conditions. Future improvements will likely require larger gloss corpora, stronger structural constraints, and more robust reranking methods capable of preserving gloss syntax and semantics simultaneously.

4.6. Limitations

Several limitations should be acknowledged. First, the available parallel corpus is relatively small compared to standard neural machine translation benchmarks, which restricts the ability of large multilingual models to fully adapt to the characteristics of Mexican Sign Language glosses. This limitation is particularly evident in the Spanish-to-MSLG direction, where models frequently produced malformed glosses or mixed-language outputs.

Second, gloss representations are not fully standardized and may exhibit lexical and structural variation across examples. As a result, multiple valid gloss realizations may exist for the same Spanish sentence, while automatic evaluation metrics compare system outputs against only a single reference translation.

Third, although heuristic reranking improved output quality in many cases, the proposed scoring functions rely on manually designed rules and may not generalize well to unseen domains or alternative gloss conventions. More sophisticated reranking methods based on learned quality estimation or large language models could potentially provide more robust candidate selection.

Fourth, the ensemble experiments revealed substantial complementarity between models, as demonstrated by the oracle ensemble analysis. However, the practical ensemble methods explored in this work were unable to fully exploit this complementarity, suggesting that more advanced ensemble strategies remain an open research direction.

Finally, the experiments focused primarily on text-based gloss representations and did not incorporate visual sign language information. Future work could investigate multimodal approaches that jointly leverage gloss annotations and sign language video data.

5. Conclusions

This work presented the SignNexus submission to the MSLG-SPA 2026 shared task on bidirectional translation between Mexican Sign Language glosses and Spanish. We explored multiple multilingual transformer architectures, parameter efficient adaptation strategies, reranking techniques, and data augmentation methods in order to address the challenges of low resource gloss translation.

Our experiments showed that multilingual pretrained sequence to sequence models can successfully adapt to the MSLG-SPA task when combined with lightweight fine tuning approaches such as LoRA. Among the evaluated systems, NLLB achieved the strongest performance for gloss to Spanish translation, while mBART consistently produced the best results for Spanish to gloss generation. These findings suggest that different multilingual architectures exhibit complementary strengths depending on translation direction and output structure.

We also demonstrated that decoding strategies play an important role in translation quality. Beam search decoding combined with heuristic reranking improved fluency and output consistency compared to naive generation methods. In contrast, sampling based decoding and few shot prompting strategies produced unstable or noisy outputs in our experiments. Ensemble analysis further revealed that combining different multilingual systems has significant potential, as evidenced by the large gap between oracle and practical ensemble performance.

Several challenges remain open for future research. The Spanish to gloss task continues to be substantially harder than gloss to Spanish translation due to the symbolic and compressed nature of gloss sequences. In addition, the relatively small size of the available corpus limits the ability of large neural architectures to generalize robustly. Future work may explore larger multilingual language models, stronger constrained decoding mechanisms, improved synthetic data generation, and LLM based reranking approaches capable of preserving gloss semantics more accurately.

Overall, the results obtained in this work confirm that multilingual pretrained models combined with parameter efficient adaptation techniques provide a promising direction for sign language gloss translation under low resource conditions.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-5.3 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] 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).

[2] A. Bonet-Jover, J. A. 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.

[3] 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 (NeurIPS), volume 30, 2017. URL: https://papers.nips.cc/paper/7181-attention-is-all-you-need.

[4] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Transactions of the Association for Computational Linguistics 8 (2020) 726–742. URL: https://aclanthology.org/2020.tacl-1.47/.

[5] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Jain, et al., No language left behind: Scaling human-centered machine translation, arXiv preprint arXiv:2207.04672 (2022). URL: https://arxiv.org/abs/2207.04672.

[6] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of NAACL, 2021, pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41/.

[7] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, C. Raffel, Byt5: Towards a token-free future with pre-trained byte-to-byte models, Transactions of the Association for Computational Linguistics 10 (2022) 291–306. URL: https://aclanthology.org/2022.tacl-1.17/.

[8] H. W. Chung, L. Hou, S. Longpre, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022). URL: https://arxiv.org/abs/2210.11416.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: International Conference on Learning Representations (ICLR), 2022. URL: https://arxiv.org/abs/2106.09685.

[10] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of ACL, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/.

[11] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures

for Machine Translation and/or Summarization, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

[12] M. Popovi'c, chrf: character n-gram f-score for automatic mt evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049/.