Gloss-to-Text Translation for Libras and Portuguese: Evaluating Pretrained and Fine-Tuned Encoder-Decoder Models

João Pedro Tomaszewski[1], **Brenda S. Santana**[1], **Antonielle Martins**[1], **Guilherme Corrêa**[1]

[1] Federal University of Pelotas (UFPel) – Pelotas – RS – Brazil

jptomaszewski@inf.ufpel.edu.br, brendasantana@inf.ufpel.edu.br

an.cantarellim@gmail.com,gcorrea@inf.ufpel.edu.br

Abstract. *We evaluate encoder-decoder models for Gloss-to-Text translation from Brazilian Sign Language (Libras) glosses into Portuguese using a corpus derived from Libras-UFPel. The evaluated models are mT5-small, mT5-base, Flan-T5-base, and PTT5-v2-base. Experiments were conducted with 5-fold cross-validation and evaluated using BLEU and chrF. All models improved after supervised fine-tuning, with PTT5-v2-base achieving the best overall performance. The results suggest that Portuguese-specialized encoder-decoder models are a promising direction for Gloss-to-Text translation in low-resource settings.*

1. Introduction

Sign languages are essential for communication, education, and social inclusion of deaf communities. However, the automatic translation of sign language representations into spoken or written language remains a challenge in natural language processing and sign language translation research [1, 22, 23]. One important formulation of this problem is *Gloss-to-Text* translation, in which a sequence of glosses is converted into a well-formed sentence in a spoken language.

A central difficulty of Gloss-to-Text translation lies in the structural gap between gloss sequences and natural language sentences [1, 2, 5]. Glosses are simplified and linearized representations of signs rather than full linguistic equivalents of spoken-language words. They often omit grammatical markers, function words, and inflectional morphology, while also following syntactic and annotation conventions that differ from those of spoken languages [8, 15]. As a result, translating glosses into fluent text requires not only lexical mapping, but also grammatical reconstruction and contextual inference. This challenge is particularly relevant for Brazilian Sign Language (Libras) and Portuguese, which remain comparatively underexplored in the literature [5, 7].

Previous studies have investigated sign language translation using neural sequence models, transformer-based architectures, and more recently large language models [1, 22, 23, 2]. Although these studies demonstrate the potential of modern neural architectures for sign language processing, most existing work focuses on English, German, or other better-resourced settings. In contrast, Gloss-to-Text translation for Libras and Portuguese remains less studied, especially under low-resource conditions.

Recent advances in text-to-text transformer models have created new opportunities for specialized translation tasks [20, 4, 19, 21, 3]. In particular, encoder-decoder architectures are naturally suited to Gloss-to-Text generation since they explicitly model the mapping between an input sequence and an output sentence. This makes them a promising alternative to earlier decoder-only baselines, which are less aligned with the structural demands of gloss translation.

The objective of this work is to evaluate encoder-decoder text-to-text models for Gloss-to-Text translation from Libras into structured Portuguese. To do so, we compare mT5-small, mT5-base, Flan-T5-base, and PTT5-v2-base under the same preprocessing, prompting, and evaluation setup.

The experiments use a gloss–Portuguese corpus derived from the Libras-UFPel dataset [13] and are evaluated through 5-fold cross-validation using BLEU, chrF, and qualitative analysis of generated outputs.

The main contributions of this study are threefold: (1) an evaluation of encoder-decoder models for Gloss-to-Text translation from Libras glosses into Portuguese; (2) a comparison between multilingual and Portuguese-specialized architectures under the same low-resource setting; and (3) an analysis of translation quality through cross-validation, automatic metrics, and qualitative inspection of generated outputs.

2. Related Work

This section reviews prior work most directly related to Gloss-to-Text translation, with emphasis on sign language translation, gloss-conditioned text generation, and low-resource settings involving Libras and Portuguese.

2.1. Automatic Sign Language Translation

Automatic sign language translation (SLT) aims to convert sign language input into natural language text. Early work by [1] showed that gloss-based intermediate representations can help bridge the gap between sign input and spoken-language output in neural sequence-to-sequence translation. Later, [22] demonstrated that transformer-based architectures can further improve translation quality on sign language benchmarks. More recently, [23] proposed a unified model for sign-to-gloss, gloss-to-text, and sign-to-text learning, showing that cross-task information can improve translation fluency and consistency.

Together, these studies show that sign language translation is an important problem in both NLP and multimodal learning. They also indicate that gloss representations often remain a useful intermediate layer, which makes the Gloss-to-Text stage a relevant problem in its own right rather than only a component of end-to-end SLT systems.

2.2. Gloss-to-Text Translation

Gloss-to-text translation focuses specifically on transforming gloss sequences into coherent sentences in natural language. This is particularly challenging because gloss sequences are typically concise, structurally vague, and do not preserve all the grammatical information necessary for generating fluent text. In a recent study directly centered on this stage, [5] explored the use of large language models for gloss-to-text generation and

proposed semantically aware label smoothing to better handle ambiguity in gloss annotations. Their results show that pretrained language models can be effectively adapted to the task when the training objective is designed to reflect gloss variation.

Another relevant line of work investigates ways of improving fluency and diversity in gloss-conditioned generation. For example, [14] proposed a diffusion-based framework for sign language translation that uses pseudo-gloss guidance to improve output diversity and sentence quality. In a related staged translation setting, [12] employed a transformer-based pipeline with an explicit Gloss2Text component fine-tuned with BART, demonstrating that intermediate gloss representations can remain useful even when the entire translation process starts from a visual input.

Taken together, these studies reinforce the idea that gloss-to-text translation is not a trivial decoding step. Instead, it requires models capable of reconstructing grammatical structure, resolving ambiguities, and producing fluent sentences from compact and often incomplete gloss sequences.

2.3. Low-Resource and Brazilian Sign Language Contexts

In addition to advances in reference datasets for other sign languages, recent work has highlighted the importance of adapting sign language translation methods to resource-limited and domain-specific conditions. In this context, *low-resource* refers primarily to limited linguistic and annotated data, such as the scarcity of large parallel corpora for training and evaluation, rather than to computational hardware constraints alone. For example, [2] showed that large language models can be integrated into gloss-free sign language translation pipelines through factorized learning strategies, while [9] demonstrated the value of visual-language tuning for adapting multimodal large models to sign language translation tasks. These studies suggest that pretrained models can be useful even in specialized settings, provided that they are carefully adapted to the characteristics of the task.

From the perspective of Brazilian Sign Language, scarcity of publicly available linguistic resources remains a major obstacle. A recent effort in this direction is VLibrasBD, introduced by [10], which provides a large bilingual corpus of Libras (Brazilian Portuguese) based on gloss notation for research in neural machine translation. This type of resource is particularly important because translation involving Libras (Brazilian Sign Language) and Portuguese remains much less explored than translation involving American Sign Language, German Sign Language, or benchmark datasets such as RWTH-PHOENIX-Weather 2014 [6].

2.4. Summary and Research Gap

Despite substantial advances in sign language translation, important gaps remain. First, most previous work has focused on feature-rich sign languages or end-to-end sign-to-text translation systems, rather than the gloss-to-text translation step itself. Second, while recent studies have demonstrated the potential of large pre-trained models, there is still little evidence on how multilingual, instruction-tuned architectures perform in feature-limited gloss-to-text translation contexts. Third, research involving Libras (Brazilian Sign Language) and Portuguese is still relatively scarce, especially in experiments that directly compare pre-trained models before and after fine-tuning in a controlled environment.

These gaps motivate the present work. Our study focuses on the translation of gloss sequences in Libras into structured Portuguese and evaluates four pre-trained encoder-decoder models (mT5-small, mT5-base, and Flan-T5-base) before and after fine-tuning. In doing so, we aim to clarify the extent to which task-specific adaptation improves translation quality and which options are most promising in this context.

3. Model Background and Selection

This work focuses on encoder-decoder text-to-text models, since Gloss-to-Text translation is naturally formulated as a sequence-to-sequence task. Compared with decoder-only architectures, such models provide a more suitable inductive bias for learning mappings between compressed gloss sequences and target sentences in Portuguese.

The four evaluated models are **mT5-small**, **mT5-base**, **Flan-T5-base**, and **PTT5-v2-base**. They were selected because they represent different trade-offs between multilingual coverage, Portuguese specialization, computational cost, and instruction-following ability, while remaining compatible with the same preprocessing and evaluation pipeline.

The **mT5-small** and **mT5-base** are multilingual extensions of T5 [19, 21]. The small variant provides a lightweight baseline, whereas the base variant offers greater modeling capacity and serves as the main model of the study. The **Flan-T5-base** is an instruction-tuned version of T5 [3], included to assess whether instruction tuning provides a better starting point for Gloss-to-Text generation than multilingual pretraining alone.

All models were evaluated under the same gloss–Portuguese dataset, prompt structure, preprocessing pipeline, and automatic metrics. In addition, both the original pre-trained checkpoints and their fine-tuned versions were evaluated, allowing the study to measure not only absolute performance, but also the contribution of task-specific fine-tuning for each architecture.

4. Methodology

Figure 1 presents an overview of the experimental pipeline adopted in this study, from raw data preprocessing to model fine-tuning and evaluation.

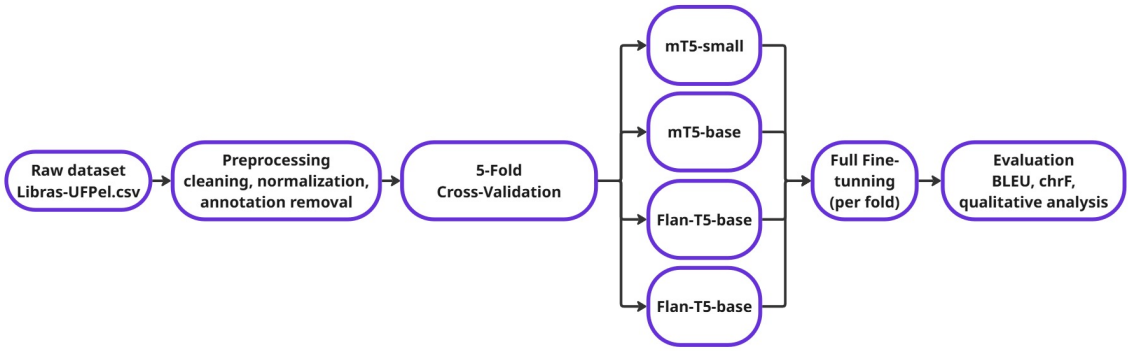

Figure 1. Overview of the experimental pipeline used for Gloss-to-Text translation.

This section describes the methodological framework adopted in this study. We present the dataset and preprocessing pipeline, the full fine-tuning setup used to adapt

pretrained text-to-text models to the Gloss-to-Text task, and the evaluation protocol used to assess translation quality. Together, these components define a controlled and reproducible experimental setting for analyzing the feasibility and limitations of translating Libras gloss sequences into structured Portuguese under data-scarce conditions.

4.1. Dataset and Preprocessing

The dataset used in this study is based on the Libras-UFPel corpus, a parallel resource created to support multimodal and linguistic research involving Brazilian Sign Language (Libras) and Portuguese [13]. The corpus was originally designed to support linguistic and lexicographic documentation rather than machine translation. Each entry is associated with one or more sign variants and includes a gloss sequence, a contextualized example sentence in Portuguese, and a corresponding video reference [13]. The gloss sequences were produced in elicited lexical documentation settings, in which signers generated contextualized examples around lexical items.

For the purposes of this work, which focuses exclusively on Gloss-to-Text translation, only the aligned gloss–sentence pairs were retained. Video links and additional linguistic metadata were excluded during preprocessing, although they remain relevant for future multimodal studies. This use of a corpus originally designed for linguistic documentation is consistent with broader efforts in sign language research to repurpose annotated resources for computational modeling [8, 1].

Table 1. Main statistics of the gloss–Portuguese dataset after preprocessing and split generation. OOV rates are computed with respect to the training vocabulary.

Statistic	Value
Examples before preprocessing	1170
Examples after preprocessing	793
Removed examples	377
Training examples	635
Validation examples	79
Test examples	79
Average gloss length (tokens)	7.49
Average target sentence length (tokens)	10.19
Gloss vocabulary size	1230
Target vocabulary size	1968
Gloss TTR	0.2070
Target TTR	0.2434
Validation gloss OOV rate	10.31%
Validation target OOV rate	14.84%
Test gloss OOV rate	13.63%
Test target OOV rate	18.70%

Table 1 summarizes the main characteristics of the processed corpus. The reduction from 1170 raw examples to 793 final aligned pairs indicates that preprocessing removed a substantial amount of noisy, incomplete, or non-usable material. The resulting dataset is relatively small, with 635 training examples and only 79 examples for validation and test, which reinforces the data-scarce nature of the task.

The statistics also highlight the structural difficulty of the problem. On average, target Portuguese sentences are longer than gloss sequences, indicating that models must

perform not only lexical mapping but also structural expansion. In addition, the target side presents both a larger vocabulary and higher TTR (Type–Token Ratio) than the gloss side, suggesting greater lexical diversity in Portuguese output. Finally, the OOV (Out Of Vocabulary) rates computed with respect to the training vocabulary show that both validation and test sets contain a meaningful proportion of unseen tokens, especially on the target side. This indicates that the models must generalize beyond simple memorization of training pairs.

A key characteristic of this corpus is that the relationship between gloss sequences and Portuguese sentences is not strictly literal or word-by-word. The target sentences often function as dictionary-style usage examples rather than direct translations of the glosses, a property already noted as relevant in gloss-based sign language resources and translation pipelines [8, 5]. As a result, there is a structural mismatch between input and output: gloss sequences encode compressed lexical content and frequently omit grammatical markers, whereas the Portuguese sentences may introduce tense, agreement, function words, and additional contextual material. This makes the task more challenging than standard sentence-level translation and helps explain why model performance must be interpreted in light of the corpus design rather than architecture alone.

The preprocessing phase aimed to standardize and clean the data before model training. The main steps were as follows:

1. **Column selection:** Only the columns corresponding to glosses and structured Portuguese example sentences were preserved.
2. **Data cleaning:** Rows containing missing, empty, or incomplete entries were removed.
3. **Removal of annotation markers:** Internal linguistic markers and variant-specific suffixes, such as tokens beginning with “cl:”, were discarded.
4. **Text normalization:** All text was lowercased and redundant whitespace was removed.
5. **Numerical normalization:** Numeric expressions in Portuguese sentences were converted into written textual form to reduce tokenization inconsistency.
6. **Final dataset generation:** The cleaned corpus was saved as a parallel dataset of gloss–Portuguese pairs and split into training, validation, and test sets.

This preprocessing ensured that the corpus was linguistically cleaner, more consistent, and suitable for supervised Gloss-to-Text learning.

4.2. Full Fine-Tuning Setup

The experiments adopt *full fine-tuning* of encoder-decoder text-to-text models, meaning that all model parameters are updated during supervised training on the gloss–Portuguese parallel corpus. This choice was motivated by both methodological and practical reasons. Full fine-tuning provides a direct adaptation setup for text-to-text transformers [19, 21, 3], remained feasible for the selected model sizes under the available GPU resources, and enabled a cleaner comparison across architectures by keeping the training strategy consistent throughout the study.

4.2.1. Training Configuration

All models were trained through supervised learning on the same gloss/Portuguese parallel corpus. Each input consists of a normalized gloss sequence prefixed by a task-oriented prompt, and each target corresponds to its structured Portuguese sentence. Tokenization follows the native tokenizer of each pretrained model, preserving compatibility with the original text-to-text formulation of T5-based architectures [19, 21, 3].

To ensure comparability across models, we kept the training procedure as consistent as possible. The main experiments used the same general optimization and decoding configuration: maximum source length of 192 tokens, maximum target length of 64 tokens, learning rate of 1×10^{-4}, weight decay of 0.01, warmup ratio of 0.05, beam size of 4 during generation, early stopping patience of 2 evaluation rounds, and random seed 42. Optimization was performed with AdamW [11]. Batch size and gradient accumulation were chosen to balance stability and GPU memory usage, while preserving comparable effective batch sizes across the main runs.

Most main experiments were configured for up to 8 epochs, with early stopping used to prevent unnecessary training after validation performance plateaued. However, in the case of mT5-base, the 8-epoch run stopped only near the upper training limit (epoch 7.95), suggesting that the model might still benefit from a slightly longer schedule. For this reason, we conducted an additional run with a maximum of 12 epochs. This second run produced only marginal gains in chrF and a slight decrease in BLEU, indicating that extending training beyond the original schedule yielded limited practical improvement.

4.2.2. Shared Hyperparameters

Whenever possible, the same hyperparameter configuration was maintained across models. This design reduces confounding factors and makes performance differences easier to attribute to architecture, model capacity, and pretraining strategy rather than to idiosyncratic training choices. Since the goal of the study is comparative rather than hyperparameter-optimization-oriented, controlled consistency across experiments was preferred over aggressive model-specific tuning.

At the same time, the selected values were chosen to remain realistic for the available hardware and for the size of the corpus. Moderate sequence lengths were used to preserve the relevant information in gloss inputs and Portuguese outputs without excessive padding or memory overhead. The learning rate and warmup strategy were chosen to enable stable adaptation in a small-data regime, while early stopping served as a safeguard against unnecessary overtraining.

4.3. Cross-Validation and Evaluation Setup

Since the dataset is relatively small, the experiments were conducted using 5-fold cross-validation to obtain more reliable performance estimates. The dataset was divided into five folds. In each iteration, one fold was used for testing, while the remaining data was used for training and validation.

The same preprocessing steps, prompt structure, optimization settings, and decoding configuration were maintained across all folds. Final results are reported as the

average performance across folds together with the standard deviation.

In addition to mT5-small, mT5-base, and Flan-T5-base, we also evaluated PTT5-v2-base [17], a Portuguese-specialized T5 model. Since the target side of the task is entirely in Portuguese, this model provides an important comparison with multilingual architectures.

4.3.1. Evaluation Metrics

To quantitatively assess translation quality, we employ BLEU (Bilingual Evaluation Understudy) [16] and chrF [18], two standard metrics in machine translation evaluation. BLEU measures n-gram overlap between generated outputs and reference translations, providing an interpretable estimate of lexical and local syntactic similarity. chrF complements this analysis by measuring character-level overlap, which is often particularly informative in morphologically rich languages and in cases where outputs are partially correct but differ at the token level.

Among these metrics, chrF is treated as the primary model-selection criterion in our experiments, since it is more tolerant of surface variation and better captures partial overlap in structured Portuguese generation. BLEU is reported as a complementary indicator. Together, these metrics provide a more informative view of model behavior than BLEU alone, especially in a task where multiple valid realizations may exist for the same gloss sequence.

5. Results and Discussion

This section presents the main quantitative findings and a qualitative inspection of selected translation outputs. Table 2 summarizes the cross-validation results obtained for the evaluated models. Qualitative examples are discussed afterwards to highlight common translation patterns, strengths, and limitations observed across architectures.

5.1. Quantitative Results

Table 2. Cross-validation results for Gloss-to-Text translation. Values are reported as mean ± standard deviation across folds.

Model	BLEU	chrF
mT5-small	5.96 ± 1.19	35.16 ± 1.02
mT5-base	11.97 ± 6.89	38.42 ± 18.39
Flan-T5-base	13.82 ± 1.46	50.42 ± 0.59
PTT5-v2-base	$\mathbf{21.70 \pm 1.65}$	$\mathbf{51.83 \pm 1.27}$

Overall, the results show that all evaluated models benefited from supervised fine-tuning and were able to learn meaningful aspects of the Gloss-to-Text task.

PTT5-v2-base achieved the best overall performance in both BLEU and chrF. This suggests that Portuguese-specialized pretraining can be particularly useful when the target side of the task is fully written in Portuguese. Flan-T5-base also showed strong and stable results across folds.

The multilingual mT5 models improved after fine-tuning, but their results were less stable, especially in the case of mT5-base. Although mT5-base achieved competitive

performance in some folds, its higher standard deviation indicates greater sensitivity to data variation.

Taken together, the results indicate that encoder-decoder text-to-text models are a viable approach for Gloss-to-Text translation in low-resource settings, especially when combined with task-specific fine-tuning and evaluation protocols designed for small datasets.

5.2. Qualitative Analysis

While quantitative metrics provide an overall comparison across models, qualitative inspection is necessary to understand how each architecture reconstructs meaning and grammatical structure from gloss sequences. Tables 3 and 4 present representative examples generated by the evaluated models. Gloss sequences preserve annotation conventions such as "@", indicating gender variation or underspecified forms.

Table 3. Qualitative comparison of evaluated models on Example 1.

Field	Text
Gloss	mesa gat@ abaixo
Reference	o gato está abaixo da mesa
mT5-small	a mesa é abaixo
mT5-base	eu tenho uma mesa de gat@ abaixo
Flan-T5-base	mesa gata está abaixo
PTT5-v2-base	o gato está abaixo da mesa

In Example 1 (Table 3), the task requires recovering a simple spatial relation from a compressed gloss sequence. mT5-small fails to preserve the core meaning, while mT5-base introduces unrelated content. Flan-T5-base partially captures the structure but produces an ungrammatical output. In contrast, PTT5-v2-base correctly reconstructs both the semantic relation and grammatical structure, matching the reference.

Table 4. Qualitative comparison of fine-tuned models on Example 2.

Field	Text
Gloss	então meu prim@ convidar me andar bicicleta mas eu ter medo cair
Reference	meu primo me convidou para andar de bicicleta, mas eu tenho medo de cair
mT5-small	eu convidou meu primo para me andar na bicicleta
mT5-base	eu convidei meu primo para andar na bicicleta
Flan-T5-base	meu primo me convidou em andar um bicicleta, mas eu tenho medo de cair
PTT5-v2-base	meu primo me convidou para andar de bicicleta, mas eu tenho medo de cair

In Example 2 (Table 4), the models must reconstruct a more complex sentence involving multiple events and syntactic expansion. mT5-small produces incorrect structure and agreement errors. mT5-base generates a fluent sentence but alters the meaning. Flan-T5-base better preserves the overall structure but introduces grammatical inconsistencies. PTT5-v2-base again produces the most accurate output, closely matching the reference.

These examples show that the evaluated models are able to extract relevant semantic information from gloss sequences, but differ in their ability to reconstruct fluent and grammatically correct Portuguese sentences. PTT5-v2-base produced the most consistent outputs overall, while Flan-T5-base also generated competitive results. The mT5 variants showed greater variability, especially in more compressed gloss structures.

5.3. Discussion and Error Analysis

The results show that Gloss-to-Text translation remains a difficult task, mainly because gloss sequences omit grammatical and contextual information that must later be reconstructed in Portuguese. This difficulty becomes even more relevant in low-resource settings such as Libras and Portuguese.

The comparison across models also highlights the importance of task-specific adaptation. While multilingual pretraining provides a useful starting point, supervised fine-tuning was necessary for all evaluated architectures.

Among the evaluated systems, PTT5-v2-base achieved the best overall balance between semantic preservation and sentence fluency. This result suggests that Portuguese-specialized pretraining can provide advantages when the target language is entirely Portuguese. Flan-T5-base also remained highly competitive, while the mT5 variants showed greater sensitivity to dataset variation across folds.

Overall, the experiments indicate that encoder-decoder models are a promising direction for Gloss-to-Text translation, although the task still depends strongly on corpus quality and data availability.

6. Conclusion and Future Work

This study investigated encoder-decoder text-to-text models for Gloss-to-Text translation from Libras gloss sequences into structured Portuguese. Using a gloss–Portuguese parallel corpus derived from the Libras-UFPel dataset, we evaluated mT5-small, mT5-base, Flan-T5-base, and PTT5-v2-base under the same preprocessing and evaluation setup.

The results showed that all evaluated models were able to learn meaningful aspects of the task after supervised adaptation. Among them, PTT5-v2-base achieved the best overall performance, suggesting that Portuguese-specialized pretraining can provide important advantages when the target language is entirely Portuguese. Flan-T5-base also produced competitive and stable results across folds.

The experiments also reinforce the difficulty of Gloss-to-Text translation in low-resource settings. Since gloss sequences omit grammatical and contextual information, the models must reconstruct important parts of Portuguese syntax and morphology during generation. In addition, the small size of the dataset and the presence of unseen vocabulary increase the difficulty of the task.

The use of 5-fold cross-validation provided a more reliable evaluation protocol by reducing dependence on a single train/test split. Overall, the results indicate that encoder-decoder fine-tuning is a promising direction for Gloss-to-Text translation involving Libras and Portuguese.

Future work may explore larger gloss–Portuguese corpora, additional Portuguese-focused architectures, and multimodal approaches that reconnect gloss representations with sign-language videos. Other evaluation strategies and prompting approaches may also help improve generation quality in low-resource scenarios.

Acknowledgments

This research was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Finance Code 001, and the Fundação de Amparo à Pesquisa

do Estado do Rio Grande do Sul (FAPERGS), grant No. 25/2551-0000817-4.

References

[1] Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7793.

[2] Chen, Z., Zhou, B., Li, J., and Wan, J. (2024). Factorized learning assisted with large language model for gloss-free sign language translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 7071–7081. ELRA and ICCL.

[3] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

[4] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pages 4171–4186.

[5] Fayyazsanavi, P., Anastasopoulos, A., and Košecká, J. (2024). Gloss2text: Sign language gloss translation using llms and semantically aware label smoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16162–16171.

[6] Forster, J., Schmidt, C., Koller, O., Bellgardt, M., and Ney, H. (2014). Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1911–1916.

[7] Guo, J., Li, P., and Cohn, T. (2025). Bridging sign and spoken languages: Pseudo gloss generation for sign language translation. *arXiv preprint arXiv:2505.15438*.

[8] Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):104–129.

[9] Liang, H., Huang, C., Xu, Y., and Tang, C. (2024). Llava-slt: Visual language tuning for sign language translation. *arXiv preprint arXiv:2412.16524*.

[10] Lima, M. A., Cruz, D., Silva, D. R., Albuquerque, D. D., Lacerda, D. F., Costa, R., Souza Filho, G. L. d., and Araújo, T. M. d. (2025). Vlibrasbd: A brazilian portuguese–brazilian sign language (libras) bilingual text dataset designed to support neural machine translation. *Data in Brief*, 62:111911.

[11] Loshchilov, I. and Hutter, F. (2018). Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

[12] Maia, W. F., Lopes, A. M., and David, S. (2025). Automatic sign language to text translation using mediapipe and transformer architectures. *Neurocomputing*, 642:130421.

[13] Martins, A., Santana, B. S., Martins, F., Lebedeff, T., Nunes, D., and Bohm, L. (2026). Libras-ufpel corpus: A parallel dataset of brazilian sign language and portuguese for multimodal research and processing. In *Proceedings of the 17th International Confer-*

ence on Computational Processing of Portuguese (PROPOR 2026), Salvador, Brazil. Association for Computational Linguistics.

[14] Moon, J., Park, J., Kim, J., and Bae, J. (2024). Diffslt: Enhancing diversity in sign language translation via diffusion model. *arXiv preprint arXiv:2411.17248*.

[15] Padden, C. and Sandler, W. (2015). Lexicalization and variation in sign languages. In *The Oxford Handbook of Deaf Studies in Language*, pages 210–229. Oxford University Press.

[16] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

[17] Piau, M., Lotufo, R., and Nogueira, R. (2024). ptt5-v2: A closer look at continued pretraining of t5 models for the portuguese language. *arXiv preprint arXiv:2406.10806*.

[18] Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

[19] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30:5998–6008.

[21] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

[22] Yin, K., Zhang, Y., and Bowden, R. (2021). Better sign language translation with stmc-transformer. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2845–2854.

[23] Zhang, B., Müller, M., and Sennrich, R. (2023). Sltunet: A simple unified model for sign language translation. *arXiv preprint arXiv:2305.01778*.