ReGeo ISSN: 2177-3246

A PLATAFORMA DA DIVERSIDADE LINGUÍSTICA BRASILEIRA: TREINAR IA EM PORTUGUÊS E LÍNGUAS MINORITÁRIAS SEM REPRODUZIR O VIÉS DE DADOS DO INGLÊS

THE BRAZILIAN LINGUISTIC DIVERSITY PLATFORM: TRAINING AI IN PORTUGUESE AND MINORITY LANGUAGES WITHOUT REPRODUCING ENGLISH DATA BIAS

LA PLATAFORMA DE LA DIVERSIDAD LINGÜÍSTICA BRASILEÑA: ENTRENAR EL IA EN PORTUGUÉS Y EN LENGUAS MINORITARIAS SIN REPRODUCIR EL SESGO DE LOS DATOS DEL INGLÉS

https://doi.org/10.56238/revgeov17n9-026

Data de submissão: 19/07/2026 **Data de publicação:** 02/09/2026

Karen Lowhany Costa da Silva
Mestranda em Propriedade Intelectual e Transferência de Tecnologia para a Inovação
Instituição: Universidade Federal do Tocantins (PROFNIT- UFT).
E-mail: karen_lowhany@mail.uft.edu.br
Lattes: http://lattes.cnpq.br/9733562596206371

Viviane Martins de Oliveira
Mestra em Linguagem e Letramento
Instituição: Universidade Estadual do Ceará-UECE
E-mail: vivianemartins789@gmail.com
Lattes: http://lattes.cnpq.br/2113718002145892

RESUMO

O desenvolvimento de grandes modelos de linguagem tem sido dominado por corpora em inglês e por variedades escritas padrão, o que produz sistemas de inteligência artificial que reconhecem como legítima apenas uma fração mínima das formas de falar existentes no mundo. No Brasil, país que abriga, além do português e de suas múltiplas variedades regionais e sociais, mais de duzentas e cinquenta línguas indígenas, de imigração e de sinais, esse viés de dados ameaça converter a inclusão digital em novo vetor de apagamento linguístico. Este artigo analisa a Plataforma da Diversidade Linguística Brasileira, consórcio de laboratórios e grupos de pesquisa que se propõe a preparar dados linguísticos representativos da diversidade nacional para o treinamento de modelos de linguagem, articulando protocolo comum de coleta de campo, procedimentos de transcrição, alinhamento e etiquetagem e constituição de datasets. A pesquisa, de natureza qualitativa e documental, examina o desenho da iniciativa à luz da literatura sobre desigualdade de recursos no processamento de linguagem natural, da política patrimonial inaugurada pelo Inventário Nacional da Diversidade Linguística e do Plano Brasileiro de Inteligência Artificial. Os resultados indicam que a plataforma desloca a questão do viés do plano corretivo, de ajustes posteriores em modelos prontos, para o plano constitutivo, da formação dos próprios dados, mas que seu êxito depende de financiamento estável, de salvaguardas éticas de soberania de dados das comunidades falantes e da integração efetiva entre a agenda patrimonial e a agenda tecnológica.

Palavras-chave: Diversidade Linguística. Inteligência Artificial. Línguas Minoritárias. Viés de Dados. Português Brasileiro.

ABSTRACT
The development of large language models has been dominated by English corpora and by standard written varieties, producing artificial intelligence systems that recognize as legitimate only a minimal fraction of the world's ways of speaking. In Brazil, a country that is home, beyond Portuguese and its multiple regional and social varieties, to more than two hundred and fifty Indigenous, immigration and sign languages, this data bias threatens to turn digital inclusion into a new vector of linguistic erasure. This article analyzes the Brazilian Linguistic Diversity Platform, a consortium of laboratories and research groups that proposes to prepare linguistic data representative of national diversity for the training of language models, articulating a common field collection protocol, transcription, alignment and labeling procedures, and dataset constitution. The research, qualitative and documental in nature, examines the design of the initiative in light of the literature on resource inequality in natural language processing, of the heritage policy inaugurated by the National Inventory of Linguistic Diversity, and of the Brazilian Artificial Intelligence Plan. The results indicate that the platform shifts the question of bias from the corrective plane, of later adjustments to finished models, to the constitutive plane, of the formation of the data themselves, but that its success depends on stable funding, on ethical safeguards of data sovereignty for speaker communities, and on the effective integration between the heritage agenda and the technological agenda.

Keywords: Linguistic Diversity. Artificial Intelligence. Minority Languages. Data Bias. Brazilian Portuguese.

RESUMEN
El desarrollo de grandes modelos de lenguaje ha sido dominado por corpus en inglés y por variedades escritas estándar, lo que produce sistemas de inteligencia artificial que reconocen como legítima solo una fracción mínima de las formas de hablar existentes en el mundo. En Brasil, país que alberga, además del portugués y sus múltiples variedades regionales y sociales, más de doscientas cincuenta lenguas indígenas, de inmigración y de signos, este sesgo de datos amenaza con convertir la inclusión digital en un nuevo vector de borrado lingüístico. Este artículo analiza la Plataforma de la Diversidad Lingüística Brasileña, consorcio de laboratorios y grupos de investigación que se propone preparar datos lingüísticos representativos de la diversidad nacional para el entrenamiento de modelos de lenguaje, articulando protocolo común de recolección de campo, procedimientos de transcripción, alineamiento y etiquetado y constitución de datasets. La investigación, de naturaleza cualitativa y documental, examina el diseño de la iniciativa a la luz de la literatura sobre desigualdad de recursos en el procesamiento de lenguaje natural, de la política patrimonial inaugurada por el Inventario Nacional de la Diversidad Lingüística y del Plan Brasileño de Inteligencia Artificial. Los resultados indican que la plataforma desplaza la cuestión del sesgo del plano correctivo, de ajustes posteriores en modelos listos, al plano constitutivo, de la formación de los propios datos, pero que su éxito depende de financiamiento estable, de salvaguardas éticas de soberanía de datos de las comunidades hablantes y de la integración efectiva entre la agenda patrimonial y la agenda tecnológica.

Palabras clave: Diversidad Lingüística. Inteligencia Artificial. Lenguas Minoritarias. Sesgo de Datos. Portugués Brasileño.

1 INTRODUÇÃO

Os grandes modelos de linguagem que hoje medeiam buscas, traduções, assistentes virtuais e ferramentas educacionais foram treinados sobre uma base de dados profundamente desigual: das mais de sete mil línguas faladas no mundo, apenas um punhado dispõe de recursos digitais suficientes para sustentar tecnologias de linguagem maduras, e o inglês concentra parcela desproporcional dos corpora, dos benchmarks e dos investimentos de pesquisa (JOSHI *et al.*, 2020).

Essa concentração não é um detalhe técnico, mas uma escolha com consequências: modelos treinados sobre dados massivamente coletados da internet tendem a codificar as visões de mundo, os padrões discursivos e as hierarquias sociais dominantes nos textos que ingerem, funcionando como papagaios estocásticos que repetem, em escala industrial, os vieses de sua dieta de dados (BENDER *et al.*, 2021).

O Brasil ocupa posição singular nesse cenário. De um lado, o português é uma das línguas mais faladas do mundo e dispõe de presença digital considerável; de outro, essa presença é dominada pela variedade escrita padrão, enquanto as falas regionais e populares que constituem o cotidiano linguístico da maioria da população permanecem sub-representadas. Além disso, o país abriga mais de duzentas e cinquenta línguas além do português: línguas indígenas de dezenas de famílias, línguas de imigração enraizadas em comunidades centenárias, línguas de sinais e falares de comunidades tradicionais, quase todas ausentes dos corpora que alimentam a inteligência artificial contemporânea.

É nesse contexto que emerge a Plataforma da Diversidade Linguística Brasileira, consórcio de laboratórios e grupos de pesquisa em linguística que se propõe a preparar dados linguísticos para o treinamento de grandes modelos de linguagem considerando a diversidade linguística brasileira, com protocolo conjunto de coleta de campo, procedimentos de transcrição, alinhamento e etiquetagem e constituição de datasets representativos (FREITAG *et al.*, 2025a).

A pergunta que orienta este artigo é dupla: em que medida a Plataforma da Diversidade Linguística Brasileira representa uma resposta estrutural, e não apenas paliativa, ao viés de dados do inglês; e quais condições institucionais, éticas e financeiras precisam ser satisfeitas para que a iniciativa cumpra sua promessa sem reproduzir, em escala nacional, a mesma lógica extrativa que critica no plano internacional.

A relevância do tema é reforçada pela convergência de três agendas públicas: a agenda patrimonial, inaugurada pelo Inventário Nacional da Diversidade Linguística em 2010; a agenda internacional da Década das Línguas Indígenas, que exige dos Estados ações concretas de revitalização; e a agenda tecnológica do Plano Brasileiro de Inteligência Artificial, que elegeu a

soberania de dados e de infraestrutura como eixos estratégicos do investimento público em IA (BRASIL, 2024).

2 FUNDAMENTAÇÃO TEÓRICA

2.1 A DESIGUALDADE DE RECURSOS NO PROCESSAMENTO DE LINGUAGEM NATURAL

A literatura de processamento de linguagem natural descreve uma taxonomia da desigualdade: no topo, línguas hiperdocumentadas, com corpora anotados, benchmarks e comunidades de pesquisa consolidadas; na base, a imensa maioria das línguas do mundo, para as quais inexistem sequer dados brutos digitalizados em volume utilizável. A trajetória de uma língua nessa hierarquia não reflete seu número de falantes nem sua vitalidade cultural, mas a história econômica e política de sua documentação (JOSHI *et al.*, 2020).

O problema não se resolve com modelos multilíngues treinados sobre dados desbalanceados: quando a fração de uma língua no corpus é mínima, o modelo transfere para ela padrões estruturais da língua dominante, produzindo textos gramaticalmente plausíveis, porém culturalmente vazios ou distorcidos. Os riscos documentados incluem a homogeneização discursiva, o apagamento de variedades estigmatizadas e a amplificação de estereótipos presentes nos dados de origem (BENDER *et al.*, 2021).

A experiência brasileira com modelos monolíngues ilustra o valor de dados próprios: o BERTimbau, treinado sobre corpus de páginas web brasileiras, superou o modelo multilíngue de referência em todas as tarefas avaliadas de similaridade semântica, inferência textual e reconhecimento de entidades, demonstrando que investir em dados nacionais produz ganhos mensuráveis de desempenho (SOUZA; NOGUEIRA; LOTUFO, 2020).

O BERTimbau, contudo, também expõe o limite da primeira geração de modelos brasileiros: seu corpus de treinamento é a web em português, isto é, majoritariamente a escrita padrão urbana. O modelo fala português, mas fala um único português. A fronteira seguinte não é ter modelos na língua nacional, e sim ter modelos que conheçam a língua nacional em sua extensão real, falada, variável e plural.

Figura 1. A pirâmide dos dados linguísticos que alimentam a inteligência artificial



Fonte: elaboração própria com base em Joshi *et al.* (2020) e Freitag *et al.* (2025a).

2.2 SOCIOLINGUÍSTICA, VIÉS E O CASO BRASILEIRO

A sociolinguística oferece o arcabouço para compreender por que o viés de dados é também um viés social: as tecnologias de linguagem treinadas sobre a norma escrita tendem a tratar as variedades populares como erro, reproduzindo no ambiente digital a mesma hierarquização entre falares de prestígio e falares estigmatizados que opera na escola e no mercado de trabalho, e os estudos da linguagem têm papel central no desenho de sistemas de processamento de linguagem natural mais inclusivos (SILVA JÚNIOR, 2026).

No plano educacional, a adoção acrítica de ferramentas generativas ameaça aprofundar desigualdades: pesquisas recentes sobre os impactos éticos e sociais da inteligência artificial generativa apontam riscos de exclusão sociotecnológica e de enfraquecimento da mediação pedagógica quando a tecnologia chega antes da reflexão sobre seus pressupostos culturais e linguísticos (SILVA *et al.*, 2025).

Ao mesmo tempo, a literatura sobre inteligência artificial e inclusão escolar reconhece o potencial dessas ferramentas para personalizar a aprendizagem e apoiar estudantes com perfis diversos, desde que seu desenvolvimento incorpore as realidades linguísticas e culturais dos usuários a que se destinam, o que exige exatamente o tipo de dado que hoje falta (SANTOS *et al.*, 2025).

A dimensão territorial da língua completa o quadro teórico: estudos sobre povos indígenas brasileiros, como o realizado com o povo Akwẽ-Xerente do Tocantins, mostram que língua, memória

e território formam um contínuo inseparável, de modo que documentar uma língua minoritária não é apenas gravar palavras, mas registrar formas de dizer enraizadas em modos de vida, o que impõe à coleta de dados para IA exigências metodológicas e éticas que os raspadores de internet ignoram por completo (SANTANA *et al.*, 2026).

2.3 DA POLÍTICA PATRIMONIAL À POLÍTICA TECNOLÓGICA

O Brasil dispõe, desde 2010, de um instrumento pioneiro de reconhecimento da pluralidade linguística: o Decreto nº 7.387 instituiu o Inventário Nacional da Diversidade Linguística como mecanismo oficial de identificação, documentação e valorização das línguas portadoras de referência à identidade brasileira, permitindo que línguas indígenas e de imigração recebam o título de Referência Cultural Brasileira (BRASIL, 2010).

A agenda internacional reforça essa direção: a Década Internacional das Línguas Indígenas, proclamada para o período de 2022 a 2032, convoca os Estados a garantir o direito dos povos indígenas de preservar, revitalizar e transmitir suas línguas, incluindo expressamente o acesso dessas línguas às tecnologias digitais e aos meios de comunicação (UNESCO, 2022).

No plano tecnológico, o Plano Brasileiro de Inteligência Artificial para o período de 2024 a 2028 mobiliza investimentos da ordem de vinte e três bilhões de reais em cinco eixos, incluindo infraestrutura de supercomputação, nuvem soberana e desenvolvimento de modelos em língua portuguesa, sob o lema de uma IA voltada ao bem de todos (BRASIL, 2024).

Entre a política patrimonial, que reconhece as línguas como patrimônio, e a política tecnológica, que financia a infraestrutura de IA, abre-se, entretanto, uma lacuna: nenhum dos dois instrumentos define quem prepara os dados linguísticos que conectariam uma agenda à outra. É precisamente nessa lacuna que a Plataforma da Diversidade Linguística Brasileira se posiciona.

3 METODOLOGIA

Esta pesquisa adota abordagem qualitativa, de caráter exploratório e descritivo, estruturada como revisão integrativa da literatura combinada com análise documental, método adequado para sintetizar o conhecimento disponível sobre um fenômeno emergente e identificar lacunas que orientem pesquisas futuras (SOUZA; SILVA; CARVALHO, 2010).

O corpus documental compreende três conjuntos: primeiro, os documentos fundadores da Plataforma da Diversidade Linguística Brasileira, incluindo o registro do projeto publicado em periódico da Associação Brasileira de Linguística e sua versão preprint em português; segundo, os marcos normativos e programáticos brasileiros pertinentes, o Decreto nº 7.387 de 2010 e o Plano

Brasileiro de Inteligência Artificial; terceiro, a literatura científica nacional e internacional sobre viés de dados, desigualdade de recursos linguísticos e ética da coleta de dados de comunidades minoritárias (FREITAG *et al.*, 2025b).

A seleção das fontes priorizou publicações recentes, concentradas entre 2020 e 2026, em periódicos brasileiros e em veículos internacionais de referência da área de processamento de linguagem natural. A análise seguiu três eixos: o desenho institucional da plataforma, sua resposta ao problema do viés de dados e as condições de sustentabilidade e de ética da iniciativa.

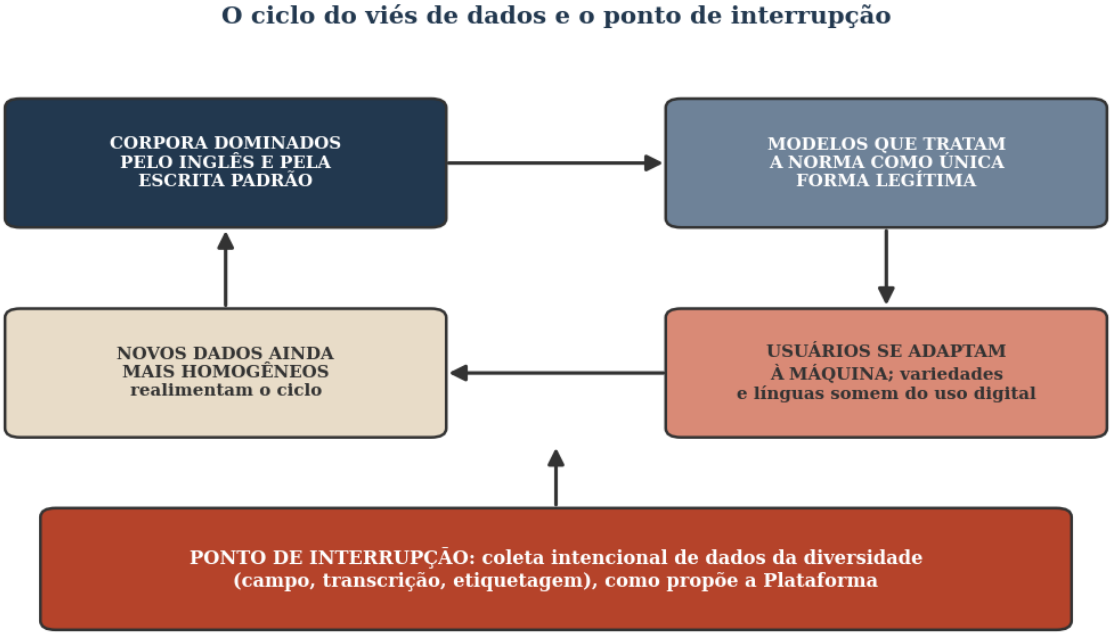
Por tratar-se de estudo teórico-documental, a pesquisa não envolveu seres humanos nem dados pessoais, dispensando apreciação por comitê de ética. As figuras e tabelas apresentadas são elaborações conceituais próprias, construídas a partir da literatura citada, sem reprodução de material protegido por direitos autorais.

4 RESULTADOS

4.1 O PROBLEMA CONSTITUTIVO: O CICLO DO VIÉS DE DADOS

O primeiro resultado da análise é a caracterização do viés de dados como um ciclo que se retroalimenta. Corpora dominados pelo inglês e pela escrita padrão geram modelos que tratam a norma como única forma legítima de linguagem; usuários, para serem compreendidos pelas máquinas, adaptam sua escrita e sua fala ao que o sistema reconhece; essa adaptação produz novos dados ainda mais homogêneos, que alimentam a próxima geração de modelos. A cada volta do ciclo, as variedades não padrão e as línguas minoritárias perdem espaço no ambiente digital, no exato momento histórico em que a presença digital se torna condição de vitalidade linguística.

Figura 2. O ciclo do viés de dados e o ponto de interrupção proposto pela Plataforma



Fonte: elaboração própria com base em Bender *et al.* (2021) e Freitag *et al.* (2025a).

A resposta corretiva convencional, que consiste em ajustar modelos prontos por meio de filtros, afinações e instruções posteriores, não interrompe esse ciclo, porque atua sobre o sintoma e não sobre a causa: o modelo não pode representar bem aquilo que nunca viu. A contribuição central da plataforma é deslocar a intervenção para o momento constitutivo, o da formação dos dados, mediante coleta intencional e metodologicamente controlada da diversidade linguística real (FREITAG *et al.*, 2025a).

4.2 A ARQUITETURA DA PLATAFORMA

A plataforma organiza-se como consórcio nacional de laboratórios e grupos de pesquisa em linguística, articulados por um protocolo conjunto de coleta de dados de campo, concebido para ser replicado longitudinalmente pelos grupos participantes, ao qual se somam procedimentos padronizados de transcrição, alinhamento e etiquetagem, com vistas à constituição de datasets que representem a diversidade linguística brasileira em suas múltiplas dimensões (FREITAG *et al.*, 2025a).

Essa arquitetura tem três méritos analíticos. Primeiro, aproveita capacidade instalada: os laboratórios de sociolinguística, dialetologia, fonética e línguas indígenas já dominam as técnicas de coleta e tratamento que a IA agora demanda, e o consórcio as converte em infraestrutura de dados. Segundo, distribui territorialmente a produção: em vez de um centro coletor único, uma rede que alcança as variedades onde elas são faladas. Terceiro, padroniza sem uniformizar: o protocolo comum garante interoperabilidade dos dados sem apagar as especificidades de cada situação de coleta.

Figura 3. Da coleta ao modelo: a arquitetura em rede da Plataforma



Fonte: elaboração própria com base em Freitag *et al.* (2025a).

A Tabela 1 sistematiza as camadas da diversidade linguística brasileira que a plataforma se propõe a representar, com a situação típica de dados de cada uma.

Tabela 1. Camadas da diversidade linguística brasileira e situação de dados para IA

Camada da diversidade	Exemplos	Situação típica de dados
Português escrito padrão	Textos jornalísticos, acadêmicos e institucionais da web	Abundante; domina os corpora existentes e os modelos atuais
Variedades faladas do português	Falares regionais, sociais e populares; fala espontânea	Escassa; acervos sociolinguísticos dispersos, pouco estruturados para IA
Línguas indígenas	Mais de duzentas línguas de dezenas de famílias linguísticas	Crítica; documentação heterogênea, raramente digitalizada em formato utilizável
Línguas de imigração	Talian, hunsrückisch, pomerano, japonês e outras línguas comunitárias	Muito escassa; registros comunitários e acadêmicos pontuais
Línguas de sinais	Libras e línguas de sinais emergentes	Escassa; exige dados em vídeo com anotação especializada

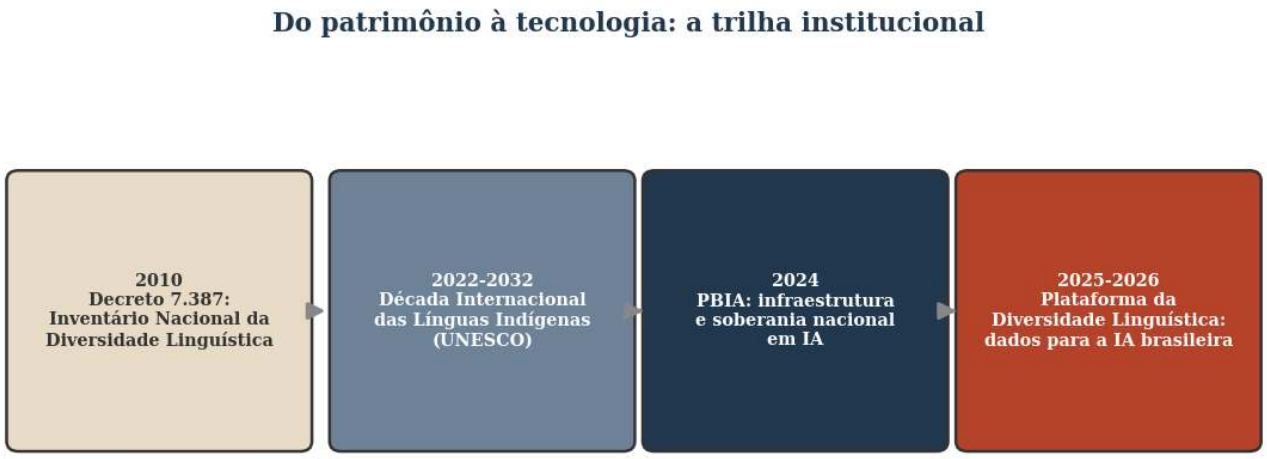
Fonte: elaboração própria com base em Freitag *et al.* (2025a) e Brasil (2010).

4.3 A TRILHA INSTITUCIONAL E O TESTE DO FINANCIAMENTO

A plataforma não surge no vazio: ela é o elo mais recente de uma trilha institucional que começa no Inventário Nacional da Diversidade Linguística, passa pela adesão brasileira à Década Internacional das Línguas Indígenas e desemboca na agenda de soberania tecnológica do Plano Brasileiro de

Inteligência Artificial, que prevê infraestrutura de supercomputação e desenvolvimento de modelos em português (BRASIL, 2024).

Figura 4. Do patrimônio à tecnologia: a trilha institucional que desemboca na Plataforma



Fonte: elaboração própria com base em Brasil (2010; 2024), UNESCO (2022) e Freitag *et al.* (2025a).

O teste decisivo dessa trilha, contudo, é o financiamento. A proposta da plataforma foi submetida à chamada pública dos Institutos Nacionais de Ciência e Tecnologia de 2024 e, embora aprovada quanto ao mérito, não foi contemplada com recursos naquela rodada. O dado é analiticamente revelador: o país que decidiu investir bilhões em infraestrutura de IA ainda não assegurou o financiamento da camada de dados linguísticos que daria a essa infraestrutura um conteúdo genuinamente nacional. Sem essa camada, o supercomputador corre o risco de processar, em território brasileiro, os mesmos dados enviesados de sempre.

4.4 AS SALVAGUARDAS ÉTICAS COMO CONDIÇÃO DE LEGITIMIDADE

O quarto resultado diz respeito à ética da coleta. Dados de línguas minoritárias não são commodity: são expressão de memória, território e identidade coletiva, como evidencia a pesquisa com povos como os Akwẽ-Xerente, e sua incorporação a modelos de linguagem sem consentimento e sem retorno configuraria nova modalidade de extrativismo, agora linguístico (SANTANA *et al.*, 2026).

Figura 5. Salvaguardas éticas para a coleta de dados de línguas minoritárias



Fonte: elaboração própria com base em Freitag *et al.* (2025a) e UNESCO (2022).

Cinco salvaguardas emergem da análise como condição de legitimidade da plataforma: consentimento livre, prévio e informado das comunidades; protagonismo dos falantes como parceiros da pesquisa, e não como fontes a serem extraídas; soberania de dados, com governança comunitária sobre uso, acesso e benefício; retorno concreto, na forma de produtos úteis à educação e à revitalização linguística; e transparência na documentação dos corpora e de seus usos nos modelos. A Tabela 2 contrasta o paradigma dominante de treinamento com o paradigma orientado à diversidade que a plataforma propõe.

Tabela 2. Dois paradigmas de treinamento de modelos de linguagem

Dimensão	Paradigma dominante (viés do inglês)	Paradigma orientado à diversidade
Origem dos dados	Raspagem massiva da internet, sem curadoria linguística	Coleta de campo com protocolo científico e curadoria especializada
Variedades representadas	Escrita padrão das línguas hegemônicas	Fala espontânea, variedades regionais e sociais, línguas minoritárias
Relação com falantes	Anonimato extrativo; falante ignora o uso de seus dados	Consentimento, parceria e retorno às comunidades
Tratamento do viés	Correção posterior sobre modelo pronto	Prevenção constitutiva na formação dos datasets
Avaliação do modelo	Benchmarks traduzidos do inglês	Tarefas e referências construídas para a realidade linguística nacional
Resultado esperado	IA que fala uma norma única	IA que reconhece o país como ele fala

Fonte: elaboração própria com base em Bender *et al.* (2021), Joshi *et al.* (2020) e Freitag *et al.* (2025a).

5 DISCUSSÃO

Os resultados dialogam diretamente com a literatura internacional sobre desigualdade de recursos: a taxonomia das línguas segundo sua dotação de dados mostra que a posição de uma língua na hierarquia tecnológica é historicamente construída e, portanto, politicamente reversível, desde que haja investimento deliberado na produção de recursos, exatamente o que a plataforma operacionaliza para o caso brasileiro (JOSHI *et al.*, 2020).

A advertência sobre os papagaios estocásticos ganha, no contexto brasileiro, um contorno específico: não se trata apenas do risco de modelos repetirem estereótipos, mas do risco de a própria infraestrutura digital do país consolidar a hierarquização entre variedades de prestígio e variedades estigmatizadas que a sociolinguística denuncia há décadas, agora com a aparência de neutralidade técnica que os sistemas computacionais conferem às suas saídas (BENDER *et al.*, 2021).

Nesse sentido, a contribuição da sociolinguística ao combate dos vieses da inteligência artificial não é acessória, mas estrutural: é ela que fornece os conceitos de variação, prestígio e estigma sem os quais nem sequer é possível formular o que conta como viés linguístico em um sistema computacional, e é ela que domina as metodologias de coleta capazes de capturar a língua em uso real (SILVA JÚNIOR, 2026).

O caso do BERTimbau oferece à discussão um precedente encorajador e um alerta: encorajador porque demonstrou que dados nacionais melhoram objetivamente o desempenho dos modelos; alerta porque mostrou que a mera nacionalidade do corpus não garante sua representatividade, já que a web em português continua sendo um retrato parcial e socialmente enviesado da língua (SOUZA; NOGUEIRA; LOTUFO, 2020).

No plano educacional, as consequências são imediatas: se as ferramentas generativas que chegam às escolas tratarem as variedades populares como erro, a promessa de personalização da aprendizagem se converterá em mais um mecanismo de correção normativa dos estudantes de origem popular, aprofundando as desigualdades que a literatura sobre IA e inclusão escolar adverte que a tecnologia pode tanto mitigar quanto amplificar (SANTOS *et al.*, 2025; SILVA *et al.*, 2025).

A discussão institucional revela uma assimetria que merece destaque: a agenda patrimonial brasileira reconheceu as línguas como patrimônio em 2010, mas com instrumentos de documentação pensados para o registro cultural, não para o treinamento de modelos; a agenda tecnológica de 2024 financiou a infraestrutura de processamento, mas sem definir a política de dados linguísticos. A plataforma é a primeira iniciativa a conectar explicitamente as duas pontas, e o fato de ter sido aprovada no mérito sem obter financiamento sugere que o sistema de fomento ainda não dispõe de categoria adequada para infraestruturas de dados que são, simultaneamente, ciência linguística e insumo tecnológico.

Por fim, a dimensão ética discutida nos resultados encontra respaldo na agenda da Década Internacional das Línguas Indígenas, que condiciona a digitalização das línguas ao protagonismo de seus falantes: a inclusão de uma língua em um modelo de IA só constitui revitalização se a comunidade controlar os termos dessa inclusão; caso contrário, é apropriação que esvazia o patrimônio que alega preservar (UNESCO, 2022).

6 LIMITAÇÕES DA PESQUISA

Este estudo apresenta limitações que devem ser explicitadas. Em primeiro lugar, trata-se de análise documental de uma iniciativa em estágio inicial: a plataforma foi registrada formalmente em 2025 e seus resultados empíricos, os datasets e os modelos deles derivados, ainda não existem em escala avaliável, de modo que as conclusões dizem respeito ao desenho da proposta, não a seu desempenho.

Em segundo lugar, a análise baseia-se em fontes públicas, sem entrevistas com os pesquisadores do consórcio ou com as comunidades falantes envolvidas, o que limita o alcance das inferências sobre a dinâmica interna da rede e sobre a percepção das comunidades quanto às salvaguardas propostas.

Em terceiro lugar, o cenário tecnológico e regulatório da inteligência artificial evolui com rapidez incomum: números de investimento, chamadas de fomento e marcos normativos citados refletem o estado das fontes até o momento da escrita, e desenvolvimentos posteriores podem alterar as condições institucionais aqui analisadas.

7 PERSPECTIVAS FUTURAS

A agenda de pesquisa que decorre deste estudo é ampla. No plano empírico, será necessário avaliar, quando os primeiros datasets da plataforma estiverem disponíveis, seu impacto mensurável sobre o desempenho de modelos em tarefas com variedades não padrão e línguas minoritárias, replicando para a diversidade interna do português o tipo de comparação que o BERTimbau estabeleceu entre modelos nacionais e multilíngues.

No plano ético-institucional, estudos futuros devem acompanhar a implementação das salvaguardas de soberania de dados, documentando como consentimento, governança comunitária e retorno se materializam em contratos, licenças de uso e produtos concretos, e examinar mecanismos de financiamento adequados a infraestruturas de dados de longo prazo, que não se encaixam nos ciclos curtos dos editais convencionais.

No plano educacional, abre-se a perspectiva de investigar o uso de modelos treinados com dados da diversidade em contextos escolares concretos, verificando se sistemas que reconhecem as variedades dos estudantes alteram as dinâmicas de estigmatização linguística documentadas pela sociolinguística educacional, e de estender essa investigação à educação escolar indígena, onde a tecnologia linguística pode servir tanto ao ensino do português quanto à revitalização das línguas originárias.

8 CONCLUSÃO

Este artigo analisou a Plataforma da Diversidade Linguística Brasileira como resposta estrutural ao viés de dados que o domínio do inglês e da escrita padrão imprime aos grandes modelos de linguagem. A análise mostrou que a originalidade da iniciativa está no deslocamento do enfrentamento do viés: do plano corretivo, dos ajustes sobre modelos prontos, para o plano constitutivo, da formação dos próprios dados, mediante um consórcio nacional que converte a capacidade instalada da linguística brasileira em infraestrutura de dados para a inteligência artificial.

Três conclusões sintetizam o estudo. Primeira: o viés de dados é um ciclo que se retroalimenta, e só a coleta intencional da diversidade, com protocolo científico e curadoria especializada, interrompe esse ciclo no ponto certo. Segunda: a plataforma conecta duas agendas públicas que corriam em paralelo, a patrimonial e a tecnológica, mas essa conexão permanecerá promessa enquanto o sistema de fomento não reconhecer infraestruturas de dados linguísticos como investimento estratégico, lacuna evidenciada pela aprovação de mérito sem financiamento. Terceira: a legitimidade do projeto depende de salvaguardas éticas que garantam às comunidades falantes o controle sobre seus dados, sob pena de

a inclusão digital reeditar, em novo formato, o extrativismo que historicamente incidiu sobre os povos e suas línguas.

Se essas condições forem satisfeitas, o resultado ultrapassa a técnica: um país que treina sua inteligência artificial com a totalidade de suas vozes não está apenas melhorando modelos, está decidindo que nenhuma forma brasileira de falar é erro a ser corrigido. Essa decisão, mais do que qualquer supercomputador, é o que distingue uma IA meramente instalada no Brasil de uma IA efetivamente brasileira.

CONFLITOS DE INTERESSE

Os autores declaram não haver conflitos de interesse de natureza financeira, institucional ou pessoal relacionados a este estudo. A pesquisa não recebeu financiamento específico de agências de fomento dos setores público, privado ou sem fins lucrativos.

REFERÊNCIAS

BENDER, E. M.; GEBRU, T.; McMILLAN-MAJOR, A.; SHMITCHELL, S. On the dangers of stochastic parrots: can language models be too big?. In: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT '21), 2021, Nova York. Proceedings [...]. Nova York: ACM, 2021. p. 610-623. DOI: 10.1145/3442188.3445922.

BRASIL. Decreto nº 7.387, de 9 de dezembro de 2010. Institui o Inventário Nacional da Diversidade Linguística e dá outras providências. Diário Oficial da União: seção 1, Brasília, DF, 10 dez. 2010.

BRASIL. Ministério da Ciência, Tecnologia e Inovação. Plano Brasileiro de Inteligência Artificial (PBIA) 2024-2028: IA para o bem de todos. Brasília, DF: MCTI, 2024.

FREITAG, R. *et al.* Brazilian Linguistic Diversity Platform: linguistic data for a Brazilian AI. Cadernos de Linguística, Campinas, v. 6, n. 4, 2025b.

FREITAG, R. M. K. *et al.* Plataforma da Diversidade Linguística Brasileira: dados linguísticos para uma IA brasileira. SciELO Preprints, 2025a. DOI: 10.1590/SciELOPreprints.11957.

JOSHI, P.; SANTY, S.; BUDHIRAJA, A.; BALI, K.; CHOUDHURY, M. The state and fate of linguistic diversity and inclusion in the NLP world. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 58., 2020, on-line. Proceedings [...]. Stroudsburg: ACL, 2020. p. 6282-6293.

SANTANA, L. S. B. *et al.* Entre palavra e território: língua, memória e formas de dizer do povo amazônico tocantinense Akwẽ-Xerente. Aracê, São José dos Pinhais, v. 8, n. 3, e12729, 2026. DOI: 10.56238/arev8n3-156.

SANTOS, D. S. dos *et al.* Inteligência artificial e inclusão escolar: desafios e possibilidades na educação do século XXI. Aracê, São José dos Pinhais, v. 7, n. 10, e9488, 2025. DOI: 10.56238/arev7n10-306.

SILVA JÚNIOR, E. S. da. O papel da sociolinguística no combate aos vieses da inteligência artificia1: contribuições dos estudos da linguagem e do processamento de linguagem natural para tecnologias mais inclusivas. Lumen et Virtus, São José dos Pinhais, v. 17, n. 63, e13946, 2026. DOI: 10.56238/levv17n63-012.

SILVA, D. S. da; GOMES, H. P.; SOUSA, D. B.; MIRANDA, M. da S. Inteligência artificial generativa: impactos éticos e sociais. Lumen et Virtus, São José dos Pinhais, v. 16, n. 54, e10511, 2025. DOI: 10.56238/levv16n54-169.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS), 9., 2020, Rio Grande. Intelligent Systems: LNCS 12319. Cham: Springer, 2020. p. 403-417. DOI: 10.1007/978-3-030-61377-8_28.

SOUZA, M. T. de; SILVA, M. D. da; CARVALHO, R. de. Revisão integrativa: o que é e como fazer. Einstein (São Paulo), São Paulo, v. 8, n. 1, p. 102-106, 2010.

UNESCO. Global action plan of the International Decade of Indigenous Languages (2022-2032). Paris: UNESCO, 2022.