# CONSTRUÇÃO DE CORPUS DE LIBRAS E APRENDIZAGEM DE MÁQUINA: DESAFIOS E PERSPECTIVAS

LUÍSA CARVALHO BÖHM[1]; ANTONIELLE CANTARELLI MARTINS[2]; FRANCIELLE CANTARELLI MARTINS

[1]*UFPel – luisacarvalhobohm@gmail.com*
[2]*UFPel – an.cantarellim@gmail.com*
[3]*UFPel – franciellecantarellim@gmail.com*

## 1. INTRODUÇÃO

A construção de corpora de línguas de sinais constitui um desafio particular para a comunidade acadêmica. Diferentemente das línguas orais, que contam com registros escritos consolidados, as línguas de sinais exigem documentação visual-espacial em vídeo. Um corpus é formado por produções sinalizadas alinhadas a glosas na modalidade escrita da língua oral (no caso da Libras, o português), geralmente por meio do software ELAN (CRASBORN; SLOETJES, 2008).

Além de vídeos tradicionais, a captura de movimento (MoCap) vem sendo incorporada como técnica complementar, fornecendo dados tridimensionais que podem ser processados por algoritmos de aprendizado de máquina e ampliam a precisão de análises linguísticas e aplicações computacionais (BAUMGÄRTNER et al., 2020). Entretanto, um corpus não se resume à coleta: a anotação é etapa indispensável e bastante trabalhosa, podendo demandar de 20 a 40 horas para cada hora de gravação (JOHNSTON, 2019). Fenômenos como classificadores, apontamentos e expressões não manuais, ainda sem padronização ampla, tornam essa tarefa ainda mais complexa (MARTINS et al., 2023).

Corpora também são fundamentais para o Processamento de Linguagem Natural (PLN) aplicado às línguas de sinais, alimentando sistemas de tradução e reconhecimento automático, sobretudo em abordagens de Tradução Neural (BRAGG et al., 2019; DE MARTINO et al., 2023). Tais recursos podem ampliar a inclusão da comunidade surda brasileira, especialmente diante da escassez de intérpretes e das barreiras persistentes de acesso em diferentes contextos sociais (KAYO et al., 2021).

A construção desses recursos exige trabalho interdisciplinar entre linguistas, pesquisadores surdos e especialistas em computação, para garantir tanto a representatividade da Libras quanto a aplicabilidade tecnológica (DE MARTINO et al., 2023). No Brasil, embora iniciativas como o Corpus de Libras/UFSC e o Inventário Nacional de Libras (INL) estejam em andamento (QUADROS et al., 2022), ainda há carência de bases amplas, anotadas e acessíveis.

Este trabalho analisa o INL – núcleo Pelotas, com o propósito de torná-lo adequado tanto para investigações linguísticas quanto para aplicações em aprendizado de máquina. Busca-se mapear os desafios e as potencialidades do corpus, considerando que reúne língua espontânea em entrevistas e diálogos, e avaliar sua viabilidade para pesquisas em PLN e no desenvolvimento de sistemas de tradução automática.

## 2. METODOLOGIA

A metodologia foi estruturada para tornar o Inventário Nacional de Libras – núcleo Pelotas um corpus adequado tanto à análise linguística da Libras quanto ao uso em experimentos de aprendizado de máquina.

### 2.1 Coleta de dados

O corpus é composto por entrevistas e diálogos entre pessoas surdas em contextos espontâneos, sem ouvintes, garantindo naturalidade linguística (QUADROS et al., 2022). Os participantes incluem diferentes faixas etárias e gêneros, assegurando diversidade sociolinguística. As interações são gravadas em quatro ângulos (plano amplo, individuais e superior), possibilitando documentar sinais, uso do espaço e expressões não manuais.

### 2.2 Anotação linguística

Os dados são importados para o software ELAN (CRASBORN; SLOETJES, 2008; WITTENBURG et al., 2006), onde recebem anotações multilayer, incluindo:

- Glosa padronizada e lematizada em português (MARTINS et al., 2023);
- Tradução alinhada ao português escrito, constituindo corpus paralelo bilíngue;
- Classificadores, apontamentos e expressões não manuais, elementos complexos, mas indispensáveis (JOHNSTON, 2019).

Essa etapa é manual, realizada por bolsistas com revisão colaborativa. A análise e transcrição deve ser minuciosa, tornando o processo trabalhoso, porém essencial para a qualidade dos dados.

### 2.3 Mapeamento de práticas para aprendizado de máquina

Embora o objetivo final seja o uso em sistemas automáticos, esse processamento exige infraestrutura robusta e equipes especializadas. Neste estágio, a pesquisa foca no mapeamento de melhores práticas para corpora aplicados ao PLN em línguas de sinais, analisando protocolos de iniciativas como o Corpus de Libras/UFSC, o BSL Corpus (CORMIER et al., 2012) e o Auslan Corpus (JOHNSTON, 2019).

O mapeamento inclui:

- Identificação de camadas de anotação relevantes para treinamento supervisionado;
- Discussão sobre padronização de glosas e lematização (MARTINS et al., 2023);
- Análise das dificuldades na anotação de classificadores, apontamentos e expressões não manuais;
- Estudo de métricas de avaliação automática e humana em projetos internacionais.

## 3. RESULTADOS E DISCUSSÃO

A construção de um corpus da Língua Brasileira de Sinais (Libras) é um pilar essencial tanto para a pesquisa linguística quanto para o desenvolvimento de tecnologias de acessibilidade. Os resultados parciais desta investigação não se concentram em experimentos computacionais diretos, mas no mapeamento dos principais desafios e boas práticas relacionados à preparação do INL – núcleo Pelotas para aplicações em aprendizado de máquina.

Um dos entraves mais relevantes refere-se à alta demanda de tempo e mão de obra especializada para a anotação linguística. Diferentemente das línguas orais, a Libras requer uma anotação multimodal detalhada. Essa complexidade torna o processo exaustivo e pouco padronizado (MARTINS et al., 2023).

Outro ponto crítico é a falta de consenso sobre formalismos de anotação. Embora existam iniciativas internacionais consolidadas, como o Auslan Corpus (JOHNSTON, 2019) e o BSL Corpus (CORMIER et al., 2012), no Brasil a diversidade de propostas dificulta comparações e a reusabilidade dos dados. O mapeamento realizado evidencia a necessidade de protocolos mais consistentes, capazes de lidar com fenômenos característicos da Libras, como simultaneidade e iconicidade.

Do ponto de vista tecnológico, mesmo que as redes neurais ofereçam grande potencial para reconhecimento de padrões (DE MARTINO et al., 2023; BRAGG et al., 2019), os desafios de generalização permanecem altos. A variação regional, social e individual da Libras, somada à ausência de correspondência direta com o português, limita a eficácia dos modelos. Nesse sentido, os resultados reforçam a importância de parcerias interinstitucionais, como a cooperação com o CCD-TAAL/UNICAMP, para integrar o corpus de Pelotas a iniciativas de tradução automática em maior escala.

Assim, os resultados obtidos até aqui apontam que o impacto imediato da pesquisa está em oferecer um diagnóstico das lacunas, das potencialidades e das boas práticas de anotação, contribuindo para que o INL-Pelotas se consolide como uma base robusta para estudos futuros em linguística de sinais e processamento de linguagem natural.

## 4. CONCLUSÕES

A construção de um corpus de Libras é um empreendimento complexo, mas estratégico para a pesquisa linguística e o desenvolvimento de tecnologias de acessibilidade. Os resultados desta investigação indicam que o INL – núcleo Pelotas tem grande potencial para se consolidar como base robusta de dados.

Entre os principais achados, destacam-se:

- a necessidade de mão de obra especializada para a anotação, sobretudo em fenômenos como classificadores, apontamentos e expressões não manuais;
- a urgência de padronização de protocolos de anotação e lematização, assegurando comparabilidade entre corpora;

- a constatação de que, mesmo com limitações de infraestrutura, é possível avançar no mapeamento de boas práticas para a construção de corpora aplicáveis ao aprendizado de máquina.

A pesquisa também mostra que a consolidação de tradutores automáticos Libras–Português depende da articulação entre linguistas, pesquisadores surdos e especialistas em inteligência artificial, além de parcerias interinstitucionais — como com o CCD-TAAL/UNICAMP — para viabilizar o processamento em larga escala.

Em síntese, a principal contribuição deste trabalho está em preparar o INL-Pelotas para integrar iniciativas de PLN aplicadas às línguas de sinais, reforçando que a acessibilidade da comunidade surda só será alcançada com a união entre rigor metodológico, interdisciplinaridade e inovação tecnológica.

## 5. REFERÊNCIAS BIBLIOGRÁFICAS

BAUMGÄRTNER, L. et al. Automated sign language translation: the role of artificial intelligence now and in the future. In: INT. CONF. ON COMPUTER-HUMAN INTERACTION RESEARCH AND APPLICATIONS (CHIRA), 4., 2020, Budapest. *Proceedings…* [S. l.]: SCITEPRESS, 2020.

BRAGG, D. et al. Sign language recognition, generation, and translation: an interdisciplinary perspective. In: **ASSETS '19**, 2019, Pittsburgh. *Proceedings…* Pittsburgh: ACM, 2019.

CORMIER, K.; SCHEMBRI, A.; VINSON, D. The British Sign Language Corpus Project: open access archives for BSL linguistics research. *Sign Language Studies*, v. 12, n. 4, p. 362–379, 2012.

CRASBORN, O.; SLOETJES, H. Enhanced ELAN functionality for sign language corpora. In: WORKSHOP ON THE REPRESENTATION AND PROCESSING OF SIGN LANGUAGES, 3., 2008, Paris. *Proceedings…* Paris: ELRA, 2008.

DE MARTINO, J. M. et al. Neural machine translation from text to sign language. *Universal Access in the Information Society*, v. 22, p. 1–15, 2023.

JOHNSTON, T. *Auslan Corpus Annotation Guidelines*. Sydney: Macquarie Univ.; Melbourne: La Trobe Univ., 2019.

KAYO, Y. et al. Including signed languages in natural language processing. In: ANN. MEET. OF THE ASSOC. FOR COMPUTATIONAL LINGUISTICS (ACL), 59., 2021, Bangkok. *Proceedings…* [S. l.]: ACL, 2021. p. 7347–7360.

MARTINS, A. C. et al. Construindo critérios de lematização para a Língua de Sinais Brasileira. *TradTerm*, v. 42, p. 1–30, 2023.

QUADROS, R. M. et al. Inventário nacional de Libras. *Fórum Linguístico*, v. 17, n. 4, p. 5457–5474, 2022.