Exploring Pre-Trained Models From the Transformer T5 Family for Bidirectional Spanish-LSM Gloss Translation in a Low-Resource Scenario

Lizeth Hernández[1,*], Ansel Rodriguez[1], Agustín Almaraz[1] and Lino Rodriguez[1]

[1]*Centro de Investigación Científica y de Educación Superior de Ensenada, Andador 10, No. 110, Ciudad del Conocimiento, C.P. 63173, Tepic, Nayarit.*

Abstract

Machine Translation of Mexican Sign Language (MSL) in gloss format constitutes a relevant challenge in natural language processing, especially in low-resource scenarios where existing parallel corpora are scarce, limited in size and not standardized. This work addresses the development of a bidirectional translation tool between Spanish and MSL gloss using a corpus of 489 sentence pairs. The objective is to evaluate the viability of deep learning models in a limited-data context. To achieve this, different versions of the transformer T5 family were compared using 5-fold cross-validation and 5 training epochs. The evaluation was conducted using metrics such as BLEU, METEOR and chrF for both translation directions, as well as COMET for the gloss to text task. Additionally, different text preprocessing strategies were analyzed to measure their impact on performance. The results show that the ByT5-small model achieved the best overall performance for both translation tasks. It is concluded that meaningful results can be achieved even with a reduced corpus, providing evidence for the future development of more complex translation systems for MSL.

Keywords

Bidirectional Translation, Natural Language Processing, Neural Machine Translation, Mexican Sign Language, Sign Language Translation, Transformers

1. Introduction

Machine Translation between natural language and sign language represents an important area within natural language processing due to its potential to improve communication between deaf and hearing people [1]. In this context, gloss functions as a direct transcription of sign language [2] allowing this problem to be addressed as a machine translation task. However, according to [3], gloss-based machine translation presents significant challenges such as the scarcity and limited standardization of parallel corpora.

In the case of Mexican Sign Language, available resources for machine translation tasks continue to be limited, which makes it difficult to train deep learning-based models. Despite these challenges, recent work suggests that it is possible to achieve competitive results by fine-tuning pre-trained models, even in low-resource scenarios [4, 5, 6].

Based on this hypothesis, the present work proposes a bidirectional translation tool between Spanish and LSM gloss using pre-trained models from the Transformer T5 family. The aim of this research is to evaluate the performance of these different models in translation tasks from text to gloss using a corpus of 489 parallel Spanish-LSM gloss pairs. Additionally, different text preprocessing strategies are analyzed in order to identify configurations that can improve the tool's performance.

The content presented in this work forms part of the MSLG-SPA 2026 shared task [7] where the proposed system achieved 17th place out of 19 teams in the global competition ranking.

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

lizeth.hernandez@cicese.edu.mx (L. Hernández); ansel@cicese.edu.mx (A. Rodriguez); almaraz@cicese.edu.mx (A. Almaraz); lrodrigu@cicese.edu.mx (L. Rodriguez)

0000-0001-9971-0237 (A. Rodriguez); 0000-0002-9132-9973 (A. Almaraz); 0000-0002-7541-4772 (L. Rodriguez)

2. Literature Review

Recent approaches to machine translation between text and gloss have focused on the use of deep learning models, particularly encoder-decoder architectures, sequence-to-sequence (Seq2Seq) models, and attention mechanism [8]. However, the scarcity of parallel corpora remains the central challenge for this type of task [3]. [4] define gloss translation as a low-resource machine translation task. In this context, training complex neural architectures from scratch with large volumes of data is impractical. While authors such as [9, 10, 11] propose addressing this problem through synthetic data augmentation strategies, other works, such as [6], demonstrate that fine-tuning pre-trained models can outperform neural architectures trained from scratch. The results reported in these studies indicate that Transformer-based architectures, particularly models such as T5, Flan-T5, mBART, and Llama 3, present significant advantages in scenarios with limited data availability.

In the case of Spanish and Mexican Sign Language, the available research remains limited. [12] compared a rule-based approach with a deep learning model for Spanish-to-LSM gloss translation, observing that rule-based methods can maintain competitive performance when the corpus is small. On the other hand, [13] developed a Spanish-LSM gloss corpus containing 3000 parallel pairs and evaluated pre-trained Transformer models, reporting BLEU scores above 90.

Despite these advances, significant limitations remain regarding the size and availability of parallel corpora, especially in bidirectional tasks and very low-resource scenarios. Furthermore, the comparison between different variants of pre-trained models and the impact of preprocessing strategies remain underexplored in LSM, particularly when using small corpora.

3. Methodology

Based on the previous analysis of the state of the art, it was identified that recent approaches to translation between text and gloss often employ sequence-to-sequence neural models [8, 14]. For the present work, an approach inspired by previous proposals that combine pre-trained neural machine translation models with data transformation strategies to mitigate scarcity of parallel corpora was adopted as a methodological reference [6, 15]. However, because these works do not provide a complete implementation, an adaptation of the pipeline was developed using tools available in the HuggingFace Transformers library within the Visual Studio Code environment. All experiments were conducted on a Lenovo LOQ laptop equipped with an NVIDIA GeForce RTX 5070 GPU, 12 GB of RAM and an i7 Intel processor.

The methodological objective was to construct and evaluate a bidirectional translation tool between Spanish and LSM gloss, divided into the following tasks:

- Text-to-gloss translation (T2G)
- Gloss-to-text translation (G2T)

To achieve this, an experimental approach was implemented, focusing on comparing different pre-trained models from the T5 family, as well as analyzing the effect of different preprocessing and simplification strategies on the input text.

3.1. Evaluated Models

Following the work of [6], four pre-trained models from the Transformer T5 family were evaluated: mT5-small, Flan-T5-small, T5-small, and ByT5-small. These models were selected because they approach natural language processing tasks using a text-to-text architecture (Seq2Seq models), facilitating the translation task. Furthermore, the T5 model family offers versions with sufficiently significant differences in the way they represent and process input text to allow for a controlled comparison between them without completely modifying the methodological framework. The purpose of this comparison was to identify which model was best suited for the translation tasks. Smaller versions of these models were used due to technical limitations.

Table 1
Examples of Spanish-LSM gloss pairs within the corpus used.

Text in Spanish	Text in LSM Gloss
Esa persona es astuta, ten cuidado.	ESA PERSONA COLMILLO TÚ CUIDAR
¿Le pusiste atención a tu maestra?	¿TUYA MAESTRO+MUJER TÚ PONER-ATENCIÓN ELLA?
Luis fue de vacaciones a Panamá.	VACACIONES dm-LUIS IR PANAMÁ
Cristóbal Colón descubrió América hace años.	HACE-PASADO CRISTÓBAL-COLÓN DESCUBRIR AMÉRICA
Sofía se aprovechó del apoyo de sus papás.	dm-SOFÍA SU PAPÁ+MAMÁ APROVECHARSE-DE APOYO

3.2. Evaluation

The performance of the models was evaluated using the following metrics commonly employed in machine translation tasks:

- **BLEU [16]:** A standard metric in machine translation evaluation. It calculates the similarity between the predicted translation and the reference translation based on matches between words (unigrams) and sequences of words (n-grams).
- **METEOR [17]:** Unlike BLEU, this metric considers not only exact matches, but also words with similar forms or meanings, providing more semantic awareness.
- **chrF [18]:** This metric calculates the similarity between the machine-generated translation and the reference using charater n-grams instead of word n-grams.
- **COMET [19]:** Based on pre-trained neural models, it assesses the semantic quality of the result and its closeness to the expected meaning.

The first three metrics were used to evaluate both ranslation directions, while COMET was additionally incorporated for the gloss-to-text task due to its sensitivity to the overall meaning of the sentence, which complements the other metrics based on more superficial matching criteria.

Because cross-validation was implemented in the comparison experiments, this work reports for each configuration, the average metric values obtained across the different folds, as well as the standard deviation. This approach allowed for a more robust representation of the model's performance under each configuration.

3.3. Dataset

The corpus used for both translation subtasks consists of 489 pairs of Spanish sentences and their corresponding LSM (Mexican Sign Language) glosses. This dataset was constructed from the *Diccionario de la Lengua de Señas Mexicana de la Ciudad de México*, provided by [20], from which the sentence pairs were manually collected and selected. Table 1 presents some examples of parallel Spanish-gloss sentences with the corpus.

The annotations presented in the glosses (dashes, signs, prefixes) describe how the person performed the sign or gesture attributed to each word. For example, the symbol '+' between words is used to represent a word composed of two signs. In the second example, the sign for 'MAESTRO' (Teacher) followed by the sign for 'MUJER' (Woman) represents the word 'Maestra' (female teacher). These conventions provide glosses with their characteristic linguistic structure; therefore, it was decided not to modify them during the corpus preprocessing.

3.4. Experimental Design

To ensure a fair comparison, all models were trained under a homogeneous configuration consisting of five training epochs, a 5e-5 learning rate [4], and a maximum input and output sequence length of 32 tokens for most experiments. Two additional experiments explored a maximum sequence length of 64 tokens. Furthermore 5-fold cross-validation was implemented to obtain a more robust estimate of each

model's performance given the small corpus size. From the results obtained for each fold, the average value of each evaluation metric was calculated. Finally, for each translation direction, the configuration with the best average performance was selected and used for qualitative testing.

The experiments were structured in two stages:

- **First Stage: Comparison between architectures.** Experiments were conducted using each T5 family-based model in both translation directions, employing only the corpus with basic cleaning. This stage aimed to identify the model best suited for the translation task.
- **Second Stage: Adjustments to the experiment configuration.** The model that achieved the highest values across the evaluation metrics during the first stage was selected for evaluation under different text preprocessing configurations, as well as variations in the maximum input and output length mentioned above.

3.5. Data Preprocessing

For each data type (Spanish text or gloss), a certain degree of preprocessing (cleaning) was applied depending on the configuration being evaluated. In the case of the Spanish text, a basic cleaning was applied to reduce superficial errors without altering the semantic of the sentences. This included conversion to lowercase, whitespace normalization, and removal of non-alphabetic characters while retaining question marks and exclamation points. From this base version, additional configurations were tested as data augmentation strategies, with the objective of analyzing the effect of these transformations on the model's performance. These strategies included the removal of articles ('el', 'la', 'los', etc.), simplification of certain verbs, and accent removal. The purpose of these modifications was to reduce the grammatical variability of the input text (in this case, Spanish) and approximate its structure to the way the glosses are represented, while also slightly increasing the size of the corpus.

On the other hand, the preprocessing applied to the gloss sentences was minimal. It was decided to retain the existing annotations, as they were considered linguistically relevant, and therefore were preserved during both training and prediction generation. In general, only whitespace normalization was applied, and in some specific configurations, accent removal.

It is important to note that the explored transformations were not always applied simultaneously, but rather evaluated as independent configurations to explore the model's response to each one. Therefore, preprocessing was considered a central component of the experimental design.

4. Results

Tables 2 and 3 present the average results obtained for both translation directions, while Table 4 reports the corresponding standard deviations.

In general, experiments 7 and 8 in which the ByT5-small model was used, obtained the highest values in the evaluated metrics for both translation directions. These results are consistent with previous studies reporting advantages of character or byte level representation for similar tasks [4, 6].

One possible explanation for the superior performance of ByT5-small is, as stated before, due to its byte-level input representation. Unlike T5, Flan-T5, and mT5 which rely on subword tokenizarion, ByT5 processes text directly as sequences of bytes which tend to be longer. This may function better for LSM glosses, since they contain compound signs, prefixes, etc. In this context, ByT5 may preserve more of the original information present in the gloss representation.

Regarding preprocessing strategies, the configurations related to verbal simplification and article removal performed better (higher values in the metrics) in both translation tasks, possibly because they approximate the grammatical structure of the Spanish text to that of the LSM gloss. Under this configuration, the best T2G results were obtained with an average BLEU of 1.874, a METEOR score of 0.156 and a chrF score of 33.89. On the other hand, those strategies based on the elimination of accents did not obtain consistent results. The effect of increasing the maximum input and output sequence length from 32 to 64 tokens was also evaluated using the same experimental settings, but no consistent

Table 2
Results of Text-to-Gloss (T2G) translation experiments

Stage	Experiment	Model	Configuration	Long Max	BLEU↑	METEOR↑	chrF↑
First Etapa	exp1	mT5-small	Basic cleaning	32	0.065	0.018	0.135
	exp3	Flan-T5-small	Basic cleaning	32	0.392	0.132	6.115
	exp5	T5-small	Basic cleaning	32	0.367	0.093	10.379
	exp7	ByT5-small	Basic cleaning	32	1.316	0.121	28.758
Second Stage	exp9	ByT5-small	Basic cleaning + removal (articles)	32	1.368	0.131	31.439
	exp10	ByT5-small	Limpieza + eremoval (articles) y simplificación (verbos)	32	**1.874**	**0.156**	**33.890**
	exp11	ByT5-small	Basic cleaning + removal (articles)	64	1.802	0.144	30.022
	exp12	ByT5-small	Basic cleaning + removal (articles) and simplification (verbs)	64	1.431	0.143	31.033
	exp13	ByT5-small	Basic cleaning + removal (articles and accents)	32	1.093	0.127	27.912

Table 3
Results of Gloss to Text (G2T) translation experiments

Stage	Experiment	Model	Configuration	Long Max	BLEU↑	METEOR↑	chrF↑	COMET↑
First Stage	exp2	mT5-small	Basic cleaning	32	0.075	0.009	4.061	0.285
	exp4	Flan-T5-small	Basic cleaning	32	0.809	0.084	15.432	0.446
	exp6	T5-small	Basic cleaning	32	1.158	0.096	19.673	0.476
	exp8	ByT5-small	Basic cleaning	32	**1.967**	**0.135**	**31.044**	**0.527**
Second Stage	exp14	ByT5-small	accents removal	32	1.747	0.137	31.400	0.530

improvement was observed. It was concluded that, given the size and type of sequences present in the corpus, increasing the maximum sequence length may introduce greater variability between folds, since the sentences are already short.

The removal of accents from both glosses and text was also explored. Although this strategy did not improve performance in the T2G direction, an improvement was observed in the G2T direction compared to the baseline evaluation from the first stage.

Finally, Table 4 reports the standard deviations obtained during both stages of experimentation. It can be observed that the METEOR and COMET metrics showed minimal variability between folds, while BLEU and especially chrF presented relatively high deviations in some of the configurations, particularly in those experiments where ByT5-small was used. This high variability may be attributed to the sensitivity of these metrics to small lexical differences.

4.1. Qualitative Analysis

The best performing text-to-gloss translation model (experiment 10) was compared with the qualitative results reported in the work of [13], as this study represents the closest reference to the present work. In that study, two pre-trained Transformer models for Spanish-to-LSM gloss translation were evaluated using a considerably larger corpus. Because this reference study only addresses the text-to-gloss translation task, the comparison is limited to this direction. Table 5 presents some of the results.

For the Gloss-to-text translation task, Table 6 presents a small sample of the qualitative results obtained from experiment 8.

Examples from both Tables 5 and 6, indicate that the model is generally capable of identifying the main semantic content of the text. However, frequent error include omission of auxiliary verbs, grammatical

Table 4
Standard deviations obtained for each experiment.

Stage	Experiment	Model	Direction	BLEU	METEOR	chrF	COMET
First Stage	exp1	mT5-small	T2G	0.009	0.009	0.029	
	exp2	mT5-small	G2T	0.011	0.006	0.396	0.01
	exp3	Flan-T5-small	T2G	0.094	0.014	0.654	
	exp4	Flan-T5-small	G2T	0.162	0.011	0.869	0.01
	exp5	T5-small	T2G	0.129	0.012	1.020	
	exp6	T5-small	G2T	0.273	0.01	0.753	0.01
	exp7	ByT5-small	T2G	0.385	0.007	4.409	
	exp8	ByT5-small	G2T	0.312	0.03	1.339	0.011
Second Stage	exp9	ByT5-small	T2G	0.386	0.013	1.127	
	exp10	ByT5-small	T2G	0.475	0.02	1.804	
	exp11	ByT5-small	T2G	0.534	0.011	1.806	
	exp12	ByT5-small	T2G	0.527	0.008	1.113	
	exp13	ByT5-small	T2G	0.242	0.011	3.689	
	exp14	ByT5-small	G2T	0.929	0.022	1.822	0.013

Table 5
Qualitative comparison between the models presented [13] and the best t2g model obtained in this work.

Source	Model 1	Model 2	Proposed Model	Reference Translation
Él estudia español	El español estudiar	El español estudiar	TÚ ESTUDIAR SPAÑOL	El español estudiar
Quiero aprender lsm	Yo lsm aprender querer	Yo lsm aprender	LSM QUIERO APRENDIR	Yo lsm aprender querer
Ayer estuve ocupada	Ayer yo ocupada estar	Ayer yo ocupada estar	AYER ESTUVE OCUPADA	Ayer yo ocupada estar
A mi mamá le gusta cocinar	Mamá mía cocinar gustar	Mamá mía cocinar gustar	MI MAMÁ LE GUSTAR	Mamá mía cocinar gustar
La niña es sorda	Niño mujer sorda	Niño mujer sorda	NIÑA SORDAR	Niño mujer sorda

Table 6
Qualitative results obtained in this work for the g2t subtask.

Source in LSM gloss	Proposed Model prediction	Reference Translation
TÚ YA DEBER RESPONSABLE TÚ YA ADULTO	el responsable tú ya deberí	tú ya debes ser responsable ya eres un adulto
PERRO MUCHO SED	el perro muchos sed	el perro tiene mucha sed
ELLA TENER CABELLO-LACÍO	lacío tiene cabello	ella tiene cabello lacio
dm-LUZ BABY-SHOWER MAÑANA	luz baby shower mañana	mañana será el baby shower de luz

and spelling errors and difficulting generating the correct word order.

5. Discussion

The qualitative results show that the proposed model is capable of preserving the main meaning of the sentences for both translation directions. First, for the gloss-to-text subtask, it can be observed that

it presents differences in structure with respect to the reference translations as illustrated by the first example in Table 5. Similarly, in the second example, 'Quiero aprender lsm', the model maintains the key concepts 'LSM', 'quiero', and 'aprender' but modifies their order. Furthermore, it struggles with translations that require more specific grammatical transformations, such as 'La niña es sorda', where the generated translation 'NIÑA SORDAR' reflects these construction errors. This behaviour may be attributed to the limited number of samples available during training.

For the gloss-to-text subtask, the examples illustrated in Table 6 shows several recurring error patterns such as the model omiting auxiliary verbs or grammatical elements as observed in the second example 'PERRO MUCHO SED', where the verb 'tener' is missing or the third example where the prediction omits the subject 'ella'. Another common error is the unnatural word order, such as in 'ELLA TENER CABELLO-LACÍO', where the generated sentence 'lacío tiene cabello' preserves the main concept of the phrase but produces an incorrect structure. Other predictions correspond to nearly literal gloss conversions, as illustrated in 'dm-LUZ BABY-SHOWER MAÑANA', where the structure of the sentence lacks the grammatical transformations necessary for a natural Spanish sentence.

The Quantitive results obtained in this worked were compared with those reported by [13], who obtained BLEU scores of 91.13 and 94.23 when evaluating a Spanish-LSM Gloss corpus of 3000 pairs, a considerable difference can be observed. This variation may be attributed to several factors. First, the corpus used in that study is approximately six times larger than the one used in the present work. Second, their dataset was constructed in a controlled manner based on the grammatical rules of both Spanish and LSM. In fact, the authors explicitly state that their objective was not to reflect real or everyday language use, but rather to create a structured and standardized dataset version suitable for NLP tasks.

In this context, although the absolute BLEU scores obtained in the present work are lower, the experimental setting is also more restrictive. Therefore, the results of this work could be interpreted as an approximation under very low-resource conditions, rather than a direct comparison of performance with a larger and more structured corpus.

In the official evaluation of the MSLG-SPA 2026 shared task, the proposed system achieved 17th place out of 19 participating teams in the overall ranking. For the gloss-to-text subtask, experiments 8 and 14 ranked 43rd and 44th respectively out of 48 submited runs. For the text-to-gloss subtask, experiment 11 and 10 ranked 46th and 47th respectively out of 52 total submited runs.

These results indicate that the proposed approach achieved a relatively low ranking compared with other participants systems. One possible explanation could be the use of more compact neural models. In this work, smaller versions of Transformer-based models were employed due to various computational and technical limitations. However, based on previously reviewed work, it is reasonable to expect that the use of larger models, combined with more sophisticated data augmentation and/or processing techniques, could significantly improve the performance of the translation tool.

Even if the system did not reach the performance levels of the top ranked submissions, which achieved a global score of 1.5489 compared with the global score of -1,2713 obtained in this work, the results demostrate that meaningful translation performance can be achieved even under low-resource conditions.

6. Conclusion

This work addressed the problem of bidirectional machine translation between Spanish and Mexican Sign Language (LSM) glosses in a low-resource scenario using a corpus of 489 parallel sentence pairs. An experimental system based on pre-trained sequence-to-sequence models was implemented and evaluated to identify the architectures and preprocessing strategies best suited for each translation subtask. The results showed that the choice of model significantly influences system performance.

In the initial comparison between architectures, ByT5-small achieved the highest scores across the evaluation metrics in both translation directions, suggesting that byte-level representation are particularly useful for handling the specific characteristics of LSM glosses, such as compound signs,

prefixes, and other special markers present in the corpus. Furthermore, the preprocessing of the Spanish text was found to have a significant effect on model performance.

Simplification based on the removal of items and the partial reduction of verbal variability produced consistent improvements in performance, suggesting that, in low-resource scenarios, data preparation decisions can be as important as the architecture used.

Preprocessing strategies based on article removal and the partial reduction of verbal variability produced consistent performance improvements, suggesting that in low-resource scenarios data preprocessing decisions can be as important as the architecture itself.

When comparing these results with previous work on Spanish-to-LSM gloss translation, it is observed that the absolute metric values obtained in this study are lower. However, this difference should be interpreted considering that the present work was developed with a significantly smaller corpus and under more restrictive experimental conditions.

Future work, should focus on increasing the size of the corpus and exploring higher capacity models, as well as additional data processing and adaptation strategies. Despite these limitations, the present work represents a contribution to the problem of bidirectional translation between Spanish and LSM glosses and provides a useful foundation for the development of more robust systems in the future.

Declaration on Generative AI

During the preparation of this work, the author used Google Translate in order to check grammar and spelling. After using these tool, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] U. Farooq, M. Shafry, M. Rahim, N. Khan, A. Hussain, A. Abid, Advances in machine translation for sign language: approaches, limitations, and challenges, Neural Computing and Applications (2021) 14357–14399. doi:10.1007/s00521-021-06079-3.

[2] T. Ananthanarayana, P. Srivastava, A. Chintha, A. Santha, B. Landy, J. Panaro, A. Webster, N. Kotecha, S. Sah, T. Sarchet, R. Ptucha, I. Nwogu, Deep learning methods for sign language translation, ACM Trans. Access. Comput. 14 (2021). URL: https://doi.org/10.1145/3477498. doi:10.1145/3477498.

[3] B. Haddow, R. Bawden, A. V. Miceli Barone, J. Helcl, A. Birch, Survey of low-resource machine translation, Computational Linguistics 48 (2022) 673–732. URL: https://aclanthology.org/2022.cl-3.6/. doi:10.1162/coli_a_00446.

[4] X. Zhang, K. Duh, Approaching sign language gloss translation as a low-resource machine translation task, in: D. Shterionov (Ed.), Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL), Association for Machine Translation in the Americas, Virtual, 2021, pp. 60–70. URL: https://aclanthology.org/2021.mtsummit-at4ssl.7/.

[5] X. Zhang, K. Duh, Improving Sign Language Gloss Translation with Low-Resource Machine Translation Techniques, Springer Nature Switzerland, Cham, 2024, pp. 219–246. URL: https://doi.org/10.1007/978-3-031-47362-3_9. doi:10.1007/978-3-031-47362-3_9.

[6] A. Kermani, H. Irani, V. Metsis, Finetuning pre-trained language models for bidirectional sign language gloss to text translation, in: M. Hasanuzzaman, F. M. Quiroga, A. Modi, S. Kamila, K. Artiaga, A. Joshi, S. Singh (Eds.), Proceedings of the Workshop on Sign Language Processing (WSLP), Association for Computational Linguistics, IIT Bombay, Mumbai, India (Co-located with IJCNLP–AACL 2025), 2025, pp. 73–81. URL: https://aclanthology.org/2025.wslp-main.11/.

[7] 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.

[8] N. Shahin, L. Ismail, From rule-based models to deep learning transformers architectures for natural language processing and sign language translation systems: survey, taxonomy and performance evaluation, Artificial Intelligence Review 57 (2024). URL: http://dx.doi.org/10.1007/s10462-024-10895-z. doi:10.1007/s10462-024-10895-z.

[9] A. Moryossef, K. Yin, G. Neubig, Y. Goldberg, Data augmentation for sign language gloss translation, in: D. Shterionov (Ed.), Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL), Association for Machine Translation in the Americas, Virtual, 2021, pp. 1–11. URL: https://aclanthology.org/2021.mtsummit-at4ssl.1/.

[10] J. Y. Jang, H.-M. Park, S. Shin, S. Shin, B. Yoon, G. Gweon, Automatic gloss-level data augmentation for sign language translation, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 6808–6813. URL: https://aclanthology.org/2022.lrec-1.734/.

[11] J. Ye, W. Jiao, X. Wang, Z. Tu, Scaling back-translation with domain text generation for sign language gloss translation, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 463–476. URL: https://aclanthology.org/2023.eacl-main.34/. doi:10.18653/v1/2023.eacl-main.34.

[12] J. C. Hernández-Cruz, C. E. Rose-Gómez, S. González-López, Translation of spanish text to mexican sign language glossed text using rules and deep learning, in: Resilience and Future of Smart Learning, Springer Nature Singapore, Singapore, 2022, pp. 233–243.

[13] 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. URL: https://doi.org/10.1038/s41597-025-04871-7. doi:10.1038/s41597-025-04871-7.

[14] E. McGill, L. Chiruzzo, H. Saggion, Bootstrapping pre-trained word embedding models for sign language gloss translation, in: Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), European Association for Machine Translation (EAMT), Sheffield, UK, 2024, pp. 116–132. URL: https://aclanthology.org/2024.eamt-1.13/.

[15] S. M. S, G. R. R, Text-to-isl gloss translation: Bridging language accessibility through nlp, International Journal on Science and Technology 16 (2025). URL: https://doi.org/10.71097/ijsat.v16.i2.5092. doi:10.71097/ijsat.v16.i2.5092.

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

[17] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

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

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

[20] 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).

A. Resources

The complete implementation and code are available via:

- https://drive.google.com/drive/folders/1V4GGOfqYaGs9gXvaw3ZC0xq7506GTB0q?usp=sharing