CICESE-LCDAA at MSLG-SPA 2026: Bidirectional Translation Between Mexican Sign Language Glosses and Spanish Using Seq2Seq Models and LLM Prompting

Jesús Antonio Navarrete-López[1,†], Joan Manuel Raygoza-Romero[1,†], Cielo Aholiva Higuera-Gutierrez[1,†], Juliette Ninnie Lazo-Vazquez[2,†] and Irvin Hussein Lopez-Nava[1,*,†]

[1]*Data Science and Machine Learning Lab, Centre for Scientific Research and Higher Education of Ensenada (CICESE), Baja California, Mexico*

[2]*Faculty of Sciences, Autonomous University of Baja California (UABC), Ensenada, Baja California, Mexico*

Abstract

Automatic translation between Mexican Sign Language (LSM) and written Spanish presents significant challenges due to the structural differences between both modalities and the limited availability of linguistic resources. This paper presents the systems developed for the MSLG-SPA 2026 shared task using both *seq2seq* approaches and *prompting*-based strategies with Large Language Models (LLMs). Several preprocessing techniques, linguistic normalization procedures, and *special tokens* were applied, together with adaptation strategies such as *fine-tuning*, LoRA, and data augmentation on models including mBART-50, T5-Flan, and BART-Base. Experimental results showed that mBART-50 achieved the best performance among the *seq2seq* models, while the *prompt*-based approach using Gemma 4 obtained the best overall results. In the official evaluation, the *prompt*-based system achieved a BLEU score of 54.03 and a COMET score of 0.8971 for the MSLG2SPA subtask, whereas for SPA2MSLG, the best configuration based on LoRA and data augmentation reached a BLEU score of 21.19. These findings highlight the potential of LLMs and lightweight adaptation strategies for sign language machine translation tasks under low-resource scenarios.

Keywords

Mexican Sign Language, Machine Translation, Large Language Models, Seq2Seq Models, Prompt Engineering

1. Introduction

Automatic translation between Mexican Sign Language (MSL) and Spanish represents a significant technical challenge due to several factors, including the differences between their modalities, as MSL is a non-written language [1], as well as their distinct grammatical structures and multiple technical issues related to occlusion, illumination, and the accurate capture of the signing space. These aspects must be considered within a complete translation pipeline, whose complexity further increases in the context of bidirectional translation, where additional challenges associated with sign production must also be addressed.

One approach to tackle translation in both directions is to assume that complex tasks such as automatic MSL recognition achieve adequate performance, allowing the translation process to operate directly on the annotations generated by these systems, commonly referred to as glosses [2]. Glosses constitute a linguistic representation used to describe signs through textual labels.

The MSLG-SPA 2026 task[3], organized within the framework of IberLEF 2026[4], is the first shared task focused on bidirectional translation between MSL glosses and Spanish. This initiative aims to

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

jnavarrete@cicese.edu.mx (J. A. Navarrete-López); jraygoza@cicese.edu.mx (J. M. Raygoza-Romero); aholiva@cicese.edu.mx (C. A. Higuera-Gutierrez); juliette.lazo@uabc.edu.mx (J. N. Lazo-Vazquez); hussein@cicese.edu.mx (I. H. Lopez-Nava)

0009-0009-2890-4822 (J. A. Navarrete-López); 0000-0003-3085-5678 (J. M. Raygoza-Romero); 0009-0005-1172-2679 (C. A. Higuera-Gutierrez); 0009-0002-6865-4468 (J. N. Lazo-Vazquez); 0000-0003-3979-9465 (I. H. Lopez-Nava)

address a critical gap in the development of language technologies for the Deaf community in Mexico.

Our proposal is based on two main approaches. The first leverages the capabilities of a causal Large Language Model (LLM) through prompt engineering and few-shot learning strategies to generate translations in the target language. The second approach employs a multilingual *seq2seq* pretrained model, which is adapted through fine-tuning to learn translation in both directions.

2. Task and Dataset

The MSLG-SPA 2026 shared task focuses on bidirectional machine translation between Mexican Sign Language glosses (MSLG) and written Spanish. Unlike translation between spoken languages, sign languages pose unique challenges derived from their visual-gestural nature, morphological richness, and the absence of a standardized writing system. In this context, glosses serve as an intermediate symbolic representation that enables computational processing and the automatic evaluation of translation systems.

The task is divided into two complementary subtasks. The first, referred to as MSLG2SPA, consists of generating Spanish sentences from sequences of MSL glosses, aiming to produce fluent and semantically accurate translations. The second subtask, SPA2MSLG, addresses the inverse problem, where the objective is to translate Spanish sentences into gloss sequences that adequately represent the grammatical structure and semantic content of MSL.

The primary objective of this shared task is to establish the first benchmark for automatic translation between MSL glosses and Spanish. Additionally, it seeks to provide a common evaluation protocol that facilitates the comparison of different approaches and promotes the development of methods based on both machine learning techniques and linguistic knowledge.

2.1. Dataset Description

The dataset provided for the MSLG-SPA 2026 shared task was designed to support reproducible and comparable research in bidirectional machine translation between MSL glosses and written Spanish. The corpus consists of sentence-level aligned pairs composed of MSL gloss sequences and their corresponding Spanish translations.

Each pair represents semantically equivalent expressions across both modalities, enabling the two translation directions considered in the task: MSLG2SPA and SPA2MSLG. However, the grammatical structure of MSL differs substantially from that of Spanish. Gloss sequences do not necessarily follow Spanish syntactic order, nor do they explicitly preserve verbal inflections, agreement markers, or functional elements. As a result, translation becomes a problem that requires reordering, inference, and semantic abstraction beyond simple word-to-word correspondence.

Although the initial design of the dataset considered removing auxiliary gloss annotations in order to simplify the representation and promote uniformity among participating systems, this edition of the shared task retained a limited set of conventions due to their linguistic relevance and their frequent use in current MSL glossing practices.

These conventions appear directly within the gloss sequences and must therefore be treated as part of the token stream during both training and evaluation. The preserved conventions included in the corpus are described below.

- Hyphen (-): When two glosses are connected by a hyphen, both correspond to a single sign.
- Plus sign (+): Used to represent compound signs or lexical contractions.
- Number sign (#): Indicates that the sign corresponds to a lexical borrowing produced through fingerspelling.
- Prefix dm-: Used to explicitly mark names or lexical elements produced through fingerspelling.

The dataset is divided into three disjoint subsets: a training set composed of 490 translation pairs (50%), used for model training and internal validation; a test set for the MSLG2SPA subtask containing

244 pairs (25%); and a test set for the SPA2MSLG subtask with 244 pairs (25%). This partitioning allows both subtasks to share the same training data while being evaluated on independent test sets, thereby preventing information leakage between translation directions. Table 1 presents several examples extracted from the corpus.

Table 1
Examples of parallel sentence pairs from the MSLG-SPA 2026 corpus.

MSLG	Spanish
dm-MIGUEL PECHO ÉL DOLER	Miguel tiene dolor en el pecho. (*Miguel has chest pain.*)
#TV PUBLICIDAD HABER MUCHO	En la TV hay mucha publicidad. (*There is a lot of advertising on TV.*)
JOVEN HOMBRE CORTEJA ELLA SU NOVIO+MUJER	El muchacho corteja a su novia. (*The young man courts his girlfriend.*)
TODOS-SÁBADOS MI AMIGO+MUJER SUYO NOVIO CASINO IR	Mi amiga y su novio van los sábados al casino. (*My female friend and her boyfriend go to the casino every Saturday.*)
RASURADORA YO QUERER COMPRAR	Quiero comprar una rasuradora. (*I want to buy a razor.*)

3. Methods

For this shared task, two main machine translation approaches were explored. The first approach consists of a pipeline based on fine-tuning pretrained models, including both Spanish monolingual and multilingual architectures. This approach was further complemented by the incorporation of an external dataset as a data augmentation strategy to improve model generalization under low-resource conditions.

The second approach is based on prompting techniques using a LLM. In this setting, task-oriented prompts were designed by incorporating examples selected through a stratified sampling process, with the aim of guiding the generation process toward accurate translations in the target language.

3.1. Seq2Seq Translation

The first approach explored in this work is based on *seq2seq* translation. To this end, several Transformer-based architectures were internally evaluated, including both multilingual models and models specialized in Spanish-language tasks. Given the limited amount of available training data, two complementary strategies were proposed to mitigate this constraint. The first consisted of fine-tuning the evaluated pretrained architectures on the shared task dataset, while the second involved data augmentation through the incorporation of an external corpus adapted to match the format and characteristics of the data provided by the shared task. The overall pipeline of the proposed *seq2seq*-based approach is illustrated in Figure 1.

3.1.1. Base Models

The evaluated models included several widely used generative architectures for machine translation and text generation tasks, such as BART-Base [5], FLAN-T5 Base [6], and mBART-50 [7]. In addition, pretrained models specialized in Spanish were also explored, including BETO-Seq2Seq [8]—a BERT-based model trained on large-scale Spanish corpora—and Helsinki-NLP/opus-mt-es-es [9], which was selected as a reference architecture to evaluate neural translation models previously optimized for the Spanish language.

The selection of these models aimed to compare the behavior of monolingual and multilingual architectures under similar training conditions, as well as to analyze their ability to adapt to bidirectional translation tasks between gloss sequences and written language.

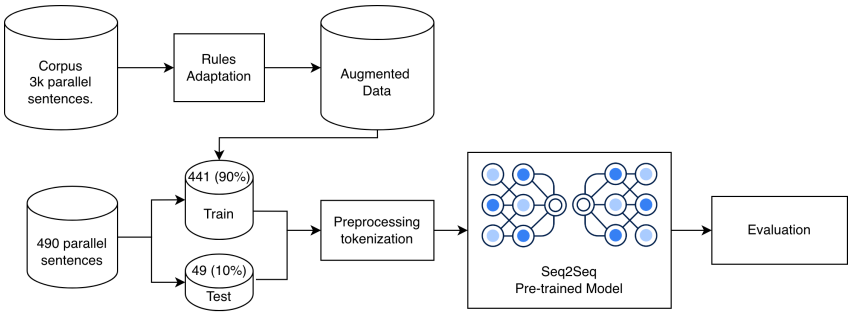

Figure 1: Sequence to Sequence methodology diagram

3.1.2. Fine-tuning strategies

To analyze model behavior under different adaptation schemes, two training strategies were evaluated: full fine-tuning and parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA). In the full fine-tuning approach, all model parameters were updated during training using the MSLG-SPA 2026 parallel corpus. This strategy enables the internal representations of the model to be fully adapted to the bidirectional translation task between gloss sequences and Spanish, although at the expense of higher computational cost.

Additionally, LoRA was explored as a parameter-efficient adaptation strategy. This method introduces low-rank matrices into specific attention components of the model, enabling the adaptation of large-scale architectures while significantly reducing the number of trainable parameters.

For the LoRA experiments, a configuration tailored for *seq2seq* models was employed, using a rank of $r = 16$, a scaling factor of $\alpha = 32$, and a dropout rate of 0.1. The adaptations were applied to the attention projection layers corresponding to q_{proj}, k_{proj}, v_{proj}, and out_{proj}, with the goal of modifying only the layers associated with the multi-head attention mechanism.

3.1.3. Data Augmentation

Inspired by the work of [10], which proposes a data augmentation strategy based on linguistically motivated rules rather than heuristic pseudo-gloss generation, we employed the dataset introduced by [11] and adapted it to the conventions defined for the MSLG-SPA 2026 shared task. To achieve this, several linguistic transformations and normalization procedures were applied to the gloss sequences, including the unification of compound structures and the simplification of lexical repetitions. Representative examples of these transformations are presented in Table 2. The main adaptations applied to the corpus are described below:

- **Normalization of repetitive classifier markers.** Repeated sequences associated with classifiers were replaced with simplified pluralized forms, reducing structural redundancy within gloss sequences.
- **Lexical normalization of possessives.** Replacement rules were applied to certain possessive forms in order to maintain consistency with the conventions observed in the official corpus.
- **Simplification of lexical repetitions.** Sequences composed of consecutive repetitions of the same gloss were transformed into equivalent pluralized representations to homogenize morphological patterns across the dataset.
- **Unification of gender-related variants.** Some lexical variants associated with gender distinctions were normalized into a canonical representation to reduce lexical sparsity during training.
- **Transformation of compound structures.** Expressions composed of multiple glosses were converted into compound structures using the "+" symbol, following the conventions established by the shared task.

- **Format standardization.** Finally, all gloss sequences were converted to uppercase in order to maintain orthographic consistency throughout the extended corpus.

Table 2
Examples of corpus adaptations applied to the external dataset in order to align it with the conventions of the MSLG-SPA 2026 shared task.

SPA	MSL	Adapted
mi suegra estaba leyendo un libro ayer (*my mother-in-law was reading a book yesterday*)	ayer suegro mujer mía leer	AYER SUEGRO+MUJER MÍA LEER
los perros beben agua (*the dogs drink water*)	perro cl+cl+cl agua beber	PERROS AGUA BEBER
algunas puertas (*some doors*)	puerta algunas	PUERTA ALGUNAS
novia (*girlfriend*)	novio mujer	NOVIO+MUJER
la ropa está mojada (*the clothes are wet*)	ropa mojada	ROPA MOJADA
ayer no comí (*yesterday I did not eat*)	ayer yo comer nada	AYER YO COMER NADA
tu abuelo era soldado (*your grandfather was a soldier*)	abuelo tuyo soldado	ABUELO TUYO SOLDADO

3.1.4. Preprocessing

Prior to training, the parallel sequences of glosses and Spanish sentences were subjected to a preprocessing and tokenization pipeline compatible with the evaluated seq2seq architectures. For the MSLG2SPA direction, gloss sequences were converted to lowercase before encoding, whereas the target Spanish sentences were kept in their original form.

Tokenization was performed using the tokenizer associated with each pretrained model. Both input and target sequences were truncated or padded to a maximum length of 128 tokens, ensuring homogeneous sequence lengths during training.

Additionally, padding tokens in the target labels were replaced with the value -100, in order to exclude these positions from the computation of the loss function during supervised training. This procedure prevents artificial padding tokens from influencing the model optimization process.

3.1.5. Special Tokens

In order to preserve relevant linguistic information present in the gloss sequences, a special token scheme was implemented based on the conventions described in Section 2.1, as well as on the explicit detection of negation structures within the input text. This strategy is motivated by the hypothesis that certain grammatical rules and MSL-specific conventions may induce distributional representations that deviate from the linguistic context originally learned by pretrained models, thereby hindering their correct semantic interpretation during the translation process.

To mitigate this issue, specific linguistic structures were transformed into explicit symbolic representations using special tokens. In particular, compound signs, simple signs composed of multiple lexical units, lexical borrowings produced through fingerspelling, sequences associated with *fingerspelling*, and negative constructions were explicitly annotated.

Compound signs represented using the "+" symbol were encapsulated using the tokens <COMPOUND> and </COMPOUND>, while structures joined by "-" were represented using the tokens <SINGLE> and </SINGLE>. Similarly, sequences associated with fingerspelling and lexical borrowings were transformed using the tags <DM> and <LOAN>, respectively.

Additionally, an explicit negation detection mechanism was implemented. Both structures prefixed with "no-" and equivalent lexical patterns within gloss sequences were encapsulated using the tokens <NEG> and </NEG>, allowing the model to explicitly represent semantic negation relations during training.

Finally, all special tokens were added to the tokenizer vocabulary, and the model embedding matrix was subsequently resized to enable the learning of dedicated representations associated with these linguistic structures.

3.2. Prompt-based Translation

This approach leverages the ability of LLMs to adapt to different tasks through in-context learning strategies and guided reasoning using Chain-of-Thought prompting. For this purpose, prompts were designed to perform bidirectional translation between MSL glosses and Spanish without requiring any explicit additional training process. The prompts incorporate example pairs selected through a stratified sampling procedure, with the aim of providing representative contextual information to the model during generation.

The model used in this approach was Gemma 4 with 31 billion parameters, selected due to its recent capabilities in generation and multilingual reasoning tasks. The overall methodology employed in this approach is illustrated in Figure 2.

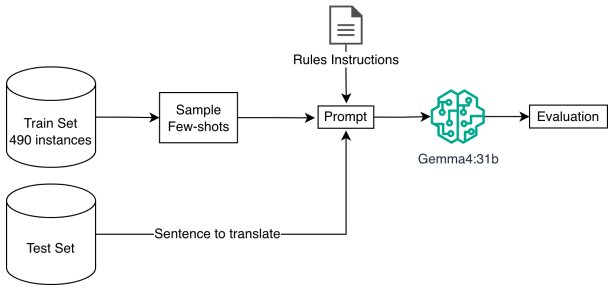

Figure 2: Prompt-based Methodology diagram

3.2.1. Prompt Design

The prompts used for both subtasks were developed and iteratively refined with the aim of providing the model with clear instructions regarding the translation task and the linguistic constraints associated with the MSLG-SPA 2026 corpus. Throughout this process, we sought to maximize the structural consistency of the generated outputs while reducing conversational responses or semantic deviations produced by the model.

For the MSLG2SPA subtask, the prompt was designed to guide the model in translating gloss sequences into grammatically correct Spanish sentences. To this end, explicit descriptions of the linguistic conventions present in the glosses were included, such as the use of the symbols “-”, “+”, “#” and the prefix “dm-” indicating how the model should interpret these structures during generation. In addition, strict instructions were incorporated to restrict the output exclusively to the translated sentence, avoiding additional explanations or conversational text. The prompt is presented below:

```
Role: You are a specialized Mexican Sign Language (MSL) to Spanish translator. Your task
    is to accurately translate MSL gloss sequences into grammatically correct Spanish
    sentences while strictly adhering to linguistic notation conventions.

MSL Notation Conventions:
1. Hyphen (-): Words joined by a hyphen represent a single, unified sign (e.g., "YA-VEO"
    means "Ya veo"). Do not split these into separate concepts.
2. Plus Sign (+): Indicates compound signs or contractions (e.g., "MAMÁ+PAPÁ" translates
    to "padres"). These must be treated as a single semantic unit.
3. Hash (#): Indicates a fingerspelled loanword (e.g., "#OK"). These are direct lexical
    borrowings.
```

```
4. "dm-" Prefix: Indicates manual fingerspelling for specific names or terms (e.g., "dm-
    LUIS").

Strict Rules:
- Preserve these annotations exactly as they appear in the source glosses during
    processing.
- Output ONLY the translated Spanish sentence between double quotes.
- Do not provide explanations, notes, or additional commentary.
- Ensure the resulting Spanish is natural and follows standard grammar rules.

### Examples
{few_shots_block}

### Task:
Translate the following MSL gloss sequence into Spanish.
```

On the other hand, for the SPA2MSLG subtask, the prompt was designed to generate gloss sequences following the structural conventions of MSL used in the corpus. In this case, the instructions emphasized aspects related to uppercase formatting, the use of verbs in infinitive form, the removal of unnecessary functional words, and syntactic reordering according to common MSL structures, such as Topic-Comment or Time-Subject-Object-Verb. Additionally, the correct use of the glossing conventions employed in the dataset was specified. The prompt is presented below:

```
Role: You are a specialized Spanish to Mexican Sign Language (MSL) translator. Your task
    is to accurately convert Spanish sentences into MSL Gloss sequences (GLOSSES) while
    strictly adhering to linguistic notation and structural conventions.

MSL Glossing Rules:
1. Capitalization: All glosses must be in ALL CAPS.
2. Verb Form: Use verbs in their INFINITIVE form only.
3. Syntax: Remove unnecessary articles, pronouns, and prepositions that do not exist in
    MSL.
4. Word Order: Follow the standard MSL structure (usually Topic-Comment or Time-Subject-
    Object-Verb).

Notation Conventions (CRITICAL):
- Hyphen (-): Use for words that represent a single unified sign (e.g., "YA-VEO", "CON-
    PERMISO").
- Plus Sign (+): Use for compound signs or contractions (e.g., "MAMÁ+PAPÁ" for the concept
     of 'parents').
- Hash (#): Use for fingerspelled loanwords (e.g., "#OK", "#WEB").
- "dm-" Prefix: Use for fingerspelling specific names or proper nouns (e.g., "dm-LUIS", "
    dm-MÉXICO").

Strict Rules:
- Apply the appropriate symbols (-, +, #, dm-) whenever the semantic context requires it
    based on the dataset conventions.
- Output ONLY the gloss sequence.
- Do not provide explanations, notes, or any conversational text.

### Examples
{few_shots_block}

### Task:
Translate the following Spanish sentence into MSL gloss sequence.
```

In both subtasks, the prompts incorporated translation examples selected through stratified sampling (few-shot prompting), providing the model with representative instances of the different linguistic

structures present in the corpus. This strategy enabled the inclusion of additional contextual information during inference without requiring any parameter fine-tuning.

3.2.2. Stratified Few-Shot Sampling

In order to provide representative examples to the model during the in-context learning process, a stratified sampling strategy was implemented for the selection of few-shot examples. Rather than randomly selecting translation pairs, the examples were grouped using the pseudo-label `type_LSM`, which categorizes gloss sequences according to the linguistic conventions present in the sentence, such as compound structures formed with "+", unified signs marked with "-", lexical borrowings represented with "#", and sequences associated with fingerspelling indicated by the prefix "dm-".

From each group, up to four examples were randomly selected using a fixed seed ($random\_state = 42$) to ensure reproducibility. This strategy enabled the construction of balanced example sets with respect to the different glossing conventions present in the corpus, avoiding biases toward dominant linguistic structures within the training data.

The selected examples were then directly incorporated into the prompt using an Input-Output format, providing the model with explicit references to the expected structure of the translations. An example of the format used is shown below:

```
### Examples

Input: dm-MIGUEL PECHO ÉL DOLER
Output: "Miguel tiene dolor en el pecho."

Input: dm-JUAN ASUNTO APARTE
Output: "Con Juan, es un asunto aparte."

Input: dm-DAVID SU PAPÁ BOMBERO
Output: "El papá de David es bombero."

Input: dm-NOEMÍ YA ULTRASONIDO PARA SABER SUYO BEBE NIÑO NIÑO+MUJER
Output: "Noemí se hizo un ultrasonido para saber si su bebé era niño o niña."
```

The incorporation of stratified examples exposed the model to different linguistic structures and annotation conventions during inference, aiming to improve the consistency and adequacy of the generated translations.

3.3. Internal Validation

To enable a controlled comparison among the evaluated models, an internal validation scheme based on stratified splits of the official training set was implemented. Initially, the corpus was divided using a 90/10 ratio for training and internal testing, preserving the distribution of linguistic categories present in the data. Subsequently, an additional 10% was extracted from the training subset to construct the validation set used during hyperparameter tuning and model selection.

3.3.1. Fine-Tuning Approach Experimental Setup

The models were trained using supervised learning for a maximum of 10 epochs with the AdamW optimizer and a linear learning rate scheduler. During training, padding tokens were masked using the value -100 in order to exclude these positions from the loss computation. Sequence generation during evaluation was performed using beam search with a beam size of 4.

The best-performing model obtained during the initial experiments was subsequently subjected to the adaptation strategies described in Section 3.1.2. For this stage, training was conducted using a learning rate of 1×10^{-4}, a batch size of 16 samples for both training and evaluation, and a maximum of 20 epochs.

Additionally, a warmup ratio of 0.1 and a weight decay of 0.01 were employed. During sequence generation, beam search with a beam size of 4 and a maximum generation length of 128 tokens was used. Models were evaluated and saved at the end of each epoch, with the best-performing checkpoint selected automatically according to the BLEU score on the validation set. Furthermore, training was performed using mixed precision (FP16) in order to optimize memory usage and accelerate the training process.

3.3.2. Prompt-Based Approach Experimental Setup

The experiments corresponding to the prompt-based approach were developed iteratively, evaluating different variants of the prompts described above as well as multiple few-shot configurations. Prompt evaluation was conducted under two main scenarios. In the first scenario, the generated predictions were directly compared against the internal test set previously used for the seq2seq models, enabling a homogeneous comparison between both approaches. In the second scenario, experiments were conducted using the full training set in order to analyze the general behavior of the model across different linguistic structures present in the corpus. For local model execution, Ollama was used as the inference environment, employing an OpenAI-compatible API interface for the integration and automation of experimental runs.

All experiments were executed on a workstation equipped with an Intel Core i9-14900KF processor, 64 GB of RAM, and an NVIDIA RTX 5090 GPU with 32 GB of VRAM.

3.4. Shared Test Validation

For each subtask of the shared task, a total of six runs were submitted. Run `run0` corresponds to the best system obtained under the prompt-based approach, selected based on the results observed during internal validation.

The remaining runs correspond to the best-performing seq2seq model from the preliminary experiments, which was subsequently subjected to different fine-tuning and data augmentation strategies in order to analyze the impact of each configuration on the final system performance.

The evaluated configurations are as follows:

- **run1:** fine-tuning using only the official base training set.
- **run2:** LoRA-based adaptation using only the official base training set.
- **run3:** fine-tuning combined with data augmentation using the adapted corpus.
- **run4:** LoRA-based adaptation combined with data augmentation.
- **run5:** LoRA-based adaptation using data augmentation and incorporating transformations based on special tokens.

This experimental setup allowed us to independently analyze the impact of data augmentation, parameter-efficient adaptation strategies, and the incorporation of explicit linguistic information through special tokens on the bidirectional translation task between MSL glosses and Spanish.

4. Results

In this section, we present the results obtained by the different approaches evaluated on the MSLG2SPA and SPA2MSLG subtasks of the MSLG-SPA 2026 shared task. First, we report the results obtained during internal validation, which allowed us to analyze the behavior of the different seq2seq architectures, fine-tuning strategies, and prompting-based configurations. Subsequently, we present the official results obtained on the shared task evaluation set.

4.1. Subtask MSLG2SPA

Table 3 presents the results obtained during internal validation for the MSLG2SPA subtask using different seq2seq architectures trained on raw, unprocessed data. The results indicate that mBART-50 achieved the best overall performance across all evaluated metrics, reaching a BLEU-4 score of 14.59 and a METEOR score of 0.4236. The second-best performance was obtained by T5-Flan, which yielded competitive results relative to mBART-50, particularly in terms of BLEU-1 and BLEU-2. However, a notable decline is observed in BLEU-3 and BLEU-4, suggesting greater difficulty in maintaining structural coherence in longer sequences or in capturing complex contextual dependencies.

In contrast, BART-Base exhibited moderate performance, although consistently below that of the aforementioned models. Meanwhile, Helsinki-ES-ES and BETO-Seq2Seq showed significantly lower performance across all evaluated metrics. In particular, BETO-Seq2Seq achieved the worst overall results, with BLEU-3 and BLEU-4 values close to zero.

Based on these findings, mBART-50 was selected as the base model for subsequent experiments involving fine-tuning strategies, data augmentation, and the incorporation of special tokens.

Table 3
Results obtained during internal validation of seq2seq models for the MSLG2SPA subtask using raw (unprocessed) data.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	chrF
mBART-50	48.91	29.91	20.81	14.59	0.4236	42.91
T5-Flan	39.71	23.93	13.89	7.19	0.3520	37.68
BART-Base	33.97	19.24	11.67	6.32	0.3333	33.17
Helsinki-ES-ES	22.21	9.24	3.25	0.00	0.2579	14.47
BETO-Seq2Seq	3.00	0.75	0.00	0.00	0.0853	7.41

The results indicate that the incorporation of different adaptation strategies led to a substantial improvement in the initial performance of the mBART-50 model compared to experiments conducted on raw data (see Table 4). The *LoRA + Data Aug.* configuration achieved the best performance in terms of BLEU-1, BLEU-2, BLEU-3, and BLEU-4, reaching a BLEU-4 score of 16.23. The inclusion of *special tokens* also produced highly competitive results, remaining very close to the best overall configuration. Although the *LoRA + Data Aug. + Special Tokens* variant showed a slight decrease in BLEU-4 compared to *LoRA + Data Aug.*, it achieved the highest METEOR (0.5893) and chrF (59.04) scores, suggesting better semantic preservation and higher character-level similarity in the generated translations. The strategy based solely on *Fine-tuning + Data Aug.* resulted in a noticeable degradation across all evaluated metrics.

Table 4
Performance evolution of the mBART-50 model under different adaptation strategies evaluated during internal validation for the MSLG2SPA subtask using preprocessed data.

Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	chrF
Fine-tuning	56.98	29.77	18.08	9.95	0.5514	53.25
LoRA	59.22	30.74	18.85	12.80	0.5728	58.44
Fine-tuning + Data Aug.	54.52	25.85	13.47	7.14	0.5013	49.33
LoRA + Data Aug.	61.54	35.29	22.50	16.23	0.5642	57.53
LoRA + Data Aug. + Special Tokens	60.71	33.33	22.18	13.36	0.5893	59.04

Table 5 shows a progressive improvement as the level of specificity and contextualization in the prompts used by the prompting-based approach increases. Overall, each modification introduced into the prompt design leads to consistent gains across evaluation metrics, suggesting that causal LLMs are highly sensitive to how the task and its associated linguistic constraints are formulated.

The initial configuration, based solely on role definition, already achieves competitive performance compared to seq2seq models, reaching a BLEU-4 score of 45.20. This indicates that the model already possesses strong translation capabilities without task-specific adaptation. From this point, the gradual inclusion of explicit rules describing glossing conventions leads to consistent improvements in BLEU-1,

BLEU-2, and BLEU-3, with smaller variations in METEOR and chrF, suggesting increased structural consistency in the generated outputs. This trend continues when evaluating on the full dataset (Full Data Eval.), where further improvements are observed, reaching a BLEU-4 of 47.80. The best overall performance is obtained when few-shot examples are incorporated, achieving a BLEU-4 of 49.99 and a METEOR score of 0.7788. These results show a clear trend: higher levels of detail, constraints, and contextual information within the prompt lead to better model understanding of the task and more coherent translations.

Table 5
Evolution of the prompts evaluated during internal validation of the prompting-based approach for the MSLG2SPA subtask using raw data.

Prompt Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	chrF
Role Definition	69.90	59.60	51.60	45.20	0.7620	72.68
Role + Rules	71.30	61.20	53.00	45.40	0.7080	71.43
Role + Rules (Full Data Eval.)	73.40	62.80	54.60	47.80	0.7440	73.42
Role + Rules + Few-Shots (Full Data Eval.)	**76.07**	**65.50**	**57.04**	**49.99**	**0.7788**	**74.52**

The official results reported in Table 6 for the MSLG2SPA subtask show a clear superiority of the prompting-based approach over supervised learning strategies. In particular, run0 (prompt-based) achieves the best overall performance across all evaluated metrics, reaching a BLEU score of 54.03, METEOR of 0.7191, chrF of 76.56, and COMET of 0.8971, confirming the strong capability of the language model to perform translation in this domain without parameter fine-tuning.

The seq2seq-based approaches, however, exhibit substantially lower performance, with gradual improvements observed as techniques such as LoRA, data augmentation, and special tokens are incorporated. Among these configurations, the most complete setup (run5) achieves the best results within the supervised group, reaching a BLEU score of 27.58 and a COMET score of 0.8045, although it remains significantly below the performance obtained by the prompting-based approach.

These results suggest that, in the context of translation between MSL glosses and Spanish, foundation models with instruction-following capabilities exhibit a clear advantage over models adapted through traditional fine-tuning approaches, even when the latter incorporate additional adaptation strategies and corpus enrichment techniques.

Table 6
Official results obtained in the MSLG2SPA subtask of the MSLG-SPA 2026 shared task.

Run	BLEU	METEOR	chrF	COMET
run0 (Prompt-based)	**54.03**	**0.7191**	**76.56**	**0.8971**
run1 (Fine-tuning)	23.40	0.4642	55.07	0.7665
run2 (LoRA)	25.01	0.4689	57.45	0.7915
run3 (Fine-tuning + Data Aug.)	20.53	0.4273	52.54	0.7325
run4 (LoRA + Data Aug.)	23.57	0.4552	57.17	0.7947
run5 (LoRA + Data Aug. + Special Tokens)	27.58	0.4871	60.04	0.8045

4.2. Subtask SPA2MSLG

The results presented in Table 7 for the SPA2MSLG subtask show that the best overall performance was achieved by mBART-50, reaching the highest BLEU-4 score (8.30) as well as the best METEOR score (0.4209) among the evaluated models. This behavior further confirms the strength of pretrained multilingual models for bidirectional translation tasks between Spanish and MSL glosses.

BART-Base ranks second in performance, showing competitive results particularly in BLEU-1 and BLEU-2, although it exhibits a significant drop in BLEU-3 and BLEU-4, indicating difficulties in maintaining coherence when generating longer sequences. T5-Flan, on the other hand, achieves intermediate performance but shows a more pronounced degradation in BLEU-4, which decreases to 0.00.

Helsinki-ES-ES and BETO-Seq2Seq obtain the lowest performance across all evaluated metrics, highlighting significant limitations in modeling the inverse translation direction (from Spanish to glosses), particularly in settings that require strong generative capabilities and syntactic reordering. Based on these results, mBART-50 was selected as the base model for the SPA2MSLG subtask in subsequent experiments.

Table 7
Results obtained during internal validation of seq2seq models for the SPA2MSLG subtask using raw (unprocessed) data.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	chrF
mBART-50	46.51	31.36	15.60	8.30	0.4209	43.36
BART-Base	41.00	24.13	13.42	7.75	0.3668	43.60
T5-Flan	34.47	19.73	6.57	0.00	0.3342	33.80
Helsinki-ES-ES	11.52	3.51	0.00	0.00	0.1528	10.91
BETO-Seq2Seq	1.52	0.60	0.00	0.00	0.0395	5.99

The mBART-50 model under different adaptation strategies for the SPA2MSLG subtask using preprocessed data shows that configurations based on LoRA and data augmentation improve performance compared to basic fine-tuning, particularly in surface-level similarity metrics such as BLEU-1, BLEU-2, and chrF (see Table 8).

The best configuration in terms of BLEU-1, BLEU-2, and chrF corresponds to LoRA + Data Aug. + Special Tokens, achieving 71.91 in BLEU-1 and 67.82 in chrF, which indicates a higher overall similarity to the reference translations. However, this same configuration does not obtain the best performance in BLEU-4, where Fine-tuning + Data Aug. achieves the highest score of 11.90, suggesting a better ability to capture more complete structural patterns in the generated outputs.

The strategy based solely on LoRA does not show consistent improvements over standard fine-tuning, particularly in BLEU-4, where it even exhibits a decline in performance. This suggests that the effectiveness of LoRA strongly depends on its combination with other adaptation techniques.

Table 8
Performance evolution of the mBART-50 model under different adaptation strategies evaluated during internal validation for the SPA2MSLG subtask using preprocessed data.

Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	chrF
Fine-tuning	58.56	26.01	12.90	10.53	0.5297	61.09
LoRA	56.68	25.00	14.29	6.76	0.4829	58.37
Fine-tuning + Data Aug.	49.57	21.55	15.15	11.90	0.4242	53.31
LoRA + Data Aug.	59.56	26.70	14.96	10.00	0.5216	63.91
LoRA + Data Aug. + Special Tokens	71.91	35.48	14.60	7.87	0.6283	67.82

The results in Table 9 show the evolution of the prompting-based approach for the SPA2MSLG subtask using raw data. Overall, a progressive improvement is observed as prompt complexity increases, particularly when few-shot examples are incorporated, indicating a greater ability of the model to adapt to the task through instructions and contextual examples.

The initial configuration, based solely on role definition, shows moderate performance, although with a BLEU-4 score of 0.00, indicating difficulties in generating sequences that are fully aligned with the reference translations. The incorporation of explicit rules partially improves lower-order n-gram metrics, but does not lead to consistent gains in BLEU-4.

When evaluation is performed using the full dataset, a recovery in BLEU-4 performance is observed (5.70), suggesting better model generalization under a broader context. Finally, the few-shot prompting configuration achieves the best results across all metrics, reaching a BLEU-4 score of 11.69 and the highest METEOR value (0.4271), confirming the importance of examples as structural guidance for gloss generation.

Los resultados oficiales de la Tabla 10 para la subtarea SPA2MSLG muestran un desempeño general

Table 9
Evolution of the prompts evaluated during internal validation of the prompting-based approach for the SPA2MSLG subtask using raw data.

Prompt Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	chrF
Role Definition	51.30	26.20	11.50	0.00	0.3910	53.16
Role + Rules	52.60	23.30	9.70	0.00	0.3760	56.98
Role + Rules (Full Data Eval.)	47.30	22.30	11.60	5.70	0.3730	55.65
Role + Rules + Few-Shots (Full Data Eval.)	**54.95**	**30.12**	**18.14**	**11.69**	**0.4271**	**58.21**

superior del enfoque basado en prompting, específicamente en el run0, el cual alcanza el mayor valor de chrF (63.20) y un BLEU competitivo (20.23), consolidándose como uno de los sistemas más robustos en términos de similitud global con las referencias. The official results in Table 10 for the SPA2MSLG subtask show an overall superior performance of the prompting-based approach, particularly run0, which achieves the highest chrF score (63.20) and a competitive BLEU score (20.23), establishing it as one of the most robust systems in terms of overall similarity to the reference translations.

Among the supervised learning approaches, the best BLEU performance is obtained by run4 (LoRA + Data Aug.), with a score of 21.19, along with the highest METEOR score (0.5104) within this group, indicating better semantic alignment compared to the other configurations. In contrast, run3 (Fine-tuning + Data Aug.) and run5 (LoRA + Data Aug. + Special Tokens) yield the lowest BLEU scores, suggesting that data augmentation without an adequate adaptation strategy may negatively affect generation quality. The results confirm that LoRA-based strategies combined with data augmentation provide a more stable balance within the seq2seq framework, while the prompting-based approach maintains a clear advantage in terms of global quality and structural similarity of the outputs.

Table 10
Official results obtained in the SPA2MSLG subtask of the MSLG-SPA 2026 shared task.

Run	BLEU	METEOR	chrF
run0 (Prompt-based)	20.23	0.4963	**63.20**
run1 (Fine-tuning)	18.94	0.4780	57.74
run2 (LoRA)	17.08	0.5015	59.43
run3 (Fine-tuning + Data Aug.)	12.11	0.4119	51.24
run4 (LoRA + Data Aug.)	**21.19**	**0.5104**	60.37
run5 (LoRA + Data Aug. + Special Tokens)	11.49	0.4781	55.87

4.3. Qualitative Evaluation

The qualitative examples presented in Table 11 allow us to analyze the behavior of the best-performing configurations obtained for both subtasks of the shared task. The results show that both the prompting-based approach for MSLG2SPA and the LoRA + data augmentation configuration for SPA2MSLG are able to adequately preserve the main semantic content of the sentences, even when there are substantial structural differences between the references and the generated outputs.

In the MSLG2SPA subtask, the prompting-based model produces natural and grammatically correct translations in most cases. Examples such as “ÁFRICA ADENTRO GHANA” and “CANCÚN HABER CRUCERO” are translated in a nearly identical manner to the references. Similarly, in cases like “HACE-PASADO CRISTÓBAL-COLÓN DESCUBRIR AMÉRICA”, the model correctly reinterprets the implicit temporal expression in the gloss through the phrase “Hace mucho tiempo”, demonstrating semantic abstraction beyond literal translation. However, slight structural variations are also observed, as in “MAÑANA BARCO YA LLEGAR”, where the system generates “Mañana llega el barco” instead of “El barco llegó en la mañana”, partially altering the verb tense and the temporal interpretation of the original sentence.

On the other hand, in the SPA2MSLG subtask, the results show that the model is able to generate gloss sequences that are coherent with the typical structure of MSL. In several cases, the predictions adequately

preserve the expected semantic ordering, as in “CANCÚN CRUCERO HABER” or “PARACAÍDAS DM-AIMÉ SALTAR”. The system also demonstrates the ability to incorporate some domain-specific linguistic conventions, including the use of prefixes such as “DM-” in proper names. Structural and lexical inconsistencies also appear, particularly in more complex or ambiguous expressions such as “YO APENAS VER TI (FRENTE POCO)”, where the model introduces additional elements that are not explicitly present in the reference. These qualitative examples show that the proposed systems are able to capture a large portion of the semantic relationships between glosses and Spanish, although challenges remain in generating temporally precise structures, handling specific glossing conventions, and modeling syntactic variations inherent to MSL.

Table 11
Qualitative examples obtained during internal validation using the best-performing configurations of the shared task for the MSLG2SPA (*run0*) and SPA2MSLG (*run4*) subtasks.

MSLG	SPA	run0 (MSLG2SPA)	run4 (SPA2MSLG)
PAYASO ZAPATO SUYO COLOR AMARILLO	Los zapatos del payaso son amarillos. (*The clown's shoes are yellow*)	El zapato del payaso es de color amarillo.	PAYASO ZAPATOS AMARILLOS
ÁFRICA ADENTRO GHANA	Ghana está en África. (*Ghana is in Africa*)	Ghana está en África.	DM-GHANA ESTAR EN ÁFRICA
APENAS YO VER TI	Apenas te vi (hace poco). (*I just saw you (a short time ago)*)	Apenas te vi.	YO APENAS VER TI (FRENTE POCO)
dm-AIMÉ SALTAR PARACAIDAS	Aimé saltó en paracaídas. (*Aimé jumped with a parachute*)	Aimé saltó en paracaídas.	PARACAÍDAS DM-AIMÉ SALTAR
CANCÚN HABER CRUCERO	En Cancún hay un crucero. (*There is a cruise ship in Cancun*)	En Cancún hay un crucero.	CANCÚN CRUCERO HABER
HACE-PASADO CRISTÓBAL-COLÓN DESCUBRIR AMÉRICA	Cristóbal Colón descubrió América hace años. (*Christopher Columbus discovered America years ago*)	Hace mucho tiempo, Cristóbal Colón descubrió América.	AÑOS ATRÁS CRISTÓBAL COLÓN DESCUBRIR-AMERICA
MAÑANA BARCO YA LLEGAR	El barco llegó en la mañana. (*The ship arrived in the morning*)	Mañana llega el barco.	MAÑANA BARCO LLEGAR

5. Disscusion

In the MSLG2SPA subtask, the results obtained by the seq2seq models show that LoRA-based configurations tend to outperform traditional fine-tuning, particularly in higher-order BLEU metrics. This behavior suggests that parameter-efficient adaptation mechanisms enable better preservation of knowledge acquired during pretraining, thereby improving the modeling of complex sequential dependencies even under low-resource settings. Furthermore, configurations combining LoRA with data augmentation achieve the best overall performance among supervised approaches, indicating that both strategies can effectively complement each other.

Data augmentation applied in isolation through standard fine-tuning does not consistently lead to performance improvements. In fact, some configurations exhibit a noticeable degradation in evaluation metrics, suggesting that data augmentation may introduce additional noise into the training space when not supported by stronger regularization or more robust adaptation mechanisms. The results also indicate that the preprocessing applied to gloss sequences has a significant impact on model performance, particularly in the case of mBART-50. A comparison between the results reported in Tables 3 and 4 shows that models trained on preprocessed data consistently outperform those trained on raw data across most metrics. In particular, the use of normalized lowercase gloss representations appears to facilitate adaptation in pretrained models, consistent with the findings of [11], who report that lowercase textual representations lead to better alignment with the representation spaces learned

during pretraining.

Regarding the prompting-based approach, the results show that the incorporation of few-shot examples plays a fundamental role in the quality of the generated translations. The examples act as an effective contextual guide for inducing domain-specific structural and semantic patterns in MSL glosses, enabling the model to dynamically adapt its behavior without requiring additional training. This suggests that instruction-tuned language models exhibit strong contextual generalization capabilities, particularly in low-resource translation tasks.

In the SPA2MSLG subtask, the seq2seq configurations exhibit a different behavior compared to that observed in MSLG2SPA. In particular, the strategy based solely on LoRA does not provide consistent improvements over standard fine-tuning, especially in BLEU-4, where a decrease in performance is even observed. This behavior suggests that the effectiveness of LoRA strongly depends on its combination with other adaptation techniques, such as data augmentation or specialized preprocessing, and that its isolated application is not sufficient to adequately capture the syntactic structures and conventions specific to MSL glosses.

Prompting-based approach show a strong dependence on the quality and amount of context provided to the model. The initial configuration, based solely on role definition, achieves moderate performance, although with a BLEU-4 score of 0.00, reflecting difficulties in generating complete sequences that are structurally aligned with the reference translations. The incorporation of explicit rules leads to partial improvements in lower-order n-gram metrics; however, these gains do not translate into consistent improvements in BLEU-4, indicating that isolated instructions are insufficient to induce complex gloss structures. The inclusion of few-shot examples yields substantial improvements across all evaluation metrics, particularly in BLEU-3 and BLEU-4. This demonstrates that examples serve as a more effective contextual signal than purely descriptive instructions for inducing task-specific structural patterns and linguistic conventions. These findings indicate that LoRA-based configurations combined with data augmentation provide a more stable balance within the seq2seq framework, while the prompting-based approach maintains a clear advantage in terms of overall quality and structural similarity of the generated translations.

The incorporation of special tokens show that, although improvements in BLEU are relatively small compared to other configurations, more consistent gains are observed in metrics such as METEOR and chrF. This behavior is particularly relevant since both metrics more flexibly capture aspects related to semantic similarity and structural correspondence beyond exact n-gram matches. In this sense, the results suggest that the explicit representation of linguistic conventions through special tokens helps better preserve semantic information and the structural organization of gloss sequences during the translation process. Similarly, the results obtained by mBART-50 in both subtasks suggest that the multilingual capabilities acquired during pretraining facilitate the learning of semantic and syntactic relationships between MSL glosses and Spanish, even in low-resource scenarios. This is particularly relevant given the limited size of the corpus used in the shared task, where models with prior multilingual pretraining are able to generalize more effectively than monolingual architectures or models specialized in contextual representation.

6. Conclusion

Automatic translation between Mexican Sign Language (MSL) glosses and Spanish is a particularly challenging problem due to the structural, grammatical, and semantic differences between both linguistic modalities, especially under low-resource settings. This work addressed this problem within the context of the MSLG-SPA 2026 shared task, exploring both supervised approaches and instruction-based generative models for the MSLG2SPA and SPA2MSLG subtasks. The evaluated approaches included multiple pretrained seq2seq architectures, standard fine-tuning strategies, and parameter-efficient adaptation using LoRA, as well as data augmentation techniques and preprocessing mechanisms based on special tokens. In addition, a prompting-based approach was explored using causal LLMs, incorporating explicit instructions and few-shot examples selected through stratified sampling.

The results show that multilingual models, particularly mBART-50, exhibit stronger capabilities in modeling semantic and syntactic relationships between MSL glosses and Spanish compared to the other evaluated architectures. LoRA-based configurations also demonstrate consistent improvements within the supervised setting, especially when combined with data augmentation. The prompting-based approach achieves the best overall performance in the MSLG2SPA subtask, reaching a BLEU score of 54.03 and substantially outperforming seq2seq models. In the SPA2MSLG subtask, the best configuration corresponds to LoRA combined with data augmentation, achieving a BLEU score of 21.19. Additionally, the results suggest that the incorporation of special tokens contributes to better preservation of semantic and structural information in gloss sequences, particularly reflected in METEOR and chrF scores.

Future work will focus on integrating more complex linguistic annotations through explicit special token representations, incorporating morphosyntactic and discourse-level information specific to sign languages. It is also relevant to explore instruction-tuning strategies tailored specifically to sign language translation tasks, with the goal of better adapting generative models to this domain. Finally, the development and evaluation of sign language–specific architectures remains a promising direction, particularly for capturing the unique linguistic and structural properties of this form of representation.

Declaration on Generative AI

Generative AI model Gemini Flash 2.5 (developed by Google) was used during the preparation of this manuscript exclusively for language-related assistance, including English translation, grammar revision, and stylistic refinement of the text. The authors carefully reviewed and edited all generated content and take full responsibility for the final version of the manuscript, including its scientific accuracy, interpretations, and conclusions.

References

[1] M. Cruz Aldrete, Grammar of Mexican Sign Language, Phd thesis in linguistics, Center for Linguistic and Literary Studies, El Colegio de México, 2008.

[2] J. A. Navarrete López, Automatic Recognition of Dynamic Signs of Mexican Sign Language, Master's thesis, Center for Scientific Research and Higher Education at Ensenada (CICESE), Ensenada, Baja California, Mexico, 2025.

[3] 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).

[4] 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.

[5] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. rahman Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019, pp. 7871–7880. doi:10.18653/v1/2020.acl-main.703.

[6] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, D. Valter, S. Narang, G. Mishra, A. W. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, ArXiv abs/2210.11416 (2022). doi:10.48550/arxiv.2210.11416.

[7] 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. doi:10.1162/tacl_a_00343.

[8] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. P'erez, Spanish pre-trained bert model and evaluation data, ArXiv abs/2308.02976 (2023). doi:10.48550/arxiv.2308.02976.
[9] J. Tiedemann, S. Thottingal, Opus-mt – building open translation services for the world (2020) 479–480.
[10] A. Moryossef, K. Yin, G. Neubig, Y. Goldberg, Data augmentation for sign language gloss translation, in: Proceedings of the 1st international workshop on automatic translation for signed and spoken languages (AT4SSL), 2021, pp. 1–11.
[11] V. Lara-Ortiz, R. Q. Fuentes-Aguilar, I. Chairez, Spanish to mexican sign language glosses corpus for natural language processing tasks, Scientific Data 12 (2025) 702.