Semantic Vector Space Mapping between Brazilian Portuguese and Libras Glosses Using Siamese Models

Lucas dos Santos Pereira[1], **Brenda S. Santana**[1], **Antonielle Martins**[1], **Guilherme Corrêa**[1]

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

lspereira@inf.ufpel.edu.br, bssalenave@inf.ufpel.edu.br

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

Abstract. *This work presents a study towards constructing a shared semantic vector space between Brazilian Portuguese sentences and sequences of glosses representing utterances in Libras (Brazilian Sign Language). We propose a siamese model based on BERTimbau, trained via contrastive learning, employing mean pooling and controlled generation of negative pairs. We evaluate the model on a retrieval task between Portuguese and glosses. Preliminary results indicate that semantic alignment between textual modalities is feasible even with limited corpus size, constituting a first step towards automatic translation systems and semantic indexing for Libras.*

1. Introduction

Research on Sign Language technologies in Brazil has grown significantly over the past two decades, driven by advances in linguistic studies of Libras and by increasing demands for digital accessibility [Costa et al. 2025b]. Several Brazilian studies highlight the importance of textual resources for Libras, particularly noting challenges related to the lack of gloss standardization and the scarcity of large, publicly available annotated corpora [Junior 2024]. These limitations directly hinder the development of robust computational tools, as inconsistent annotations prevent models from learning stable semantic patterns, while data scarcity leads to overfitting and poor generalization across different signers [Bezerra et al. 2024].

To mitigate these data bottlenecks, glosses are widely used in national and international academic contexts as an approximate symbolic transcription of sign languages [Sincan et al. 2025]. Although they do not constitute a formal writing system, these representations function as a practical bridge between sign languages and written languages, enabling initial computational experimentation even under resource-constrained conditions [Xie et al. 2025]. In Brazil, recent studies have leveraged glosses for tasks such as recognition, alignment, and translation [De Martino et al. 2023], demonstrating their utility as intermediate representations in data-driven approaches. These initiatives suggest that gloss-based resources can support the creation of more scalable translation and retrieval models for Libras, especially when video data is limited, expensive, or time-consuming to annotate.

In this context, our work investigates the construction of a semantic vector space between Portuguese and Libras gloss sequences, mapping both modalities into a common distributed representation. The proposal aligns with ongoing Brazilian research efforts

to bridge written/spoken language with sign languages for translation, accessibility, and education [Bezerra et al. 2024]. Establishing a shared embedding space is viewed here as a foundational step for several downstream tasks, including semantic search, modality alignment, and pre-training for sign language translation systems.

To operationalize this mapping, we adopt a siamese neural architecture inspired by classical contrastive learning approaches [Schroff et al. 2015, Khosla et al. 2020] and embedding generation strategies similar to Sentence-BERT [Reimers and Gurevych 2019]. These models are known for their ability to learn robust representations preserving similarity, effectively separating positive and negative pairs in the embedding space, which is essential for modeling semantic relationships between Portuguese text and gloss sequences.

This paper is organized as follows. Section 2 presents related work on sign language technology and contrastive learning; Section 3 describes the dataset and preprocessing procedures; Section 4 details the strategy for generating positive and negative contrastive pairs; Section 5 presents the proposed siamese architecture; Section 6 describes the training setup and optimization details; Section 7 reports the evaluation protocol and results; Section 8 discusses the findings and outlines directions for future work; and Section 9 concludes the paper.

2. Related Work

Brazilian research in computing applied to Libras has progressed in areas such as sign recognition, transcription, linguistic alignment, and the use of glosses for computational modeling [De Martino et al. 2023]. Despite these advances, several studies highlight persistent challenges, including the lack of gloss standardization and the scarcity of annotated corpora, which limit the development of robust systems [Paiva and Costa 2024].

In Portuguese-language NLP, embedding-based approaches are well established, and models like BERTimbau [Souza et al. 2020] capture semantic similarity effectively. However, few Brazilian studies explicitly focus on learning joint representations between Portuguese and Libras glosses, motivating the adoption of contrastive learning methods such as Contrastive Loss [Hadsell et al. 2006], Supervised Contrastive Learning [Khosla et al. 2020], and Sentence-BERT [Reimers and Gurevych 2019], here applied within a siamese architecture tailored to the bilingual scenario.

A key limitation of existing approaches lies in their reliance on either purely textual or purely visual representations, without explicitly modeling the semantic relationship between written language and gloss sequences. While recognition systems focus on mapping video to glosses, and translation systems on mapping glosses to spoken language, relatively little attention has been given to learning shared semantic representations that directly align these modalities at the sentence level. This gap is particularly relevant in low-resource settings, where the lack of large multimodal datasets makes end-to-end approaches difficult to scale [Grossi and Ferreira Filho 2024].

In this context, gloss sequences play an important role as an intermediate representation. By abstracting from the raw visual signal while preserving core semantic content, glosses provide a structured, linguistically grounded bridge between sign languages and written languages. This characteristic makes them especially suitable for representation

learning approaches, where the goal is to capture semantic similarity rather than reconstruct full linguistic form [Silva et al. 2025].

More broadly, recent research in Sign Language Processing (SLP) has increasingly explored multimodal architectures that combine visual and textual representations of sign languages [Krishna et al. 2021]. Many of these approaches rely on pose estimation frameworks such as OpenPose or MediaPipe [Maia et al. 2025] to extract skeletal representations from sign language videos, which are then encoded by deep neural models for recognition or translation tasks [Costa et al. 2025a]. In this context, gloss-based textual representations often serve as an intermediate linguistic layer between visual input and written language output.

A shared semantic embedding space between Portuguese sentences and Libras glosses could therefore function as a bridge between textual NLP models and visual encoders, enabling future multimodal systems that integrate linguistic and visual information. In particular, learning such representations in a purely textual setting provides a controlled baseline for evaluating semantic alignment before introducing additional sources of complexity, such as visual features. In this way, the main contribution of this study is to propose a shared semantic vector space linking Portuguese sentences and their corresponding glosses using a locally produced dataset created by professional translators [Rego et al. 2025].

3. Corpus *Libras-UFPel Corpus*

The dataset used in this study originates from a large-scale bilingual resource developed within an interdisciplinary project involving Libras translators and computer science researchers. The base resource, titled *Libras-UFPel Corpus* [Martins et al. 2026], is a controlled dataset that documents the semantic relationship between Brazilian Portuguese and Libras glosses.

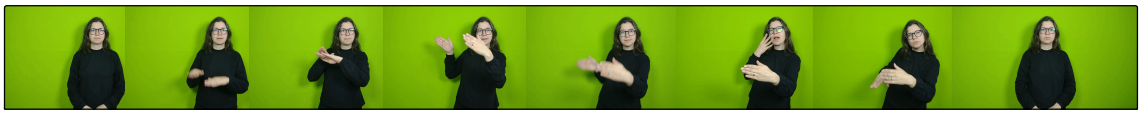

Figure 1. Visual representation of the dataset structure and sign sequences

Unlike purely textual corpora, each record in this collection integrates multiple modalities. As illustrated in Figure 1, the resource provides a direct mapping between visual sign production and linguistic notation. The general quantitative composition of the controlled subset is summarized in Table 1.

Table 1. Quantitative composition of the Libras-UFPel Corpus.

Dataset Component	Quantity / Type
Portuguese Sentences	2,400
Libras Gloss Sequences	2,400
Signers (Native Deaf)	4
Audiovisual Records (Videos)	2,400
Unique Lexical Identifiers (ID-Gloss)	1,913

The original structure of the *Libras-UFPel Corpus* resource, prior to preprocessing, contains essential metadata for both linguistic analysis and computer vision tasks. Table 2 exemplifies the primary columns and the nature of the raw data.

Table 2. Structure and sample instances of the *Libras-UFPel Corpus* resource.

Attribute	Description	Example Entry
GLOSS_ID	Unique lemma identifier	BELOW.1
EXAMPLE	Portuguese and English source sentence	*O gato está abaixo da mesa.* The cat is under the table.
GLOSS	Full Libras gloss sequence	*MESA GATO ABAIXO.1* TABLE CAT BELOW.1
SIGN LINK	Sign video reference	*URL/video_path*
EXAMPLE LINK	Contextual video reference	*URL/example_path*

3.1. Preprocessing

For this experiment, the dataset underwent extraction and filtering. The original resource (internally identified as *Libras-UFPel Corpus*) contained administrative metadata and audiovisual links necessary for linguistic documentation but unnecessary for textual semantic modeling [Khosla et al. 2020].

The preprocessing stage consisted of selecting only the linguistic pairing columns: the Portuguese sentence and its respective gloss translation. Also, preprocessing filters for removal of inconsistencies and normalization were applied. Table 3 compares the data structure before and after this process. The preprocessing stage included a filtering procedure aimed at improving the linguistic consistency and semantic reliability of the dataset.

Table 3. Comparison of data structure between the raw resource and the final dataset.

Attribute	Original Dataset (Libras-UFPel Corpus)	Final Dataset
PT Sentence	Present	Present (Normalized)
Libras Gloss	Present (Multilayer)	Present (Manual Sign)
ID-Gloss	Present (Auxiliary)	Removed
Video Link	Present (URL)	Removed

During manual inspection of the original *Libras-UFPel Corpus* resource, several types of inconsistencies were identified and removed. Entries with incomplete annotations were excluded, including cases where either the Portuguese sentence or the corresponding gloss sequence was missing or truncated. Gloss sequences containing irregular tokens and duplicated identifiers.

Very short or semantically underspecified examples were filtered out, since such instances tend to produce unstable embeddings and provide limited contextual information. Additionally, pairs in which the semantic correspondence between the Portuguese sentence and the gloss sequence was unclear or only partially represented were excluded.

After applying these criteria, the dataset was reduced to a curated subset of 809 validated sentence-gloss pairs. Although smaller, this filtered corpus provides higher semantic consistency, which is essential for training contrastive models that rely on accurate alignment between positive pairs. The resulting curated dataset consists of simplified parallel rows. Table 4 illustrates the final format of the data used for the embedding alignment task.

Table 4. Examples of the validated 809-pair dataset after preprocessing.

Portuguese/English Sentence (Target)	**Libras Gloss Sequence (Source)**
O gato está abaixo da mesa. The cat is under the table.	*MESA GATO ABAIXO* TABLE CAT UNDER
Eu abandonei o curso de inglês. I dropped out of the English course.	*EU ABANDONAR CURSO INGLÊS* I DROP-OUT COURSE ENGLISH
Meu aniversário é no mês de abril. My birthday is in April.	*MEU ANIVERSÁRIO MÊS ABRIL* MY BIRTHDAY MONTH APRIL
Hoje a loja abre às 9:00. Today the store opens at 9:00.	*HOJE LOJA ABRIR HORA NOVE* TODAY STORE OPEN HOUR NINE

3.2. Linguistic Representation

The gloss annotations follow the *ID-gloss* convention, in which lexical signs are represented by stable uppercase labels (e.g., `CASA`, `ESCOLA`). In this format, each Portuguese sentence is aligned with a complete gloss sequence representing the corresponding Libras utterance, rather than isolated lexical items. This structure makes the dataset suitable for sentence-level semantic alignment tasks.

The dataset used in this study consists of bilingual pairs of Portuguese sentences and gloss sequences, forming a textual representation of Libras focused on semantic correspondence. In contrast to token-level representations, this formulation preserves contextual dependencies across the entire utterance, allowing the model to learn relationships at the sentence level rather than relying on isolated lexical matches.

The broader corpus includes richer annotation layers, such as auxiliary identifiers, audiovisual links, and independent tiers for non-manual markers. However, in the present study, only the manual gloss tier was used. Therefore, the model operates on a simplified symbolic representation of Libras, restricted to textual gloss sequences aligned with Portuguese sentences [Martins et al. 2026].

This representation choice is methodologically intentional. By focusing on sentence–gloss sequence pairs, we isolate the problem of semantic alignment between textual modalities before incorporating additional dimensions of sign language structures, such as facial expressions, gaze, body movement, and temporal structure [Camgoz et al. 2020]. This controlled setup allows us to evaluate whether a shared semantic embedding space can be learned from bilingual textual alignments under constrained conditions.

Although this simplified representation does not capture the full multimodal complexity of Libras, it offers important advantages. In particular, it reduces the variability as-

sociated with visual features and annotation inconsistencies, enabling more stable training and clearer interpretation of the learned embedding space. At the same time, it establishes a foundation upon which future multimodal extensions can be built, integrating visual and non-manual information into a unified representation.

4. Contrastive Pair Generation

A central component of contrastive learning is the construction of meaningful training instances that guide the model to organize the embedding space according to semantic similarity. In the context of Portuguese sentences and their corresponding Libras gloss sequences, each training instance is represented as a labeled tuple composed of a sentence, a gloss sequence, and a binary similarity label [Hadsell et al. 2006]. Formally, we define each instance as:

$$(\text{pt}_i, \text{gloss}_j, y),$$

Formally, each instance is defined as $(\text{pt}_i, \text{gloss}_j, y)$, where pt_i denotes a sentence in Brazilian Portuguese, gloss_j corresponds to a sequence of glosses representing a Libras expression, and $y \in \{0, 1\}$ indicates whether the two elements are semantically aligned.

Positive instances correspond to aligned sentence–gloss pairs extracted directly from the dataset:

$$(\text{pt}_i, \text{gloss}_j, 1)$$

where the label 1 indicates semantic correspondence between the Portuguese sentence and the gloss sequence.

Negative instances are constructed by combining sentences and gloss sequences that do not correspond semantically:

$$(\text{pt}_i, \text{gloss}_j, 0), \quad i \neq j.$$

where the label 0 indicates the absence of semantic alignment.

Negatives are selected from within the batch, ensuring diversity and computational efficiency. Although simple, this strategy produces moderately challenging examples, as glosses within the same batch often share contextual similarities [Reimers and Gurevych 2019].

5. Proposed Pipeline for Semantic Vector Space Mapping

The architecture adopted in this work follows the siamese network paradigm, a design widely employed in international research for tasks involving verification, matching, and semantic similarity [Hadsell et al. 2006, Schroff et al. 2015, Reimers and Gurevych 2019, Khosla et al. 2020].

Figure 2 illustrates the main stages of the proposed pipeline. During training, paired Portuguese sentences and Libras gloss sequences are provided as inputs and processed through a shared BERTimbau transformer, which maps both modalities into a common embedding space. The pipeline begins with a preprocessing stage, where sentence–gloss pairs are organized and combined with dynamically generated negative examples. These inputs are then encoded into fixed-size embeddings, which are compared using cosine similarity.

This similarity is optimized through a contrastive learning objective, in which semantically aligned pairs are encouraged to have higher similarity, while non-aligned pairs are pushed apart. The resulting gradients are propagated back through the shared encoder, allowing the model to learn representations that preserve semantic relationships across modalities. After training, the learned encoder is used in a semantic retrieval setting, where a Portuguese sentence is matched against a set of gloss candidates, and performance is evaluated using ranking-based metrics such as Recall@K and Mean Reciprocal Rank (MRR).

After training, the model is used in a semantic retrieval task where Portuguese sentences are matched with gloss embeddings. System performance is evaluated using Recall@1, Recall@5, Recall@10, and Mean Reciprocal Rank (MRR).

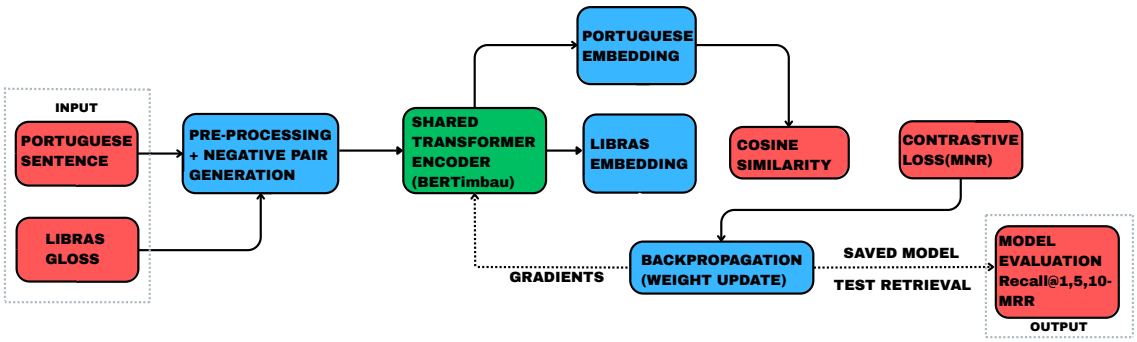

Figure 2. Siamese Model Architecture for Portuguese-Libras Semantic Alignment utilizing BERTimbau

Siamese models are particularly suitable when the objective is to learn a representation space in which semantically related pairs are mapped closer together while unrelated pairs are pushed apart. This property is especially relevant for the Portuguese–Libras scenario, where the goal is to align two modalities with different structures but shared semantic content.

For the encoder, we adopt BERTimbau [Souza et al. 2020], a transformer pretrained on large Brazilian Portuguese corpora. Using a language-specific pretrained model helps preserve the linguistic characteristics of Portuguese during embedding generation. The same encoder is applied to gloss sequences. Although gloss tokens are not part of the original BERTimbau vocabulary, they function as simplified symbolic units that can be effectively modeled through contrastive learning, allowing both modalities to be projected into a shared semantic space.

Sentence-level embeddings are obtained through mean pooling, a strategy widely used in Sentence-BERT models and shown to produce stable representations for short or structurally irregular segments [Reimers and Gurevych 2019]. This is particularly suitable for gloss sequences, which typically follow a telegraphic grammatical structure.

Training uses the *MultipleNegativesRankingLoss* objective, which treats all other examples in the batch as implicit negative samples. This approach improves training efficiency and encourages stronger separation between positive and negative pairs without requiring explicit construction of complex negative pairs [Khosla et al. 2020].

6. Training Setup

The model was trained using supervised contrastive learning with the *MultipleNegativesRankingLoss* objective, which brings aligned sentence–gloss pairs closer in the embedding space while treating other batch examples as implicit negatives.

Fine-tuning was performed on the pretrained BERTimbau encoder using mean pooling to obtain sentence-level embeddings. Optimization was carried out with AdamW using a learning rate of 2×10^{-5}, mini-batches of 16 sentence–gloss pairs, and 3 training epochs. Table 5 summarizes the main training parameters.

Table 5. Training configuration

Parameter	Value
Encoder	BERTimbau
Pooling	Mean pooling
Loss	MultipleNegativesRankingLoss
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	16
Epochs	3
Hardware	RTX 4050 (6GB)

7. Evaluation

The evaluation follows a semantic retrieval protocol commonly adopted in contrastive learning models. Given a Portuguese sentence from the test split, its embedding is compared against all Libras gloss embeddings using cosine similarity. The glosses are ranked according to similarity, and the position of the correct gloss is used to compute the evaluation metrics.

We report Recall@1, Recall@5, Recall@10, and Mean Reciprocal Rank (MRR), which measure how effectively the model retrieves the correct gloss among its closest semantic neighbors. These metrics are widely used to assess the quality of shared embedding spaces in retrieval tasks [Reimers and Gurevych 2019].

Under this protocol, the model achieved Recall@1 of 0.9877, Recall@5 and Recall@10 of 1.0, and an MRR of 0.9938. These results indicate that the siamese model successfully learned a shared embedding space in which semantically aligned Portuguese–gloss pairs are positioned close to each other.

However, these results must be interpreted in light of the training strategy. Negative instances are generated implicitly within each batch, resulting in relatively easy negatives. Consequently, the reported scores likely overestimate the model's ability to discriminate between semantically similar candidates. In this sense, the current experiment should be understood as a baseline demonstrating the feasibility of learning a shared semantic space between Portuguese sentences and Libras gloss sequences.

From a modeling perspective, the learned alignment operates primarily at a lexical-semantic level. The gloss representation used in this dataset corresponds to a simplified, telegraphic encoding of meaning, where many content words are preserved

Table 6. Examples of semantic retrieval between Portuguese sentences and Libras glosses using the trained model.

Portuguese/English sentence	Correct gloss	Top-1	Top-2	Top-3
Um amigo sofreu acidente de carro, por falta de atenção.	*AMIG@ ACIDENTE CARRO CAUSA FALTAR ATENÇÃO*	*AMIG@ ACIDENTE CARRO CAUSA FALTAR ATENÇÃO*	*EU PREOCUPAR PORQUE MEU AV@ CARRO ACIDENTE EL@ TRABALHAR TAXI*	*ALGUNS TER CARRO ALGUNS NÃO TER CARRO*
A friend had a car accident because of inattention.	FRIEND CAR ACCIDENT CAUSE LACK ATTENTION	FRIEND CAR ACCIDENT CAUSE LACK ATTENTION	I WORRY BECAUSE MY GRANDPARENT CAR ACCIDENT HE/SHE WORK TAXI	SOME HAVE CAR SOME NOT HAVE CAR
Os surdos vão no campo de futebol neste sábado.	*BREVE SÁBADO TER SURD@ VIR CAMPO JOGO*	*BREVE SÁBADO TER SURD@ VIR CAMPO JOGO*	*FACULDADE AMPLIAR CURSO ACESSIBILIDADE PESSOAS SURD@*	*AMANHÃ NOITE TER AULA DANÇAR VEM*
Deaf people are going to the soccer field this Saturday.	SOON SATURDAY HAVE DEAF PEOPLE COME FIELD GAME	SOON SATURDAY HAVE DEAF PEOPLE COME FIELD GAME	COLLEGE EXPAND COURSE ACCESSIBILITY DEAF PEOPLE	TOMORROW NIGHT HAVE DANCE CLASS COME
Eu briguei no banco porque estão roubando dinheiro.	*EU BRIGAR BANCO PORQUE ROUBAR DINHEIRO*	*EU BRIGAR BANCO PORQUE ROUBAR DINHEIRO*	*POLICIA PEGAR PESSOA LADR@ EL@ TENTAR ASSALTAR ME*	*ENTÃO ANO PASSADO MEU EX*
I argued at the bank because they are stealing money.	ARGUE BANK BECAUSE STEAL MONEY	ARGUE BANK BECAUSE STEAL MONEY	POLICE CATCH PERSON THIEF HE/SHE TRY ROB ME	THEN LAST YEAR MY EX

but grammatical structure is reduced. As a result, the task resembles semantic matching between two textual modalities rather than full modeling of the linguistic structure of Libras.

Table 6 presents examples from the test set together with the top-3 retrieved glosses. In most cases, the correct gloss appears in the first position, while the remaining neighbors are semantically related rather than random. This behavior suggests that the embedding space captures broader semantic neighborhoods beyond exact lexical overlap. Although Recall@1 is very high, a small number of errors were observed. In these cases, the correct gloss typically appears in the second position rather than the first. A qualitative inspection shows that these errors are not random: the retrieved glosses preserve the overall semantic context but differ in finer aspects such as agency, temporal structure, or discourse focus. This indicates that the model is sensitive to coarse semantic similarity but still struggles with fine-grained distinctions between semantically related candidates.

Figure 3 shows the cosine similarity distribution for positive and negative pairs. Positive pairs concentrate in higher similarity ranges, while negative pairs are mostly associated with lower similarity values, indicating that the contrastive objective successfully separates aligned and non-aligned pairs in the embedding space.

To further analyze the structure of the embedding space, we project the representations into two dimensions using t-SNE (Figure 4). The visualization shows that Portuguese sentences and gloss embeddings occupy overlapping regions, rather than forming

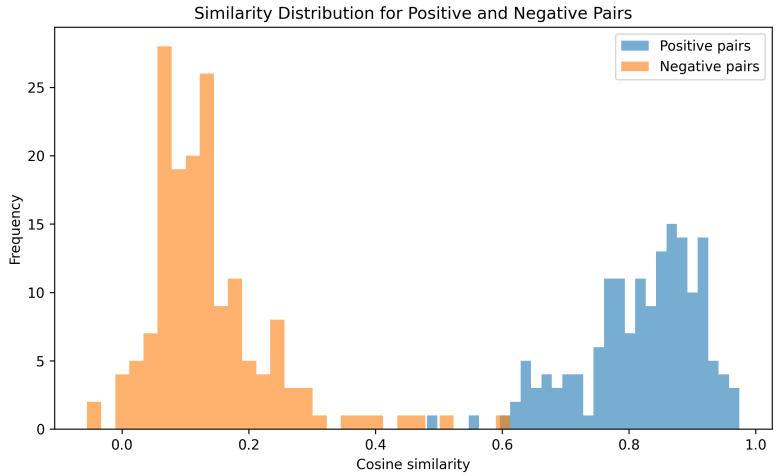

Figure 3. Cosine similarity distribution for positive and negative Portuguese–Libras gloss pairs.

clusters based on modality. This suggests that the model organizes representations according to shared semantic structure rather than language identity.

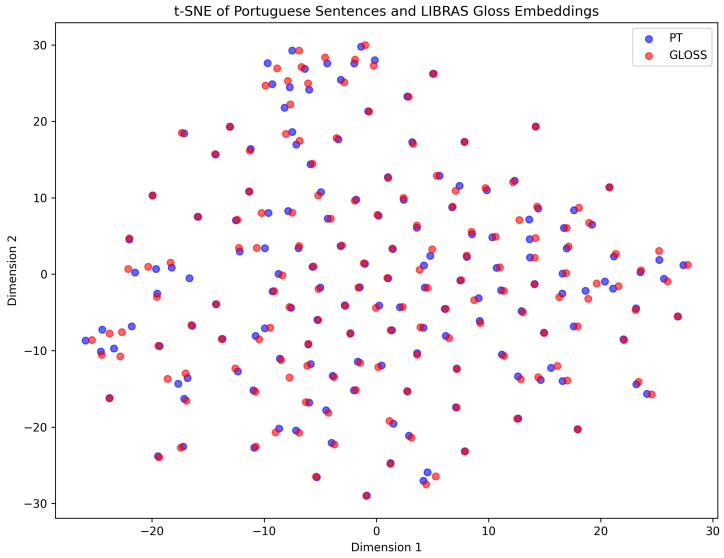

Figure 4. t-SNE projection of Portuguese sentence embeddings (PT) and Libras gloss embeddings (GLOSS).

Finally, it is important to note that the gloss representation used in this subset does not include explicit markers such as classifiers (`CL:`) or pointing signs (`PT:`). As a result, the current evaluation reflects primarily lexical-semantic alignment, rather than deeper structural aspects of Libras. Future work should incorporate richer annotation schemes in order to evaluate the contribution of such linguistic features to semantic representation.

8. Discussion

The results indicate that semantic alignment between Portuguese sentences and Libras glosses can be effectively learned using a siamese architecture based on BERTimbau and

contrastive learning. The high Recall and MRR values show that the model successfully organizes both modalities into a shared embedding space, where semantically aligned pairs are consistently positioned among the closest neighbors.

However, this alignment should be interpreted primarily as lexical-semantic rather than structural. The gloss representation used in the dataset corresponds to a simplified, telegraphic encoding of meaning, preserving key lexical items while reducing grammatical complexity. As a result, the model learns to associate Portuguese sentences and gloss sequences based on shared semantic content, without explicitly modeling deeper linguistic structures of Libras, such as spatial reference, classifiers, or non-manual markers.

Despite this limitation, the learned representation has practical relevance. It can support semantic search in annotated corpora [Paiva and Costa 2024], assist in indexing and retrieval of sign language videos [Baltrušaitis et al. 2019], and serve as an intermediate representation for translation systems [Koller et al. 2019, Camgoz et al. 2020]. In future work, these embeddings may be integrated with visual encoders, enabling multimodal systems that better capture the visuospatial nature of sign languages.

Some limitations of the current setup must be highlighted. First, the dataset presents a long-tail distribution, with many glosses appearing only a few times, which makes it difficult for the model to learn stable representations for rare lexical items . Second, the training strategy relies on in-batch negative sampling, which tends to generate relatively easy negatives. In such cases, the model can often rely on coarse lexical differences to distinguish correct and incorrect pairs, which may inflate evaluation metrics.

To address this limitation, future work should incorporate explicit hard negative mining strategies. One possible approach is to generate controlled negative pairs by minimally modifying semantically aligned examples. For instance, given a positive pair such as

(“I bought a house”, “I buy house”),

a hard negative can be constructed by altering a key semantic component while preserving most of the lexical structure, as in

(“the teacher explained the lesson”, “teacher explain lesson”),

Or by modifying argument structure:

(“I bought a house”, “I sell house”),
(“the student explained the lesson”, “teacher explain lesson”).

Such transformations preserve lexical overlap while introducing semantic inconsistencies, forcing the model to move beyond surface-level matching and capture deeper relational properties such as agency, event structure, and participant roles. Additionally, semantically similar but non-equivalent glosses within the dataset can be used as hard negatives by selecting candidates with high cosine similarity but incorrect alignment. This can be implemented by periodically retrieving the nearest neighbors of each sentence and treating incorrect but highly similar glosses as adversarial training examples.

Together, these strategies would increase the difficulty of the learning task, encouraging the model to learn more discriminative representations that go beyond surface-level lexical matching. This is particularly relevant for sign language processing, where

meaning often depends on structural and visuospatial cues that are not fully captured by simplified textual glosses. From a broader perspective, this work supports Portuguese–Libras semantic alignment in computational systems and contributes to future accessible technologies for the Deaf community, including information retrieval, education, and assistive communication.

9. Conclusion

This work presented an initial study on learning a shared semantic embedding space between Brazilian Portuguese sentences and Libras gloss sequences using a siamese architecture based on BERTimbau and contrastive learning. The results demonstrate that, even with a relatively small curated dataset, the model is able to align both modalities in a coherent representation, enabling effective cross-modal retrieval between Portuguese sentences and gloss sequences.

These findings indicate that semantic alignment between spoken/written language and sign language representations can be achieved through contrastive learning, even when relying on simplified gloss-based encodings. Rather than modeling full linguistic structure, the approach captures core semantic correspondences, providing a practical and scalable solution under conditions of limited data availability.

Beyond its technical contribution, this research is directly connected to broader efforts toward digital accessibility for the Deaf community in Brazil. Establishing shared semantic representations between Portuguese and Libras is a fundamental step toward the development of tools for semantic search, accessible information retrieval, educational support, and future translation systems adapted to the linguistic characteristics of Brazilian sign language. In this sense, the work contributes not only to Natural Language Processing, but also to the advancement of inclusive technologies.

The success of this initial alignment is strongly linked to the interdisciplinary nature of the project, combining linguistic expertise with computational modeling. The dataset used in this study was curated by professional Libras translators, ensuring semantic consistency and preserving key aspects of the language. The results also indicate that the model captures semantic relations beyond simple lexical matching, although it does not explicitly model structural elements of Libras such as classifiers or pointing markers.

Future work should focus on expanding the dataset, incorporating harder negative samples, and integrating visual encoders to move toward multimodal representations that better capture the visuospatial nature of Libras. Such advances are essential for developing more robust, generalizable, and linguistically informed models, ultimately contributing to more effective and inclusive AI technologies.

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

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Bezerra, E. T., Oliveira, I. d. S., Fonseca, J. R. M. d., Celestino, E. M., Caitano, T. F., Vieira, A. J. F., Nascimento, R. d. S., Chaves, J. P. d. A., Monteiro, A. L., Pinheiro, J. C. M., et al. (2024). Tecnologias assistivas para o ensino de libras: soluções inovadoras para a educação inclusiva. *Revista Foco*, 17(11):e6576.

Camgoz, N., Hadfield, S., Koller, O., and Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.

Costa, V., Tavares, L., Conceição, R., Agostini, L., Santana, B., Lebedeff, T., and Corrêa, G. (2025a). A lightweight i3d-based approach for real-time brazilian sign language recognition. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web*, pages 501–505, Porto Alegre, RS, Brasil. SBC.

Costa, V., Tomaszewski, J. P., Pereira, L., Santana, B. S., and Corrêa, G. (2025b). Ai-based approaches for brazilian sign language recognition: A systematic literature review: Insights on methods, metrics, and resources for libras recognition. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web (WebMedia 2025)*, WebMedia 2025, page 585–597. Sociedade Brasileira de Computação - SBC.

De Martino, J. M. et al. (2023). Neural machine translation from text to sign language. *Universal Access in the Information Society*, pages 1–12.

Grossi, V. S. and Ferreira Filho, B. S. (2024). Aplicação de técnicas de reconhecimento de imagens na classificação de sinais em libras (linguagem brasileira de sinais) para tradução em texto.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1735–1742.

Junior, A. K. (2024). *K-LIBRAS: SISTEMA DE CONHECIMENTO PARA A TRADUÇÃO DA LÍNGUA BRASILEIRA DE SINAIS (LIBRAS/GLOSA)*. PhD thesis, UNIVERSIDADE FEDERAL DE SANTA CATARINA.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673.

Koller, O., Ney, H., and Bowden, R. (2019). Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5969–5978.

Krishna, S., P, V. V., and J, D. B. (2021). Signpose: Sign language animation through 3d pose lifting. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2640–2649.

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

Martins, A. C. et al. (2026). Corpus libras-ufpel: A parallel dataset of brazilian sign language and portuguese for multimodal research and processing. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR 2026)*, Salvador, BA, Brazil. To appear in ACL Anthology.

Paiva, R. and Costa, O. (2024). Avaliação do uso de modelos de aprendizagem profunda na tradução automática de línguas de sinais.

Rego, R. C., de Morais, L. M., and Almeida, W. M. (2025). Brazilian sign language recognition using deep learning based on fast fourier transform and kinematic features. *IEEE Access*, 13:202875–202892.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP*, pages 3982–3992.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823.

Silva, A., Araújo, S., Farias, M., Santos, I., Pessoa, A., and Junior, G. B. (2025). Reconhecimento inteligente de sinais em libras: Um modelo computacional assistivo para educação inclusiva em ambientes educacionais. *RENOTE*, 24(2):547–558.

Sincan, O. M., Low, J. H., Asasi, S., and Bowden, R. (2025). Gloss-free sign language translation: An unbiased evaluation of progress in the field. *Computer Vision and Image Understanding*, 261:104498.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417.

Xie, Y., Su, W., Zhong, C., Cai, C., and Yuan, Y. (2025). Gloss-free sign language translation based on fusion attention. *Applied Soft Computing*, 185:113848.