Low-Resource Sign Language Glossing Profits From Data Augmentation

Vania Lara-Ortiz
Tecnológico de Monterrey
School of Engineering and Sciences
Zapopan, Jalisco, Mexico
A01798098@tec.mx

Sebastian Padó
Institute for Natural Language Processing
University of Stuttgart, Germany
pado@ims.uni-stuttgart.de

Abstract

Glossing is the task of translating from a written language into a sequence of *glosses*, i.e., textual representations of signs from some sign language. While glossing is in principle 'just' a machine translation (MT) task, sign languages still lack the large parallel corpora that exist for many written language pairs and underlie the development of dedicated MT systems. In this work, we demonstrate that glossing can be significantly improved through data augmentation. We fine-tune a Spanish transformer model both on a small dedicated corpus 3,000 Spanish–Mexican Sign Language (MSL) gloss sentence pairs, and on a corpus augmented with an English–American Sign Language (ASL) gloss corpus. We obtain the best results when we oversample from the ASL corpus by a factor of 4, achieving a BLEU increase from 62 to 85 and a TER reduction from 44 to 20. This demonstrates the usefulness of combining corpora in low-resource glossing situations.

1 Introduction

Sign languages (SLs) are visual-gestural languages and the primary means of communication for Deaf communities (Schönström, 2021). Although they serve as a crucial bridge between hearing and deaf people, they remain a understudied area in natural language processing (NLP), which represents an obstacle to diversity and equity (UNESCO General Conference, 2003).

SLs do not have a standardized form in the written modality. Researchers often represent signs through *glosses*: is a notation system used to translate sign language into written form. It is written in uppercase letters and aims to represent the syntactic structure and functioning of the sign language without the interference of spoken language grammar (see Table 1). Glossing helps to describe the semantic, syntatic, and morphological characteristics of SL (Burad, 2008).

Spanish ↔ MSL	Yo quiero comer manzana (I want to eat apple)	YO MANZANA COMER QUERER (I APPLE EAT WANT)
English ↔ ASL	She is studying at the library today	TODAY SHE STUDY LIBRARY

Table 1: Examples of utterances in written language (left) and glossed sign language (right). MSL: Mexican Sign Language, ASL: American Sign Language.

Sign Language Translation (SLT) is the task of translating between sign languages and written or spoken languages. Some approaches translate directly between the two modalities (Camgoz et al., 2018; Hamidullah et al., 2024) but many approaches use glosses as an intermediate representation that breaks up the difficult task into more manageable steps (Chen et al., 2022; Mesch and Wallin, 2015). Glosses, due to their textual nature, also fit naturally into the framework of machine translation for written languages (Müller et al., 2023).

SLT represents a major challenge in the NLP community due to the scarcity of high quality data, particularly parallel corpora. According to (Sennrich and Zhang, 2019), around 1 million parallel sentences are required to effectively train a typical Neural Machine Translation (NMT) model. Existing SL corpora fall far short of this scale: The widely used RWTH-PHOENIX-Weather 2014T (Camgoz et al., 2018) corpus, based on German Sign Language (DGS) weather forecast broadcasts, contains 8,257 parallel German to DGS glosses sentences. ASLG-PC12 (Othman and Jemni, 2012) is an English-American Sign Language (ASL) gloss parallel corpus, containing approximately 87,710 aligned sentence pairs.

In this work, we consider Mexican Sign Language (MSL), the main language used throughout

Proceedings of the Workshop on Sign Language Processing (WSLP), pages 18–23

Mexico among a large segment of the deaf population (Bickford, 1991). It has its own grammatical structure and lexicon, distinct from spoken Spanish. In the NLP community, it remains highly underrepresented due to the absence of corpora. To date, only one parallel corpus is publicly available (Lara-Ortiz et al., 2025). It contains 3000 aligned sentence pairs of Spanish and glossed MSL. Our research questions (RQs) in this situation are:

RQ1 What performance do we achieve for Spanish-to-MSL glossing with a standard NMT model in this low-resource situation?

RQ2. How does this compare to a knowledge-based baseline translation model?

RQ3. Can we improve this setup with data augmentation based on other language pairs?

RQ1 simply establishes the state of affairs for fine-tuning a Transformer-based NLM on the task, using what is currently considered the standard setup for dedicated Machine Translation. Under the assumption that this does not work well, we investigate two different directions. RQ2 asks whether NMT is the most reasonable approach and compares it a simple knowledge-based approach that observes that the use of symbols as glosses (cf. Table 1) can be approximated by lemmatization. RQ3 stays with the NMT paradigm and combines the Spanish-MSL corpus with a larger English-ASL corpus. In doing so, we build on a recent body of work which shows that data augmentation can be crucial in low-resource scenarios to improve the performance of NN models in general (Li et al., 2022) and of NMT models specifically (Haddow et al., 2022).

2 Background: Sign Languages

Sign languages are human languages that arise in Deaf communities and are transmitted primarily through the visual-gestural modality. They exhibit the full range of linguistic structure found in spoken languages, including phonology (e.g., handshape, location, movement, orientation), morphology (e.g., classifier constructions), and syntax (e.g., topicalization). Crucially, sign languages are not derived from or subordinate to the spoken languages of their surrounding communities; for example, American Sign Language (ASL) is genealogically unrelated to English. Cross-linguistic variation among sign languages provides rich data for typological and theoretical inquiry.

Due to the difficulty of reducing sign languages to a written representation, corpora for SLs are much sparser than for spoken/written languages: Most major corpora contain only thousands to tens of thousands of sentences (Kopf et al., 2022).

SLs used in this study. In our study, we focus on the languages already shown in Table 1, namely Mexican Sign Language (MSL) and American Sign Language (ASL). Although ASL and MSL and ASL are distinct SLs with different lexicons and grammars, they are both derived from French Sign Language (LSF) and share structural similarities, such as SOV word order – in contrast to Spanish and English SVO word order. For example, the sentence *Yo como manzana (I eat apple)* in MSL becomes *YO MANZANA COMER (I APPLE EAT)* and its counterpart in ASL gloss stays the same. Due to these similarities, an MSL glossing model should learn generalizable patterns when the training data is augmented with English-ASL glosses.

For MSL, we use the the first and only publicly available parallel corpus for Spanish and MSL glosses (Lara-Ortiz et al., 2025). The corpus consists of 3000 aligned sentence pairs and features simplified gloss annotations. The MSL side is characterized by very short and highly compressed gloss sequences. MSL utterances cluster strongly around 1–5 tokens, with a median close to 3 tokens. This contrasts with the Spanish side, which shows a broader and slightly longer distribution (median $\approx$ 4 tokens). For ASL, we use the ASLG-PC12 corpus (Othman and Jemni, 2012) due to its size (87,710 sentence pairs). It is also widely used for gloss translation tasks in NLP (Cao et al., 2022). Specifically, we used a subset of ASL-PCG12 considering sentences with less than 7 tokens per sentence in the ASL glosses part to reduce the distributional mismatch with our SPA–MSL data. This filtering serves as a normalization step that ensures that the augmented training data are more consistent with the linguistic properties of the MSL sequences present in our primary dataset. Under this condition, 16,900 pairs of English-ASL are left over. A manual inspection of the added ASL samples shows that they come from institutional proceedings (e.g., *"opening of the sitting"*, *"documents received"*, *"there were two further issues raised"*).

3 Methodology

3.1 Base Model for Spanish-Mexican Sign Language Glossing (RQ1)

Our base model for translation from Spanish to MSL glosses (**Base Model** below) is based on BARTO (Bidirectional and Auto-Regressive Transformer for Paraphrasing in Spanish) (Araujo et al., 2024), a Transformer pre-trained on large-scale Spanish dataset. BARTO uses the BART architecture (Lewis et al., 2020), an Encoder-Decoder sequence-to-sequence model trained for paraphrasing. In the absence of NMT models for sign languages, paraphrasing models like BART(O) represent a reasonable starting point for glossing, since they are trained to reformulate sentences while preserving their core meaning. BARTO in particular is well suited for our task, given the lexical similarity between Spanish and MSL glosses, since it is pre-trained on Spanish corpora.

However, Spanish and MSL differ significantly in syntax: while Spanish typically follows a Subject-Verb-Object (SVO) structure, MSL often adopts Subject-Object-Verb (SOV) patterns. Morphologically, MSL glosses do not encode verb conjugations, gender, or number. For these reasons, we fine-tune BARTO using a parallel corpus of 3,000 Spanish–MSL sentence pairs. During fine-tuning, BARTO learns to suppress inflectional morphology and produce outputs that conform to MSL syntax.

3.2 Knowledge-Based Baseline (RQ2)

We also consider a baseline (**Baseline (Lem)** below). It builds on the observation (cf. Table 1) that the symbols used for glossing are generally lemmas of the written language used in the same language community as the sign language. This suggests that simple lemmatization should represent an informed baseline for 'translating' the written language into glosses that can account for the change in lexical material but not the reordering that also takes place.

Concretely, we employ the Spanish lemmatizer provided by decision tree-based TreeTagger (Schmid, 1995) package which computes both part-of-speech tags and lemmas. We employ a small number of postprocessing steps to make the lemma sequence more like sign language glosses: (a), we remove all articles, auxiliaries, reflexive pronouns, and prepositions (which are generally omitted in the gloss sequences); (b) we replace feminine nouns with ending *a* by the masculine noun followed by *mujer (woman)*, again following the MSL conventions, such as *abuela (grandmother)* → *abuelo (grandfather)* + *mujer (woman)*; (c) plurals like *niños (boys)* → *niño ellos (they)*, following a similar convention for plurals; (d) we restore all adjectives, which the lemmatizer changes to masculine, to their original forms. As stated above, we do not adjust word order, and the output still has predominantly the Spanish default SVO structure. The lemmatizer can be improved with rule-based reordering, but this would re- quire a full handcrafted grammar for MSL (e.g., systematic SOV reordering) beyond the scope of this study

3.3 Data Augmentation (RQ3)

Finally, we experiment with a family of models that take the Base Model (fine-tuned on 3k sentence pairs in Spanish – MSL glosses) and incrementally incorporate parallel samples from the English-ASL dataset. As stated above, this experiment tests whether knowledge from a different SL can benefit the translation of another (cross-lingual transfer) despite syntactic differences.

Specifically, we add 3000, 6000 and 9000 ASL gloss sentence pairs – i.e., the same amount as for the original language pair and 2 and 3 times as much, respectively. This leads us to data-augmented models we designate as **DA Model (3+3k), (3+6k), (3+9k)**. The MSL and ASL datasets were simply concatenated.

4 Experimental Setup

To ensure robust evaluation, we partitioned each dataset into 10 equally sized subsets and carried out a variant of 10-fold cross validation, varying the combination of training (80%), validation (10%), and test (10%) subsets across five runs. For example, in the base model using 3000 Spanish-MSL gloss pairs, 2400 samples were used for training and 300 each for validation and testing in each run.

Before training, all data were preprocessed: text was lowercased, extra spaces were removed, punctuation was preserved, and input sequences were tokenized using Sentence Piece tokenization, provided by BARTO. For all experiments, we used the Hugging Face with the following training configuration: a learning rate of 10^{-4}, weight decay of 0.01, and a total of 30 training epochs. The batch size was set to 32 for training and 64 for evaluation. An evaluation was performed at the end of each epoch. We enabled half-precision training to reduce memory usage and speed up computation.

Generation was performed during the evaluation.

We evaluate our models with the standard MT metrics BLEU-1 through BLEU-4 (higher is better) and TER (lower is better). This enables us to measure both glossing quality at the level of individual tokens as well as overall quality.

Our experimental results are shown in Table 2, and example translations in Table 3. We now reconsider the research questions from Section 1.

5 Results

RQ1: Performance of the Base Model. The base model achieves a BLEU-1 score of 0.62, indicating a reasonable unigram performance. The BLEU-n scores however decline substantially for higher n (e.g., BLEU-4=0.35), indicating that the model struggles with longer phrases. This is also shown by the pretty high TER score of 44.2. The examples in Table 3 confirm that the parallel corpus that is available for Spanish–MSL glosses is not sufficient for the model to acquire the syntactic patterns of MSL, neither regarding function words nor word order – the output still looks largely Spanish.

RQ2: Performance of the Baseline. The lemmatizer-based baseline achieves a BLEU-1 score of 0.79, surpassing the base model considerably in unigram precision. This demonstrates again the proximity of glosses in MSL to Spanish lemmas – and in fact, the Baseline outperforms the Base model also substantially on the TER metric. However, the Baseline is basically unable to produce correct longer n-grams, which is expected, since it does not even attempt to capture the word order differences between Spanish and MSL glosses. Table 3 confirms that the Baseline does a fair job for very short sentences (such as pair 2) but not otherwise.

RQ3: Performance of the Data-Augmented Model. The DA model now also outperforms the Lemmatizer baseline on all metrics. However, we observe a clear behavior of diminishing returns: increasing the training corpus to 3+6=9k yields a smaller improvement, and a final increase to 3+9=12k sees essentially unchanged performance. The TER results mirror the behavior we find for BLEU, as do the example translations in Table 3: there is a clear improvement from the base model to the DA mode in terms of syntactic pattern, but then little further adaptation. This is expected, since mixing in more ASL data ultimately causes the model to optimize more towards ASL glossing. Indeed, we consider it a positive result that the AD model's performance on MSL glossing remains stable: Other studies on multi-lingual MT using a comparable setup found that results on a language pair can suffer when too much data for another language pair is added (Johnson et al., 2017).

6 Conclusions

In this paper, we have considered the translation from a written language into sign language glosses, a task that is both important from an equity point of view and difficult to capture with our current standard neural models due to the lack of large corpora. Indeed, our base model does worse than a lemmatizer baseline according to some metrics. Augmenting the training data with a gloss corpus for another (closely related) sign language yields a fair increase in glossing quality, but with diminishing returns for the addition of more data. These findings are largely in agreement with findings of data augmentation methods across a range of tasks (Li et al., 2022) but still do not yield a satisfactory answer to the question of how glossing can be further improved in such low-resource scenarios. Avenues for future research include the creation of synthetic data (see Perea-Trigo et al. (2024) for a rule-based approach) with the challenge of achieving a natural distribution, or alternatively the combination of a lemmatization-based approach—which is very good at generating the correct lexical material—with a reordering strategy to match the sign language's syntactic patterns, e.g., inspired by traditional statistical MT (Durrani et al., 2011). Exploring augmentation with an unrelated language pair, such as German–DGS, also represents a promising direction. Moreover, evaluating additional sequence-to-sequence architectures such as mT5, mBART, and other multilingual pretrained models remains an open line of research.

Acknowledgments. We express our sincere gratitude to the *Grupo Promotor de la LSM* for their guidance and support during the development of this project. We also acknowledge the financial support provided by the German Academic Exchange Service (DAAD) through the program *Research Grants – One-Year Grants for Doctoral Candidates, 2024/25*, funding number 57693452.

7 Limitations

In our study, we considered only two sign languages (with a focus on one of them, namely Mex-

Metric	Base Model (3k)	Baseline (Lem)	DA Model (3+3k)	DA Model (3+6k)	DA Model (3+9k)
BLEU-1	0.62 ± 0.072	0.79 ± 0.068	0.78 ± 0.076	0.84 ± 0.044	0.84 ± 0.045
BLEU-2	0.53 ± 0.106	0.39 ± 0.103	0.69 ± 0.104	0.76 ± 0.063	0.76 ± 0.061
BLEU-3	0.44 ± 0.132	0.17 ± 0.098	0.60 ± 0.105	0.67 ± 0.084	0.67 ± 0.080
BLEU-4	0.35 ± 0.1432	0.08 ± 0.140	0.48 ± 0.097	0.55 ± 0.098	0.55 ± 0.097
TER	44.2 ± 9.56	34.5 ± 10.19	28.7 ± 12.19	21.0 ± 6.76	21.0 ± 6.37

Table 2: Results for Mexican Sign language glossing with the Base, Baseline and Data-Augmented Models

Original (Spanish)	Ellas viven en México (They live in Mexico)	La niña está loca (The girl is crazy)	Tu amiga es distraída (Your friend (female) is distracted)
Original (MSL gloss)	MÉXICO ELLAS VIVIR (Mexico they live)	NIÑO MUJER LOCA (Boy woman crazy)	AMIGO MUJER TUYA DISTRAÍDA ASÍ (Friend woman yours distracted [PARTICLE])
Base Model	Ellas viven en México	La niña be loca	Tu amiga es distraída
Baseline (Lem)	Ellas vivir México	Niño mujer loca	Tuya amigo mujer distraída
DA Model (3+3k)	México ellas vivir	Niño mujer loca	Amiga tuya distraída así
DA Model (3+6k)	México ellas vivir	Niña loca	Amigo mujer tuya distraída
DA Model (3+9k)	México ellas vivir	Niño mujer loca	Amigo mujer tuya distraída así

Table 3: Three example sentence pairs with translations by the different models

ican Sign Language), and only a single neural language model. It remains to be tested to what extent these results generalize to other sign languages and to other NLMs.

8 Ethical Considerations

This project was carried out with the awareness and support of the *Grupo Promotor de la LSM*, a group of Mexican Deaf people and MSL interpreters, whose participation ensured alignment with community perspectives. However, the group cannot represent the full diversity of Mexican Sign Language (MSL), and the dataset may not capture all regional or sociolinguistic variations. Moreover, glossing inherently simplifies the grammatical richness of MSL. Finally, it is important to note that this dataset and any translation systems built from it should complement, but never replace, the work of professional interpreters, since misuse could negatively impact accessibility and the rights of the Deaf community.

References

Vladimir Araujo, Maria Mihaela Trusca, Rodrigo Tufiño, and Marie-Francine Moens. 2024. Sequence-to-sequence Spanish pre-trained language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14729–14743, Torino, Italia. ELRA and ICCL.

J. Albert Bickford. 1991. Lexical variation in Mexican Sign Language. *Sign Language Studies*, 72:241–276.

Viviana Burad. 2008. La glosa: Un sistema de notación para la lengua de señas. *Cultura sorda*.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7793, Salt Lake City, USA. IEEE.

Yong Cao, Wei Li, Xianzhi Li, Min Chen, Guangyong Chen, Long Hu, Zhengdao Li, and Kai Hwang. 2022. Explore more guidance: A task-aware instruction network for sign language translation enhanced with data augmentation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2679–2690, Seattle, United States. Association for Computational Linguistics.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5120–5130.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, Oregon, USA. Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. 2024. Sign language translation with sentence embedding supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–434, Bangkok, Thailand. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. The Sign Language Dataset Compendium: Creating an overview of digital linguistic resources. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association (ELRA).

Vania Lara-Ortiz, Rita Fuentes-Aguilar, and Isaac Chairez. 2025. Spanish to Mexican Sign Language glosses corpus for Natural Language Processing tasks. *Scientific Data*, 12.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Johanna Mesch and Lars Wallin. 2015. Gloss annotations in the Swedish sign language corpus. *International Journal of Corpus Linguistics*, 20.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Achraf Othman and Mohamed Jemni. 2012. English-ASL gloss parallel corpus 2012: ASLG-PC12. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon at LREC*.

Marina Perea-Trigo, Celia Botella-López, Miguel Ángel Martínez-del Amor, Juan Antonio Álvarez García, Luis Miguel Soria-Morillo, and Juan José Vegas-Olmos. 2024. Synthetic corpus generation for deep learning-based translation of Spanish Sign Language. *Sensors*, 24(5).

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.

Krister Schönström. 2021. Sign languages and second language acquisition research: An introduction. *Journal of the European Second Language Association*, 5:30–43.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

UNESCO General Conference. 2003. Recommendation concerning the promotion and use of multilingualism and universal access to cyberspace. Technical report, UNESCO.