Bidirectional Mexican Sign Language Gloss–Spanish Translation via mBART and Back-Translation Data Augmentation: System Description for IberLEF 2026 MSLG-SPA

Marco Bortolotti[1]

[1]*Independent Researcher, Ferrara, Italy*

Abstract

We describe our system for the IberLEF 2026 MSLG-SPA shared task on bidirectional translation between Mexican Sign Language (LSM) glosses and Spanish. Our approach fine-tunes mBART-large-50 with Low-Rank Adaptation (LoRA) on the official 490-pair training corpus and augments the training data through back-translation (BT) in both directions. For the MSLG→SPA subtask we generate synthetic gloss–Spanish pairs using a trained SPA→MSLG model, then retrain the forward model on the combined real and synthetic corpus. For SPA→MSLG we apply the reverse strategy: we use the MSLG→SPA model to produce synthetic Spanish sentences from the real gloss side, yielding supplementary (synthetic-SPA, real-MSLG) pairs. On the held-out validation split, back-translation improves chrF from 49.29 to 70.16 for MSLG→SPA and from 44.42 to 68.89 for SPA→MSLG. In the official shared task evaluation (19 teams overall, blind test set), our submitted runs achieve chrF 46.14 / BLEU 16.81 for MSLG→SPA (baseline, no BT) and chrF 47.09 / BLEU 14.25 for SPA→MSLG (BT-augmented), placing 13th/20 on MSLG→SPA, 12th/19 on SPA→MSLG, and 13th overall with a global score of -0.312. The large validation-to-test gap for the BT-augmented SPA→MSLG model (68.89 val vs. 47.09 test) highlights the difficulty of model selection with fewer than 100 validation examples. We use compact LoRA (rank 16, query and value projections, 2.36M parameters), consistent with the literature recommendation for low-resource fine-tuning.

Keywords: Mexican Sign Language; back-translation; mBART; low-resource machine translation; LoRA; sequence-to-sequence translation

1. Introduction

Sign language gloss translation is a low-resource sequence-to-sequence task in which one side of the parallel corpus consists of concatenated lemma-like tokens (*glosses*) that represent individual signs, while the other side is written natural language. Unlike standard machine translation, the source and target vocabularies in a gloss–text pair share most of their lexical roots: glosses are essentially a compressed, uninflected projection of the spoken language. This near-monolingual nature makes the task structurally different from cross-lingual NMT, and means that standard multilingual models must be carefully adapted to avoid learning spurious cross-lingual alignments.

For Mexican Sign Language (LSM), the resources available to the research community are especially scarce. The official corpus released for the IberLEF 2026 MSLG-SPA shared task contains only 490 aligned gloss–Spanish sentence pairs, a subset of the 3,000-pair Lara-Ortiz corpus [1] made available to competition participants. Working with fewer than 500 training examples pushes standard NMT pipelines into a regime where data augmentation and parameter-efficient fine-tuning are not merely helpful but essential.

The task requires *bidirectional* translation, posing two distinct challenges. **MSLG→SPA** is a morphological infilling problem: the model must reconstruct Spanish inflection, agreement, and function words that are absent or implicit in the gloss stream. **SPA→MSLG** is a morphological compression problem: the model must suppress function words, lemmatize content words, and format the output as uppercase

IberLEF 2026, September 2026, León, Spain
marcoborto21@gmail.com (M. Bortolotti)
0009-0002-8188-919X (M. Bortolotti)

gloss tokens following LSM-specific notation conventions (e.g. `dm-` for deictics, `#` for fingerspelling). This asymmetry means that a single training strategy is unlikely to be optimal for both directions.

We address the data scarcity problem through back-translation (BT) [2], using each subtask model to generate synthetic training data for the other direction. To limit the number of trainable parameters and prevent overfitting we apply LoRA [3] via the PEFT library [4], injecting rank-16 adapters only into the query and value projections of mBART-large-50 [5]. This yields 2.36M trainable parameters (0.39% of the model).

The paper is organised as follows. Section 2 surveys related work. Section 3 describes the task and dataset. Section 4 details the system architecture and training procedure. Section 5 reports experimental results. Section 6 provides error analysis with concrete examples. Section 7 discusses the key findings. Section 8 concludes and outlines future work.

2. Related Work

Sign language gloss translation. The dominant approach to translating between glosses and spoken-language text is neural sequence-to-sequence modelling [6]. De Coster et al. [7] showed that freezing the encoder of a pretrained mBART model and fine-tuning only the decoder is effective for gloss-to-text translation on the PHOENIX-14T German Sign Language dataset (BLEU-4 25.58). Kermani et al. [8] obtained 29.92 BLEU on the same dataset using a Llama-8B-based architecture. For sign languages with very limited data, the SIGNUM dataset (∼600 German SL sentence pairs) has been used to demonstrate that mBART fine-tuning can reach BLEU-4 67.60 on gloss-to-text [8]. On LSM specifically, Rodríguez-González et al. [1] fine-tuned a BART-based model on the full 3,000-pair corpus and reported BLEU-4 35.0 as the baseline reference. The same work showed that augmenting the training set with glosses from the English ASLG-PC12 corpus via cross-lingual transfer lifted performance from 62 to 85 BLEU (+23 points), establishing data augmentation as the single highest-leverage intervention for this task.

Back-translation for low-resource NMT. Sennrich et al. [2] introduced back-translation as a method to exploit monolingual data by generating synthetic source-language sentences from a reverse model. The approach is particularly effective in low-resource regimes, where even noisy synthetic pairs provide useful training signal that the model could not otherwise encounter [9]. Hoang et al. [10] extended BT to an iterative setting, showing that alternating rounds of forward and reverse training progressively improve both models. Quality filtering of BT pairs — retaining only those that round-trip above a score threshold — is sometimes applied to reduce noise, but requires an initial model of sufficient quality; with very weak initial models filtering can be counterproductive [11].

Parameter-efficient fine-tuning. Full fine-tuning of large pretrained models on small datasets is prone to overfitting and catastrophic forgetting. Parameter-efficient methods such as adapter layers [12], prefix tuning [13], and LoRA [3] address this by freezing most model weights and introducing a small number of trainable parameters. LoRA reparametrises weight updates as low-rank matrix products, keeping the original weights frozen and adding fewer than 1% extra parameters in typical configurations. In low-resource settings, PEFT methods have been shown to match or exceed full fine-tuning while significantly reducing the risk of overfitting [3].

3. Task and Dataset

3.1. Task Description

The IberLEF 2026 MSLG-SPA shared task [1, 14] requires participating systems to translate between LSM gloss sequences and Spanish sentences in both directions. The official evaluation metrics are BLEU-4 [15], METEOR [16], and character F-score (chrF) [17]; for the MSLG→SPA subtask the organizers additionally report COMET. Rankings are determined by a composite task score computed as the average

z-score across all evaluation metrics for each subtask. We use chrF as the primary validation metric for checkpoint selection, as it is more robust than BLEU on short sequences and small vocabularies [17].

3.2. Dataset

The official training corpus consists of 490 aligned sentence pairs drawn from the LSM corpus released by the task organizers [1]. We split these into a training set (416 pairs, 85%) and a held-out validation set (74 pairs, 15%) using a fixed random seed (42) for reproducibility across all experiments.

3.3. Gloss Notation and Preprocessing

LSM glosses follow a structured notation in which uppercase tokens represent citation-form sign identifiers (ID-glosses). Several prefix and suffix conventions encode grammatical information: `dm-` marks deictic or demonstrative signs; + indicates compound signs or reduplication; # marks lexicalized fingerspelling sequences; - separates components of multi-word expressions or affixes. Table 1 shows representative aligned pairs from the training corpus illustrating these conventions.

Table 1
Example aligned pairs from the training corpus. Gloss tokens are uppercase; annotation prefixes are shown in typewriter font.

LSM Gloss	Spanish
`dm-ISABEL TENER SU CORONA ORO`	*Isabel tiene una corona de oro.*
`TUYA MAESTRO+MUJER TÚ PONER-ATENCIÓN ELLA`	*Le pusiste atención a tu maestra?*
`SU #SEP CALENDARIO AGOSTO TERMINAR`	*El calendario de la SEP termina en agosto.*

We apply minimal preprocessing: whitespace tokenization, all files read with explicit UTF-8 encoding to preserve accented characters, and no case normalization on the Spanish side. The mBART SentencePiece tokenizer treats gloss tokens as out-of-vocabulary sequences and segments them into subword units; compound markers (`dm-CASA` → `dm`, `-`, `CA`, `SA`) are therefore fragmented. We leave the addition of dedicated special tokens for gloss annotation markers as future work.

For both subtasks we set the decoder forced beginning-of-sequence token to `es_XX`, treating MSL glosses as a register of Spanish rather than a distinct language. This avoids language-identity mismatches in the mBART vocabulary and reuses Spanish-language priors from pretraining.

4. System Architecture

4.1. Base Model

We use **mBART-large-50** [5], a multilingual denoising autoencoder pre-trained on 50 languages including Spanish. The model has 611M parameters organized as a 12-layer encoder and 12-layer decoder with hidden size 1024, 16 attention heads, and a shared vocabulary of 250,054 SentencePiece tokens. mBART-large-50 extends the original mBART [18] with additional language coverage and has demonstrated strong transfer to low-resource translation tasks. We use beam search decoding with beam width $k = 5$ and length penalty 1.0 at inference.

4.2. Parameter-Efficient Fine-Tuning via LoRA

Full fine-tuning of a 611M-parameter model on 416 examples causes severe overfitting: early experiments (not reported here) showed validation loss increasing from epoch 5 onward while training loss continued to decrease. We apply LoRA [3] via the PEFT library [4] to inject trainable low-rank matrices into the

query (q_proj) and value (v_proj) projection layers of every Transformer attention block. All other weights, including the encoder and decoder embeddings, remain frozen.

Formally, for a frozen weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA adds a trainable perturbation $\Delta W = BA$ where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ with rank $r \ll \min(d, k)$. The forward pass computes $h = W_0 x + \frac{\alpha}{r} BA\, x$, where α is a scaling hyperparameter. We set $r = 16$ and $\alpha = 32$, yielding **2.36M trainable parameters** (0.39% of the total), consistent with the LoRA finding that low ranks suffice for small datasets [3].

4.3. Training Details

Table 2 summarises all hyperparameters used in the final submitted system.

Table 2
Hyperparameter settings for the final system.

Hyperparameter	Value
Base model	mBART-large-50
LoRA rank r	16
LoRA α	32
LoRA target modules	q_proj, v_proj
LoRA dropout	0.1
Trainable parameters	2.36M (0.39%)
Optimizer	AdamW
Learning rate	5×10^{-4}
LR schedule	Linear decay
Warmup steps	50
Weight decay	0.01
Gradient clip (max norm)	1.0
Batch size (physical)	8 (Colab T4) / 4 (RTX 2050)
Gradient accumulation	1 (Colab) / 4 (RTX 2050)
Effective batch size	8
Max epochs	40 (baseline) / 60 (BT models)
Early stopping patience	10 (val chrF)
Best model criterion	Highest val chrF
Beam width	5
Length penalty	1.0
Max source / target length	128 tokens
Decoder forced BOS	es_XX
Random seed	42
Hardware	NVIDIA T4 16 GB (main); RTX 2050 4 GB (debug)

4.4. Back-Translation Pipeline

Back-translation is implemented as a two-phase procedure that exploits the bidirectional nature of the task. Figure 1 gives a concise description.

The monolingual Spanish pool $\mathcal{E}$ used as BT source in Phase 1 consists of 344 sentences: 99 manually authored sentences in the same register and domain as the training corpus (everyday topics, family, school, nature), combined with the 245 Spanish sentences from the official SPA→MSLG test input file. Using test-set inputs as unlabelled monolingual data is legitimate because the test inputs are observable at submission time and no gold-standard glosses are used.

Back-translation quality filtering. A common practice is to retain only synthetic pairs whose round-trip chrF exceeds a threshold τ [11]. We experimented with $\tau = 0.1$: only 7 out of 344 candidate

Algorithm 1: Two-Phase Back-Translation

Input: real pairs $\mathcal{D} = \{(\text{MSLG}_i, \text{SPA}_i)\}_{i=1}^{416}$, monolingual Spanish $\mathcal{E} = \{e_j\}_{j=1}^{344}$ (99 manual + 245 test inputs)

Phase 1 — Augmenting MSLG→SPA:

1. Train initial $M_{s2m}^{(0)}$ on $\mathcal{D}$ (SPA→MSLG direction)
2. Generate $\hat{\text{MSLG}}_j = M_{s2m}^{(0)}(e_j)$ for each $e_j \in \mathcal{E}$
3. Filter: retain all 344 pairs (quality threshold = 0.0; see Section 4.4)
4. $\mathcal{D}^{+} = \mathcal{D} \cup \{(\hat{\text{MSLG}}_j, e_j)\}$ $\quad (|\mathcal{D}^{+}| = 760)$
5. Retrain M_{m2s}^{*} on $\mathcal{D}^{+}$ (MSLG→SPA direction)

Phase 2 — Augmenting SPA→MSLG:

6. Generate $\hat{\text{SPA}}_i = M_{m2s}^{*}(\text{MSLG}_i)$ for each pair in $\mathcal{D}$
7. $\mathcal{D}^{++} = \mathcal{D} \cup \{(\hat{\text{SPA}}_i, \text{MSLG}_i)\}$ $\quad (|\mathcal{D}^{++}| = 832)$
8. Train final M_{s2m}^{*} on $\mathcal{D}^{++}$ (SPA→MSLG direction)

Figure 1: Two-phase back-translation procedure.

pairs passed this filter, rendering the augmented set statistically useless. The weak initial SPA→MSLG model (baseline val chrF 44.42) produces noisy synthetic glosses, so round-trip scores are uniformly low regardless of translation quality. Setting $\tau = 0.0$ (no filtering) retains all 344 pairs and leads to the large performance gains reported in Table 3, indicating that even noisy back-translated data provides sufficient training signal to outweigh its noise in this regime.

5. Experiments

5.1. Main Results

Table 3 reports internal validation performance for the baseline and BT-augmented systems, used for checkpoint selection. Table 4 reports the official test set scores and our ranking among participating teams [1].

Table 3
Internal validation results (74 pairs, seed 42). Used for checkpoint selection only; not comparable to official evaluation.

Subtask	System	BLEU-4	chrF
MSLG→SPA	Baseline (no BT)	17.28	49.29
	+ Back-translation	**55.10**	**70.16**
SPA→MSLG	Baseline (no BT)	9.70	44.42
	+ Reverse BT	**49.09**	**68.89**

Table 4
Official test set results for the submitted runs and subtask rankings. Task Score is the average z-score across evaluation metrics (BLEU, METEOR, chrF, and COMET for MSLG→SPA; BLEU, METEOR, and chrF for SPA→MSLG). Global rank (19 teams) is determined by the average of both subtask scores.

Subtask	Submitted run	BLEU	METEOR	chrF	COMET	Task Score
MSLG→SPA	baseline (no BT)	16.81	0.371	46.14	0.699	−0.464
SPA→MSLG	baseline + rev. BT	14.25	0.361	47.09	—	−0.161
Subtask ranks: 13th/20 (MSLG→SPA), 12th/19 (SPA→MSLG)						
Global rank: 13 / 19						−0.312

On the internal validation set, back-translation provides consistent and large improvements on both

subtasks (+20.87 chrF for MSLG→SPA, +24.47 for SPA→MSLG). On the blind test set, both submitted runs achieve below-median performance. Notably, the MSLG→SPA submission was the *baseline without BT*: despite the large validation improvement (+20.87 chrF), the BT-augmented checkpoint was not selected for submission because its validation gains were concentrated on a handful of the 74 held-out sentences, and the model exhibited higher output variance on out-of-split examples during manual inspection. For MSLG→SPA, the submitted baseline generalises modestly (49.29 val vs. 46.14 test, a 3.15-point drop); the higher BT validation score (70.16) was not used for the final submission. For SPA→MSLG, the BT-augmented model shows a larger validation-to-test gap (68.89 val vs. 47.09 test). Both effects are analysed in Section 7.

5.2. Ablation Studies

Learning rate and regularisation. Reducing the learning rate from 5×10^{-4} to 1×10^{-4} with tighter gradient clipping (max norm 0.3) produced results at parity with or marginally worse than the original baseline on both subtasks (MSLG→SPA chrF 48.38 vs. 49.29; SPA→MSLG 42.75 vs. 44.42). An earlier variant that also added label smoothing 0.1 showed further regression and was discontinued. The first training run at 1×10^{-4} also revealed a gradient overflow issue (gradient norm NaN during the second and third epochs), indicating that our chosen 5×10^{-4} schedule with light warm-up is more numerically stable on this architecture and dataset size.

6. Error Analysis

We manually inspected a sample of test set system outputs ($N \approx 40$ per subtask) and categorised errors following the taxonomy of [8]; percentages are approximate estimates from this sample, normalised to sum to 100% for readability. Table 5 and Table 6 show representative examples for each error category.

6.1. MSLG→SPA Errors

Table 5
MSLG→SPA error examples from the blind test set. Errors in system output are underlined. Gold references were not publicly released; error categories are identified by manual inspection.

Input (MSLG)	System output	Category
LICENCIATURA ESTUDIOS #LSM MI HERMANO+MUJER ESTUDIAR	*Mi hermano estudia la licenciatura.*	Agreement (38%)
dm-JOSÉ SU CARÁCTER FUERTE ÉL	*Joshua fue asesinado por su carácter.*	Hallucination (27%)
RÍO dm-NILO ESTE GRANDE ADENTRO ÁFRICA	*El río de Nicolás es muy grande en África.*	Proper noun / dm- (20%)
PRECIO GASOLINA AUMENTAR	*El precio de gallina es muy barato.*	Semantic div. (15%)

Morphological agreement errors (38%) arise most often with compound glosses such as HERMANO+MUJER (literally "male-sibling + female marker"), where the model fails to resolve the compound to its feminine lemma *hermana*, producing incorrect gender agreement downstream. Hallucinations (27%) are triggered by deictic proper-name tokens: dm-JOSÉ is misread as a different name (*Joshua*) and the model then fabricates entirely false narrative content (*fue asesinado*) from its Spanish language prior. Proper-name errors under the dm- prefix (20%) account for a distinct failure mode in which the model recognises the deictic frame but substitutes a phonetically similar wrong referent (*Nicolás* for *Nilo*).

Table 6
SPA→MSLG error examples from the blind test set. Errors in system output are underlined. Gold references were not publicly released; error categories are identified by manual inspection.

Input (Spanish)	System output	Category
El gobierno no autorizó la ley ciudadana.	LEGENDA CIUDADANA EL GOBIERNO NO AUTORIR	Stopword (42%)
Mi cuñada tiene su cabello chino.	MI CUÑA TENER SU CABELLO chino	Case norm. (25%)
Ángela le gusta el ballet desde niña.	dm-AQUELA ELLA GUSTAR BALLET desde NIÑO	Deictic dm- (18%)
El fin de semana mis primos y yo vimos muchas panteras sueltas en Africam Safari.	FIN SEMANA MÍ PRIMO+MUJER ELLA VER MUCHA PANTERA SUELTA	Over-compression (15%)

6.2. SPA→MSLG Errors

Stopword retention (42%) is the most pervasive error: the model copies Spanish function words (*el*, *la*) into the gloss stream and also produces garbled tokens (AUTORIR, LEGENDA) absent from any correct gloss translation, suggesting that both suppression failures and hallucinated tokens co-occur in this category. Case normalisation failures (25%) are systematic and amenable to trivial post-processing (forced uppercasing), representing low-hanging fruit for future improvement; the example also shows partial truncation of inflected forms (CUÑA instead of CUÑADA). Deictic marker errors (18%) manifest differently from the MSLG→SPA direction: the model does use the dm- prefix but attaches the wrong proper-name token (AQUELA for ANGELA) and additionally retains function words and inverts the gender of an adjacent noun (NIÑO instead of NIÑA).

7. Discussion

Why does back-translation help so much? A gain of more than 20 chrF points from adding 344 synthetic pairs (MSLG→SPA) or 416 (SPA→MSLG) to a 416-example training set is larger than typically reported in medium-resource NMT [9]. We attribute this to three compounding factors. First, 416 pairs is below the minimum required for mBART's cross-attention to learn the structural regularity of the gloss language; every additional example, even a noisy synthetic one, substantially reduces this deficit. Second, for MSLG→SPA, the synthetic pairs provide real Spanish sentences as targets, exploiting mBART's strong Spanish generation prior and grounding the model in fluent output that it would not see from 416 training examples alone. Third, for SPA→MSLG, reverse BT provides real gloss sequences as labels, which are the hardest-to-generate output in this direction. The model is thus exposed to many more unique gloss patterns without requiring additional human annotation.

SPA→MSLG as morphological compression. A key qualitative insight from error analysis is that SPA→MSLG is structurally closer to *keyword extraction followed by lemmatization* than to NMT. The core gloss vocabulary largely consists of lemmatized Spanish content words; the primary transformation is the suppression of function words and the normalization of inflectional morphology. LSM-specific annotations (dm- deictics, # fingerspelling) extend beyond the Spanish lexicon, but these are a minority of tokens in the corpus. This means that rule-based post-processing (forced uppercase, stopword removal) could substantially reduce the two most frequent error categories (42% + 25% of identified errors), with the caveat that the stopword category also includes hallucinated tokens (e.g. AUTORIR, LEGENDA) that require model-level correction. Conversely, for MSLG→SPA, the task is genuinely generative: the model must hallucinate inflectional morphology absent from the input, making rule-based approaches insufficient.

BT quality filtering on weak initial models. The failure of the $\tau = 0.1$ quality filter (7/344 pairs retained) has a general implication for BT pipelines: when the initial model is weak (baseline chrF 44.42), round-trip scores are uniformly low and a threshold-based filter discards nearly everything. In this regime, filtering harms rather than helps, and unfiltered noisy BT is the correct choice. This finding is consistent with the recommendation of [11], who proposed tagged BT (marking synthetic sentences with a special token) as a more robust alternative to quality-threshold filtering, particularly when initial model quality is low.

Validation-to-test gap and submission choice. The two submitted runs tell different stories about generalisation. For MSLG→SPA, the submitted baseline (no BT) generalises well: 49.29 val vs. 46.14 test (a 3.15-point drop). The BT-augmented checkpoint reached 70.16 val chrF but was not selected for submission, as manual inspection showed higher output variance on out-of-split examples; in hindsight this conservative decision may have been justified, though the BT checkpoint was never evaluated on the blind test set so a definitive conclusion is not possible. For SPA→MSLG, the submitted BT-augmented model (`bortolotti_baseline_bt_SPA2MSLG`) shows a genuine generalisation gap: 68.89 val vs. 47.09 test (21.8 points). We attribute this gap to the small validation split: 68.89 chrF on 74 examples is driven by a handful of short or structurally simple sentences, and the BT-augmented model appears to have overfit to their specific distribution rather than learning the general SPA→MSLG mapping. These results underscore the well-known difficulty of model selection in extremely low-resource settings [2]: with fewer than 100 validation examples, any metric-based selection criterion carries substantial noise.

7.1. Limitations

- **Small validation set**: all reported results are on a 74-sample validation split. Statistical confidence is limited; a 1–2 chrF change on this set can arise from a handful of examples. Results should be interpreted as directional rather than precise.
- **Training data subset**: the competition released 490 of the 3,000 available Lara-Ortiz pairs. Performance on the full corpus and on out-of-domain data is unknown.
- **Validation leakage in BT models**: the 74 held-out validation glosses appear as targets in the synthetic ($\text{SPA}_{\text{synth}}$, $\text{MSLG}_{\text{real}}$) pairs used to train the BT-augmented SPA→MSLG model (Phase 2 of the BT pipeline, Section 4.4). This subtle leakage likely contributes to the inflated validation scores of the BT models and partially explains the validation-to-test gap.
- **No human evaluation**: all metrics are automatic. Automatic MT metrics are known to correlate imperfectly with human judgment on low-resource and morphologically rich languages [17].
- **Tokenizer-gloss mismatch**: mBART's SentencePiece tokenizer fragments compound gloss tokens. Adding dedicated special tokens for `dm-`, `+`, `#`, and `-` may improve gloss comprehension but was not tested in this work.
- **Gloss-level translation only**: this system takes manually annotated glosses as input. Integration with a video-to-gloss recognition stage is outside the scope of this work.

8. Conclusion

We presented a bidirectional LSM gloss–Spanish translation system based on mBART-large-50 with compact LoRA fine-tuning, submitted to the IberLEF 2026 MSLG-SPA shared task (19 teams overall). Working with only 490 official training pairs, back-translation augmentation yields large internal validation improvements (+20.87 chrF for MSLG→SPA, +24.47 chrF for SPA→MSLG). On the official blind test set our submitted runs achieve chrF 46.14 / BLEU 16.81 (MSLG→SPA, baseline; 13th/20 on that subtask) and chrF 47.09 / BLEU 14.25 (SPA→MSLG, BT-augmented; 12th/19), placing 13th overall by composite z-score. The gap between validation and test performance underscores the instability of metric-based model selection with fewer than 100 validation examples. We adopt compact LoRA (rank 16, q+v modules, 2.36M parameters) and find that quality-threshold filtering of BT pairs is harmful

when the initial model is weak. Error analysis reveals that SPA→MSLG is structurally a morphological compression task, suggesting that rule-based post-processing could correct a majority of errors at zero cost.

Several high-return directions remain for future work:

- **Rule-based post-processing for SPA→MSLG**: forced uppercase normalisation and stopword removal could substantially reduce the two most frequent error categories (case normalisation 25%; stopword retention 42%, partially) at zero additional training cost.
- **Iterative back-translation** [10]: alternating BT rounds between the two subtask models are expected to progressively improve synthetic pair quality.
- **Tagged back-translation** [11]: prepending a `<BT>` token to synthetic source sentences allows the model to distinguish real from synthetic training signal, which may reduce noise amplification.
- **Cross-lingual transfer from ASL**: intermediate pretraining on the ASLG-PC12 English ASL gloss corpus has been reported to lift performance by +23 BLEU on LSM [1].
- **Dedicated gloss special tokens**: adding `dm-`, `+`, `#`, and `-` as vocabulary entries would prevent SentencePiece fragmentation of LSM-specific morphemes.
- **Comparative analysis of competing systems**: the working notes of the other participating teams, once published, will provide insight into which architectures and data strategies drove the top results. In particular, it remains unclear whether the leading systems relied on large language models with few-shot prompting, on fine-tuned NMT models trained on larger corpora, or on hybrid approaches combining neural and rule-based components. A systematic comparison across system descriptions would be valuable for establishing best practices on this low-resource task.

Acknowledgements

The author thanks the IberLEF 2026 MSLG-SPA task organisers for releasing the dataset and evaluation infrastructure. The source code for this work is publicly available at https://github.com/marcoBorto2921/mslg-spa-2026.

Declaration on Generative AI

During the preparation of this work the author used Claude Code (Anthropic) to assist with Python implementation, experiment tracking, and paper drafting. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of this publication.

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, and C. H. Martínez-Rodríguez, "Overview of MSLG-SPA at IberLEF 2026: Bidirectional Translation between Mexican Sign Language Glosses and Spanish," *Procesamiento del Lenguaje Natural*, vol. 77, 2026.

[2] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," in *Proc. 54th Annual Meeting of the ACL*, Berlin, Germany, 2016, pp. 86–96.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. ICLR 2022*, 2022.

[4] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan, "PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods," 2022. [Online]. Available: https://github.com/huggingface/peft

[5] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual Translation from Denoising Pre-Training," in *Findings of ACL-IJCNLP 2021*, 2021, pp. 726–742.

[6] N. C. Camgöz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation," in *Proc. CVPR 2020*, Seattle, WA, USA, 2020, pp. 10023–10033.

[7] M. De Coster, M. Van Herreweghe, and J. Dambre, "Frozen Pretrained Transformers for Neural Sign Language Translation," in *Proc. 1st Int. Workshop on Automatic Translation for Signed and Spoken Languages*, 2021, pp. 1–6.

[8] A. Kermani, H. Irani, and V. Metsis, "Finetuning Pre-trained Language Models for Bidirectional Sign Language Gloss to Text Translation," in *Proc. Workshop on Sign Language Processing (WSLP)*, Mumbai, India, 2025, pp. 73–81. https://doi.org/10.18653/v1/2025.wslp-main.11

[9] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding Back-Translation at Scale," in *Proc. EMNLP 2018*, Brussels, Belgium, 2018, pp. 489–500.

[10] V. C. D. Hoang, P. Koehn, G. Haffari, and T. Cohn, "Iterative Back-Translation for Neural Machine Translation," in *Proc. 2nd Workshop on Neural Machine Translation and Generation*, Melbourne, Australia, 2018, pp. 18–24.

[11] I. Caswell, C. Chelba, and D. Grangier, "Tagged Back-Translation," in *Proc. 4th Conf. on Machine Translation (WMT19)*, Florence, Italy, 2019, pp. 53–63.

[12] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-Efficient Transfer Learning for NLP," in *Proc. ICML 2019*, Long Beach, CA, USA, 2019, pp. 2790–2799.

[13] X. L. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," in *Proc. ACL-IJCNLP 2021*, 2021, pp. 4582–4597.

[14] A. Bonet-Jover, J. Á. González-Barba, and L. Chiruzzo, "Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages," in *Proc. IberLEF 2026 (co-located with SEPLN 2026)*, CEUR-WS.org, 2026.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proc. 40th Annual Meeting of the ACL*, Philadelphia, PA, USA, 2002, pp. 311–318.

[16] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT*, Ann Arbor, MI, USA, 2005, pp. 65–72.

[17] M. Popović, "chrF: character n-gram F-score for Automatic MT Evaluation," in *Proc. 10th Workshop on Statistical Machine Translation*, Lisbon, Portugal, 2015, pp. 392–395.

[18] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual Denoising Pre-training for Neural Machine Translation," *Trans. of the ACL*, vol. 8, pp. 726–742, 2020.