

Estado da publicação: O preprint não foi publicado em outro meio.

Plataforma da Diversidade Linguística Brasileira: repositório de dados linguísticos para uma inteligência artificial soberana

Marcia dos Santos Machado Vieira, Raquel Meister Ko Freitag, Juliana Bertucci Barbosa

<https://doi.org/10.1590/SciELOPreprints.18232>

Submetido em: 2026-09-30

Postado em: 2026-10-02 (versão 1)

(AAAA-MM-DD)

A moderação deste preprint recebeu o(s) endosso(s) de:

- Carolina Felicissimo (ORCID: <https://orcid.org/0000-0001-5065-2689>)

Plataforma da Diversidade Linguística Brasileira: repositório de dados linguísticos para uma inteligência artificial soberana

Brazilian Linguistic Diversity Platform: a repository of linguistic data for sovereign artificial intelligence

Marcia dos Santos Machado Vieira

Universidade Federal do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brasil. ORCID: <https://orcid.org/0000-0002-2320-5055>

Raquel Meister Ko Freitag

Universidade Federal de Sergipe, Aracaju, Sergipe, Brasil. ORCID: <https://orcid.org/0000-0002-4972-4320>

Juliana Bertucci Barbosa

Universidade Federal do Triângulo Mineiro, Uberaba, Minas Gerais, Brasil. ORCID: <https://orcid.org/0000-0002-1510-633X>

RESUMO

Objetivo: Apresentar a proposta da Plataforma da Diversidade Linguística Brasileira, iniciativa da Associação Brasileira de Linguística (ABRALIN) e do Grupo de Trabalho de Sociolinguística da Associação Nacional de Pós-Graduação e Pesquisa em Letras e Linguística (ANPOLL), e discutir sua contribuição para o desenvolvimento de modelos de linguagem em português previsto no Plano Brasileiro de Inteligência Artificial (PBIA) 2024-2028.

Método: Ensaio teórico-propositivo, fundamentado na análise documental do PBIA e em estudos sobre vieses linguísticos em modelos de linguagem e sobre a gestão de coleções de dados sociolinguísticos no Brasil.

Resultado: Os modelos de linguagem disponíveis são treinados com dados em que predomina o inglês e em que o português aparece sobretudo em registros escritos, formais e urbanos, o que repercute na reprodução de vieses, estereótipos e na invisibilização de variedades e falantes sub-representados; paralelamente, as coleções de dados linguísticos produzidas por especialistas permanecem fragmentadas e concentradas regional e institucionalmente. A Plataforma é proposta como repositório nacional de coleções de dados linguísticos, orientado pelos princípios FAIR e sustentado por uma rede colaborativa de laboratórios.

Conclusões: A Plataforma constitui ponte estratégica entre a ciência linguística, a ciência de dados, a ciência da informação e o desenvolvimento tecnológico, com impactos para o ensino, para as políticas linguísticas e para a soberania digital brasileira.

Palavras-chave: Repositórios digitais, Curadoria de dados, Diversidade linguística, Ciência aberta, Inteligência artificial.

ABSTRACT

Objective: To present the proposal of the Brazilian Linguistic Diversity Platform, an initiative of the Brazilian Association of Linguistics (ABRALIN) and the Sociolinguistics Working Group of the National Association of Postgraduate Studies and Research in Letters and Linguistics (ANPOLL), and to discuss its contribution to the development of Portuguese language models foreseen in the Brazilian Artificial Intelligence Plan (PBIA) 2024-2028.

Methods: Theoretical and propositional essay, based on documentary analysis of the PBIA and on studies of linguistic bias in language models and of the management of sociolinguistic data collections in Brazil.

Results: The language models currently available are trained on data dominated by English, in which Portuguese appears mostly in written, formal, and urban registers, which results in the reproduction of biases, stereotypes, and the invisibility of underrepresented varieties and speakers; meanwhile, linguistic data collections produced by specialists remain fragmented and regionally and institutionally concentrated. The Platform is proposed as a national repository of linguistic data collections, guided by the FAIR principles and supported by a collaborative network of laboratories.

Conclusions: The Platform represents a strategic bridge between linguistic science, data science, information science, and technological development, with impacts on education, language policies, and Brazilian digital sovereignty.

Keywords: Digital repositories, Data curation, Linguistic diversity, Open science, Artificial intelligence.

INTRODUÇÃO

O Plano Brasileiro de Inteligência Artificial (PBIA) 2024-2028, lançado pelo Ministério da Ciência, Tecnologia e Inovação, tem entre seus objetivos centrais o desenvolvimento de modelos avançados de linguagem em português, fundamentados em dados nacionais que contemplem nossa diversidade cultural, social e linguística, com vistas a fortalecer a soberania em IA (Brasil, 2025). Essa formulação política é uma janela de oportunidade para articular a pauta tecnológica à realidade linguística de um país com mais de duzentos milhões de falantes do português e mais de 250 outras línguas – indígenas, de imigração, de sinais, crioulas e afro-brasileiras –, segundo o Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN, 2010), cada uma com suas variedades e transformações sociais em curso.

O português do Brasil é, do ponto de vista quantitativo, uma das grandes línguas da internet e do mundo digital. Contudo, do ponto de vista qualitativo, os modelos de linguagem hoje disponíveis em português são desenvolvidos com conjuntos de dados em que predomina o inglês e em que o português aparece sobretudo em registros escritos, formais e urbanos, próximos das variedades de maior prestígio, em geral associadas ao eixo Rio de Janeiro – São Paulo. Parte do conteúdo multilíngue da web, sobretudo em línguas com menos recursos, resulta ainda de tradução automática (Thompson *et al.*, 2024). Essa configuração implica que sistemas de IA reproduzam vieses, estereótipos e invisibilizações que já estão presentes nas concepções gramaticais e didático-pedagógicas tradicionais, agora potencializados pela escala dos modelos computacionais.

Um exemplo desse fenômeno é o estudo de Kerche, Zook e Graham (2026), que demonstra a tendência de LLMs a associar falantes nordestinos a perfis de menor escolaridade ou menor capacidade cognitiva, replicando e amplificando preconceitos sociais já documentados pelos estudos sociolinguísticos brasileiros (Freitag; Gois, 2024). Estados mais ricos e visíveis na mídia, como São Paulo, Distrito Federal e Minas Gerais, obtêm pontuações mais altas, enquanto estados do norte, como Amazonas e Maranhão, são classificados muito abaixo. O modelo amplifica hierarquias socioeconômicas existentes porque termos associados a “inteligente” (como “melhor escolaridade”, “descoberta inovadora” e “aluno genial”) aparecem com mais frequência ao lado de locais de elite do Sudeste e do Sul, em vez de medir inteligência real. Essas regiões são excluídas das associações semânticas, e preconceitos de padrão reforçam a dissociação semântica em classificações expostas. O mapa é moldado mais pela forma como o conteúdo digital discute diferentes regiões do que pela realidade (Kerche; Zook; Graham, 2026).

O que antes era um estereótipo local, reproduzido em livros didáticos ou no senso comum, adquire, nos sistemas de IA, uma projeção global e uma aparência de objetividade algorítmica, tornando-se ainda mais difícil de contestar. A sub-representação de variedades linguísticas nos dados de treinamento de LLMs intensifica assimetrias históricas, pois o que não é ouvido não é lembrado – ou tem até mesmo sua existência questionada.

Paralelamente, a comunidade de pesquisa sociolinguística brasileira acumulou, ao longo de décadas, acervos de dados de fala, escrita e línguas sinalizadas, constituídos com rigor metodológico e representatividade regional. No entanto, esses acervos permanecem fragmentados, armazenados em sites de pesquisadores individuais, em servidores institucionais heterogêneos ou em formatos inacessíveis para leitura automatizada. Seu potencial para o treinamento de modelos de IA é imenso, mas permanece amplamente inexplorado pela ausência de uma infraestrutura nacional de armazenamento, curadoria e interoperabilidade.

É nesse cenário que se insere a proposta da Plataforma da Diversidade Linguística Brasileira, iniciativa da Associação Brasileira de Linguística (ABRALIN) e do Grupo de Trabalho (GT) de Sociolinguística da Associação Nacional de Pós-Graduação e Pesquisa em Letras e Linguística (ANPOLL). A Plataforma da Diversidade Linguística Brasileira é uma proposta de repositório nacional de coleções de metadados e objetos digitais correlacionados de pesquisas de línguas e variedades brasileiras, respondendo simultaneamente às demandas da ciência aberta, da educação linguística, da política e da soberania tecnológica.¹

Este artigo apresenta os fundamentos, a proposta e as implicações do projeto da Plataforma da Diversidade Linguística Brasileira, argumentando que ela representa uma contribuição estratégica na articulação entre ciência linguística, ciência de dados, ciência da informação e

¹ A proposta da Plataforma vem sendo discutida em diferentes momentos. Foi apresentada, desde 2021, em eventos científicos e de divulgação, como o Festival do Conhecimento da Universidade Federal do Rio de Janeiro, o Congresso Internacional da Abralín e o Seminário Internacional Viagens da Língua, do Museu da Língua Portuguesa (Barbosa; Machado Vieira, 2022); teve seus fundamentos expostos em Machado Vieira *et al.* (2021) e sua relação com o ensino de português discutida em Machado Vieira e Barbosa (2022); foi registrada como projeto voltado a dados linguísticos para uma IA brasileira em Freitag *et al.* (2025); teve os bancos de dados (socio)linguísticos brasileiros mapeados em Machado Vieira *et al.* (2026b); e teve os desafios para sua implementação como repositório científico temático discutidos em Machado Vieira *et al.* (2026a). O presente artigo retoma essa discussão com outro recorte, centrado na relação entre dados linguísticos e o desenvolvimento de modelos de linguagem.

tecnologia. A infraestrutura dessa plataforma, ainda a ser construída com base numa rede transdisciplinar de (inter)ações, visa a acolher coleções de dados representativas de variedades e variações das línguas em uso no Brasil, além das do idioma português. O objeto focalizado neste artigo é a relação entre elementos prospectados para o desenvolvimento do projeto da plataforma e ganhos na área de processamento de linguagem via soluções de inteligência artificial que possam operar com diversas coleções de dados representativas do português das diferentes regiões brasileiras.

O ESTADO DA ARTE: LLMs EM PORTUGUÊS E A REPRESENTAÇÃO DA DIVERSIDADE LINGUÍSTICA

Os modelos de língua em larga escala (*Large Language Models* – LLMs) são sistemas de inteligência artificial treinados com volumes massivos de texto, geralmente coletados da internet sem critérios rigorosos quanto à representatividade linguística, à qualidade dos dados ou ao consentimento dos falantes. Modelos como GPT, Gemini e LLaMA declaram capacidades multilíngues, mas a presença de textos em inglês e de contextos culturais ocidentais nos conjuntos de treinamento distorce seus resultados em outras línguas (Freitag, 2026a).

No caso do português, essa distorção é especialmente relevante. Embora o Brasil seja um dos maiores mercados digitais do mundo, os dados em português disponíveis para treinamento de LLMs são, em grande parte, produções formais escritas, concentradas nos registros de maior prestígio e de circulação urbana, e incluem também conteúdo traduzido, inclusive automaticamente (Thompson *et al.*, 2024). Os modelos aprendem, assim, o português a partir de um filtro cultural e varietal estreito, que não representa a diversidade dos mais de duzentos milhões de falantes brasileiros.

Estudos recentes demonstram que LLMs reproduzem e amplificam preconceitos contra variedades linguísticas minoritárias. Fleisig *et al.* (2024) demonstraram que GPT-3.5 Turbo e GPT-4 reforçam discriminação linguística contra variedades não padrão do inglês. Mengesha *et al.* (2021) documentaram confusões sistemáticas de modelos em relação ao inglês afro-americano vernacular. Shrawgi *et al.* (2024) identificaram viés de gênero, nacionalidade, raça e religião em múltiplos LLMs. No Brasil, padrões análogos se manifestam na sub-representação dos falares regionais, especialmente das variedades nordestinas e amazônicas, historicamente associadas a perfis sociais estigmatizados (Freitag; Gois, 2024). Na apresentação do projeto da Plataforma, esse diagnóstico foi formulado nos seguintes termos:

Uma IA eticamente sensível para a soberania nacional requer que a diversidade linguística seja considerada de maneira plena, com amostras linguísticas diversificadas para o treino de LLMs. Sem isso, a reprodução de uma IA que considera apenas o português e uma de suas variedades tem efeito na conformação de padrões linguísticos hegemônicos, invisibilizando e marginalizando ainda mais as variedades linguísticas subrepresentadas (Freitag *et al.*, 2025, p. 21).

Essa marginalização alcança, conseqüentemente, os falantes dessas variedades. Para compreender onde dados diversificados podem intervir, é preciso distinguir as etapas de desenvolvimento de um modelo de linguagem, porque os dados linguísticos cumprem funções diferentes em cada uma delas. Na primeira, o pré-treinamento, o modelo aprende a

prever a continuação de sequências de texto a partir de volumes massivos de dados não anotados; trata-se de um aprendizado autossupervisionado, em que o próprio texto fornece o sinal de treino, e é nessa etapa que se forma a maior parte do conhecimento linguístico e factual do modelo. Na segunda, o ajuste fino supervisionado, o modelo pré-treinado é exposto a exemplos anotados ou estruturados, como pares de instrução e resposta, textos classificados ou transcrições alinhadas ao áudio, para especializar seu comportamento em determinadas tarefas, domínios ou formas de interação. A terceira etapa, a avaliação, mede o desempenho do modelo em conjuntos de teste construídos para esse fim.

Duas evidências ajudam a dimensionar o papel de cada etapa. De um lado, o prosseguimento do pré-treinamento de modelos abertos com textos em português, com uma pequena fração do orçamento computacional original, produziu modelos com desempenho superior ao de modelos centrados no inglês e de modelos multilíngues em tarefas em português (Pires *et al.*, 2023). De outro, experimentos de ajuste fino mostram que um número reduzido de exemplos cuidadosamente selecionados basta para orientar o comportamento de um modelo, mas indicam que quase todo o seu conhecimento é adquirido no pré-treinamento (Zhou *et al.*, 2023). A consequência para a diversidade linguística é direta: dados de qualidade, ainda que pouco volumosos, são eficazes para ajustar o comportamento de um modelo, mas não bastam para suprir, na etapa de ajuste, a ausência de variedades e registros no pré-treinamento.

É nesse quadro que se situa a contribuição das coleções sociolinguísticas. Seu volume, medido em horas de gravação ou em milhões de palavras, é modesto diante dos corpora de pré-treinamento, mas elas reúnem o que esses corpora raramente oferecem: amostras de fala espontânea estratificadas por região, idade, escolaridade e outras variáveis sociais, com metadados sobre os falantes e as condições de coleta, transcritas e anotadas segundo protocolos explícitos. Esses atributos as tornam adequadas a pelo menos três usos: o ajuste fino de modelos, inclusive de reconhecimento automático de fala, para variedades pouco representadas; a construção de conjuntos de avaliação estratificados, capazes de revelar diferenças de desempenho entre variedades que uma medida agregada ocultaria, como já se fez com o perfilamento dialetal de variedades do português brasileiro por LLMs (Freitag; Gois, 2024); e, a longo prazo, a composição de corpora de pré-treinamento mais representativos da diversidade do português brasileiro. O campo da pesquisa sociolinguística acumula décadas de tradição na produção de dados com esse rigor, um ativo estratégico ainda pouco aproveitado pelo desenvolvimento tecnológico.

COLEÇÕES DE DADOS LINGÜÍSTICOS NO BRASIL: FRAGMENTAÇÃO E INVISIBILIDADE

A pesquisa sociolinguística no Brasil tem longa e produtiva tradição. Desde o Projeto Norma Urbana Culta (NURC), definido em 1968, passando pelos bancos de dados que emergem nos anos 1980 e 1990, como o Programa de Estudos sobre o Uso da Língua (PEUL), originado no Projeto Censo da Variação Linguística do Estado do Rio de Janeiro (1982), e o projeto Variação Linguística Urbana da Região Sul do Brasil (VARSUL), até o Atlas Linguístico do Brasil (ALiB), iniciado em 1996, pesquisadores de todo o país têm construído bancos de dados de fala, escrita e línguas sinalizadas, com representatividade regional e social cuidadosamente planejada (Freitag, 2021, 2026b). A Plataforma da Diversidade Linguística Brasileira é uma continuidade e um aprimoramento desses esforços.

Contudo, um mapeamento realizado pelo GT de Sociolinguística da ANPOLL revelou que a maioria dessas coleções permanece fragmentada e assistemática do ponto de vista de sua gestão e compartilhamento. Machado Vieira *et al.* (2021) identificaram que grande parte dos acervos está vinculada a programas de pós-graduação e concentrada na faixa litorânea do Brasil. O levantamento prosseguiu: registrava ao menos 28 bancos em março de 2022 (Barbosa; Machado Vieira, 2022) e, em sua atualização com respostas recebidas até fevereiro de 2024, reúne 30 bancos sob responsabilidade de 17 instituições (Machado Vieira *et al.*, 2026b, p. 2-3). A distribuição confirma a concentração: 20 bancos têm sede no Sudeste, 7 no Nordeste e 3 no Sul, nenhum no Norte ou no Centro-Oeste, e uma única instituição, a Universidade Federal do Rio de Janeiro, responde por 12 deles (Machado Vieira *et al.*, 2026b, p. 6-7). Predominam os registros de fala: 17 bancos são exclusivamente de fala, 6 combinam fala e escrita, 6 são apenas de escrita e 1 é multimodal (Machado Vieira *et al.*, 2026b, p. 5). Os autores advertem que o mapa representa a sede das instituições, e não necessariamente os locais de coleta, e que as lacunas podem refletir também o alcance da divulgação do levantamento (Machado Vieira *et al.*, 2026b). Os dados, quando disponíveis, são acessíveis em sites individuais de pesquisadores, em formatos não padronizados e sem metadados que permitam buscas automatizadas. Sousa e Freitag (2024) observam que a sistematização de corpora sociolinguísticos tem se dado no nível individual, com dados armazenados e compartilhados em sites dos próprios pesquisadores, e que essa situação se repete no Brasil, onde ainda falta padronização tanto na coleta quanto na disponibilização dos dados, uma situação que compromete a cumulatividade dos dados e a possibilidade de generalizações em larga escala.

Essa situação tem ao menos duas consequências graves. A primeira é a invisibilidade científica: o trabalho de coleta, transcrição, anotação e curadoria de dados linguísticos não é adequadamente valorizado como produção técnica, não conta com sistemas de citação específicos e não encontra espaço nos repositórios institucionais existentes. Na Ciência da Informação, esse problema é tratado no âmbito da curadoria digital, entendida como o conjunto de atividades de gestão, preservação e agregação de valor aos dados ao longo de seu ciclo de vida, capaz de reposicionar os dados como parte do próprio ciclo da comunicação científica (Sales; Sayão, 2012). No centro dessa infraestrutura estão os repositórios digitais de dados de pesquisa, sistemas voltados à seleção, descrição, arquivamento, acesso e compartilhamento de dados, que deixam de ser subprodutos da pesquisa para se tornarem recursos informacionais em si (Sayão; Sales, 2016). O Deposita, repositório do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), por exemplo, foi concebido para o compartilhamento de produção científica convencional (teses, dissertações, artigos), e não oferece a infraestrutura adequada para a tipologia de objetos digitais que caracterizam as coleções de dados linguísticos: arquivos de áudio e vídeo, transcrições alinhadas, anotações multimodais, ontologias linguísticas. Repositórios abrangentes também não favorecem práticas de análise linguística que implicam constituição de amostras de usos a partir de diversos extratos de coleções de dados depositadas. Assim, um repositório focado na temática linguística favorece tanto indexação personalizada para as necessidades de descrição em metadados quanto recuperação de informação e objetos digitais de modo preciso, sem ruído de extração advinda de documentos não relacionados.

A segunda consequência é a vulnerabilidade patrimonial. A precariedade com que são armazenados os produtos de documentações linguísticas coloca em risco a salvaguarda da diversidade linguística brasileira. O incêndio do Museu Nacional em 2018, que destruiu um

acervo inestimável de documentação de línguas indígenas brasileiras, é o exemplo mais trágico dessa fragilidade. Acervos em fitas magnéticas, em servidores sem backup, em computadores pessoais de pesquisadores aposentados: esse é o cenário atual de parte dos dados sociolinguísticos brasileiros (Freitag, 2021, 2022b).

Do lado oposto, o desenvolvimento acelerado de tecnologias de linguagem cria uma demanda crescente por dados de qualidade. Empresas de tecnologia, laboratórios de pesquisa e órgãos governamentais buscam ativamente coleções de dados linguísticos estruturados para o treinamento, ajuste fino e avaliação de modelos. A ausência de uma infraestrutura nacional articulada impede que esse encontro aconteça de forma sistemática e eticamente orientada.

A Plataforma da Diversidade Linguística Brasileira posiciona-se exatamente nesse espaço: como a ponte entre a comunidade científica produtora de dados e os agentes do desenvolvimento tecnológico que demandam esses dados. Mais do que um repositório passivo, a Plataforma propõe uma mudança de cultura na forma como a comunidade sociolinguística compreende, gerencia e disponibiliza seus dados.

O diagnóstico de fragmentação não é exclusivo do Brasil. Na Europa, a dispersão de recursos linguísticos por instituições, projetos e formatos incompatíveis motivou a criação do CLARIN (*Common Language Resources and Technology Infrastructure*), infraestrutura de pesquisa distribuída concebida para superar essa fragmentação e tornar recursos e tecnologias da linguagem visíveis e acessíveis de forma persistente (Wittenburg, 2010; Krauwer; Maegaard, 2022). Constituído em 2012 como consórcio europeu de infraestrutura de pesquisa (ERIC), o CLARIN consolidou-se como uma das principais infraestruturas para as humanidades (Fišer; Witt, 2022) e serve de inspiração para a Plataforma da Diversidade Linguística Brasileira, como se discute na próxima seção.

A PLATAFORMA DA DIVERSIDADE LINGUÍSTICA BRASILEIRA

Proposta e objetivos

A Plataforma da Diversidade Linguística Brasileira é definida como um repositório de dados de e para dados/objetos digitais de pesquisas de línguas e variedades, para ciência aberta, com computação de alta performance – de armazenamento, processamento e acessibilidade escaláveis –, numa infraestrutura interconectada nacional e internacionalmente, visível como referência para depósito, partilha e reutilização. Desde sua formulação inicial, o projeto foi descrito nos seguintes termos:

Em linhas gerais, o projeto Plataforma da Diversidade Linguística Brasileira consiste na organização de um repositório digital de amostras de fala e escrita, bem como de amostras de língua(s) sinalizada(s), presentes no território brasileiro. Essa plataforma deve ter uma arquitetura tecnológica e arquivística potente e plenamente capaz não só de catalogar e abrigar coleções de dados e metadados sociolinguísticos e histórico-culturais, mas também de potencializar o acesso a informações diversas por diferentes públicos e incentivar múltiplos sujeitos a cooperarem e interoperarem em prol da contínua atualização, reformulação, ampliação, curadoria e revitalização desse repositório (Machado Vieira; Barbosa, 2022, p. 338).

A proposta orienta-se pelos princípios FAIR (*Findable, Accessible, Interoperable, Reusable*) de gestão de dados científicos, que norteiam a ciência aberta internacionalmente e que, no

Brasil, vêm sendo difundidos e aplicados no âmbito da iniciativa GO FAIR Brasil (Sales *et al.*, 2021). Na Ciência da Informação, esses princípios têm sido articulados aos estudos de encontrabilidade, que complementam a orientação dos princípios FAIR para os sistemas com a atenção ao sujeito que busca e reutiliza os dados (Campos; Dias; Sousa, 2023). A Plataforma articula-se à rede GO FAIR Brasil e à Rede de Humanidades, envolvendo um consórcio inclusivo, colaborativo e aberto a adesões, autogerido, liderado pela comunidade, e implicando diferentes domínios de saber, disciplinas, indivíduos, organizações, instituições, comunidades e regiões do país.

A experiência do CLARIN oferece referências concretas para a arquitetura da Plataforma. Três de seus componentes interessam diretamente ao caso brasileiro: uma rede de centros interconectados, com financiamento de longo prazo e organizados em base nacional; um serviço de identificadores persistentes que garante a estabilidade das referências aos recursos e a seus fragmentos; e um domínio comum de metadados que torna visível, em um catálogo unificado, o que está depositado em diferentes instituições (Wittenburg, 2010). Para a descrição, o CLARIN adotou um modelo de metadados baseado em componentes, a *Component Metadata Infrastructure* (CMDI), em que cada comunidade pode construir perfis adequados a seus tipos de recurso, desde que reutilize elementos de registros de conceitos compartilhados (Windhouwer; Goosen, 2022). Essa solução é especialmente pertinente para coleções sociolinguísticas, que precisam descrever perfis de falantes, condições de coleta e camadas de anotação sem abrir mão da interoperabilidade. Para a língua portuguesa, o modelo já tem uma realização: o PORTULAN CLARIN, nó português do CLARIN ERIC, reúne um repositório para depósito e arquivamento de longo prazo de recursos e metadados e um ambiente em que ferramentas de tecnologia da linguagem são disponibilizadas aos usuários (Branco *et al.*, 2020). A Plataforma pode, assim, nascer interoperável com uma infraestrutura já existente para o português europeu, ao mesmo tempo em que responde a uma lacuna específica: a documentação da diversidade do português brasileiro e das demais línguas faladas no país.

O processo de alimentação da Plataforma é cíclico e contínuo, uma vez que as línguas estão sempre em mudança, premissa central da Sociolinguística. Isso implica que a gestão da Plataforma deve prever mecanismos de atualização permanente, com protocolos de coleta periódica de novas amostras e de reprocessamento de acervos existentes, assegurando que os dados disponibilizados reflitam a dinâmica real da diversidade linguística brasileira em cada momento histórico.

Impactos educacionais, políticos e tecnológicos

Do ponto de vista educacional, a Plataforma tem potencial transformador para a educação linguística no Brasil. A Base Nacional Comum Curricular (BNCC) reconhece o direito de aprendizagem relacionado à diversidade linguística brasileira (BRASIL, 2018), mas os materiais didáticos disponíveis continuam a reproduzir uma visão normativa da língua portuguesa centrada no eixo Rio de Janeiro – São Paulo, perpetuando preconceitos e estigmas contra variedades regionais. O acesso a amostras autênticas de fala de todas as regiões do país permitirá que professores da Educação Básica produzam atividades didáticas que mobilizem seus alunos em favor da valorização das riquezas culturais e linguísticas de sua região.

O impacto sobre o preconceito linguístico é uma das dimensões mais relevantes da proposta. Pesquisas sociolinguísticas documentam há décadas que variedades sub-representadas sofrem discriminação linguística, que, ao ser incorporada pelos sistemas de IA, ganha escala global e aparência de neutralidade algorítmica. A Plataforma pode reverter esse ciclo ao ampliar sistematicamente a documentação de variedades de todas as regiões brasileiras, fornecendo dados para que os modelos de IA sejam ajustados de forma a reduzir estereótipos e discriminações. Materiais didáticos que hoje perpetuam o preconceito linguístico contra falantes nordestinos, amazônicos e de periferias urbanas poderiam ser revistos à luz de evidências empíricas robustas, provenientes de dados coletados em campo por especialistas (Machado Vieira; Barbosa, 2022; Freitag, 2022a). Tais materiais têm o potencial de reverberar em olhares e práticas sobre o uso da língua portuguesa além do Brasil, haja vista o interesse na língua em diferentes instituições de promoção do idioma no mundo.

Do ponto de vista da política linguística, a Plataforma da Diversidade Linguística Brasileira responde ao desafio de assegurar, de forma sistemática, a preservação e o acesso ao patrimônio linguístico brasileiro. O Brasil conta com mais de 250 línguas, segundo o Inventário Nacional da Diversidade Linguística (INDL), mantido pelo IPHAN (IPHAN, 2010), mas grande parte delas carece de documentação digital estruturada e acessível. Uma infraestrutura nacional que valorize esse patrimônio tem implicações para políticas de revitalização de línguas ameaçadas, para a diplomacia cultural e para a promoção da diversidade brasileira no cenário internacional. Tem poder de agregar iniciativas estanques de documentação (como a anunciada pelo Museu da Língua Portuguesa em 2025 quanto a línguas e culturas indígenas) (Museu da Língua Portuguesa, 2025) e, então, promover sensibilização na sociedade para documentação e comunicação no país também centradas no conhecimento de línguas além do idioma português. No caso das línguas indígenas, a abertura dos dados precisa ser compatibilizada com os princípios CARE (benefício coletivo, autoridade para controlar, responsabilidade e ética), que complementam os princípios FAIR ao considerar o tratamento ético dos dados de povos indígenas e seu direito de decidir sobre o acesso e o uso desses dados (Torino; Monteiro; Vidotti, 2023). Essa exigência se torna mais aguda no contexto da inteligência artificial generativa, que se alimenta de grandes volumes de dados frequentemente coletados sem consentimento ou reconhecimento das fontes, e que coloca em questão a soberania informacional desses povos (Carrasco *et al.*, 2025). Uma plataforma voltada ao treinamento de modelos de linguagem deve, portanto, prever camadas de governança que permitam às comunidades estabelecer condições de acesso e reúso de seus dados.

No campo tecnológico, o conjunto de dados a ser disponibilizado pela Plataforma tem potencial de subsidiar o treinamento e o ajuste fino de LLMs em português, respondendo à Ação 9 do PBIA: “IA baseada em dados nacionais (LLM em português), com o fomento à curadoria de conjuntos de dados nacionais e apoio ao desenvolvimento de modelos fundacionais, em particular modelos de linguagem de grande escala (LLMs), especializados em português” (Brasil, 2025, p. 70). Além disso, dados linguísticos estruturados impactam diretamente a capacidade de sistemas, sites e motores de busca de processar, identificar, recuperar e traduzir informações em português de forma qualificada – incluindo a tradução do português para outras línguas e o reconhecimento automático de variedades regionais do português brasileiro –, ampliando a inclusão digital de falantes de variedades não padrão e fortalecendo a soberania linguística nacional.

Implicações organizacionais e rede colaborativa

A construção da Plataforma da Diversidade Linguística Brasileira implica uma transformação profunda na cultura de trabalho da comunidade sociolinguística brasileira. Historicamente, a pesquisa linguística opera em grupos isolados, com lógicas de trabalho individual, com protocolos de coleta de dados heterogêneos e sem incentivos para o compartilhamento aberto de dados. A Plataforma da Diversidade Linguística Brasileira demanda o movimento oposto: a constituição de uma rede de laboratórios que adotem protocolos unificados de coleta, transcrição, anotação e disponibilização de dados, operando como um sistema de agentes colaborativos com metas coletivamente gerenciadas. Esse movimento tem precedente: a construção do CLARIN exigiu, além da infraestrutura técnica, a harmonização de licenças, direitos e questões éticas, a formação de pesquisadores para criar recursos compatíveis e a definição de um modelo organizacional e de financiamento para uma federação persistente de centros (Wittenburg, 2010). O PORTULAN CLARIN, por sua vez, mostra que esse modelo pode ser implementado por uma comunidade de língua portuguesa, reunindo centros de pesquisa em linguística e em computação (Branco *et al.*, 2020).

As frentes de trabalho previstas abrangem: (i) mapeamento de repositórios e coleções de dados existentes em todo o Brasil e centralização de informações; (ii) definição de princípios, políticas, diretrizes e padrões de armazenamento e compartilhamento; (iii) desenvolvimento da plataforma digital e sua arquitetura sociotécnica; (iv) definição de condições de leitura por humanos e máquinas e interoperabilidade; e (v) curadoria técnico-científica, disseminação e gerenciamento contínuo do repositório. Todo esse percurso exige a colaboração articulada entre ciência linguística, ciência de dados, ciência da informação e tecnologia de infraestrutura, em uma agenda interdisciplinar de fato.

A comunidade sociolinguística, ao engajar-se nesse projeto, passa a mapear sistematicamente todo o território brasileiro – incluindo as regiões Norte e Centro-Oeste, sem nenhum banco cadastrado no mapeamento mais recente (Machado Vieira *et al.*, 2026b), e o interior do Nordeste, historicamente sub-representados nas pesquisas. Isso implica ampliar parcerias com instituições de ensino superior e pesquisa de todas as regiões, capacitar pesquisadores em protocolos de coleta e curadoria de dados, e construir uma infraestrutura colaborativa que distribua tanto o trabalho quanto os benefícios da produção científica. A articulação com o IBICT, com a Rede Nacional de Ensino e Pesquisa (RNP) e com outras instituições de ciência da informação e tecnologia é, portanto, indispensável para que a Plataforma alcance a escala e a sustentabilidade necessárias.

CONSIDERAÇÕES FINAIS

A língua portuguesa é, do ponto de vista da inteligência artificial, uma urgência estratégica para o Brasil. Com mais de duzentos milhões de falantes, o português figura entre as grandes línguas do mundo digital em volume de dados, mas permanece sub-representado em diversidade e variedade: as variedades regionais e sociais, os registros orais e informais e os falantes de grupos historicamente estigmatizados têm presença reduzida nos dados que alimentam os modelos de língua mais avançados. Essa sub-representação tem efeitos diretos sobre quem usa esses sistemas no Brasil e reforça a dependência do país em relação a dados e modelos produzidos fora dele.

A Plataforma da Diversidade Linguística Brasileira representa uma resposta coletiva e articulada a esse desafio. Ao agregar coleções de dados linguísticos coletados por especialistas em todo o Brasil, ao estabelecer padrões de metadados e interoperabilidade, ao construir uma infraestrutura de armazenamento segura e acessível, e ao promover uma cultura de ciência aberta e trabalho colaborativo, a Plataforma posiciona a comunidade sociolinguística brasileira como protagonista na construção de uma IA soberana e eticamente sensível.

A comunidade sociolinguística precisa definir estratégias e transformações para a defesa de um repositório nacional de dados linguísticos, articulando ciência linguística, ciência de dados, ciência da informação e tecnologia. A tipologia de objetos digitais a ser considerada – arquivos de áudio, vídeo, transcrições, anotações multimodais – implica caminhos tecnológicos específicos e parcerias com instituições como o IBICT e a Rede Nacional de Ensino e Pesquisa (RNP). O caminho é a construção de um sistema agregador de infraestruturas que direcione para uma plataforma nacional, interoperável com sistemas internacionais de documentação linguística, como o CLARIN e seu nó de língua portuguesa, o PORTULAN CLARIN.

O momento é oportuno: o Plano Brasileiro de Inteligência Artificial prevê explicitamente o desenvolvimento de modelos de linguagem com dados nacionais representativos da diversidade cultural, social e linguística brasileira. A construção de um modelo de linguagem de grande escala nacional é, contudo, um desafio de dimensão e duração muito superiores ao horizonte de 2024 a 2028 estabelecido no plano, e o ajuste fino de modelos de pesos abertos tem sido, neste momento, o caminho mais viável. Essa via não assegura autonomia: a abertura de um modelo é uma decisão de quem o desenvolve e pode ser revista. A Meta, cuja família Llama foi decisiva para a difusão de modelos de pesos abertos, lançou em abril de 2026 o Muse Spark, primeiro modelo de sua nova divisão de IA, como sistema fechado, sem disponibilizar seus pesos. Um acervo de dados linguísticos documentado, preservado e governado, por sua vez, serve ao ajuste dos modelos disponíveis hoje e continuará servindo aos que vierem, sejam eles abertos, fechados ou desenvolvidos no país. A construção da Plataforma da Diversidade Linguística Brasileira é um projeto coletivo, de longo prazo e estratégico para o presente e o futuro do Brasil e da língua portuguesa, e, por isso mesmo, precisa ser iniciada agora.

REFERÊNCIAS

BARBOSA, Juliana Bertucci; MACHADO VIEIRA, Marcia dos Santos. **Projeto Plataforma da Diversidade Linguística Brasileira**. São Paulo: Museu da Língua Portuguesa, 29 abr. 2022. Disponível em: <https://www.museudalinguaportuguesa.org.br/projeto-plataforma-diversidade-linguistica-brasileira/>. Acesso em: 30 set. 2026.

BRANCO, António; MENDES, Amália; QUARESMA, Paulo; GOMES, Luís; SILVA, João; TEIXEIRA, Andrea. Infrastructure for the science and technology of language PORTULAN CLARIN. In: INTERNATIONAL WORKSHOP ON LANGUAGE TECHNOLOGY PLATFORMS, 1., 2020, Marseille. **Proceedings** [...]. Marseille: European Language Resources Association, 2020. p. 1-7. Disponível em: <https://aclanthology.org/2020.iwlt-1.1/>. Acesso em: 30 set. 2026.

BRASIL. Ministério da Ciência, Tecnologia e Inovação. **IA para o bem de todos: Plano Brasileiro de Inteligência Artificial 2024-2028**. Brasília, DF: MCTI; CGEE, 2025. Disponível em: https://www.gov.br/mcti/pt-br/centrais-de-conteudo/publicacoes-mcti/plano-brasileiro-de-inteligencia-artificial/pbia_mcti_2025.pdf. Acesso em: 30 set. 2026.

BRASIL. Ministério da Educação. Base Nacional Comum Curricular. Brasília, DF: MEC, 2018.

CAMPOS, Arthur Ferreira; DIAS, Guilherme Ataíde; SOUSA, Marckson Roberto Ferreira de. Princípios FAIR, encontrabilidade e dados: por um diálogo teórico e uma aplicação prática. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 28, p. 1-21, 2023. DOI: <https://doi.org/10.5007/1518-2924.2023.e88060>.

CARRASCO, Lais Barbudo; TORINO, Emanuelle; CONEGLIAN, Caio Saraiva; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Soberania e governança de dados de povos indígenas em tempos de inteligência artificial generativa. **Revista Ciência Aberta Lusófona**, v. 1, 2025. DOI: <https://doi.org/10.82524/recal.2025.65>.

FIŠER, Darja; WITT, Andreas (ed.). **CLARIN**: the infrastructure for language resources. Berlin; Boston: De Gruyter, 2022. DOI: <https://doi.org/10.1515/9783110767377>.

FLEISIG, Eve; SMITH, Genevieve; BOSSI, Madeline; RUSTAGI, Ishita; YIN, Xavier; KLEIN, Dan. Linguistic bias in ChatGPT: language models reinforce dialect discrimination. *In*: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2024, Miami. **Proceedings** [...]. Miami: Association for Computational Linguistics, 2024. p. 13541-13564. DOI: <https://doi.org/10.18653/v1/2024.emnlp-main.750>.

FREITAG, Raquel Meister Ko. Gestão de acervos de documentação linguística: desafios, responsabilidades e planejamento. *In*: BRESCANCINI, Cláudia (org.). **Projeto VARSUL – Variação Linguística do Sul do Brasil – 36 anos**. Porto Alegre: Zouk, 2021. p. 13-51.

FREITAG, Raquel Meister Ko. O tratamento da variação linguística na escola para o combate à discriminação e ao preconceito. *In*: SILVA, Noadia Íris da; MINUSSI, Rafael Dias (org.). **Linguística na educação básica**. Campinas: Mercado de Letras, 2022a. p. 209-222.

FREITAG, Raquel Meister Ko. Sociolinguistic repositories as asset: challenges and difficulties in Brazil. **The Electronic Library**, v. 40, n. 5, p. 607-622, 2022b. DOI: <https://doi.org/10.1108/EL-03-2022-0071>.

FREITAG, Raquel Meister Ko. **Linguística e inteligência artificial**: por dentro da máquina de fazer língua. São Paulo: Contexto, 2026a.

FREITAG, Raquel Meister Ko. Sociolinguística no Brasil: marcos, contribuições e perspectivas. *In*: BATISTA, Ronaldo de Oliveira (org.). **Linguísticas e linguistas no Brasil: uma história dos anos 1960-1990**. São Paulo: Parábola, 2026b. p. 255-269.

FREITAG, Raquel Meister Ko; GOIS, Túlio Sousa de. Performance in a dialectal profiling task of LLMs for varieties of Brazilian Portuguese. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 15., 2024, Belém. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2024. p. 317-326. DOI: <https://doi.org/10.5753/stil.2024.241891>.

FREITAG, R. M. K.; MACHADO VIEIRA, M. S.; BARBOSA, J. B.; OLIVEIRA JR., M.; ATAÍDE, C.; CORREIA, A. S.; POST DA SILVEIRA, A.; CARVALHO, A. B.; ARAUJO, A. S.; CARDOSO, B. C. S.; SNICHELOTTO, C. A. R.; MARTINS, E. C.; PROCÓPIO, E. S.; BATTISTI, E.; TEIXEIRA, E. N.; ALTINO, F. C.; BATISTA, H. R.; MACEDO, H. T.; MONGUILHOTT, I. O. S.; CLEVESTON, I.; DICKINSON, K. V.; HÜBNER, L. C.; MIRANDA, L. S.; MOTA, M. B.; SENE, M. G.; SILVA, M. R.; SOUSA, M. D. A. F.; SAVEDRA, M. M. G.; BIN, P. R.; QUADROS, R. M.; MARENGO, S. M. D. A.; ARAÚJO, S. S. F.; GOIS, T. S.; SOUSA, V. V.; DIAS, V. C. Plataforma da Diversidade Linguística Brasileira: dados linguísticos para uma IA brasileira. **Cadernos de Linguística**, Campinas, v. 6, n. 4, e863, 2025. DOI: <https://doi.org/10.25189/2675-4916.2025.V6.N4.ID863>.

IPHAN – INSTITUTO DO PATRIMÔNIO HISTÓRICO E ARTÍSTICO NACIONAL. **Inventário Nacional da Diversidade Linguística**. Brasília, DF: IPHAN, 2010. Disponível em: <https://indl.iphan.gov.br/>. Acesso em: 30 set. 2026.

KERCHE, Francisco W.; ZOOK, Matthew; GRAHAM, Mark. The silicon gaze: a typology of biases and inequality in LLMs through the lens of place. **Platforms & Society**, v. 3, 2026. DOI: <https://doi.org/10.1177/29768624251408919>.

KRAUWER, Steven; MAEGAARD, Bente. CLARIN – how it started. In: FIŠER, Darja; WITT, Andreas (ed.). **CLARIN: the infrastructure for language resources**. Berlin; Boston: De Gruyter, 2022. p. 1-30. DOI: <https://doi.org/10.1515/9783110767377-001>.

MACHADO VIEIRA, Marcia dos Santos; BARBOSA, Juliana Bertucci. Coleções de dados brasileiras para o ensino de português. In: MEIRELES, Vanessa; MACHADO VIEIRA, Marcia dos Santos (org.). **Variação e ensino de português no mundo**. São Paulo: Blucher, 2022. p. 317-354.

MACHADO VIEIRA, Marcia dos Santos; BARBOSA, Juliana Bertucci; FREITAG, Raquel Meister Ko; BORGES, Maria Manuel; MEDEIROS, Ana Lúcia Silva. Acervos de dados abertos à sociedade: memória linguística e sociocultural e potencialidade de (re)uso. **Cadernos de Linguística**, Campinas, v. 2, n. 1, e607, 2021. DOI: <https://doi.org/10.25189/2675-4916.2021.V2.N1.ID607>.

MACHADO VIEIRA, Marcia dos Santos; FREITAG, Raquel Meister Ko; BARBOSA, Juliana Bertucci; OLIVEIRA JR., Miguel; ATAÍDE, Cleber A.; FELICÍSSIMO, Carolina Howard. Desafios para a implementação de um repositório científico temático de dados linguísticos brasileiros. **Revista da Rede Brasileira de Repositórios Digitais**, Florianópolis, v. 1, p. 1-10, 2026a. Trabalho apresentado no IV Encontro da Rede Brasileira de Repositórios Digitais, Florianópolis, 2026.

MACHADO VIEIRA, Marcia dos Santos; WIEDEMER, Marcos Luiz; BARBOSA, Juliana Bertucci; FREITAG, Raquel Meister Ko. **Mapeamento de bancos de dados (socio)linguísticos no Brasil**. [S. l.: s. n.], 2026b. 9 p. DOI: [10.5281/zenodo.21998473](https://doi.org/10.5281/zenodo.21998473).

MENGESHA, Zion; HELDRETH, Courtney; LAHAV, Michal; SUBLEWSKI, Juliana; TUENNERMAN, Elyse. “I don’t think these devices are very culturally sensitive”: impact of automated speech recognition errors on African Americans. **Frontiers in Artificial Intelligence**, v. 4, 725911, 2021. DOI: <https://doi.org/10.3389/frai.2021.725911>.

MUSEU DA LÍNGUA PORTUGUESA. **Brasil ganha centro de documentação de línguas e culturas indígenas**. São Paulo: Museu da Língua Portuguesa, 2025. Disponível em: <https://www.museudalinguaportuguesa.org.br/brasil-ganha-centro-de-documentacao-de-linguas-e-culturas-indigenas/>. Acesso em: [INSERIR DATA DE ACESSO].

PIRES, Ramon; ABONIZIO, Hugo; ALMEIDA, Thales Sales; NOGUEIRA, Rodrigo. Sabiá: Portuguese large language models. *In*: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS, 12., 2023, Belo Horizonte. **Proceedings** [...]. Cham: Springer, 2023. p. 226-240. DOI: https://doi.org/10.1007/978-3-031-45392-2_15.

SALES, Luana Farias; SAYÃO, Luís Fernando. O impacto da curadoria digital dos dados de pesquisa na comunicação científica. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 17, n. esp. 2, p. 118-135, 2012. DOI: <https://doi.org/10.5007/1518-2924.2012v17nesp2p118>.

SALES, Luana Farias; VEIGA, Viviane Santos de Oliveira; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. DOI: <https://doi.org/10.22477/9786589167242>.

SAYÃO, Luís Fernando; SALES, Luana Farias. Algumas considerações sobre os repositórios digitais de dados de pesquisa. **Informação & Informação**, Londrina, v. 21, n. 2, p. 90-115, 2016. DOI: <https://doi.org/10.5433/1981-8920.2016v21n2p90>.

SHRAWGI, Hari; RATH, Prasanjit; SINGHAL, Tushar; DANDAPAT, Sandipan. Uncovering stereotypes in large language models: a task complexity-based approach. *In*: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 18., 2024, St. Julian's. **Proceedings** [...]. [S. l.]: Association for Computational Linguistics, 2024. p. 1841-1857.

SOUSA, Marta Deysiane Alves Faria; FREITAG, Raquel Meister Ko. Bancos de dados sociolinguísticos e a Ciência Aberta: compartilhamento de dados e conhecimentos. **Caderno Revista Diálogos**, v. 12, n. 1, p. 165-187, 2024.

THOMPSON, Brian; DHALIWAL, Mehak Preet; FRISCH, Peter; DOMHAN, Tobias; FEDERICO, Marcello. A shocking amount of the web is machine translated: insights from multi-way parallelism. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 62., 2024, Bangkok. **Findings of the Association for Computational Linguistics**: ACL 2024. Bangkok: Association for Computational Linguistics, 2024. p. 1763-1775. DOI: <https://doi.org/10.18653/v1/2024.findings-acl.103>.

TORINO, Emanuelle; MONTEIRO, Elizabete Cristina de Souza de Aguiar; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Plano de gestão de dados de pesquisa de povos indígenas: considerações acerca dos princípios FAIR e CARE. **Revista Brasileira de Preservação Digital**, Campinas, v. 4, e023007, 2023.

WINDHOUWER, Menzo; GOOSEN, Twan. Component metadata infrastructure. *In*: FIŠER, Darja; WITT, Andreas (ed.). **CLARIN**: the infrastructure for language resources. Berlin; Boston: De Gruyter, 2022. p. 191-222. DOI: <https://doi.org/10.1515/9783110767377-008>.

WITTENBURG, Peter. Archiving and accessing language resources. **Concurrency and Computation: practice and experience**, v. 22, n. 17, p. 2354-2368, 2010. DOI: <https://doi.org/10.1002/cpe.1605>.

ZHOU, Chunting; LIU, Pengfei; XU, Puxin; IYER, Srinivasan; SUN, Jiao; MAO, Yuning; MA, Xuezhe; EFRAT, Avia; YU, Ping; YU, Lili; ZHANG, Susan; GHOSH, Gargi; LEWIS, Mike; ZETTLEMOYER, Luke; LEVY, Omer. LIMA: less is more for alignment. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 37., 2023, New Orleans.

Advances in Neural Information Processing Systems 36. Red Hook: Curran Associates, 2023. p. 55006-55021.

Declaração de contribuição dos autores

Todas as autoras contribuíram nos seguintes papéis da taxonomia CRediT: Conceituação; Curadoria de dados; Análise formal; Administração do projeto; Escrita – rascunho original; Escrita – revisão e edição.

Declaração de conflito de interesse

As autoras declaram não ter interesses financeiros relacionados a este trabalho. Declaram, como interesse não financeiro, que participam da concepção e da coordenação do projeto da Plataforma da Diversidade Linguística Brasileira, objeto deste artigo, e que integram as associações científicas que o promovem.

Declaração de disponibilidade de dados da pesquisa

Todo o conjunto de dados de apoio aos resultados deste estudo foi publicado no próprio artigo.

Declaração de uso de IA

Na preparação deste manuscrito, as autoras utilizaram o Claude (Anthropic, modelo Claude Opus 5.5) para adequar o texto às normas da revista e da ABNT, organizar e conferir referências e propor redação e reorganização de trechos. Todo o conteúdo foi revisado, editado e validado pelas autoras, que assumem integral responsabilidade pelo texto.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença [Creative Commons CC-BY](#).
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.