Libras-UFPel Corpus: A Parallel Dataset of Brazilian Sign Language and Portuguese for Multimodal Research and Processing

Antonielle Martins [1], Brenda S. Santana[2], Francielle Martins[1], Tatiana Lebedeff[1], Darley Nunes[1], Luisa Bohm[3],
[1]Center for Letters and Communication (CLC – UFPel), [2]PPGC – UFPel,
[3]Technology Development Center (CDTec – UFPel)
Correspondence: an.cantarellim@gmail.com

Abstract

The Libras-UFPel Corpus is a multimodal, multilayer parallel resource designed for the documentation and computational analysis of Brazilian Sign Language (Libras) in systematic alignment with written Portuguese. By integrating controlled recordings with naturalistic data from the *Inventário Nacional de Libras-Pelotas*, the corpus ensures interoperability through shared methodological standards. The dataset currently comprises 4,800 controlled audiovisual records (2,400 sentences and 2,400 isolated signs) fully paired with Portuguese translations, supplemented by approximately 10 hours of spontaneous interaction from three new naturalistic interviews, currently in the editing phase. To date, 1,200 controlled sentences have been lemmatized, gloss-annotated and translated, providing a structured parallel subset for Libras-to-Portuguese Sign Language Processing tasks such as recognition and machine translation. The annotation model follows a hierarchical structure covering lexical, partially lexical, and non-lexical signs, including independent tiers for non-manual markers. By bridging descriptive linguistics and Natural Language Processing, Libras-UFPel Corpus serves as a reference source for bilingual data-driven modeling, advancing digital inclusion and linguistic accessibility.

1 Introduction

Sign language linguistics emerged as an international field following the pioneering structural analysis of American Sign Language (ASL) by Stokoe (1960). In Brazil, research on Língua Brasileira de Sinais (Libras) was established by Brito (1984) and has since been extensively documented in the comprehensive grammars by Quadros et al. (2023a,b). While these works systematized categories like classifiers and non-manual markers, many visuospatial particularities of Libras remain under-described, especially from computationally applicable perspectives.

The field of Sign Language Processing (SLP) seeks to bridge this gap by aligning linguistic principles with machine learning (Bragg et al., 2019). As Yin et al. (2021) argue, integrating sign languages into Natural Language Processing (NLP) requires an epistemological shift to recognize them as complete, multimodal systems. Beyond this theoretical challenge, there is an urgent social demand: despite robust legal frameworks (Laws 10.436/2002, 13.146/2015) and a population of millions of signers IBGE (2010), the development of assistive tools is constrained by a primary bottleneck: the scarcity of high-quality, interoperable datasets (De Coster et al., 2024; Niu et al., 2024).

To address these challenges, we present the Libras-UFPel Corpus, an interdisciplinary dataset that adopts a hybrid architecture comprising 4,800 controlled records (2,400 sentences and 2,400 signs) and approximately 10 hours of naturalistic data from the *Inventário Nacional de Libras* (INL-Pelotas). The Libras-UFPel Corpus adopts a multilayer architecture aligned with the annotation guidelines defined by Johnston (2024) and the computational principles for sign language translation discussed by De Martino et al. (2023). By incorporating the methodological frameworks for multimodal datasets proposed by Niu et al. (2024) and De Sisto et al. (2022), this work seeks to contribute to the bridge between descriptive linguistics and NLP, providing a structured reference source for bilingual research in Libras and Portuguese.

2 Related Works

The development of structured linguistic resources is essential for advancing SLP. As noted by Bragg et al. (2019), SLP faces unique challenges compared to spoken languages due to its inherently multimodal nature, requiring the integration of computer vision and natural language processing. Yin et al. (2021) further argue that including sign lan-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1068–1073

guages in the broader NLP landscape requires an epistemological shift, by recognizing them as complete, culturally situated systems rather than lexical translations.

In Brazil, the *Inventário Nacional de Libras* (INL) stands as the primary methodological reference, documenting regional and sociolinguistic diversity through naturalistic recordings (Quadros et al., 2020). These records provide high-quality multilayer data necessary for complex linguistic analysis. However, as emphasized by Crasborn and Sloetjes (2008), video annotation remains a major bottleneck in the field, often requiring 20 to 100 hours of labor for every hour of video, depending on the granularity of the tiers (e.g., mouthings, eye gaze, and torso orientation).

The gold standard for such documentation is ELAN[1] (Wittenburg et al., 2006), a professional annotation software designed for multimodal recordings that facilitates time-aligned, hierarchical layers (tiers) of transcription. Despite the widespread use of ELAN, many current SLP datasets suffer from domain bias, being restricted to highly scripted environments like weather forecasts (Yin et al., 2021).

Such datasets often fail to capture the complexity of naturalistic signing, particularly regarding non-manual markers (NMMs). Far from being mere affective expressions, NMMs, which encompass facial expressions, head movements, and upper body leans, function as a fundamental, parallel linguistic channel that occurs simultaneously with manual signs. For instance, specific eyebrow positions signal question types, such as raised eyebrows indicating a yes/no question, while body shifts establish spatial references. These markers are essential for conveying grammatical structures and prosodic nuances of sign language that are frequently absent or oversimplified in the more rigid signing styles found in scripted datasets (Yin et al., 2021).

To bridge this gap, recent research has focused on Neural Machine Translation (NMT) from text to sign and vice-versa. De Martino et al. (2023) and De Coster et al. (2024) highlight that the lack of standardized, large-scale parallel corpora is the primary obstacle to achieving state-of-the-art results in NMT. Furthermore, the need for semi-automatic annotation tools is becoming critical. Mukushev et al. (2022) demonstrate that AI-assisted pipelines can significantly reduce manual labor by providing preliminary gloss suggestions, provided the training data follows rigorous standards like those established by Johnston (2019).

The Libras-UFPel Corpus addresses these gaps by combining controlled elicitation with naturalistic INL-Pelotas data. By employing a standardized multilayer framework (video $\leftrightarrow$ gloss $\leftrightarrow$ translation), our project integrates linguistic theory with computational applicability, ensuring that the resulting models are grounded in the actual sociolinguistic reality of Libras users.

3 Methodology

Ethical Considerations: This study was approved by the Research Ethics Committee of the *Federal University of Pelotas* (CAEE: 85511424.6.0000.5317). All participants signed an informed consent form (TCLE) authorizing their participation and the public dissemination of the recordings for research purposes.

The methodology was designed to ensure traceability, standardization, and replicability, combining corpus linguistics techniques with data engineering protocols. The collaboration between Libras and Computer Science researchers aims to create a resource for both descriptive and computational studies, significantly enhancing its robustness through the integration of data from the INL-Pelotas project.

3.1 Controlled and Naturalistic Data Collection

The Libras-UFPel Corpus integrates two complementary data streams to capture the phonological, morphological, and discursive complexity of sign language:

Controlled Elicitation: This subset consists of 2,400 sentences and 2,400 isolated trigger signs. Four native deaf signers participated in elicitation tasks where a trigger sign was presented, and the participant produced a natural, contextualized sentence. To ensure visual consistency for computer vision tasks (pose estimation and feature extraction), recordings were made in a controlled studio with a fixed frontal camera (60 fps), neutral background, and uniform three-point lighting to minimize shadows.

Naturalistic Data INL-Pelotas: Following the INL methodology (Quadros et al., 2020)—aligned with international standards such as the *DGS-Korpus* (Germany) and *Corpus NGT* (Netherlands)—this subset targets 18 interviews conducted in pairs (totaling 36 participants) of 3–

[1]Available in: https://www.mpi.nl/corpus/html/elan/

4 hours each. Data is captured via a four-camera setup (one frontal, two lateral, and one zenithal), providing crucial information for 3D pose estimation and spatial referencing. While several Brazilian universities already maintain established INL records, the INL-Pelotas expands this standardized framework. As these records are publicly available, they provide a significant open-access foundation for future large-scale data aggregation and machine learning training. Currently, 10 hours from three interviews are in the processing phase.

3.2 Multilayer Annotation and Hierarchical Model

The annotation process is currently being conducted in ELAN (Wittenburg et al., 2006), following a hierarchical tier structure and the ID-gloss strategy. This model is adapted from the guidelines of Johnston (2019, 2024) and is consistent with the methodological framework of De Martino et al. (2023). Furthermore, the process is guided by an internal INL-Pelotas document to ensure interoperability across the datasets.

This architecture ensures interoperability with global sign language datasets and allows for time-aligned, multi-modal analysis where child tiers (e.g., NMMs) inherit the temporal boundaries of parent tiers (Glosses).

The framework distinguishes three categories of signs, each with specific metadata requirements:

- **Lexical signs:** Represented by a unique, invariant label in uppercase (ID-gloss). Lexical variants are indexed (e.g., `CASA.1`, `CASA.2`) to facilitate the training of recognition models that must distinguish between different realizations of the same lemma without losing lexical consistency (Martins et al., 2023).
- **Partially lexical signs:** This category includes classifiers (`DV:`) and pointing signs (`IX:`).
  - **Classifiers:** Annotated using a semantic-descriptive approach, identifying the specific action or state depicted (e.g., *DV:ABRIR_LIVRO*). They represent a complex mapping of spatial and movement properties that are essential for high-level semantic understanding in SLP.
  - **Pointing signs (deictics):** Documented by describing their *locus* in the signing space and their grammatical function. This includes identifying specific referents, establishing pronominal points (e.g., *IX-mulher*, *IX-nós*), or indicating spatial-temporal relations. Mapping these deictic markers provides the necessary ground truth for tasks such as anaphora resolution and entity linking, bridging the gap between physical movement and linguistic reference.
- **Non-lexical items:** Spontaneous gestures or affective expressions are annotated with the prefix `GES:`. This labeling is crucial to isolate "gestural noise" from formal grammatical structures, preventing models from misclassifying spontaneous movements as linguistic units.

Beyond manual signs, the model incorporates specialized tiers for complex linguistic features. Fingerspelling is identified with the `FS:` prefix (e.g., FS:amanda), while incorporated negation is marked to capture morphological shifts (e.g., *NÃO_GOSTAR*).

The analysis of NMMs is conducted through independent, synchronized tiers for eyebrows, eye gaze, and mouthings. This granular separation allows for the study of simultaneity—a core feature of sign languages where grammatical information (such as interrogative or negative prosody) is conveyed through facial expressions while the hands execute lexical items. Finally, dedicated tiers for body and torso movements capture the use of signing space and perspective changes role shift, providing essential data for studying spatial reference and discursive cohesion in Libras.

3.3 Processing, Alignment, and Quality Assurance

The processing pipeline follows a collaborative curation workflow. Initial manual transcription and translation are performed in spreadsheets by pairs of linguists, including at least one deaf native signer. This ensures that Portuguese translations maintain semantic and pragmatic equivalence with the source sign language, avoiding literal translations that might obscure its natural syntax.

Data is then migrated to ELAN for temporal alignment. Following Crasborn and Sloetjes (2008), we define the onset as the first frame of the relevant hand configuration and the offset as the

return to a neutral position or the beginning of the next sign.

To ensure internal consistency and dataset reliability, this rigorous manual validation cycle focuses on the agreement of Gloss-IDs and the precision of temporal boundaries. Currently, 1,200 sentences have completed this process, providing a high-confidence gold standard. Although this careful peer review ensures data quality, it renders the expansion of the corpus to a massive scale both slow and costly. To mitigate this limitation in future work, we intend to explore semi-automatic annotation strategies, such as human-in-the-loop approaches, to accelerate the growth of the data set without compromising linguistic accuracy (Monarch, 2021).

3.4 Potential Applications in SLP Tasks

The multilayer architecture of the Libras-UFPel Corpus is intended to serve as a reference source for various SLP tasks, supporting the development of models that explore linguistic features beyond video-to-text translation. The controlled subset is designed to assist in Continuous Sign Language Recognition (CSLR) by providing temporal alignment, which may facilitate the use of Connectionist Temporal Classification (CTC) losses and spatial feature extraction through frameworks such as MediaPipe (Bragg et al., 2019). Regarding NMT, the alignment between glosses and Portuguese translations seeks to provide a basis for addressing structural divergences between the two languages (De Martino et al., 2023; De Coster et al., 2024). Furthermore, by including independent tiers for NMM such as eyebrow movement and eye gaze, this resource aims to complement studies focused on manual features, potentially assisting in the training of classifiers for grammatical facial expressions.

Finally, integrating these records into semi-automatic, human-in-the-loop workflows streamlines the curation process by providing model-based suggestions for manual review (Mukushev et al., 2022). This high-confidence dataset subsequently serves as the foundation for training fully automatic annotation pipelines, significantly reducing the labor required for large-scale linguistic documentation and continuous corpus expansion.

4 Preliminary Results

The initial phase of the Libras-UFPel Corpus project demonstrates the feasibility of its hybrid architecture, resulting in a high-quality parallel subset suitable for both SLR and NMT tasks. The current status of the records and the progress of the manual validation cycle are quantitatively summarized in Table 1. This dataset provides a solid foundation for feature extraction, particularly due to the visual consistency maintained in the controlled subset

Table 1: Current status of the Libras-UFPel Corpus records and annotations.

Dataset Component	**Type**	**Quantity**
Controlled Sentences	Audiovisual	2,400
Isolated Trigger Signs	Audiovisual	2,400
Naturalistic Interaction (INL)	Audiovisual	10 hours*
Annotation Progress (Controlled)	**Status**	**Total**
Gloss-annotated	Completed	1,200
ELAN Migration/Segmentation	In Progress	600
Lemmatized	Completed	1,200
Portuguese Translation	Completed	1,200

**Note: Naturalistic data is currently in the editing and segmentation phase.*

The controlled subset, consisting of 4,800 records, provides the visual consistency necessary for feature extraction and pose estimation tasks. Preliminary analysis of the 1,200 fully annotated sentences reveals a rich morphological landscape, characterized by complex verb agreement patterns and a high frequency of descriptive classifiers. The inter-annotator review process confirmed high consistency in ID-gloss selection, establishing this subset as a reliable reference source for training baseline SLR models.

A significant outcome of the validated subset is the successful synchronization of independent tiers for NMMs including eyebrows, eye gaze, and head orientation with the manual glosses. This multi-tier alignment, illustrated in the hierarchical model, is crucial for capturing the simultaneity of Libras. It allows computational models to process grammatical and prosodic information that occurs in parallel with lexical production, addressing a common gap in single-channel datasets.

To quantitatively characterize the lexical diversity of the validated subset (1,200 sentences), we computed basic statistics regarding vocabulary and sentence structure. The subset comprises 1,913 unique gloss types across 6,616 total tokens, indicating a high lexical density and a rich morpholog-

ical landscape. This diversity is further evidenced by the presence of complex verb agreement patterns and a high frequency of descriptive classifiers. With an average sentence length of 6.3 signs ($\sigma = 3.7$) , the dataset captures intricate spatial-temporal structures that move beyond simple lexical mapping. Such characteristics ensure that future SLP tasks, particularly NMT, will more accurately reflect the spontaneous sociolinguistic reality of Libras users

In the current semi-controlled subset, the distribution of grammatical categories indicates a significant presence of verbs and indexical signs (pointing). While a higher density of classifiers is anticipated in the naturalistic data, this distribution already provides a basis for modeling the spatial-temporal structures of Libras, moving beyond simple lexical mapping.

Simultaneously, the integration of 10 hours of naturalistic data from the INL-Pelotas represents a key component of the project. This multi-view infrastructure is essential for capturing spatial referencing and role shift, features that are often occluded in single-camera datasets. While this subset is currently in the editing phase, its inclusion ensures that future SLP tasks, particularly NMT, will reflect the spontaneous sociolinguistic reality of the Brazilian deaf community.

5 Discussion and Conclusions

The Libras-UFPel Corpus aims to demonstrate the feasibility of reconciling descriptive linguistic precision with the technical requirements of computational research. By integrating controlled elicitations with naturalistic data, the corpus offers a framework for studies in phonology, syntax, and discourse, ensuring methodological alignment with established national datasets (Quadros et al., 2020) and incorporating current analytical perspectives on dataset construction discussed by Niu et al. (2024) and De Sisto et al. (2022).

By providing parallel resources, we hope to facilitate the development of systems that can better navigate the structural differences between Libras and Portuguese. Furthermore, the active involvement of deaf researchers at every stage ensures that the corpus is an ethically sourced and culturally representative resource.

6 Future Work

As an ongoing collaborative effort, immediate future work will focus on the final editing and alignment of the naturalistic data gathered from the INL-Pelotas to deepen linguistic understanding and improve the training of specific interactional features, such as pointing and spontaneous dialogue markers. Concurrently, we aim to refine the temporal alignment of non-manual features, including eyebrows, eye gaze, and torso, to support 3D pose estimation and affective computing tasks. Furthermore, leveraging the precise annotation of pointing signs and classifiers will provide critical ground truth for complex NLP challenges, such as anaphora resolution and spatial dependency parsing in sign languages. Finally, once the current validation cycle is fully completed, the initial subset of 1,200 lemmatized sentences will be released through a tiered-access repository that balances open science principles with data sovereignty and the protection of signers' image rights. By bridging the gap between linguistics, accessibility, and Artificial Intelligence, the Libras-UFPel Corpus contributes to the digital inclusion of sign languages and provides a replicable model for collaborative research in the Global South.

Acknowledgements

This research is supported by the Federal University of Pelotas (UFPel), the Rio Grande do Sul Research Foundation (FAPERGS) under grant No. 25/2551-0000817-4, and PAPIN/UFPel. This study was also financed in part by the São Paulo Research Foundation (FAPESP) under grant No. 2021/02365-2.

The authors extend their gratitude to the deaf community of Pelotas and the researchers of the *Inventário Nacional de Libras* (INL) for their invaluable contributions. Special thanks are also due to Jose Mario De Martino for his guidance and supervision during the research activities funded by FAPESP.

References

Danielle Bragg and 1 others. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pages 16–31. ACM.

Lucinda Ferreira Brito. 1984. Similarities & differences

in two Brazilian sign languages. *Sign Language Studies*, (42):45–56.

Onno Crasborn and Han Sloetjes. 2008. Enhanced ELAN functionality for sign language corpora. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*, pages 39–43, Marrakech, Morocco.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2024. Machine translation from signed to spoken languages: state of the art and challenges. *Universal Access in the Information Society*, 23(3):1305–1331.

José Mario De Martino and 1 others. 2023. Neural machine translation from text to sign language. *Universal Access in the Information Society*, pages 1–12.

Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and translation. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 2478–2487, Marseille, France. European Language Resources Association.

IBGE. 2010. Censo demográfico 2010. Instituto Brasileiro de Geografia e Estatística. Accessed: 2024-05-20.

Trevor Johnston. 2019. Auslan corpus annotation guidelines. Technical report, Macquarie University, Sydney.

Trevor Johnston. 2024. Auslan corpus annotation guidelines. Technical report, Macquarie University, Sydney.

Antonielle Cantarelli Martins, José Mario De Martino, Janice Gonçalves Temoteo Marques, and Francielle Cantarelli Martins. 2023. Construindo critérios de lematização para a língua de sinais brasileira. *TradTerm*, 45:147–179. Número Especial – Libras, Lexicografia e Cultura.

Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning Publications.

Medet Mukushev, Arman Sabyrov, Madina Sultanova, Vadim Kimmelman, and Anara Sandygulova. 2022. Towards semi-automatic sign language annotation tool: SLAN-tool. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages*, pages 159–164, Marseille, France. European Language Resources Association.

Zhe Niu, Ronglai Zuo, Brian Mak, and Fangyun Wei. 2024. A Hong Kong sign language corpus collected from sign-interpreted TV news. *arXiv preprint arXiv:2405.00980*.

Ronice Müller de Quadros, Jair Barbosa da Silva, Rodrigo Nogueira Machado, and Carlos Roberto Ludwig. 2020. Inventário nacional de libras. *Fórum Linguístico*, 17(4):5457–5474.

Ronice Müller de Quadros, Jair Barbosa da Silva, Miriam Royer, and Vinícius Rodrigues da Silva, editors. 2023a. *Gramática da Libras, Vol. 1*. Instituto Nacional de Educação de Surdos (INES), Rio de Janeiro.

Ronice Müller de Quadros, Jair Barbosa da Silva, Miriam Royer, and Vinícius Rodrigues da Silva, editors. 2023b. *Gramática da Libras, Vol. 2*. Instituto Nacional de Educação de Surdos (INES), Rio de Janeiro.

William C. Stokoe, Jr. 1960. Sign language structure: An outline of the visual communication systems of the american deaf. *The Journal of Deaf Studies and Deaf Education*, 10(1):3–37. Original work published 1960.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy. ELRA.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7347–7360. Association for Computational Linguistics.