



Libras e inteligência artificial: desafios lexicográficos

Libras and Artificial Intelligence: lexicographic challenges

Eduardo Lucas Sousa ENÉAS*

Bárbara Tayná Santos Eugênios da Silva DANTAS**

Thereza Sophia Jacome PIRES***

RESUMO: Este artigo discute os desafios teórico-metodológicos do tratamento lexicográfico da Língua Brasileira de Sinais (Libras) na era da inteligência artificial, com atenção aos riscos e às possibilidades abertas por tecnologias de reconhecimento, tradução, anotação e organização automatizada de dados multimodais. Parte-se da compreensão da Libras como língua visual-espacial, dotada de organização gramatical própria, variação linguística e modos específicos de constituição lexical e terminológica. Nesse sentido, problematiza-se a tendência de reduzir sinais a equivalentes em português ou a glosas, sobretudo em contextos digitais e computacionais nos quais a língua sinalizada passa a ser convertida em dado processável. Metodologicamente, o estudo caracteriza-se como uma reflexão teórico-crítica, fundamentada em pesquisas sobre lexicografia e terminologia da Libras, dicionários e glossários digitais, produção de sinais-termos, processamento computacional de línguas de sinais, *datasets*, anotação, vieses algorítmicos e participação surda em projetos de inteligência artificial. A discussão evidencia que dicionários, glossários e plataformas digitais em Libras desempenham papel relevante na documentação do léxico comum e especializado, mas ainda enfrentam problemas como dispersão dos registros, heterogeneidade de formatos, fragilidade de acesso, insuficiência de metadados, lacunas de validação e dependência da língua portuguesa como eixo principal de busca. Argumenta-se que a inteligência artificial pode contribuir para a organização, indexação, recuperação e comparação de sinais, desde que não substitua a análise linguística, a curadoria humana e a validação comunitária. O artigo destaca, ainda, que *datasets* pouco representativos, modelos centrados na glosa e métricas descontextualizadas podem reforçar apagamentos, padronizações indevidas e desigualdades epistêmicas. Como encaminhamento, propõe-se uma lexicografia da Libras multimodal, ética e participativa, orientada pela descrição linguística da língua sinalizada, pelo registro da variação, pela governança ética dos dados e pela centralidade da comunidade surda na produção, validação e circulação dos recursos lexicográficos. Conclui-se que o uso da

* Doutorando pelo Programa de Pós-Graduação em Linguística da Universidade Federal da Paraíba (PROLING/UFPB). Professor do Departamento de Língua Portuguesa e Linguística da UFPB. João Pessoa, PB – Brasil. eduardolucas.vip@gmail.com

** Doutoranda pelo Programa de Pós-Graduação em Linguística da Universidade Federal da Paraíba (PROLING/UFPB). Professora do curso de Fonoaudiologia da Faculdade Maurício de Nassau, em Campina Grande. João Pessoa, PB – Brasil. barbarataynaseds@gmail.com

*** Doutora em Linguística pelo Programa de Pós-Graduação em Linguística da Universidade Federal da Paraíba (PROLING/UFPB). Professora adjunta do curso de Psicopedagogia da UFPB. João Pessoa, PB – Brasil. jacomethereza@gmail.com

inteligência artificial nesse campo deve ser compreendido como apoio sociotécnico situado, e não como solução autônoma para os desafios lexicográficos da Libras.

PALAVRAS-CHAVE: Libras. Lexicografia. Inteligência Artificial. Glosa. Comunidade surda.

ABSTRACT: This article discusses the theoretical and methodological challenges of the lexicographic treatment of Brazilian Sign Language (Libras) in the age of artificial intelligence, with attention to the risks and possibilities opened by technologies for recognition, translation, annotation, and automated organization of multimodal data. The discussion is based on the understanding of Libras as a visual-spatial language, endowed with its own grammatical organization, linguistic variation, and specific modes of lexical and terminological constitution. In this sense, the article problematizes the tendency to reduce signs to Portuguese equivalents or glosses, especially in digital and computational contexts in which the signed language is converted into processable data. Methodologically, the study is characterized as a theoretical-critical reflection grounded in research on Libras lexicography and terminology, digital dictionaries and glossaries, the production of term-signs, computational processing of sign languages, *datasets*, annotation, algorithmic biases, and Deaf participation in artificial intelligence projects. The discussion shows that dictionaries, glossaries, and digital platforms in Libras play a relevant role in documenting both general and specialized lexicon, but still face problems such as dispersion of records, heterogeneity of formats, fragile access, insufficient metadata, gaps in validation, and dependence on Portuguese as the main search axis. It is argued that artificial intelligence may contribute to the organization, indexing, retrieval, and comparison of signs, provided that it does not replace linguistic analysis, human curation, and community validation. The article also emphasizes that poorly representative *datasets*, gloss-centered models, and decontextualized metrics may reinforce erasures, undue standardization, and epistemic inequalities. As a way forward, the article proposes a multimodal, ethical, and participatory lexicography of Libras, guided by the linguistic description of the signed language, the documentation of variation, the ethical governance of data, and the centrality of the Deaf community in the production, validation, and circulation of lexicographic resources. It concludes that the use of artificial intelligence in this field should be understood as situated sociotechnical support, rather than as an autonomous solution to the lexicographic challenges of Libras.

KEYWORDS: Brazilian Sign Language. Lexicography. Artificial Intelligence. Gloss. Deaf community.

Artigo recebido em: 29.06.2026

Artigo aprovado em: 04.09.2026

1 Introdução

A consolidação da Língua Brasileira de Sinais (Libras) como objeto legítimo de investigação linguística, educacional e sociopolítica tem produzido deslocamentos importantes nos modos de compreender acessibilidade, educação bilíngue, tradução, interpretação e produção de conhecimento em comunidades surdas. Para além de seu

reconhecimento jurídico como meio legal de comunicação e expressão no Brasil, a Libras deve ser compreendida como língua natural de modalidade visual-espacial, dotada de organização gramatical própria e de recursos fonológico-querológicos, morfológicos, sintáticos, semânticos, pragmáticos e discursivos que não se reduzem à estrutura da língua portuguesa (Brasil, 2002; Brasil, 2005). Essa compreensão é decisiva para o tratamento lexicográfico da Libras, pois o registro de unidades lexicais e terminológicas em línguas de sinais demanda procedimentos distintos daqueles tradicionalmente mobilizados na descrição de línguas orais-auditivas.

Nesse cenário, dicionários, glossários, bancos de sinais e plataformas digitais em Libras assumem papel relevante como artefatos linguísticos, pedagógicos, tradutórios e políticos. A produção de sinais-termos tem respondido a lacunas terminológicas presentes em contextos escolares, acadêmicos, científicos e profissionais, ampliando as condições de acesso de pessoas surdas a conceitos especializados em diferentes áreas do conhecimento (Albres, 2024; Carvalho; Kumada; Benitez, 2025). Entretanto, a expansão desses recursos não elimina problemas teórico-metodológicos. A literatura recente evidencia heterogeneidade quanto à organização, ao formato de divulgação, à estrutura dos verbetes, aos critérios de validação e às formas de acesso. Materiais distribuídos em sites, vídeos, canais de YouTube, anexos acadêmicos, DVDs, slides e impressos dificultam a recuperação sistemática dos dados, a interoperabilidade entre recursos e a circulação social dos sinais-termos (Goettert; Bevilacqua, 2023; Carvalho; Kumada; Benitez, 2025).

É nesse ponto que o tratamento lexicográfico da Libras precisa ultrapassar a simples compilação de equivalências entre palavras em português e sinais. A entrada lexicográfica em uma língua de sinais não pode ser concebida como mera correspondência entre uma palavra escrita e um vídeo sinalizado. O verbete deve articular diferentes camadas de informação: forma sinalizada, parâmetros fonológico-querológicos, configuração de mão, localização, movimento, orientação, expressões não manuais, acepções, contextos de uso, variantes, domínio de especialidade, relações

com outros sinais e, quando pertinente, equivalentes em português. Trata-se de uma lexicografia necessariamente multimodal, na qual imagem, vídeo, descrição linguística, escrita e metadados são componentes constitutivos do registro (Goettert; Bevilacqua, 2023; Carvalho; Kumada; Benitez, 2025).

A emergência da inteligência artificial, especialmente por meio de técnicas de visão computacional, aprendizagem profunda, processamento multimodal, reconhecimento automático, tradução e geração de línguas de sinais, introduz novas possibilidades para esse campo. Sistemas computacionais podem auxiliar na indexação de acervos visuais, na recuperação de sinais por similaridade formal, na anotação semiautomática de corpora, na identificação de padrões em *datasets* e na organização de bancos de sinais. No caso da Libras, estudos recentes apontam crescimento de pesquisas sobre reconhecimento de sinais, tradução automática, tecnologias assistivas e construção de bases multimodais (Cerna *et al.*, 2021; Santos *et al.*, 2025). Esse movimento acompanha tendências internacionais de processamento computacional de línguas de sinais (Tan *et al.*, 2024; Zaineldin *et al.*, 2024).

Contudo, a incorporação da inteligência artificial ao campo das línguas de sinais não pode ser assumida como avanço neutro, inevitável ou automaticamente inclusivo. A literatura crítica tem demonstrado que muitos sistemas são desenvolvidos a partir de pressupostos redutores sobre surdez, acessibilidade e linguagem, frequentemente orientados pela ideia de superar “barreiras de comunicação” entre pessoas surdas e ouvintes. Essa perspectiva pode obscurecer a autonomia linguística das línguas de sinais e produzir soluções guiadas pela conveniência técnica ou pelo olhar de pesquisadores ouvintes e não sinalizantes, e não pelas demandas epistemológicas, culturais e políticas das comunidades surdas (Desai *et al.*, 2024; De Meulder, 2025; Kamikubo *et al.*, 2025).

Um dos riscos centrais desse processo é o uso acrítico da glosa. Embora seja útil em determinados contextos de descrição e anotação, a glosa não constitui representação plena da língua de sinais. A redução de sinais a palavras escritas pode

apagar aspectos fonológico-querológicos, morfossintáticos, espaciais, icônicos, discursivos e pragmáticos da sinalização. Quando modelos computacionais assumem a glosa como unidade central de reconhecimento, tradução ou indexação, uma representação parcial passa a funcionar como equivalente linguístico totalizante. No âmbito lexicográfico, esse risco é ainda mais sensível: se a inteligência artificial operar apenas sobre pares “glosa-sinal” ou “palavra em português-vídeo”, o glossário resultante poderá reproduzir uma concepção empobrecida da Libras, subordinando sua organização lexical à lógica da língua portuguesa (Tan *et al.*, 2024; Desai *et al.*, 2024).

Outro problema refere-se à variação linguística e à governança dos dados. Como qualquer língua natural, a Libras apresenta variação regional, social, geracional, discursiva e comunitária, que não deve ser tratada como erro, ruído ou desvio. No entanto, procedimentos automatizados frequentemente dependem de *datasets* “limpos”, categorias pré-definidas e modelos treinados para reduzir ambiguidades, o que pode induzir apagamento de variantes e seleção arbitrária de uma forma como padrão (Desai *et al.*, 2024; De Meulder, 2025). Além disso, vídeos de sinalização envolvem corpo, rosto, identidade, pertencimento comunitário, expressão visual e informações sensíveis. Por isso, corpora, *datasets* e bancos lexicográficos de Libras demandam atenção a consentimento, privacidade, autoria, uso secundário dos dados e participação comunitária (Desai *et al.*, 2024; De Meulder, 2025; Kamikubo *et al.*, 2025).

Diante desse quadro, este artigo tem como objetivo discutir os desafios teórico-metodológicos do tratamento lexicográfico da Libras na era da inteligência artificial, problematizando riscos de automatização, apagamento da variação e produção de vieses, com vistas à proposição de princípios para uma lexicografia multimodal, ética e participativa. Parte-se da tese de que a inteligência artificial pode contribuir para a organização, recuperação, anotação e circulação de recursos lexicográficos em Libras, desde que seu uso seja orientado por critérios linguísticos, lexicográficos, éticos e comunitários. Caso contrário, há o risco de reforçar vieses já presentes na pesquisa

computacional sobre línguas de sinais, reduzir a Libras à glosa e deslocar a comunidade surda da posição de autora, usuária e validadora do conhecimento sobre sua própria língua.

Para desenvolver essa discussão, o artigo organiza-se em seis seções. Após esta introdução, a segunda seção apresenta os pressupostos sobre Libras, léxico, sinais-terminos e lexicografia digital. A terceira examina contribuições e limites da inteligência artificial, dos corpora multimodais e da anotação computacional de línguas de sinais. A quarta problematiza os riscos da automatização, especialmente no que se refere à glosa, aos *datasets*, à variação linguística, aos vieses algorítmicos e aos direitos linguísticos. A quinta propõe princípios para uma lexicografia da Libras multimodal, ética e participativa. Por fim, a sexta seção apresenta as considerações finais.

2 Libras, léxico e lexicografia digital: dicionários, glossários e sinais-terminos

A discussão sobre o tratamento lexicográfico da Libras exige, como ponto de partida, o reconhecimento de que as línguas de sinais não constituem sistemas periféricos, derivados ou incompletos em relação às línguas orais. Trata-se de línguas naturais, estruturadas em modalidade visual-espacial, nas quais a produção e a recepção linguística se organizam por meio de recursos corporais, espaciais, gestuais, visuais e expressivo-faciais. No caso da Libras, seu reconhecimento legal como meio de comunicação e expressão no Brasil contribuiu para ampliar sua presença em contextos educacionais, institucionais e acadêmicos, mas sua legitimidade linguística não decorre apenas do reconhecimento jurídico: decorre de sua organização gramatical própria e de sua capacidade de produzir sentidos em diferentes esferas da vida social (Brasil, 2002; Brasil, 2005).

A unidade lexical em Libras, frequentemente denominada sinal, não corresponde de modo simples a uma palavra do português. Embora, em práticas didáticas, tradutórias e lexicográficas, seja comum estabelecer relações de equivalência entre sinais e palavras, tal equivalência é sempre parcial, situada e dependente do

contexto. Um mesmo sinal pode ser traduzido por diferentes palavras em português, assim como uma palavra em português pode demandar diferentes sinais, a depender do domínio semântico, do contexto discursivo, do registro, da variedade regional e das escolhas da comunidade sinalizante. Desse modo, a equivalência lexical entre Libras e português deve ser compreendida como operação analítica e tradutória, e não como correspondência natural, estável ou biunívoca entre sistemas linguísticos.

Essa questão é particularmente relevante para a lexicografia bilíngue Libras-português. A lexicografia, como campo voltado à elaboração e à análise de obras como dicionários, glossários e vocabulários, precisa considerar os usuários, as funções da obra, os critérios de seleção das entradas e o modo de apresentação das informações lexicais. Em línguas de sinais, tais decisões assumem complexidade particular, pois o verbete precisa lidar com uma materialidade linguística visual-espacial. A forma de citação do sinal não se realiza prioritariamente por grafia alfabética, mas por registros visuais, vídeos, fotografias, desenhos, descrições formais, sistemas de notação ou recursos multimodais. Por isso, um dicionário ou glossário de Libras que se limite a apresentar uma palavra em português acompanhada de um vídeo do sinal tende a reduzir a complexidade da língua e a empobrecer a descrição da unidade lexical (Goetttert; Bevilacqua, 2023).

O tratamento lexicográfico da Libras precisa articular dimensões formais, semânticas, discursivas e sociolinguísticas. No plano formal, é necessário descrever elementos como configuração de mão, localização, movimento, orientação da palma, número de mãos, contato corporal, direção do movimento, uso do espaço e expressões não manuais. No plano semântico, importa explicitar acepções, restrições de uso, relações de sinonímia, polissemia, especialização terminológica e possíveis diferenças entre sinal comum e sinal-termo. No plano discursivo, devem ser considerados contextos de uso, exemplos e situações comunicativas em que determinada forma é preferida. No plano sociolinguístico, torna-se relevante registrar variantes regionais,

institucionais, comunitárias, geracionais ou profissionais, evitando a naturalização de uma única forma como padrão absoluto.

A distinção entre léxico comum e léxico especializado é central para essa discussão. A lexicologia e a lexicografia ocupam-se, de modo amplo, do estudo e da descrição do léxico geral das línguas, ao passo que a terminologia e a terminografia focalizam unidades pertencentes a campos especializados do conhecimento. No âmbito da Libras, essa distinção permite compreender a especificidade dos sinais-terms, isto é, unidades lexicais criadas, estabilizadas ou mobilizadas para designar conceitos técnicos, científicos, acadêmicos ou profissionais. A produção de sinais-terms está diretamente relacionada às demandas educacionais e acadêmicas de estudantes surdos, professores, pesquisadores e tradutores-intérpretes, sobretudo porque a ausência de sinais para conceitos especializados pode interferir na mediação da aprendizagem e na participação em diferentes áreas do conhecimento (Albres, 2024).

A emergência dos sinais-terms acompanha a ampliação da presença de pessoas surdas em contextos escolares, universitários, profissionais e científicos. À medida que a Libras passa a circular em espaços institucionais cada vez mais diversificados, aumenta a necessidade de nomear conceitos que antes estavam restritos, sobretudo, à língua portuguesa escrita ou oralizada. Essa necessidade não se limita à criação de novos sinais; envolve processos de documentação, validação, circulação, ensino e uso efetivo pelas comunidades surdas e pelos profissionais que atuam em contextos bilíngues. Um sinal-termo não se legitima apenas por ter sido criado ou registrado. Sua pertinência depende de adequação conceitual, conformidade com as possibilidades morfológicas da Libras, inteligibilidade e aceitação em práticas sociais concretas.

Nesse sentido, a produção terminológica em Libras traz benefícios e riscos. Por um lado, amplia o acesso de pessoas surdas a conhecimentos especializados e fortalece a presença da Libras em ambientes acadêmicos e científicos. Por outro, quando

realizada sem critérios linguísticos consistentes e sem participação efetiva da comunidade surda, pode gerar duplicidade terminológica, inconsistências conceituais e dispersão de registros. Carvalho, Kumada e Benitez (2025) observam que a expansão de glossários terminológicos em Libras representa avanço importante, mas pode envolver problemas como erros conceituais, criação de sinais para conceitos que já possuem correspondência na língua, ausência de consonância com regras morfológicas da Libras e falta de validação pela comunidade surda. A existência de múltiplos sinais para um mesmo conceito não é, por si só, um problema, uma vez que a variação é constitutiva das línguas naturais. O problema emerge quando essa multiplicidade não é descrita, contextualizada ou validada.

A variação lexical, portanto, deve ser compreendida como objeto de descrição, e não como falha a ser eliminada. Em perspectiva sociolinguística, variações regionais, institucionais e comunitárias expressam a vitalidade da língua e a diversidade de seus usos. No campo da lexicografia, registrar variantes significa reconhecer que uma língua não se organiza por uma forma única, homogênea e descontextualizada. No caso da Libras, a ausência de registro adequado da variação pode produzir efeitos de apagamento, sobretudo quando determinado glossário ou dicionário passa a circular como referência única em contextos educacionais, tradutórios ou tecnológicos. Assim, a lexicografia da Libras precisa equilibrar duas exigências: oferecer condições de consulta, sistematização e inteligibilidade, sem converter a padronização em supressão da diversidade linguística.

Esse equilíbrio depende, em grande medida, da distinção entre as estruturas canônicas da obra lexicográfica. Em sentido estrito, a macroestrutura corresponde ao conjunto ordenado dos lemas ou entradas que formam a nomenclatura do dicionário ou glossário. Sua definição envolve decidir quantas e quais unidades serão lematizadas e segundo quais critérios serão ordenadas; escopo e domínio temático condicionam essa seleção, mas público-alvo, função e tipo de obra constituem parâmetros prévios do projeto lexicográfico que orientam tanto a macroestrutura

quanto a microestrutura. Esta última diz respeito às informações organizadas em cada verbete. Já a medioestrutura compreende o sistema de remissões e conexões entre verbetes e outras partes do recurso (Hausmann; Wiegand, 1989; Bugueño Miranda, 2013). Em Libras, essa distinção é particularmente relevante: a nomenclatura não precisa ser subordinada apenas à ordem alfabética do português, podendo ser selecionada e ordenada segundo critérios compatíveis com a estrutura visual-espacial da língua, desde que esses critérios sejam explicitados.

A microestrutura, por sua vez, refere-se à organização interna do verbete. Em um recurso lexicográfico de Libras, ela deve contemplar não apenas o equivalente em português, mas também o vídeo do sinal, a descrição de seus parâmetros formais, a definição, o contexto de uso, exemplos, variantes, marcação de área de conhecimento, fonte do registro, informações sobre autoria ou grupo de validação e possíveis remissões para sinais relacionados. Quando se trata de sinais-termos, a microestrutura também deve explicitar o conceito especializado designado, evitando que a definição seja apenas uma tradução aproximada do termo em português. A ausência desses elementos fragiliza o valor linguístico, pedagógico e científico do glossário.

A necessidade de explicitar essas decisões pode ser observada em obras já existentes. Na análise do Dicionário da Língua Brasileira de Sinais, Goetttert e Bevilacqua (2023) destacam uma microestrutura ampla, com aceção, exemplos em português e Libras, imagem, vídeo, configuração de mão, classe gramatical e origem, mas observam que a obra não explicita de modo suficiente o usuário previsto nem sua função. No Spread the Sign-Brasil, os autores identificam lacunas de outra natureza: as buscas partem do português, sem possibilidade de acesso a partir dos sinais, e faltam exemplos contextualizados em Libras e em português (Goetttert; Bevilacqua, 2024). Esses casos mostram que a presença de vídeos ou de grande quantidade de entradas não garante, por si só, qualidade lexicográfica. É necessário articular a seleção e a ordenação da nomenclatura, as rotas de acesso, a microestrutura informativa, as

remissões e a definição clara de usuários e funções, evitando que o português permaneça como única porta de entrada e eixo organizador da consulta.

Nos glossários terminológicos, essa exigência torna-se ainda mais decisiva. Diferentemente dos dicionários de língua geral, glossários especializados lidam com conceitos técnicos e científicos, que exigem precisão semântica, delimitação conceitual e validação no interior de uma área do conhecimento. A produção de sinais-termos em Libras tem se expandido em campos como Matemática, Química, Biologia, Direito, Economia, Música, Engenharia, Nutrição, Educação Física, Geografia, Tradução, Letras-Libras e contextos acadêmicos e escolares (Albres, 2024; Carvalho; Kumada; Benitez, 2025). Essa expansão revela a vitalidade da Libras e sua capacidade de responder a demandas conceituais específicas, mas também evidencia a necessidade de critérios rigorosos de registro, definição, circulação e validação.

Muitos glossários de Libras têm origem em pesquisas de mestrado, doutorado, projetos institucionais ou iniciativas vinculadas a cursos de formação. Essa vinculação acadêmica é positiva, pois indica que a produção terminológica ocorre em diálogo com investigações linguísticas, pedagógicas e tradutórias. No entanto, a circulação desses materiais ainda é fragmentada. Carvalho, Kumada e Benitez (2025) mostram que os glossários documentados na literatura aparecem em formatos variados, como sites, blogs, vídeos, materiais impressos, livros digitais, DVDs e trabalhos acadêmicos, nem sempre disponíveis ou facilmente recuperáveis. Essa dispersão dificulta a comparação entre obras, a atualização das entradas e o uso sistemático por estudantes, professores, tradutores, intérpretes e pesquisadores.

O problema, portanto, não se reduz à quantidade de sinais registrados, mas à ausência de integração entre os registros existentes. Quando os glossários permanecem isolados em plataformas distintas, sem padrões mínimos de descrição, metadados, autoria, validação e atualização, torna-se difícil identificar quais sinais-termos já foram documentados, quais variantes circulam, quais áreas estão mais contempladas e quais lacunas permanecem. Por isso, plataformas capazes de reunir, organizar ou

interoperar diferentes glossários de Libras constituem uma demanda central para o avanço da lexicografia e da terminografia da língua.

A digitalidade oferece condições favoráveis para esse tipo de tratamento. Diferentemente de materiais impressos, plataformas digitais permitem inserir vídeos em alta qualidade, múltiplas variantes de um mesmo sinal, mecanismos de busca avançada, filtros por área temática, links entre entradas, campos de metadados, atualização contínua e recursos colaborativos. Também possibilitam a inclusão de vídeos explicativos em Libras, tutoriais de uso, espaços de feedback e mecanismos de revisão. Essas possibilidades, entretanto, só se realizam plenamente quando há um projeto lexicográfico consistente. A tecnologia amplia as formas de registro, mas não substitui a fundamentação linguística, a curadoria humana e a validação comunitária.

A glosa ocupa lugar ambíguo nesse processo. Por um lado, funciona como recurso de anotação, indexação e mediação entre línguas, permitindo que pesquisadores, tradutores e usuários localizem sinais por meio de uma representação escrita. Por outro, quando tomada como substituto da língua sinalizada, produz uma simplificação problemática. A glosa não captura integralmente a estrutura da Libras, tampouco representa de forma suficiente aspectos espaciais, simultâneos, não manuais e discursivos da sinalização. Em obras lexicográficas, deve ser entendida como elemento auxiliar de acesso, e não como núcleo definidor do sinal. Sua centralidade excessiva pode produzir a falsa impressão de que a Libras se organiza como espelhamento lexical do português, problema também apontado nas discussões recentes sobre processamento computacional de línguas de sinais (Desai *et al.*, 2024; Tan *et al.*, 2024).

3 Inteligência artificial, *corpus* e anotação em línguas de sinais

A entrada da inteligência artificial no campo das línguas de sinais tem sido marcada por avanços em reconhecimento automático, tradução, geração de sinais, visão computacional e processamento multimodal. De modo geral, esse campo

envolve pelo menos três frentes principais: o reconhecimento de sinais, a tradução entre línguas sinalizadas e línguas orais/escritas, e a geração automática de sinais por meio de avatares, vídeos sintéticos ou outros recursos computacionais. Embora essas frentes estejam relacionadas, elas não devem ser confundidas. Reconhecer um sinal, traduzi-lo e gerar uma forma sinalizada são tarefas distintas, que mobilizam diferentes modelos de língua, diferentes tipos de dados e diferentes critérios de avaliação (Tan *et al.*, 2024).

O reconhecimento automático de sinais costuma ter como objetivo identificar, em imagens ou vídeos, unidades sinalizadas isoladas ou sequências de sinais. Para isso, modelos computacionais podem utilizar dados visuais, como imagens RGB, dados de profundidade, esqueleto corporal, pontos de articulação, movimentos das mãos, expressões faciais e informações temporais. A tradução automática, por sua vez, envolve a conversão de uma sequência em língua de sinais para uma sentença em língua oral/escrita, ou o caminho inverso, quando se busca gerar sinais a partir de texto ou fala. A geração de sinais é ainda mais complexa, pois demanda não apenas selecionar equivalentes lexicais, mas produzir movimentos corporais, expressões não manuais, ritmo, espacialização e articulações compatíveis com a gramática visual-espacial da língua.

Essas distinções são fundamentais para o debate lexicográfico. Um sistema capaz de classificar configurações de mão ou reconhecer sinais isolados não necessariamente compreende relações semânticas, variações regionais, usos discursivos ou diferenças entre léxico comum e terminologia especializada. Do mesmo modo, um sistema de tradução que converte vídeos em glosas ou textos em português não resolve, por si só, o problema da descrição lexicográfica da Libras. Para a lexicografia, interessa menos a promessa genérica de “tradução automática” e mais a pergunta sobre como sistemas computacionais podem auxiliar a documentação, a busca, a comparação, a anotação e a organização de unidades lexicais em sua complexidade formal, semântica e sociolinguística.

Nesse processo, os corpora ocupam lugar central. Um *corpus* de língua de sinais não é apenas um conjunto de vídeos armazenados, mas uma coleção organizada de dados linguísticos, produzida a partir de critérios de seleção, registro, segmentação, anotação, documentação e acesso. Para que possa sustentar pesquisas linguísticas, lexicográficas ou computacionais, o *corpus* precisa explicitar quem são os sinalizantes, em que contextos os dados foram produzidos, quais variedades linguísticas estão representadas, quais tarefas comunicativas foram realizadas, quais permissões de uso foram concedidas e quais camadas de anotação foram aplicadas. Sem esses elementos, os dados podem até alimentar modelos computacionais, mas dificilmente sustentarão uma descrição linguística robusta.

No caso da Libras, os desafios são significativos. A revisão de Santos *et al.* (2025) indica crescimento recente da produção científica brasileira sobre Libras e IA, especialmente em sistemas de tradução automática, tecnologias assistivas, ensino de Libras, reconhecimento de sinais e visão computacional. Esse crescimento revela um campo em expansão, mas também evidencia limitações quanto à naturalidade linguística, à acessibilidade efetiva e ao uso ético de dados sensíveis. Muitos *datasets* são construídos em contextos controlados, com poucos sinalizantes, vocabulário restrito, iluminação padronizada e tarefas pouco próximas dos usos cotidianos da língua. Essas condições podem ser úteis para treinamento inicial de modelos, mas limitam sua generalização para situações reais de interação.

Além disso, quando os *datasets* privilegiam sinais isolados, alfabetos manuais ou listas lexicais, deixam de contemplar fenômenos discursivos, coarticulatórios, espaciais e prosódicos que são constitutivos da Libras em uso. Iniciativas voltadas ao reconhecimento de datilologia, como a Tecnologia Assistiva SignTalk, demonstram a viabilidade de aplicações baseadas em aprendizagem de máquina para reconhecimento do alfabeto manual da Libras, mas também mostram que o reconhecimento de configurações manuais representa apenas uma parte do problema linguístico mais amplo (Benevides *et al.*, 2024). Para a lexicografia, a questão não é

apenas reconhecer uma forma visual, mas descrever unidades lexicais em relação a sentidos, usos, variantes, domínios e contextos.

A experiência do dataset LIBRAS-UFOP evidencia, ao mesmo tempo, o potencial e a complexidade desse campo. Ao organizar sinais com base em pares mínimos e registrar dados multimodais, como RGB-D e esqueleto corporal, esse tipo de iniciativa mostra que o reconhecimento automático precisa lidar com diferenças formais muito sutis (Cerna *et al.*, 2021). Alterações em configuração de mão, movimento, ponto de articulação, orientação ou expressões não manuais podem produzir distinções lexicais e semânticas relevantes. Para a lexicografia, esse dado é decisivo: a forma do sinal não pode ser tratada como imagem genérica ou gesto aproximado, pois pequenas diferenças paramétricas podem distinguir unidades lexicais distintas.

A anotação é outro ponto crucial. Anotar uma língua de sinais significa segmentar, nomear, descrever e relacionar fenômenos visuais, espaciais, corporais e discursivos. Isso pode envolver glosas, descrição de parâmetros, marcação de expressões não manuais, identificação de referentes espaciais, segmentação de unidades lexicais, indicação de variantes, tradução livre, notas semânticas e informações sobre contexto de uso. Cada camada de anotação responde a uma finalidade específica. Uma anotação voltada ao treinamento de modelos de reconhecimento pode não ser suficiente para uma análise lexicográfica; uma anotação feita para tradução pode não capturar contrastes fonológico-querológicos; uma anotação baseada apenas em glosas pode apagar aspectos decisivos da estrutura sinalizada.

A glosa, nesse cenário, tem função metodológica importante, mas limitada. Ela pode servir como recurso de indexação, busca e alinhamento entre língua sinalizada e língua escrita. No entanto, não deve ser confundida com a própria língua de sinais. Ao transformar sinais em palavras escritas de outra língua, a glosa reduz a simultaneidade, a espacialidade e a iconicidade da sinalização a uma representação

linear e parcial. Em sistemas de IA, esse problema se amplia quando a glosa passa a funcionar como unidade central de treinamento ou avaliação. Modelos podem ser considerados eficientes por preverem corretamente uma sequência de glosas, ainda que não capturem elementos não manuais, variações formais, relações espaciais ou nuances discursivas indispensáveis para a compreensão da língua (Tan *et al.*, 2024; Desai *et al.*, 2024).

Do ponto de vista lexicográfico, a anotação precisa ser pensada de modo mais amplo. Um banco lexical de Libras mediado por IA deveria permitir múltiplas camadas de descrição: vídeo do sinal, segmentação temporal, parâmetros formais, equivalentes em português, definição conceitual, exemplos de uso, variantes, domínio temático, metadados dos sinalizantes, fonte do registro, informações sobre validação e relações com outros sinais. A IA pode auxiliar esse processo ao sugerir agrupamentos por similaridade visual, detectar padrões de movimento, facilitar buscas por configuração de mão, organizar grandes acervos e apoiar a identificação de possíveis variantes. No entanto, essas funções devem ser compreendidas como apoio à curadoria linguística, e não como substituição da análise humana.

Outro aspecto importante diz respeito aos dados sintéticos e aos modelos generativos. A produção de sinais por avatares ou vídeos gerados computacionalmente pode ter aplicações educacionais e comunicacionais, mas apresenta riscos quando não respeita a naturalidade linguística, a expressividade corporal e as normas de uso da comunidade sinalizante. A geração automática de sinais-termos, por exemplo, seria especialmente problemática se reduzida a combinações formais produzidas por modelos estatísticos sem validação linguística e comunitária. Criar um sinal-termo não é apenas gerar uma configuração visual plausível; é construir uma unidade lexical capaz de designar um conceito especializado, circular em práticas sociais e ser reconhecida por usuários da língua.

A relação entre IA e lexicografia da Libras deve, portanto, ser concebida em chave sociotécnica. Isso significa reconhecer que modelos computacionais não operam

fora das decisões humanas que definem quais dados serão coletados, quais variedades serão representadas, quais categorias serão anotadas, quais métricas serão usadas e quais usuários serão considerados prioritários. Um sistema treinado com dados pouco diversos tenderá a reconhecer melhor os sinais de determinados grupos e pior os de outros. Um sistema baseado apenas em glosas tenderá a privilegiar equivalências com a língua portuguesa. Um sistema orientado exclusivamente por acurácia poderá ignorar fatores linguísticos e socioculturais essenciais para a validade do recurso.

Por isso, a construção de corpora e sistemas de anotação para a lexicografia da Libras precisa assumir alguns princípios mínimos: representatividade de sinalizantes e variedades; documentação clara dos critérios de coleta; consentimento informado e possibilidade de controle sobre o uso dos dados; anotação multimodal; registro de variantes; validação por pessoas surdas; participação de especialistas em Libras, lexicografia, terminologia e áreas de domínio; e revisão contínua dos recursos. Esses princípios não são externos à qualidade tecnológica; ao contrário, são condições para que a tecnologia produza resultados linguisticamente consistentes e socialmente responsáveis.

4 Riscos da automatização no tratamento lexicográfico da Libras

A incorporação da inteligência artificial ao tratamento lexicográfico da Libras não pode ser compreendida apenas como ampliação de eficiência técnica. Embora sistemas computacionais possam contribuir para indexar vídeos, organizar grandes bases de dados, sugerir agrupamentos formais e ampliar formas de busca, tais possibilidades não eliminam questões linguísticas, epistemológicas e éticas. Ao contrário, quando aplicadas a línguas de sinais, tecnologias de IA tendem a tornar mais visíveis tensões já presentes no campo: a dependência da glosa, a subordinação à língua oral majoritária, a busca por padronização, a fragilidade dos *datasets*, a baixa representatividade dos sinalizantes e a participação limitada da comunidade surda

nos processos de pesquisa e validação (Desai *et al.*, 2024; De Meulder, 2025; Kamikubo *et al.*, 2025).

Um primeiro risco diz respeito ao modo como muitos sistemas computacionais transformam a Libras em objeto processável. Para que dados sinalizados sejam utilizados por modelos de reconhecimento ou tradução, é comum que vídeos sejam segmentados, classificados e associados a rótulos escritos, frequentemente na forma de glosas. Esse procedimento pode ser metodologicamente útil em determinadas etapas de anotação, mas torna-se problemático quando a glosa passa a ocupar o lugar da própria língua. Tan *et al.* (2024) observam que abordagens baseadas em glosas são recorrentes em modelos de processamento de línguas de sinais, especialmente em sistemas que operam em etapas intermediárias entre o vídeo sinalizado e o texto em língua oral. No entanto, a glosa não preserva plenamente a estrutura visual-espacial da língua sinalizada, nem seus recursos simultâneos, não manuais, espaciais e discursivos (Tan *et al.*, 2024; Desai *et al.*, 2024).

No campo lexicográfico, essa redução tem consequências específicas. Um recurso que associa automaticamente sinais a palavras do português pode produzir a impressão de que a Libras é uma lista de equivalentes visuais para termos escritos. Essa lógica reforça uma concepção empobrecida de verbete, na qual o sinal aparece como ilustração ou tradução direta de uma palavra, e não como unidade lexical pertencente a um sistema linguístico próprio. Goetttert e Bevilacqua (2023) mostram que a análise de dicionários on-line de Libras exige atenção à função, ao público-alvo, ao acesso às entradas, à macroestrutura e à microestrutura das obras, o que indica que a organização lexicográfica não pode ser reduzida à simples associação entre palavra e vídeo. Assim, a glosa deve ser entendida como instrumento auxiliar de anotação e acesso, não como substituto da descrição lexicográfica (Goetttert; Bevilacqua, 2023; Desai *et al.*, 2024).

Outro risco importante é o apagamento da variação. Línguas naturais são constitutivamente variáveis, e a Libras apresenta variação regional, social, geracional,

institucional e discursiva. Essa variação não é ruído: é parte do funcionamento da língua. No entanto, muitos modelos computacionais são treinados para reduzir ambiguidades, eliminar inconsistências e privilegiar formas mais recorrentes nos dados. Quando essa lógica é transferida para a lexicografia, pode-se produzir uma padronização indevida, na qual uma variante é tomada como forma “correta” e outras são invisibilizadas. Desai *et al.* (2024) chamam atenção para o fato de que pesquisas em IA para línguas de sinais frequentemente operam com *datasets* pouco representativos e, em alguns casos, tratam a diversidade linguística como obstáculo à performance do modelo.

Essa tensão entre padronização computacional e variação linguística é especialmente sensível em glossários terminológicos. A produção de sinais-termos exige algum grau de estabilização, pois conceitos especializados precisam circular com inteligibilidade em contextos educacionais, científicos e profissionais. Entretanto, estabilizar não significa apagar variantes, nem substituir o uso comunitário por uma forma artificialmente selecionada. Albres (2023/2024) evidencia que a produção de sinais-termos em Libras está relacionada à ampliação do acesso de pessoas surdas a conceitos acadêmicos e especializados. Carvalho, Kumada e Benitez (2025), por sua vez, observam que glossários terminológicos podem apresentar problemas quando a criação de sinais ocorre sem critérios linguísticos consistentes, sem atenção às regras morfológicas da Libras ou sem validação comunitária. Quando mediada por IA, essa tarefa demanda cuidado adicional: sistemas automatizados podem reforçar a forma mais presente no dataset, não necessariamente a mais adequada, mais representativa ou mais legitimada pela comunidade (Albres, 2024; Carvalho; Kumada; Benitez, 2025; Desai *et al.*, 2024).

Os *datasets* constituem um dos pontos mais críticos dessa discussão. Um modelo de IA aprende a partir dos dados que lhe são fornecidos; portanto, a qualidade, a diversidade e a documentação desses dados influenciam diretamente os resultados produzidos. Em línguas de sinais, *datasets* compostos por poucos sinalizantes, vídeos

em ambiente controlado, vocabulário restrito ou sinais isolados podem gerar sistemas que funcionam bem em situações experimentais, mas falham diante da complexidade do uso real da língua. Santos *et al.* (2025) indicam que a produção brasileira recente sobre Libras e IA tem se expandido em reconhecimento de sinais, tradução automática, tecnologias assistivas e visão computacional, mas ainda enfrenta desafios quanto à naturalidade linguística, à acessibilidade efetiva e ao uso ético dos dados. Esses limites também aparecem em discussões internacionais sobre representatividade e desenho de *datasets* em línguas de sinais (Santos *et al.*, 2025; Desai *et al.*, 2024; Tan *et al.*, 2024).

A experiência do dataset LIBRAS-UFOP mostra que a complexidade formal dos sinais não pode ser subestimada. Ao trabalhar com pares mínimos e dados multimodais, como RGB-D e esqueleto corporal, Cerna *et al.* (2021) evidenciam que diferenças sutis em configuração de mão, movimento, localização ou orientação podem distinguir unidades lexicais. Esse dado é importante para o tratamento lexicográfico porque mostra que sinais não podem ser tratados como gestos aproximados ou imagens genéricas. Em uma plataforma lexicográfica mediada por IA, a identificação automática de similaridades visuais precisa ser acompanhada de análise linguística, pois a semelhança formal entre sinais pode envolver contrastes fonológico-querológicos e semânticos relevantes (Cerna *et al.*, 2021; Goettert; Bevilacqua, 2023).

O problema dos *datasets* não é apenas técnico, mas epistemológico. A escolha de quem sinaliza, o que será registrado, como o dado será anotado, quais variantes serão aceitas e quais métricas serão utilizadas define o tipo de língua que o sistema passa a reconhecer. Se o dataset privilegia sinais produzidos de forma isolada, o modelo aprenderá uma Libras descontextualizada. Se privilegia glosas, aprenderá uma Libras filtrada pelo português. Se exclui variantes regionais, aprenderá uma Libras artificialmente homogênea. Se não documenta os perfis dos sinalizantes, dificulta qualquer análise sobre representatividade. Desse modo, o viés algorítmico não surge apenas no funcionamento interno do modelo; ele começa nas decisões humanas que

estruturam a coleta, a seleção, a anotação e a validação dos dados (Desai *et al.*, 2024; De Meulder, 2025).

A automatização também pode intensificar assimetrias já existentes entre pesquisadores ouvintes, especialistas em computação e comunidades surdas. Kamikubo *et al.* (2025) mostram que a colaboração entre pesquisadores de aprendizado de máquina e pessoas surdas sinalizantes nem sempre ocorre em condições simétricas. Muitas vezes, a comunidade surda é convocada para fornecer dados, avaliar resultados ou validar produtos já desenhados, mas não participa da definição das perguntas de pesquisa, dos objetivos tecnológicos ou dos critérios de sucesso. Essa forma de participação limitada pode produzir tokenismo, isto é, uma inclusão superficial de pessoas surdas sem redistribuição efetiva de autoridade epistêmica (Kamikubo *et al.*, 2025; Desai *et al.*, 2024).

No tratamento lexicográfico da Libras, essa questão é central. Dicionários, glossários e bancos de sinais não são apenas ferramentas técnicas; são instrumentos de documentação, circulação e legitimação do léxico. Decidir quais sinais entram em uma base, como são definidos, quais variantes são registradas, que exemplos são apresentados e quais formas são consideradas equivalentes envolve poder linguístico. A literatura sobre glossários terminológicos em Libras já aponta que a ausência de validação comunitária e de critérios claros pode comprometer a legitimidade dos sinais-termos registrados (Carvalho; Kumada; Benitez, 2025). Se essas decisões forem transferidas para sistemas automatizados ou para equipes sem participação surda substantiva, a tecnologia poderá reforçar processos de apropriação e apagamento. A comunidade surda não deve ocupar apenas a posição de usuária final ou fonte de dados, mas a de agente na produção, curadoria, validação e governança dos recursos (Carvalho; Kumada; Benitez, 2025; Kamikubo *et al.*, 2025).

A crítica de De Meulder (2025) sobre tecnologias de IA e direitos linguísticos é particularmente pertinente. A autora argumenta que tecnologias linguísticas podem produzir efeitos de erosão quando passam a ser apresentadas como substitutas de

direitos, serviços ou mediações humanas. No campo das línguas de sinais, isso se manifesta, por exemplo, na expectativa de que sistemas automáticos possam substituir intérpretes, resolver barreiras comunicacionais ou universalizar o acesso sem enfrentar desigualdades estruturais. Essa lógica se aproxima do *technoableism*, isto é, de uma confiança excessiva em soluções tecnológicas que, sob o discurso da inclusão, podem reforçar perspectivas capacitistas e deslocar a centralidade das necessidades e escolhas das pessoas surdas (De Meulder, 2025; Desai *et al.*, 2024).

Transposta para a lexicografia, essa crítica permite problematizar a ideia de que plataformas automatizadas resolveriam, por si só, os desafios dos dicionários e glossários de Libras. A IA pode organizar dados, sugerir relações, facilitar buscas e identificar padrões, mas não define sozinha a validade linguística de um sinal-termo, não substitui a análise conceitual, não resolve a variação e não garante legitimidade comunitária. Um sistema que gera ou recomenda sinais sem participação surda e sem validação linguística pode produzir formas aparentemente funcionais, mas socialmente frágeis. Da mesma forma, uma plataforma que ranqueia resultados por frequência pode reforçar desigualdades de visibilidade entre variantes, regiões e grupos de sinalizantes (De Meulder, 2025; Kamikubo *et al.*, 2025; Carvalho; Kumada; Benitez, 2025).

Outro ponto sensível é a governança dos dados. Diferentemente de corpora escritos, bases de línguas de sinais envolvem registros visuais de corpos, rostos, identidades e modos de expressão. Vídeos de sinalização não são dados neutros nem plenamente anonimizáveis. Eles carregam marcas de autoria, corporeidade, pertencimento comunitário, idade, gênero, raça, região e trajetória linguística. Por isso, o uso desses dados em sistemas de IA exige consentimento informado, definição clara de finalidade, possibilidade de restrição de uso, cuidado com armazenamento, política de acesso e mecanismos de revisão. A coleta de vídeos para fins lexicográficos ou computacionais não deve ser tratada como mera obtenção de material linguístico, mas como prática ética situada (De Meulder, 2025; Desai *et al.*, 2024; Kamikubo *et al.*, 2025).

Esses riscos não significam que a IA deva ser rejeitada no campo da lexicografia da Libras. Significam, antes, que sua incorporação precisa ser condicionada por critérios linguísticos, lexicográficos, éticos e comunitários. A automação pode ser útil quando opera como apoio à curadoria humana: sugerindo agrupamentos, facilitando a busca por parâmetros formais, organizando metadados, comparando variantes, recuperando vídeos e ampliando a acessibilidade das plataformas. Contudo, quando se apresenta como substituição da análise linguística ou da validação comunitária, tende a transformar problemas complexos em tarefas simplificadas de classificação (Tan *et al.*, 2024; Desai *et al.*, 2024; Kamikubo *et al.*, 2025).

Assim, o principal risco da automatização não está apenas no erro técnico, mas na naturalização de uma determinada concepção de Libras. Se a língua for concebida como conjunto de gestos equivalentes a palavras do português, a IA tenderá a produzir recursos lexicográficos alinhados a essa visão redutora. Se a variação for entendida como obstáculo, a tecnologia favorecerá a padronização. Se a comunidade surda for tratada como fonte de dados, e não como sujeito epistêmico, a inovação tecnológica poderá reproduzir exclusões já conhecidas. Por isso, qualquer proposta de IA aplicada ao tratamento lexicográfico da Libras precisa começar por uma pergunta anterior à modelagem computacional: que língua se pretende documentar, para quem, com quem e sob quais condições de legitimidade? (Desai *et al.*, 2024; De Meulder, 2025; Kamikubo *et al.*, 2025).

5 Por uma lexicografia da Libras multimodal, ética e participativa

A discussão desenvolvida até aqui permite afirmar que a lexicografia da Libras, na era da inteligência artificial, não pode ser concebida como simples aplicação de ferramentas computacionais a bancos de sinais. Trata-se, antes, de um projeto linguístico, lexicográfico, ético e comunitário, no qual a tecnologia deve atuar como instrumento de apoio à documentação, à busca, à anotação, à organização e à circulação de recursos lexicais e terminológicos. Para que isso ocorra, é necessário

partir de uma concepção de Libras como língua visual-espacial, variável e socialmente situada. Quando a arquitetura técnica define previamente o que conta como unidade lexical, variante, equivalente ou erro, corre-se o risco de transformar a conveniência computacional em critério linguístico, problema já apontado nas críticas aos modelos de IA para línguas de sinais baseados em *datasets* restritos, glosas simplificadas e métricas descontextualizadas (Desai *et al.*, 2024; Tan *et al.*, 2024).

Um primeiro princípio para essa lexicografia é a multimodalidade. Um recurso lexicográfico em Libras não pode restringir-se à associação entre uma palavra em português e um vídeo sinalizado. A microestrutura do verbete precisa contemplar, sempre que possível, o vídeo do sinal, a descrição dos parâmetros formais, a definição conceitual, exemplos de uso, variantes, domínio temático, informações sobre contexto de circulação, fonte do registro, autoria ou equipe de validação e remissões para sinais relacionados. Estudos sobre dicionários on-line de Libras já indicam a importância de observar função, público-alvo, macroestrutura, microestrutura e formas de acesso às entradas, o que reforça a necessidade de uma arquitetura lexicográfica mais ampla que o par português-sinal (Goettert; Bevilacqua, 2023). Nos glossários terminológicos, essa exigência torna-se ainda mais relevante, pois a unidade registrada não designa apenas um item lexical geral, mas um conceito especializado que precisa ser definido, situado e validado.

Além da microestrutura multimodal, a arquitetura de acesso e a medioestrutura dos recursos digitais devem explorar as especificidades visuais da Libras. A consulta não precisa depender exclusivamente da ordem alfabética do português: podem coexistir buscas por configuração de mão, ponto de articulação, movimento, orientação, área temática, domínio de especialidade e variante regional. Remissões entre variantes, sinais semanticamente relacionados e entradas de um mesmo domínio também podem ampliar a navegação. Esses mecanismos não integram, em sentido estrito, a macroestrutura ou nomenclatura, mas funcionam como rotas de acesso e

conexões internas que tornam a consulta mais compatível com uma língua visual-espacial (Goettert; Bevilacqua, 2023).

Um segundo princípio é a explicitação dos critérios linguísticos e lexicográficos. Dicionários, glossários e bancos de sinais devem informar como as entradas foram selecionadas, quais fontes foram utilizadas, quem participou da elaboração, como os sinais foram validados, quais variantes foram registradas e quais critérios orientaram definições e exemplos. A literatura sobre glossários terminológicos em Libras mostra que muitos materiais circulam de forma dispersa, em sites, vídeos, anexos acadêmicos, DVDs, apresentações e plataformas diversas, nem sempre com informações suficientes sobre elaboração, validação e acesso (Carvalho; Kumada; Benitez, 2025). Assim, a qualificação da lexicografia digital da Libras não depende apenas da criação de novos recursos, mas da construção de padrões mínimos de documentação, descrição, transparência e atualização.

Um terceiro princípio é a valorização da variação. Uma plataforma lexicográfica robusta não deve tratar variantes como obstáculos à padronização, mas como dados linguísticos relevantes. Em línguas naturais, a variação expressa história, circulação social, pertencimento comunitário, regionalidade, escolarização, redes de contato e práticas institucionais. No caso da Libras, registrar variantes regionais, comunitárias, geracionais ou profissionais é uma forma de reconhecer a vitalidade da língua e evitar que uma obra lexicográfica assuma indevidamente o papel de autoridade normativa absoluta. A padronização pode ser necessária em contextos terminológicos específicos, mas deve ser construída com critérios explícitos, finalidades delimitadas e participação comunitária. Caso contrário, a busca por estabilidade pode converter-se em apagamento, sobretudo quando sistemas automatizados privilegiam a forma mais frequente no dataset como se ela fosse a forma mais legítima (Desai *et al.*, 2024).

Esse cuidado é decisivo no caso dos sinais-termos. A produção terminológica em Libras responde a demandas reais de acesso ao conhecimento acadêmico, científico e profissional, mas exige atenção à adequação conceitual e à circulação social dos

sinais. Albres (2023/2024) destaca a relevância da produção de sinais-termos para conceitos acadêmicos, enquanto Carvalho, Kumada e Benitez (2025) apontam que a criação e a documentação desses sinais podem envolver problemas quando não há critérios linguísticos consistentes, validação comunitária ou atenção à existência de sinais já em circulação. Por isso, uma lexicografia terminológica da Libras deve articular especialistas da área de conhecimento, pesquisadores da Libras, tradutores-intérpretes e pessoas surdas com experiência no domínio em questão.

Um quarto princípio é compreender a inteligência artificial como apoio à curadoria, e não como substituição da análise linguística. Sistemas computacionais podem contribuir para identificar semelhanças formais entre sinais, organizar metadados, sugerir agrupamentos, facilitar buscas por parâmetros visuais, recuperar vídeos e apoiar a comparação entre variantes. O dataset LIBRAS-UFOP, por exemplo, evidencia que o reconhecimento automático precisa lidar com diferenças formais sutis entre sinais, especialmente quando se consideram pares mínimos e dados multimodais (Cerna *et al.*, 2021). Esse tipo de contribuição pode ser valioso para plataformas lexicográficas, desde que a interpretação dos dados continue dependente de análise linguística especializada. A semelhança visual identificada por um modelo não equivale, por si só, à equivalência lexical, semântica ou terminológica.

Um quinto princípio é a crítica à centralidade da glosa. A glosa pode ser útil para anotação, busca e alinhamento entre línguas, mas não deve ser tomada como representação plena da Libras. Tan *et al.* (2024) mostram que muitos modelos de processamento de línguas de sinais operam com glosas como etapa intermediária, especialmente em tarefas de reconhecimento e tradução. Desai *et al.* (2024), por sua vez, problematizam a dependência da glosa e seus efeitos redutores na pesquisa em IA para línguas de sinais. Em uma perspectiva lexicográfica, isso significa que a glosa deve ocupar lugar auxiliar na microestrutura do verbete, sem substituir o vídeo, a descrição formal, a definição, os exemplos de uso e as informações sobre variação.

Uma lexicografia centrada na glosa tende a projetar sobre a Libras uma lógica lexical derivada do português.

Um sexto princípio é a representatividade dos dados. Bases lexicográficas e corpora utilizados em sistemas de IA precisam contemplar diferentes perfis de sinalizantes, variedades regionais, contextos de uso, domínios discursivos e formas de circulação da língua. Um sistema treinado com poucos sinalizantes, em ambiente controlado e com vocabulário restrito, dificilmente poderá sustentar generalizações sobre a Libras. A revisão de Santos *et al.* (2025) indica crescimento da produção brasileira sobre Libras e IA, mas também aponta desafios relacionados à naturalidade linguística, à acessibilidade efetiva e ao uso ético de dados sensíveis. Desai *et al.* (2024) reforçam que *datasets* pouco representativos e anotações sem fundamentação linguística podem reproduzir vieses sistêmicos. Assim, a qualidade de uma plataforma lexicográfica mediada por IA depende tanto da sofisticação técnica quanto da diversidade, documentação e legitimidade dos dados que a sustentam.

Um sétimo princípio é a governança ética dos dados. Registros em línguas de sinais envolvem corpos, rostos, expressões, identidades e trajetórias linguísticas. Diferentemente de dados textuais que podem ser anonimizados com maior facilidade, vídeos de sinalização carregam marcas visíveis dos sujeitos que os produzem. Isso exige consentimento informado, definição clara de finalidade, controle sobre usos futuros, políticas de armazenamento, critérios de acesso e mecanismos de revisão. De Meulder (2025) alerta para os riscos de erosão de direitos linguísticos quando tecnologias de IA passam a ser apresentadas como substitutas de mediações humanas, serviços e políticas de acessibilidade. No campo lexicográfico, esse alerta implica reconhecer que dados linguísticos de pessoas surdas não devem ser apropriados como matéria-prima neutra para desenvolvimento tecnológico.

Um oitavo princípio é a centralidade da comunidade surda. A participação surda não deve ocorrer apenas como etapa final de validação ou como fornecimento de dados para pesquisadores externos. Pessoas surdas precisam participar da

definição das perguntas, dos critérios de seleção, da organização das entradas, da validação dos sinais, da análise das variantes, da avaliação da plataforma e da governança dos dados. Kamikubo *et al.* (2025) mostram que a colaboração entre pesquisadores de aprendizado de máquina e comunidades surdas pode ser atravessada por assimetrias epistêmicas, comunicacionais e institucionais. Por isso, uma lexicografia participativa não pode reduzir a comunidade surda à condição de usuária final: ela deve reconhecê-la como produtora de conhecimento sobre a própria língua.

Esses princípios deslocam a pergunta central sobre IA e Libras. Em vez de perguntar apenas se a inteligência artificial é capaz de reconhecer, traduzir ou gerar sinais, é necessário perguntar que tipo de recurso lexicográfico se deseja construir, para quais usuários, com quais dados, sob quais critérios linguísticos e com que formas de participação comunitária. A tecnologia pode ampliar o alcance de dicionários e glossários, favorecer a interoperabilidade entre bases, facilitar a busca multimodal e apoiar a documentação de variantes. Entretanto, esses ganhos só serão relevantes se estiverem vinculados a uma concepção de lexicografia comprometida com a complexidade linguística da Libras e com os direitos da comunidade surda.

6 Considerações finais

Este artigo discutiu os desafios teórico-metodológicos do tratamento lexicográfico da Libras na era da inteligência artificial, partindo da compreensão de que a Libras é uma língua visual-espacial, variável, multimodal e socialmente situada. Ao longo da discussão, defendeu-se que dicionários, glossários, bancos de sinais e plataformas digitais não devem ser tratados apenas como instrumentos de consulta, mas como artefatos linguísticos, pedagógicos, políticos e tecnológicos que participam da documentação, circulação e legitimação do léxico da Libras.

A análise evidenciou que a lexicografia digital da Libras tem avançado, especialmente por meio da produção de dicionários on-line, glossários terminológicos

e sinais-termos em diferentes áreas do conhecimento. Esses recursos ampliam o acesso de pessoas surdas a conceitos acadêmicos, científicos e profissionais e subsidiam práticas de ensino, tradução e interpretação. No entanto, permanecem desafios importantes, como dispersão dos registros, heterogeneidade dos formatos, fragilidade de acesso, ausência de metadados, lacunas de validação e dependência do português como eixo predominante de busca.

No que se refere à inteligência artificial, argumentou-se que suas contribuições devem ser compreendidas de modo situado. Sistemas computacionais podem auxiliar na indexação de vídeos, na recuperação de sinais, na organização de bases multimodais, na identificação de similaridades formais e na comparação de variantes. Contudo, tais possibilidades não substituem a análise linguística, a curadoria humana e a validação comunitária. A IA pode apoiar o trabalho lexicográfico, mas não deve definir sozinha o que conta como sinal, variante, equivalência, erro ou forma legítima.

Um ponto central da discussão foi a crítica à redução da Libras à glosa ou à equivalência direta com o português. Embora a glosa possa funcionar como recurso auxiliar de anotação e busca, ela não representa integralmente a complexidade visual-espacial da língua sinalizada. Quando modelos computacionais ou plataformas digitais organizam seus dados prioritariamente por glosas ou palavras em português, corre-se o risco de reforçar uma concepção empobrecida da Libras, como se seus sinais fossem apenas correspondentes visuais de itens lexicais da língua majoritária.

Também se destacou que a variação linguística deve ser assumida como dimensão constitutiva da Libras, e não como obstáculo à sistematização lexicográfica. Registrar variantes, explicitar contextos de uso e documentar processos de validação são tarefas fundamentais para uma lexicografia linguisticamente responsável, especialmente no campo dos sinais-termos.

Conclui-se que o tratamento lexicográfico da Libras na era da inteligência artificial exige uma mudança de perspectiva. Não basta criar mais glossários digitais, ampliar bancos de sinais ou desenvolver sistemas automáticos de busca. É necessário

qualificar os modos de registrar, descrever, validar, integrar e disponibilizar os dados lexicais da Libras. Quando orientada por critérios linguísticos, participação surda, curadoria humana e governança ética, a IA pode contribuir para ampliar a circulação e a organização do léxico da Libras. Quando guiada apenas por métricas de eficiência, glosas simplificadas ou *datasets* pouco representativos, tende a reforçar apagamentos, padronizações indevidas e desigualdades epistêmicas.

Referências

- ALBRES, N. A. Produção de sinais-termos em Libras para conceitos acadêmicos: uma revisão sistemática e suas contribuições para surdos, professores e tradutores e intérpretes educacionais. **Revista GTLex**, Uberlândia, v. 9, e0917, 2024. DOI <https://doi.org/10.14393/Lex-v9a2023/24-17>
- BENEVIDES, M. P. *et al.* Tecnologia assistiva SignTalk: reconhecimento em tempo real do alfabeto datilológico de Libras por inteligência artificial. **Revista de Gestão Social e Ambiental**, Miami, v. 18, n. 12, p. 1-13, e010610, 2024. DOI <https://doi.org/10.24857/rgsa.v18n12-214>
- BRASIL. Decreto nº 5.626, de 22 de dezembro de 2005. Regulamenta a Lei nº 10.436, de 24 de abril de 2002, que dispõe sobre a Língua Brasileira de Sinais - Libras, e o art. 18 da Lei nº 10.098, de 19 de dezembro de 2000. **Diário Oficial da União**: seção 1, Brasília, DF, 23 dez. 2005. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/decreto/d5626.htm. Acesso em: 12 jun. 2026.
- BRASIL. Lei nº 10.436, de 24 de abril de 2002. Dispõe sobre a Língua Brasileira de Sinais - Libras e dá outras providências. **Diário Oficial da União**: seção 1, Brasília, DF, 25 abr. 2002. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/2002/110436.htm. Acesso em: 12 jun. 2026.
- BUGUEÑO MIRANDA, F. Balanço e perspectivas da lexicografia. **Cadernos de Tradução**, Florianópolis, v. 2, n. 32, p. 15-37, 2013. DOI <https://doi.org/10.5007/27761>
- CARVALHO, K. V. P.; KUMADA, K. M. O.; BENITEZ, P. Análise de glossários terminológicos em Libras documentados na literatura. **Revista Espaço**, Rio de Janeiro, edição especial, p. 79-93, 2025. DOI <https://doi.org/10.20395/revesp.2025.62.2.79-93>

CERNA, L. R. *et al.* A multimodal LIBRAS-UFOP Brazilian sign language dataset of minimal pairs using a Microsoft Kinect sensor. **Expert Systems with Applications**, v. 167, artigo 114179, 2021. DOI <https://doi.org/10.1016/j.eswa.2020.114179>

DE MEULDER, M. Deaf in AI: AI language technologies and the erosion of linguistic rights. **arXiv**, 2025. Preprint. Disponível em: <https://arxiv.org/abs/2505.02519>. Acesso em: 14 jun. 2026.

DESAI, A.; DE MEULDER, M.; HOCHGESANG, J. A.; KOCAB, A.; LU, A. X. Systemic biases in sign language AI research: a Deaf-led call to reevaluate research agendas. *In: WORKSHOP ON THE REPRESENTATION AND PROCESSING OF SIGN LANGUAGES: EVALUATION OF SIGN LANGUAGE RESOURCES*, 11., 2024, Torino. **Proceedings [...]**. Torino: ELRA; ICCL, 2024. p. 54-65. DOI <https://doi.org/10.63317/4qt6t64rxwd2>

GOETTERT, N.; BEVILACQUA, C. R. Dicionário Spread the Sign-Brasil: análise e proposta de melhoria. **Revista GTLex**, Uberlândia, v. 9, e0920, 2024. DOI <https://doi.org/10.14393/Lex-v9a2023/24-20>

GOETTERT, N.; BEVILACQUA, C. R. Lexicografia em Libras: análise de três dicionários disponíveis on-line. **ReVEL**, edição especial, v. 21, n. 20, 2023. Disponível em: <https://www.revel.inf.br/files/c35ec4679a6406d3ed515798559fdcb6.pdf>. Acesso em: 29 jun. 2026.

GOETTERT, N.; SILVEIRA, W. D.; BEVILACQUA, C. R. Estudos do léxico em Libras: pesquisas em lexicografia e terminologia. *In: GONTIJO, T. A. M. A.; CRUZ, C. R.; FLORES, V. M. (org.). A surdez e a Libras no cenário investigativo-científico: contribuições da Universidade Federal do Rio Grande do Sul e convidados*. Cuiabá: Guará Editora, 2025. p. 69-84.

HAUSMANN, F. J.; WIEGAND, H. E. Component parts and structures of general monolingual dictionaries: a survey. *In: HAUSMANN, F. J.; REICHMANN, O.; WIEGAND, H. E.; ZGUSTA, L. (ed.). Wörterbücher/Dictionnaires: an international encyclopedia of lexicography*. Berlin; New York: Walter de Gruyter, 1989. v. 1, p. 328-360. DOI <https://doi.org/10.1515/9783110095852.1.4.328>

KAMIKUBO, R. *et al.* Exploring collaboration to center the Deaf community in sign language AI. *In: INTERNATIONAL ACM SIGACCESS CONFERENCE ON COMPUTERS AND ACCESSIBILITY*, 27., 2025, Denver. **Proceedings [...]**. New York: ACM, 2025. p. 1-18. DOI <https://doi.org/10.1145/3663547.3746390>

SANTOS, I. B. *et al.* Língua Brasileira de Sinais (LIBRAS) e Inteligência Artificial (I.A): revisão da produção científica (2019–2025). **Revista DELOS**, Curitiba, v. 18, n. 74, p. 1-19, 2025. DOI <https://doi.org/10.55905/rdelosv18.n74-045>

TAN, S. *et al.* A review of deep learning-based approaches to sign language processing. **Advanced Robotics**, v. 38, n. 23, p. 1649-1667, 2024. DOI <https://doi.org/10.1080/01691864.2024.2442721>

ZAINELDIN, H. *et al.* Silent no more: a comprehensive review of artificial intelligence, deep learning, and machine learning in facilitating deaf and mute communication. **Artificial Intelligence Review**, v. 57, artigo 188, 2024. DOI <https://doi.org/10.1007/s10462-024-10816-0>