DOI: https://doi.org/10.14393/DLv20a2026-4
e020004

Artigo

Corpus lexicográfico e visão computacional: uma metodologia baseada em IA para a coleta e anotação semiautomática de sinais em larga escala e a análise de variações lexicais (regionalismos) em Libras

Lexicographical corpus and computer vision: an AI-based methodology for large-scale semi-automatic collection and annotation of signs and the analysis of lexical variation (regionalisms) in Brazilian sign language

*Bruno Jose Betti GALASSO**

RESUMO: O estudo quantitativo da variação lexical (regionalismos) na Língua Brasileira de Sinais (Libras) é metodologicamente obstaculizado pela ausência de *corpora* lexicográficos de grande porte com anotação querológica. Metodologias tradicionais de anotação manual, como o ELAN, são inviáveis para a construção de *corpora* massivos (dezenas de milhares de horas), e a práxis de glosagem conceitual (tradução) resulta na perda de informação articulatória (significante), tornando impossível o estudo de variantes querológicas sutis. Diante dessa lacuna, o objetivo geral deste artigo é propor uma arquitetura metodológica interdisciplinar que articule Linguística de *Corpus*, Visão Computacional e Sociolinguística Computacional para a construção de um *corpus* lexicográfico variacionista de Libras em larga escala. Como objetivos específicos, busca-se: (i) mapear o estado da arte na interseção entre IA e línguas de sinais por meio de revisão sistemática PRISMA; (ii) detalhar um *pipeline* bifásico tecnicamente robusto para coleta e anotação semiautomática; (iii) propor procedimentos de análise sociolinguística computacional para descoberta de variantes regionais; e (iv) estabelecer um *framework* ético centrado na Pesquisa Liderada por Surdos (*Deaf-led Research*). A metodologia inicia-se com revisão sistemática (2018-2025), identificando a lacuna central: a IA para línguas de sinais concentra-se quase exclusivamente em reconhecimento (SLR) e tradução (SLT), tratando a variação linguística como ruído a ser eliminado quando, para a Sociolinguística, essa variação constitui precisamente o objeto de estudo. A arquitetura proposta utiliza um *pipeline* bifásico. A Fase 1 (Extração de Características) emprega estimativa de pose (MediaPipe otimizado), seguindo a otimização de dos Santos et al. (2025), para converter vídeos (coletados "in-the-wild") em representações vetoriais (séries temporais de *landmarks*), substituindo a glosa por uma "transcrição" querológica quantificável. A Fase 2 (Anotação Semiautomática) utiliza modelos *Transformers*, treinados em um *corpus* semente, para gerar sugestões de rótulos lexicais. Essas sugestões são submetidas a uma interface de validação *human-in-the-loop*, onde linguistas surdos validam ou corrigem as anotações, com o modelo sendo re-treinado

* Doutor em Educação pela Universidade de São Paulo (USP). Professor associado da Universidade Federal de São Paulo (UNIFESP), São Paulo, SP – Brasil. bruno.galasso@unifesp.br

iterativamente. Como resultado, a metodologia produz um banco de dados relacional massivo que associa formas articulatórias (vetores) a metadados sociolinguísticos. Este *corpus* habilita a análise sociolinguística computacional por meio de técnicas de *clustering* não supervisionado, permitindo a descoberta de padrões e a identificação de variantes regionais que são, subsequentemente, correlacionados com dados geográficos para mapear a variação possibilitando a criação de atlas linguísticos quantitativos para a Libras. A conclusão é que esta arquitetura supera os gargalos da lexicografia tradicional e oferece um caminho viável para a documentação da variação. O artigo discute criticamente os fundamentos éticos, posicionando a pesquisa liderada por surdos" (*Deaf-led Research*) como pilar metodológico central para mitigar vieses algorítmicos (como o tecno-capacitismo) e garantir que a tecnologia funcione como ferramenta de documentação e empoderamento, e não de substituição ou erosão de direitos linguísticos.
PALAVRAS-CHAVE: Libras. Variação Lexical. Sociolinguística Computacional. Visão Computacional. Corpus Lexicográfico.

ABSTRACT: The quantitative study of lexical variation (regionalisms) in Brazilian Sign Language (Libras) is methodologically hindered by the absence of large-scale, cherologically annotated lexicographical *corpora*. Traditional manual annotation methodologies, such as ELAN, are unfeasible for building massive *corpora* (tens of thousands of hours), and the praxis of conceptual glossing (translation) results in the loss of articulatory information (signifier), making the study of subtle cherological variants impossible. This article proposes an interdisciplinary methodological architecture that solves this dual bottleneck by articulating Corpus Linguistics, Computer Vision, and Computational Sociolinguistics. The objective is to detail a technically and ethically robust pipeline for large-scale semi-automatic collection and annotation, focused specifically on the discovery and analysis of sociolinguistic variation. The methodology begins with a systematic review (PRISMA) mapping the state-of-the-art (2018-2025), identifying the central gap: AI focuses almost exclusively on recognition (SLR) and translation (SLT), treating linguistic variation as noise rather than an object of study. The proposed architecture employs a two-phase pipeline. Phase 1 (Feature Extraction) uses pose estimation (optimized MediaPipe), following the optimization by dos Santos et al. (2025) , to convert videos (collected "in-the-wild") into vector representations (time series of landmarks), replacing glossing with a quantifiable "cherological transcription". Phase 2 (Semi-Automatic Annotation) utilizes Transformer models, trained on a seed *Corpus*, to generate lexical label suggestions. These suggestions are submitted to a *human-in-the-loop* validation interface, where Deaf linguists validate or correct the annotations, with the model being iteratively retrained. As a result, the methodology yields a massive relational database that links articulatory forms (vectors) to sociolinguistic metadata. This *Corpus* enables computational sociolinguistic analysis via unsupervised clustering techniques, allowing for pattern discovery and the identification of regional variants which are subsequently correlated with geographical data to map variation, enabling the creation of quantitative linguistic atlases for Libras. The conclusion is that this architecture overcomes the bottlenecks of traditional lexicography and offers a viable path for documenting variation. The article critically discusses the ethical foundations, positioning "Deaf-led Research" as a central methodological pillar to mitigate algorithmic biases (such as techno-ableism) and ensure the technology functions as a tool for documentation and empowerment, not for replacement or the erosion of linguistic rights.

KEYWORDS: Libras. Lexical Variation. Computational Sociolinguistics. Computer Vision. Lexicographical *Corpus*.

Artigo recebido em: 17.11.2025
Artigo aprovado em: 13.01.2026

1 Introdução

A Lei nº 10.436 de 2002 representou um marco fundamental ao reconhecer oficialmente a Língua Brasileira de Sinais (Libras) como meio legal de comunicação e expressão da comunidade surda brasileira. Este reconhecimento impulsionou exponencialmente os estudos descritivos da língua. Pesquisas seminais, como as de Quadros e Karnopp (2004) e Quadros (2016) estabeleceram a Libras como língua natural completa, dotada de complexidade fonológica (querológica[1]), morfossintática, semântica e pragmática.

Como em toda língua viva, estudos sociolinguísticos identificaram vasta variação interna. A sociolinguística variacionista aplicada à Libras, conforme documentado por Santos (2020), Oliveira, Silva e Campelo (2020) e Machado (2018), tem registrado variações diacrônicas, diafásicas, diastráticas e, proeminentemente, diatópicas (regionais). A variação lexical, especificamente o regionalismo, constitui manifestação saliente. Um mesmo conceito pode ser expresso por sinais distintos em diferentes estados brasileiros, refletindo a extensão continental do país.

O problema central da pesquisa variacionista em Libras reside no profundo descompasso entre o reconhecimento qualitativo dessa riqueza linguística e a capacidade de documentá-la em larga escala. Estudos sobre regionalismos são frequentemente restritos a escopos geográficos limitados ou baseados em metodologias de amostragem reduzida. O gargalo metodológico manifesta-se na ausência de *corpora* lexicográficos de grande escala, ecologicamente válidos e anotados

1 O termo 'querológica' (do grego *kheir*, mão) refere-se à fonologia das línguas de sinais, ou seja, ao estudo das unidades mínimas distintivas na produção dos sinais (configuração de mão, orientação, locação, movimento e expressões não manuais), em analogia à fonologia das línguas orais.

para análise quantitativa. A construção de um *corpus* compreendendo dezenas de milhares de horas de vídeo, abrangendo a diversidade geográfica brasileira, representa tarefa impossível via metodologias tradicionais de anotação manual, como o uso do software ELAN (*EUDICO Linguistic Annotator*), que exige de 20 a 40 horas de trabalho humano para cada hora de vídeo anotado (cf. seção 3). É precisamente na intersecção entre a demanda por documentação em larga escala e a inviabilidade das metodologias manuais que a Inteligência Artificial (IA), especificamente a Visão Computacional, emerge como solução potencial. Contudo, o campo da IA para línguas de sinais concentra-se quase exclusivamente em reconhecimento automático (SLR) e tradução (SLT), visando comunicação entre surdos e ouvintes. O uso da Visão Computacional como ferramenta endógena para a Linguística descrever e analisar a estrutura e variação da língua permanece vastamente inexplorado.

Nesse contexto, este artigo propõe uma arquitetura metodológica que articula Linguística de *Corpus*, Visão Computacional e Sociolinguística Computacional para construção de um *Corpus* Variacionista de Libras em larga escala. Através de *pipeline* que utiliza estimativa de pose para transcrição querológica e modelos de *Deep Learning* (Transformers) para anotação semiautomática com validação humana, criamos recurso que habilita análise geossociolinguística quantitativa dos regionalismos.

Dessa forma, o objetivo geral deste artigo é propor uma arquitetura metodológica interdisciplinar para a construção de um *corpus* lexicográfico variacionista de Libras em larga escala. Os objetivos específicos são: (i) mapear o estado da arte na interseção entre IA e línguas de sinais; (ii) detalhar o *pipeline* técnico bifásico para coleta e anotação semiautomática; (iii) propor procedimentos de análise sociolinguística computacional; e (iv) estabelecer um *framework* ético centrado na Pesquisa Liderada por Surdos.

2 Mapeamento sistemático da interseção entre IA e línguas de sinais: revisão PRISMA

Para justificar a originalidade da metodologia proposta, mapeamos rigorosamente o estado da arte na interseção entre IA (Visão Computacional) e pesquisa em línguas de sinais. Conduzimos revisão sistemática seguindo o protocolo PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*).

2.1 Procedimentos da revisão sistemática

O objetivo foi identificar, avaliar e sintetizar estudos publicados entre 2018-2025 que abordam desenvolvimento de *datasets* (*corpora*) e metodologias de IA para processamento de línguas de sinais, com atenção à Libras. Buscas foram realizadas em IEEE Xplore, ArXiv, ACM Digital Library e Google Scholar. A janela temporal captura o período de expansão das arquiteturas de *Deep Learning*, particularmente após popularização dos Transformers.

Termos de busca incluíram: ("sign language recognition" OR "sign language translation") AND ("*dataset*" OR "*Corpus*" OR "*benchmark*"); ("computer vision" OR "AI") AND ("Libras" OR "Brazilian Sign Language"); ("automatic annotation" OR "semi-automatic annotation") AND "sign language".

Foram incluídos artigos que: (a) descreviam criação de novo *dataset* de língua de sinais para Visão Computacional; (b) propunham arquiteturas de IA para SLR ou SLT; (c) discutiam métodos de extração de características. Foram excluídos: (a) focados em *hardware*; (b) centrados em línguas orais; (c) puramente teóricos sem implementação; (d) sem revisão por pares (exceto *preprints* de alta relevância com citações substanciais).

A revisão revela cenário vibrante, porém com foco concentrado. A maioria esmagadora (>90%) está direcionada a SLR (reconhecimento automático) e SLT (tradução). O campo opera sob paradigma de *benchmark*: desenvolver arquiteturas de

Deep Learning que atinjam estado da arte (SOTA) em métricas de precisão em *datasets* padronizados.

Plataformas como RWTH-PHOENIX-Weather (Língua de Sinais Alemã) e AUTSL (Turca) servem como campos de prova onde diferentes arquiteturas competem. No contexto brasileiro, identificamos o MINDS-Libras (UFMG): *Corpus* multimodal tecnicamente sofisticado, capturando vídeo RGB, profundidade e *landmarks*. Contudo, contém apenas 20 sinais, 12 sinalizantes, 1.200 amostras. Embora excelente para treinar modelos de reconhecimento para esses 20 sinais, é estatisticamente insuficiente para estudar regionalismos por carecer de diversidade lexical, geográfica e naturalidade.

O quadro 1 sintetiza principais *datasets* e expõe a lacuna central.

Quadro 1 – Síntese de *Datasets* de IA para Línguas de Sinais.

Dataset	**Língua**	**Objetivo**	**Características**	**Limitações Sociolinguísticas**
MINDS-Libras	Libras	ISLR	20 sinais; 12 sinalizantes; 1.200 amostras	Vocabulário trivial; sem diversidade geográfica
AUTSL	Turca	ISLR	226 sinais; 43 sinalizantes	Foco em sinais isolados; laboratório
RWTH-PHOENIX	Alemã	CSLR/SLT	Domínio específico (previsão do tempo)	Domínio lexical restrito
Arquiteturas SOTA	Diversas	SLR/SLT	3D-CNNs e Transformers	Variação tratada como ruído

Fonte: elaborado pelo autor.

A revisão demonstra ausência completa de estudos utilizando Visão Computacional para documentação lexicográfica ou análise sociolinguística. No paradigma de reconhecimento, variação linguística (regionalismos) é tratada como ruído que diminui precisão. *Datasets* são intencionalmente "limpos", removendo a "desordem" do uso real. Para Sociolinguística, essa "desordem" não é ruído; é o dado.

A lacuna é interdisciplinar: linguistas carecem de dados em escala; engenheiros carecem de propósito linguístico. Nossa proposta posiciona-se exatamente nesta intersecção, sugerindo um uso epistemologicamente distinto das ferramentas: não reconhecer sinais em vocabulário fechado, mas documentar articulação de vocabulário aberto e variante.

3 Desafios da anotação lexicográfica: do manual ao semiautomático

A metodologia de referência para criação de *corpora* em línguas de sinais é anotação manual detalhada usando *software* especializado. O *corpus* de Libras (UFSC) representa esforço exemplar. A ferramenta padrão é o ELAN (*EUDICO Linguistic Annotator*), *software* robusto para anotações temporalmente alinhadas com vídeo, criando múltiplas "trilhas" (*tiers*) de anotação.

Em projetos de *corpus* de Libras, como o *Corpus* de Libras da UFSC (Quadros; Cruz, 2011) e projetos correlatos (Silva, 2015), anotadores criam trilhas para: Glosa-MãoD (palavra em português representando sinal da mão dominante); Glosa-MãoE (mão não-dominante); Expressão-Não-Manual (marcadores faciais); Tradução-PB (tradução livre); Comentários. Esta metodologia é poderosa para análise qualitativa fina, permitindo isolamento de fenômenos específicos.

Apesar da utilidade, possui limitação fundamental: glosagem não é transcrição, mas tradução. Transcrição fonética (para línguas orais) ou querológica (para sinais) representa a forma articulatória (significante). Tradução salta para o conceito (significado), representando-o com rótulo de outra língua. Este salto tem duas consequências problemáticas: perda de informação articulatória (variantes regionais rotuladas identicamente) e paradoxo da glosa (para estudar variação precisamos primeiro identificá-la).

Mesmo sem perda de dados, anotação manual é inviável em escala. Uma hora de vídeo requer 20-40 horas de anotação. Para línguas de sinais, com simultaneidade de múltiplos articuladores, esse fator é maior. Construir *corpus* de dezenas de milhares

de horas é impossível com método manual. Para superar o paradoxo e a inviabilidade, propomos mudança de paradigma: substituir "tradução" (glosa) por "transcrição articulatória" (vetores de *landmarks*) e "anotação manual" por "validação humana" (*human-in-the-loop*).

4 Arquitetura metodológica para o *Corpus* Variacionista de Libras

A metodologia proposta é *pipeline* bifásico para receber e processar grandes volumes de vídeo de Libras e produzir banco de dados estruturado correlacionando itens lexicais com formas articulatórias e metadados sociolinguísticos.

Na primeira fase, marcada pela extração de características, convertemos dado bruto (pixels) em representação matemática da articulação. Este processo, estimativa de pose (*pose estimation*), gera "esqueleto" digital de *landmarks* (pontos-chave) rastreando articulações do corpo, mãos e rosto, quadro a quadro.

A coleta deve ser intencionalmente diversa. Embora dados de laboratório com metadados controlados sejam valiosos, o *corpus* deve ser alimentado por vídeos "*in-the-wild*" (YouTube, Instagram, TikTok, material educacional). A mineração ética permite coleta de milhares de horas de uso autêntico.

É fundamental, contudo, adotarmos uma estratégia de coleta híbrida, pois os dados "*in-the-wild*" (de mineração aberta) e os dados sociolinguisticamente controlados cumprem funções distintas no *pipeline*.

1. **Dados "In-the-Wild":** os milhares de horas extraídos de plataformas abertas (YouTube, TikTok etc.), embora carentes de metadados sociolinguísticos confiáveis, são essenciais para o treinamento não supervisionado dos *Autoencoders* (conforme Seção 5) e para o pré-treinamento geral dos *Transformers* (Seção 4). Esta massa de dados ensina aos modelos a robustez necessária para lidar com diferentes iluminações, ângulos e ruídos de fundo.
2. **Dados dirigidos:** Para a análise final de correlação (Seção 5), é necessário um subconjunto de dados com metadados de alta confiabilidade. Esta coleta será

"dirigida", realizada em parceria com as comunidades surdas locais, universidades e associações (conforme Seção 6). Os vídeos deste conjunto serão acompanhados por um formulário de metadados detalhado (região de origem, idade, etnia, tempo de uso da Libras etc.), que será ingerido pelo banco de dados relacional.

Esta arquitetura híbrida resolve o paradoxo: usamos a massa de dados *in-the-wild* para treinar modelos robustos e usamos a qualidade dos dados dirigidos para realizar a análise sociolinguística validada.

A escolha da ferramenta de estimativa de pose é crítica. O *software* OpenPose (Carnegie Mellon) é padrão-ouro em precisão, mas altíssimo custo computacional, inviabilizando processamento massivo. Já o MediaPipe (Google) é leve, roda em tempo real em *smartphones*, fornece 468 pontos faciais, 33 corporais, 21 por mão. Inicialmente considerado inferior, a razão era *domain mismatch*: modelos não treinados para complexidade de línguas de sinais.

A viabilidade depende de descoberta recente. Dos Santos *et al*. (2025) investigaram por que MediaPipe falhava e descobriram que o problema era "ruído" de *landmarks* irrelevantes e falta de robustez a ausências. A solução foi seleção otimizada de subconjunto de *landmarks* (focando mãos, ombros, rosto) e imputação baseada em *spline* para dados ausentes. Com isso, MediaPipe superou métodos SOTA reduzindo tempo de processamento em 5×.

Nossa metodologia adota o pipeline validado por dos Santos *et al*. (2025). Processamos com MediaPipe Holistic, armazenando apenas subconjunto otimizado: 21 pontos por mão, 5 faciais chave, 11 de pose superior. O resultado não é vídeos anotados, mas banco de dados massivo de séries temporais de vetores (x, y, z) para cada *landmark*, quadro a quadro. Esta é nossa "transcrição" querológica: representação matemática, computacionalmente tratável, da articulação.

Na Fase 1 obtivemos a forma (significante). Na Fase 2 associamos com conceito (significado/glosa). Possuímos vetores de articulação, mas não sabemos a qual item

lexical correspondem. Realizar manualmente seria o mesmo gargalo do ELAN. A solução é usar *Deep Learning* para sugerir rótulos, validados por humanos.

O estado da arte em CSLR evoluiu de 3D-CNNs (operam sobre "cubos" de vídeo, aprendendo características espaço-temporais) para Transformers (adaptados da revolução em PLN oral). Transformers são superiores em modelar dependências de longo alcance, cruciais para línguas de sinais: significado de movimento pode depender de contexto ou marcador facial distante temporalmente. Transformers capturam essas relações através do mecanismo de atenção.

Propomos pipeline *human-in-the-loop* da seguinte maneira:

1. **Treinamento do *Seed Model*:** Transformer é treinado nos *landmarks* da Fase 1 usando *corpus* "semente" (menor, já anotado manualmente, como *corpus* UFSC ou MINDS-Libras). O modelo aprende associar padrões de vetores a glosas.
2. **Geração de sugestões:** dados não rotulados são processados pelo *Seed Model*. O modelo gera sugestões (segmento temporal, rótulo sugerido, confiança estatística).
3. **Interface de validação:** interface apresenta ao anotador: vídeo original, *landmarks* visualizados, sugestão automática. O anotador—obrigatoriamente linguista surdo—valida (um clique) ou corrige. Ordens de magnitude mais rápido que anotação do zero.

O design desta interface é um componente central da metodologia de Pesquisa Liderada por Surdos (Seção 6). Ela não é apenas uma ferramenta de validação, mas um ambiente de anotação linguisticamente rico, co-desenhado com os linguistas surdos. A interface deve obrigatoriamente incluir:

a. Visualizador sincronizado: exibição lado a lado do vídeo original e da renderização dos *landmarks*, permitindo ao anotador verificar se a IA "viu" a articulação corretamente. **b. Módulo de sugestão:** apresentação clara da glosa sugerida pelo Transformer e o nível de confiança estatística. **c. Módulo de correção rápida:** campos para correção imediata da glosa (se errada) ou da segmentação temporal (início/fim do

sinal). **d. Módulo de metadados linguísticos:** campos de anotação adicionais para que o linguista possa marcar fenômenos que a IA não captura, como o uso de ênfase, marcadores não manuais icônicos ou comentários querológicos (ex: "articulação atípica", "influência L2"). **e. Módulo de metadados sociolinguísticos:** (visível apenas ao lidar com "Dados Dirigidos") exibição dos metadados do sinalizante (região, idade) para contextualizar a análise do validador. Este design garante que o especialista surdo mantenha controle cognitivo total, usando a IA como um assistente que automatiza a parte braçal (sugestão), liberando o especialista para a análise linguística sofisticada.

4. **Aprendizado iterativo:** correções e validações são incorporadas continuamente. *Seed Model* é periodicamente retreinado (*fine-tuning*).

É crucial detalhar esta etapa, pois ela enfrenta o desafio metodológico do "cold start" (partida a frio). Os *corpora* existentes (como MINDS-Libras ou *corpus* UFSC), embora valiosos, apresentam um duplo *domain mismatch*: (1) são insuficientes em volume para capturar a variação que buscamos; e (2) foram anotados primariamente via glosa textual (ELAN), não com os vetores de *landmarks* que nossa arquitetura utiliza. Treinar o Seed Model diretamente nesses dados geraria sugestões de baixa acurácia. A solução reside em um processo de curadoria manual inicial e *bootstrapping*. Em vez de depender dos *corpora* legados, o *pipeline* inicia com a criação de um "*corpus* ouro" (Gold Standard) mínimo: os linguistas surdos anotam manualmente um subconjunto de alta prioridade (ex: 50-100 lemas conhecidos por alta variação regional, como "CERVEJA" ou "CARRO"), diretamente na nova interface, associando o rótulo da glosa aos dados vetoriais extraídos.

O Seed Model v 0.1 é treinado exclusivamente neste pequeno, mas metodologicamente puro, *corpus* Ouro. A interface de validação emprega, então, uma estratégia de Aprendizagem Ativa: ela prioriza a apresentação de novos vídeos que o modelo v0.1 identifica como sendo similares aos dados já conhecidos (baixa incerteza). Os validadores humanos corrigem ou validam essas sugestões fáceis, expandindo rapidamente o *corpus* de treinamento. O modelo é retreinado (v0.2), tornando-se apto

a sugerir rótulos para exemplos mais complexos. Este ciclo de *bootstrapping* é o que permite ao sistema escalar de 100 lemas para milhares, resolvendo o paradoxo da partida a frio. A cada ciclo, modelo torna-se mais preciso, reduzindo carga dos anotadores.

Este *pipeline* resolve o gargalo escalar. Esforço humano é deslocado da tarefa braçal para supervisão cognitiva linguisticamente sofisticada. Em vez de tratar reconhecimento como fim, usamos SLR como ferramenta intermediária para objetivo linguístico: *corpus* massivo, lexicalmente rotulado e validado.

5 Da anotação à análise: Sociolinguística Computacional Aplicada

Ao término do *pipeline*, o produto é artefato novo: banco de dados relacional massivo. Cada item lexical está associado a milhares de instâncias de produção, representadas por "impressão digital" articulatória (vetores de *landmarks*) e metadados sociolinguísticos (região, idade, contexto).

O método tradicional de variação exige que pesquisador observe diferença para documentá-la. A metodologia computacional inverte essa lógica: permite descoberta não supervisionada de padrões. O processo segue quatro etapas:

1. Seleção do lema: pesquisador seleciona item lexical (ex: "CERVEJA", "CARRO").
2. Recuperação do Espaço Vetorial: sistema recupera todas instâncias de vetores de *landmarks* associadas ao lema.

Antes da clusterização, contudo, é mandatório um pré-processamento vetorial para superar dois desafios: (1) a alta dimensionalidade dos dados (séries temporais com dezenas de pontos x, y, z por quadro) e (2) o comprimento variável das sequências (um sinalizante pode executar "CARRO" em 15 quadros, outro em 25). Algoritmos de *clustering* tradicionais não operam eficientemente sobre séries temporais brutas de comprimentos distintos.

A solução é transformar a série temporal de comprimento variável em uma representação latente de comprimento fixo. Propomos o uso de arquiteturas de *deep learning*, especificamente *Autoencoders* (sejam eles Variacionais, VAEs, ou baseados em Transformers), treinados de forma não supervisionada sobre todas as instâncias vetoriais do lema selecionado (ex: todas as milhares de execuções de "CERVEJA").

O *encoder* do modelo aprende a comprimir a dinâmica articulatória completa do sinal em um único vetor de dimensionalidade reduzida (um *embedding*), que captura a essência da articulação. É este espaço latente—onde cada instância do sinal é agora um único ponto—que é submetido à Análise de *Cluster*. Esta redução de dimensionalidade é o que permite que os algoritmos identifiquem matematicamente os agrupamentos (os "Clusters") que representam as variantes querológicas.

3. Análise de *cluster*: todas instâncias vetoriais são alimentadas a algoritmos de *clustering*. Técnicas incluem: Agrupamento Hierárquico (cria dendrograma mostrando como articulações se agrupam por similaridade); Escalonamento Multidimensional (MDS - "achata" espaço vetorial em mapa 2D/3D onde proximidade representa similaridade articulatória); Agrupamento por Comitês (métodos *ensemble* combinando múltiplos algoritmos). Algoritmos operam sem supervisão—sem informação prévia sobre quantas ou quais variantes existem. O resultado não é glosa única "CERVEJA", mas múltiplos *clusters*: "CERVEJA_Cluster_1", "CERVEJA_Cluster_2" etc. Estes são candidatos empíricos a variantes querológicas. É fundamental ressaltar que os *clusters* gerados computacionalmente são hipóteses estatísticas de variantes, não categorias linguísticas validadas. Por isso, um passo subsequente de validação qualitativa é obrigatório: o linguista surdo analisa os exemplares de cada *cluster*, verificando: (a) se as diferenças articulatórias capturadas são fonologicamente significativas na língua ou se representam variação livre idiossincrática; (b) se os agrupamentos correspondem a variantes já documentadas na literatura ou se constituem descobertas inéditas; (c) se há coerência sociolinguística na distribuição geográfica ou demográfica observada.

Somente após essa validação qualitativa o *cluster* é promovido a 'variante regional confirmada' no banco de dados.

4. Correlação com metadados: sistema correlaciona estatisticamente distribuição dos *clusters* com metadados sociolinguísticos (geografia, idade, contexto).

O resultado é quantificação precisa e mapeamento geográfico da variação. Sistema gera relatórios: "Lema 'CERVEJA' possui 3 variantes principais. *Cluster_1* apresenta 95% de ocorrência na Região Sul. *Cluster_2* apresenta 90% no Nordeste. *Cluster_3* é minoritário e disperso, com concentração no RJ."

Isto move o estudo de regionalismos de prática qualitativa para mapa geossociolinguístico robusto, quantitativo. Pela primeira vez, produzir atlas linguísticos de Libras comparáveis aos de línguas orais.

Mais profundamente, *clustering* não supervisionado pode revelar variações que linguistas não descobriram. Percepção humana nota variações lexicais óbvias (sinais completamente diferentes). Algoritmo é sensível a variações querológicas sutis (pequena mudança em configuração de mão, alteração mínima em ponto de articulação) imperceptíveis em observação casual. Se estas variações sutis são sistematicamente distribuídas geograficamente, constituem variações linguisticamente reais e socialmente significativas, ainda que não perceptualmente salientes. A IA não apenas coleta eficientemente o conhecido; funciona como instrumento de descoberta, revelando padrões estruturais ocultos.

5.1 Disseminação: do cluster ao atlas e dicionário

O *pipeline* não se encerra no relatório estatístico; ele deve gerar os artefatos de documentação prometidos. Esta etapa final operacionaliza a disseminação dos dados de forma acessível à comunidade.

a) **Geração do atlas (visualização geográfica):** a correlação estatística entre clusters e metadados geográficos (obtidos dos Dados Dirigidos) é usada como input para um Sistema de Informação Geográfica (SIG). O sistema plota a distribuição percentual de

cada cluster (variante) no mapa do Brasil, gerando as "isoglossas" quantitativas que formam o atlas linguístico. Os usuários podem filtrar por lema (ex: "CARRO") e visualizar o mapa de variação correspondente.

b) **Geração do dicionário (seleção de exemplar):** um cluster estatístico (ex: "CERVEJA_Cluster_1") não é um verbete. Para criar o dicionário regional, o sistema identifica o exemplar prototípico (ou centroide) de cada cluster — a instância de vídeo que está matematicamente mais próxima da média de todas as outras execuções naquele cluster. Este vídeo exemplar, validado pelo linguista surdo, torna-se o verbete visual para aquela variante regional específica no dicionário online.

Esta etapa final transforma a análise computacional abstrata em ferramentas de consulta concretas, cumprindo o objetivo de documentação.

6 Discussão crítica: ética, vieses e centralidade surda

Uma proposta envolvendo coleta massiva de dados de comunidade linguística minorizada usando IA avançada não pode ser apresentada sem discussão crítica profunda de fundamentos éticos. A mesma tecnologia que propomos para documentação é usada primariamente para reconhecimento e tradução—aplicações que carregam riscos significativos. A literatura crítica identifica vieses sistêmicos permeando o campo:

a) Viés de dados: modelos de IA aprendem exclusivamente dos dados de treinamento. Se *corpus* for enviesado, modelo refletirá e amplificará vieses. Se coletarmos dados desproporcionalmente de sinalizantes brancos, jovens, urbanos, do Sudeste, a IA aprenderá implicitamente que esta é Libras "padrão". Qualquer desvio será tratado como "erro". Modelo não apenas reflete passivamente desigualdades sociais existentes, mas as codifica algoritmicamente e amplifica.

b) Tecno-capacitismo: Desai *et al*. (2024) identificam "tecno-capacitismo" pervasivo: crença implícita de que tecnologia representa necessariamente melhor solução e ímpeto de "consertar" corpos percebidos como deficientes. Muitas ferramentas são

projetadas por ouvintes, para ouvintes, focando em "superar barreira de comunicação", tratando surdez como problema técnico em vez de reconhecer Libras como língua completa e comunidade surda como comunidade linguística legítima.

c) Erosão de direitos linguísticos: o risco mais grave. Tecnologias de tradução automática, apresentadas prematuramente como "suficientemente boas", correm risco de serem adotadas por governos/instituições primariamente por razões financeiras. Podem ser vistas como substitutos "bons o suficiente" e "mais baratos" que intérpretes humanos. Isto poderia resultar na erosão progressiva da profissão de tradutor-intérprete e minar o direito duramente conquistado da pessoa surda ao acesso através de intérprete humano qualificado, especialmente em contextos críticos onde erros têm consequências graves (médicos, legais, educacionais).

Reiteramos que a nossa proposta não é naturalmente imune a esses riscos. Para ser simultaneamente eticamente defensável e linguisticamente válida, deve ser construída sobre *framework* ético que não é apêndice, mas pilar metodológico central.

A literatura de ética em IA (IEEE P7000, ISO/IEC 42001, NIST AI RMF) e de direitos linguísticos surdos é unânime: a solução é implementação rigorosa de "Pesquisa Liderada por Surdos" (*Deaf-led Research*).

Concretamente e operacionalmente:

a) Governança de dados: recomenda-se fortemente que *corpus* e algoritmos não sejam propriedade exclusiva de laboratórios de engenharia. Governança, controle de acesso e decisões estratégicas devem ser supervisionados por comitê majoritariamente composto por pesquisadores surdos, linguistas surdos e líderes da comunidade surda brasileira. Este comitê tem poder de veto sobre aplicações prejudiciais. A governança é o primeiro passo para a soberania dos dados. Não basta a comunidade governar os dados; ela deve possuí-los. A propriedade intelectual do *corpus* resultante e dos produtos derivados (o atlas e o dicionário) deve ser da comunidade surda, não da instituição de pesquisa ou do laboratório de engenharia.

Para garantir isso, o licenciamento do *corpus* deve ser cuidadosamente estruturado. Modelos como o *Creative Commons (CC)* devem ser aplicados com cláusulas específicas (ex: **CC BY-NC-SA** – Atribuição, Não Comercial, Compartilha Igual). A cláusula "Não Comercial" (NC) é crucial, pois impede que os dados coletados e anotados pela comunidade sejam explorados comercialmente por terceiros sem o consentimento explícito do comitê de governança. Este modelo de licenciamento garante que o *corpus* permaneça um bem comum para pesquisa e educação, alinhado ao objetivo de empoderamento, ao mesmo tempo que previne a erosão de direitos linguísticos através da exploração comercial descontrolada.

Cabe ainda ressaltar que o licenciamento cuidadoso é particularmente relevante diante da proliferação de avatares e sistemas automáticos de tradução para Libras no Brasil. Embora amplamente adotados por órgãos públicos e empresas, tais sistemas são frequentemente criticados pela comunidade surda quanto à sua eficácia, naturalidade e respeito às nuances culturais da língua (DESAI *et al*., 2024). A governança comunitária do *corpus* proposto visa prevenir que os dados sejam utilizados para alimentar sistemas que não contem com a aprovação e supervisão da comunidade surda.

b) Centralidade surda na anotação: os "humanos no ciclo" não podem ser trabalhadores precarizados. É metodologicamente desejável que sejam linguistas surdos, profissionais formados e empregados como especialistas com remuneração adequada, que guiam treinamento da IA e percebem nuances que não-sinalizante ou IA jamais identificaria. Expertise linguística surda não é consultiva; é central.

c) Foco na documentação, não substituição: finalidade declarada deve ser explícita e inequivocamente linguístico-descritiva e pedagógica, não assistivo-substitutiva. Objetivo é criar dicionários respeitando variação regional, materiais didáticos culturalmente apropriados, ferramentas para teoria linguística, recursos que empoderem comunidade surda. Aplicações de tradução que substituam intérpretes são explicitamente excluídas.

d) Mitigação ativa de viés: *pipeline* de coleta não pode simplesmente minerar dados mais acessíveis (que sobre-representam populações urbanas, jovens, Sudeste). Deve haver gerenciamento ativo para buscar proativamente dados de regiões sub-representadas (Norte, Centro-Oeste), diferentes etnias, faixas etárias (incluindo idosos), contextos socioeconômicos. Este balanceamento não é opcional; é metodologicamente essencial.

e) Transparência e consentimento: todos sinalizantes cujos vídeos são incluídos devem, sempre que possível, ser contatados para consentimento informado explícito. Quando inviável, deve haver transparência total sobre existência do *corpus*, acesso aberto a pesquisadores surdos e mecanismos claros para remoção mediante solicitação.

Assim, qualquer pesquisa liderada por surdos não é apenas eticamente correta abstratamente, mas o que garante relevância linguística e validade científica. Sem expertise de linguistas surdos guiando integralmente o processo, *corpus* correria risco de ser apenas mais um *dataset* de engenharia: vasto em tamanho mas fundamentalmente pobre em compreensão linguística, ou seja, um repositório de dados que falharia em capturar a língua como fenômeno social e cultural.

7 Conclusões

Este artigo apresentou *blueprint* metodológico detalhado, tecnicamente especificado e eticamente robusto para criar novo recurso linguístico para Libras. Demonstramos que estudo sociolinguístico quantitativo em larga escala da variação lexical está paralisado por gargalo duplo: inviabilidade escalar da anotação manual e perda de informação querológica na práxis de glosas conceituais.

Propusemos arquitetura interdisciplinar que supera ambos gargalos unindo avanços de Visão Computacional com Sociolinguística Computacional: substituímos "tradução" (glosa) por "transcrição" articulatória quantificável usando estimativa de pose otimizada (MediaPipe com seleção de subconjuntos conforme dos Santos *et al*.

2025); substituímos "anotação manual completa" por "validação *human-in-the-loop*" empregando Transformers para sugestões e linguistas surdos para validação; habilitamos descoberta não supervisionada de variantes regionais através de análise de *clustering*, movendo geossociolinguística da Libras para domínio rigorosamente quantitativo.

Crucialmente, esta proposta técnica só é viável e responsável quando rigorosamente enquadrada em *framework* ético robusto. "Pesquisa Liderada por Surdos" foi apresentada não como recomendação aspiracional, mas como pilar metodológico central operacional impedindo que projeto replique vieses sistêmicos ou contribua para erosão de direitos linguísticos.

O próximo passo é a implementação-piloto focada em validar o ciclo de *bootstrapping* da Seção 4. Este piloto consistirá em:

1. **Desenvolvimento da interface:** Construção da interface de validação com os módulos de anotação e visualização de *landmarks*.
2. **Criação do "Corpus Ouro":** Parceria com linguistas surdos para a anotação manual do conjunto inicial (aprox. 50-100 lemas de alta variação), treinando o *Seed Model v0.1*.
3. **Validação do *Active Learning*:** Testar o *pipeline* de Aprendizagem Ativa em um lote de "Dados Dirigidos", medindo a redução do esforço de anotação (tempo por validação) à medida que o modelo é retreinado iterativamente (v0.2, v0.3).

Paralelamente, será formalizado o comitê de governança comunitária, envolvendo as associações de surdos e grupos de pesquisa liderados por surdos desde esta fase de concepção, garantindo a aplicação dos princípios da Seção 6.

Trabalhos futuros devem explorar extensão da metodologia além da variação lexical: análise de variação em parâmetros querológicos específicos, identificação de variações morfossintáticas regionais, criação de modelos preditivos informando materiais pedagógicos celebrando variação em vez de suprimi-la.

A metodologia oferece caminho metodologicamente sólido, tecnicamente viável e eticamente defensável para criar o maior e mais detalhado *corpus*

lexicográfico-variacionista da Libras. Este recurso teria implicações profundas para lexicografia prática (dicionários regionais), tecnologia educativa respeitando diversidade linguística, políticas públicas de planejamento fundamentadas em dados empíricos e avanço da teoria sociolinguística aplicada a línguas visuais-espaciais.

Referências

BRAGG, D. *et al*. Sign language recognition, generation, and translation: An interdisciplinary perspective. *In*: **THE 21ST INTERNATIONAL ACM SIGACCESS CONFERENCE ON COMPUTERS AND ACCESSIBILITY**, 2019. p. 16-31. DOI https://doi.org/10.1145/3308561.3353774

CAMGÖZ, N. C. *et al*. 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**, 2020. p. 10023-10033. Disponível em: https://arxiv.org/abs/2003.13830. Acesso em: 30 jan. 2025.

CAO, Z. *et al*. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. **IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE**, v. 43, n. 1, p. 172-186, 2021. DOI https://doi.org/10.1109/TPAMI.2019.2929257

DE MEULDER, M. The legal recognition of sign languages. **Sign Language Studies**, v. 15, n. 4, p. 498-506, 2015. DOI https://doi.org/10.1353/sls.2015.0018

DESAI, S. *et al*. Artificial intelligence in sign language research: Systematic review and future directions. **ACM Computing Surveys**, v. 56, n. 3, 2024.

DOS SANTOS, D. L. V. *et al*. Proper body landmark subset enables more accurate and 5X faster recognition of isolated signs in LIBRAS. **arXiv preprint** arXiv:2510.24887, 2025. Disponível em: https://arxiv.org/abs/2510.24887. Acesso em: 30 jan. 2025.

GRIEVE, J.; SPEELMAN, D.; GEERAERTS, D. A statistical method for the identification and aggregation of regional linguistic variation. **Language Variation and Change**, v. 23, n. 2, p. 193-221, 2011. DOI https://doi.org/10.1017/S095439451100007X

HIROOKA, K. *et al*. Stack Transformer based spatial-temporal attention model for dynamic sign language and fingerspelling recognition. **arXiv preprint** arXiv:2503.16855, 2025.

HOVY, D.; JOHANNSEN, A. Computational sociolinguistics. *In*: **PROCEEDINGS OF THE 15TH CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS** (EACL), 2017.

IEEE. **IEEE P7000 series - Padrões éticos para sistemas autônomos e inteligentes**. Disponível em: https://standards.ieee.org/. Acesso em: 10 jan. 2025.

LUGARESI, C. *et al*. MediaPipe: A framework for building perception pipelines. **arXiv preprint** arXiv:1906.08172, 2019. Disponível em: https://arxiv.org/abs/1906.08172. Acesso em: 30 jan. 2025.

MACHADO, V. L. V. **Análise da variação lexical em Libras**. Repositório UFSC, 2018. Disponível em: https://repositorio.ufsc.br/. Acesso em: 15 jan. 2025.

MERCANOGLU, O.; KELES, H. AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods. **arXiv preprint** arXiv:2001.08078, 2020.

NIST. **Artificial Intelligence Risk Management Framework (AI RMF 1.0)**. Disponível em: https://www.nist.gov/itl/ai-risk-management-framework. Acesso em: 10 jan. 2025. DOI https://doi.org/10.6028/NIST.AI.100-1.jpn

OLIVEIRA, L. A.; SILVA, M. P. S. C.; CAMPELO, W. N. M. Variações linguísticas na Libras: particularidades entre as formas de comunicação/sinalização. **Revista Cocar**, v. 4, 2020.

QUADROS, R. M. **Língua de sinais brasileira**: estudos linguísticos. Porto Alegre: Artmed, 2016.

QUADROS, R. M.; CRUZ, C. R. **Língua de sinais**: instrumentos de avaliação. Porto Alegre: Artmed, 2011.

QUADROS, R. M.; KARNOPP, L. B. **Língua de Sinais Brasileira**: estudos linguísticos. Porto Alegre: Artmed, 2004. DOI https://doi.org/10.18309/anp.v1i16.560

REZENDE, T. M.; ALMEIDA, S. G. M.; GUIMARÃES, F. G. Development and validation of a Brazilian sign language database for human gesture recognition. **Research on Biomedical Engineering**, v. 37, n. 4, p. 583-595, 2021. DOI

SANTOS, J. B. **A variação lexical em Libras em três municípios do Estado de Alagoas**. Dissertação (Mestrado em Linguística e Literatura) – Universidade Federal de Alagoas, Maceió, 2020.

SILVA, K. A. A transcrição de textos do *Corpus* de Libras. *In*: **ANAIS DO VIII SIMPÓSIO INTERNACIONAL DE ESTUDOS DE GÊNEROS TEXTUAIS**, 2015.