ReGeo ISSN: 2177-3246

DIVERSIDADE LINGUÍSTICA SOB AMEAÇA DA IA GENERATIVA: O VIÉS DE SELEÇÃO DE VARIEDADE NAS APLICAÇÕES TECNOLÓGICAS E O CICLO VICIOSO DE PADRONIZAÇÃO DE UMA VARIEDADE DOMINANTE

LINGUISTIC DIVERSITY UNDER THREAT FROM GENERATIVE AI: THE VARIETY SELECTION BIAS IN TECHNOLOGICAL APPLICATIONS AND THE VICIOUS CYCLE OF STANDARDIZATION TOWARD A DOMINANT VARIETY

LA DIVERSIDAD LINGÜÍSTICA AMENAZADA POR LA IA GENERATIVA: EL SESGO DE SELECCIÓN DE VARIEDAD EN LAS APLICACIONES TECNOLÓGICAS Y EL CÍRCULO VICIOSO DE LA ESTANDARIZACIÓN DE UNA VARIEDAD DOMINANTE

https://doi.org/10.56238/revgeov17n8-010

Data de submissão: 19/06/2026 **Data de publicação:** 03/08/2026

André Alves Oliveira
Mestrando em Educação
Instituição: Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM)
Lattes: http://lattes.cnpq.br/7242352701224074

Diorman Luiz Phellype Augusto Venâncio do Amaral
Mestrando em Comunicação
Instituição: Universidade Paulista (UNIP)
Lattes: http://lattes.cnpq.br/4594082602535410

Juliano Brambilla Neri
Mestre em Línguas Estrangeiras Modernas
Lattes: http://lattes.cnpq.br/5538716506297419

RESUMO

A rápida difusão das aplicações de inteligência artificial generativa, baseadas em grandes modelos de linguagem, coloca novas questões para a sociolinguística e para a preservação da diversidade linguística. Este artigo, de natureza teórica e apoiado em revisão bibliográfica, analisa o viés de seleção de variedade imposto por essas aplicações e o ciclo vicioso que tende a padronizar uma variedade dominante em detrimento das demais. Partindo dos pressupostos da sociolinguística sobre a variação e a hierarquização das variedades linguísticas, examina-se o modo como os modelos de linguagem aprendem a partir de grandes volumes de texto, reproduzindo os vieses presentes nos dados e favorecendo as variedades mais documentadas. Argumenta-se que, ao serem treinados predominantemente sobre a variedade de prestígio e sobre as línguas de maiores recursos, os modelos tendem a naturalizar essa variedade como norma e a invisibilizar as demais, num processo que amplia, em escala tecnológica, o preconceito linguístico. Esse mecanismo constitui um ciclo vicioso: as variedades e línguas mais documentadas alimentam os modelos, que as reforçam, ao passo que as menos documentadas permanecem sub-representadas e tendem a desaparecer do horizonte tecnológico. Conclui-se que a mitigação dessa ameaça depende da articulação entre os estudos sociolinguísticos e o desenvolvimento tecnológico, orientada à documentação das variedades sub-representadas e à construção de aplicações linguisticamente mais inclusivas.

Palavras-chave: Sociolinguística. Inteligência Artificial. Variação Linguística.

ABSTRACT
The rapid diffusion of generative artificial intelligence applications based on large language models raises new questions for sociolinguistics and the preservation of linguistic diversity. This article, of a theoretical nature and supported by bibliographic review, analyzes the language variety selection bias imposed by these applications and the vicious cycle that tends to standardize a dominant variety at the expense of others. Starting from sociolinguistics assumptions about variation and the hierarchization of linguistic varieties, we examine how language models learn from large volumes of text, reproducing biases present in the data and favoring the most documented varieties. It is argued that, by being trained predominantly on the prestige variety and on languages with greater resources, the models tend to naturalize this variety as a norm and to render invisible the others, in a process that amplifies, at a technological scale, linguistic prejudice. This mechanism constitutes a vicious cycle: the most documented varieties and languages feed the models, which reinforce them, while the least documented remain under-represented and tend to disappear from the technological horizon. It is concluded that mitigating this threat depends on the articulation between sociolinguistic studies and technological development, oriented toward documenting under-represented varieties and building linguistically more inclusive applications.

Keywords: Sociolinguistics. Artificial Intelligence. Linguistic Variation.

RESUMEN
La rápida difusión de las aplicaciones de inteligencia artificial generativa basadas en grandes modelos lingüísticos plantea nuevas cuestiones para la sociolingüística y la preservación de la diversidad lingüística. Este artículo, de carácter teórico y respaldado por una revisión bibliográfica, analiza el sesgo de selección de variedades lingüísticas que imponen estas aplicaciones y el círculo vicioso que tiende a estandarizar una variedad dominante a expensas de otras. Partiendo de los supuestos sociolingüísticos sobre la variación y la jerarquización de las variedades lingüísticas, examinamos cómo los modelos lingüísticos aprenden a partir de grandes volúmenes de texto, reproduciendo los sesgos presentes en los datos y favoreciendo a las variedades mejor documentadas. Se argumenta que, al entrenarse predominantemente con la variedad de prestigio y con las lenguas que cuentan con mayores recursos, los modelos tienden a naturalizar esta variedad como norma y a invisibilizar las demás, en un proceso que amplifica, a escala tecnológica, los prejuicios lingüísticos. Este mecanismo constituye un círculo vicioso: las variedades y lenguas más documentadas alimentan a los modelos, que a su vez las refuerzan, mientras que las menos documentadas siguen estando infrarrepresentadas y tienden a desaparecer del horizonte tecnológico. Se concluye que la mitigación de esta amenaza depende de la articulación entre los estudios sociolingüísticos y el desarrollo tecnológico, orientados a documentar las variedades infrarrepresentadas y a crear aplicaciones lingüísticamente más inclusivas.

Palabras clave: Sociolingüística. Inteligencia Artificial. Variación Lingüística.

1 INTRODUÇÃO

A língua não é um sistema homogêneo, mas um fenômeno intrinsecamente variável. A sociolinguística demonstrou que toda língua se realiza por meio de variedades — modos distintos de falar associados a regiões, grupos sociais, faixas etárias e contextos —, e que essa variação é constitutiva da linguagem, e não um desvio a ser corrigido (LABOV, 2008). A diversidade linguística, nesse sentido, não é acidente, mas condição normal da existência das línguas, e sua preservação é parte da preservação da própria riqueza cultural das sociedades.

Essa variação, contudo, é socialmente hierarquizada. As sociedades atribuem prestígio desigual às diferentes variedades, elegendo uma delas como norma-padrão e desqualificando as demais como incorretas ou inferiores, processo que a sociolinguística brasileira analisou sob a designação de preconceito linguístico (BAGNO, 2015). A variedade de prestígio, associada aos grupos socialmente dominantes, é apresentada como a língua legítima, ao passo que as variedades populares e regionais são estigmatizadas, num movimento que converte diferença em desigualdade.

Sobre esse cenário incide, agora, uma nova força de grande alcance: a inteligência artificial generativa, baseada em grandes modelos de linguagem. Essas aplicações, treinadas sobre imensos volumes de texto, tornaram-se capazes de produzir linguagem com fluência notável, mas carregam consigo os vieses presentes nos dados de que se alimentam (BENDER *et al.*, 2021). Ao aprenderem a língua a partir do que está escrito e disponível, os modelos incorporam não apenas o vocabulário e a gramática, mas também as hierarquias e os preconceitos que estruturam os textos de treinamento, reproduzindo-os em escala.

Essa incorporação de vieses articula-se a uma profunda desigualdade na representação das línguas. Investigações sobre o estado da diversidade linguística no campo do processamento de linguagem natural evidenciaram que, das mais de sete mil línguas do mundo, apenas um número reduzido dispõe de recursos suficientes para figurar nas tecnologias de linguagem, o que produz enorme disparidade entre as línguas mais e menos documentadas (JOSHI *et al.*, 2020). O mesmo mecanismo que opera entre línguas opera entre variedades de uma mesma língua: as mais documentadas prevalecem, as menos documentadas desaparecem do horizonte tecnológico.

No caso do português, essa desigualdade é concreta. A própria língua portuguesa, apesar de seu grande número de falantes, é ainda tratada como língua de poucos recursos no desenvolvimento das tecnologias de linguagem, o que a coloca em desvantagem frente ao inglês e a outras línguas dominantes (FINGER, 2021). Se a língua portuguesa em seu conjunto ocupa posição desfavorável, as variedades populares e regionais do português brasileiro, ainda menos documentadas, encontram-se em situação de sub-representação ainda mais acentuada.

O problema que orienta este trabalho reside na articulação entre esses fenômenos. A hipótese que se sustenta é a de que as aplicações de inteligência artificial generativa impõem um viés de seleção de variedade, ao favorecerem a variedade dominante e mais documentada, e que esse viés alimenta um ciclo vicioso: a variedade documentada alimenta os modelos, que a reforçam como norma, enquanto as variedades sub-representadas permanecem à margem e tendem a desaparecer. Compreender esse mecanismo é condição para conceber respostas que preservem a diversidade linguística diante da tecnologia.

O objetivo geral deste artigo é analisar o viés de seleção de variedade imposto pelas aplicações de inteligência artificial generativa e o ciclo vicioso de padronização da variedade dominante que dele decorre. Como objetivos específicos, busca-se recuperar os pressupostos sociolinguísticos sobre a variação; examinar o modo como os modelos de linguagem aprendem e reproduzem vieses; caracterizar o viés de seleção de variedade; e discutir o ciclo vicioso e seus caminhos de mitigação. A relevância do tema decorre da rápida difusão dessas tecnologias e de seu potencial impacto sobre a diversidade linguística.

2 METODOLOGIA

O estudo caracteriza-se como pesquisa teórica de abordagem qualitativa, desenvolvida por meio de revisão bibliográfica, orientada pelos procedimentos de reunião, análise e síntese da produção sobre o tema (SOUZA; SILVA; CARVALHO, 2010). A revisão articulou a literatura sociolinguística sobre a variação e o preconceito linguístico à produção sobre vieses e diversidade linguística no campo do processamento de linguagem natural e da inteligência artificial.

O levantamento empregou, na busca, os termos: variação linguística, diversidade linguística, preconceito linguístico, grandes modelos de linguagem, viés algorítmico, línguas de poucos recursos e processamento de linguagem natural, nos idiomas português e inglês. Foram consultadas as bases SciELO, ACL Anthology e Google Acadêmico. A análise buscou articular os referenciais da sociolinguística aos da ciência da computação, de modo a compreender o fenômeno em sua dupla dimensão, linguística e tecnológica. As proposições foram fundamentadas na literatura examinada.

3 DESENVOLVIMENTO

3.1 VARIAÇÃO E DIVERSIDADE LINGUÍSTICA: PRESSUPOSTOS SOCIOLINGUÍSTICOS

A sociolinguística estabeleceu, como princípio fundamental, que a variação é inerente à língua. Os estudos sobre os padrões sociolinguísticos demonstraram que a variação não é aleatória, mas estruturada, correlacionando-se sistematicamente a fatores sociais como a classe, a idade, o gênero e

o contexto, de modo que cada variedade constitui um sistema linguístico plenamente funcional (LABOV, 2008). Reconhecer a variação como estruturada e funcional é a base para recusar a hierarquização que apresenta uma variedade como superior às demais, pois todas cumprem, em seus contextos, as funções da linguagem.

No português brasileiro, essa variação é particularmente rica, e coexistem normas distintas — a norma-padrão idealizada, as normas cultas efetivamente praticadas e as normas populares —, cuja hierarquização reflete as desigualdades sociais (BORTONI-RICARDO, 2004). A escola, ao tomar a norma-padrão como única língua legítima, tende a desqualificar as variedades populares trazidas pelos estudantes, reproduzindo, no plano linguístico, as assimetrias sociais que estruturam a sociedade.

Essa hierarquização constitui o que se designou preconceito linguístico: a crença, socialmente disseminada, de que existe uma única forma correta de falar, em relação à qual todas as demais seriam erros. Esse preconceito, longe de ser questão meramente linguística, é forma de discriminação social, que estigmatiza os falantes das variedades desprestigiadas (BAGNO, 2015). O combate ao preconceito linguístico, empreendido pela sociolinguística, sustenta-se no reconhecimento de que todas as variedades são legítimas e de que a escolha da norma-padrão é decisão social, e não linguística. É esse patrimônio conceitual que a difusão da inteligência artificial generativa coloca sob nova tensão.

3.2 COMO OS MODELOS DE LINGUAGEM APRENDEM E REPRODUZEM VIESES

A compreensão do impacto da inteligência artificial sobre a diversidade linguística exige entender como os modelos de linguagem aprendem. Os grandes modelos são treinados sobre enormes volumes de texto, a partir dos quais extraem padrões estatísticos que lhes permitem prever e gerar sequências linguísticas. Nesse processo, os modelos capturam não apenas a estrutura da língua, mas os vieses implícitos presentes nos dados de treinamento (BENDER *et al.*, 2021). Um modelo que aprende a partir de textos enviesados reproduzirá esses vieses, não por intenção, mas por construção: ele reflete, de volta, os padrões que encontrou nos dados que o alimentaram.

Exemplos concretos ilustram esse mecanismo. Sistemas de representação de palavras exibiram estereótipos de gênero ao associar determinadas profissões a homens e outras a mulheres, refletindo os padrões culturais presentes no corpus de treinamento, e não qualquer regra linguística (FINGER, 2021). O mesmo princípio que reproduz estereótipos de gênero reproduz hierarquias linguísticas: se os textos de treinamento privilegiam a variedade de prestígio, o modelo aprenderá a tratá-la como a forma normal e correta, relegando as demais à condição de anomalia.

Essa reprodução de vieses é agravada pela escala e pela opacidade dos modelos. A literatura crítica adverte que o próprio tamanho dos grandes modelos, treinados sobre volumes cada vez maiores

de dados não curados, dificulta a identificação e o controle dos vieses que incorporam, e questiona os custos e os riscos dessa expansão indiscriminada (BENDER *et al.*, 2021). A dificuldade de auditar o que o modelo aprendeu e por que produz determinadas respostas torna os vieses linguísticos particularmente difíceis de detectar e corrigir, o que os torna mais insidiosos.

3.3 O VIÉS DE SELEÇÃO DE VARIEDADE E A PADRONIZAÇÃO DA VARIEDADE DOMINANTE

Do funcionamento dos modelos decorre o que se pode denominar viés de seleção de variedade. Como os modelos aprendem a partir do que está documentado e disponível em maior quantidade, eles tendem a privilegiar a variedade dominante — aquela associada aos grupos de prestígio, mais presente na escrita formal e nos textos que compõem os corpora de treinamento (JOSHI *et al.*, 2020). As variedades populares e regionais, menos representadas na escrita e nos dados, comparecem em proporção muito menor, quando comparecem, o que faz com que o modelo aprenda predominantemente a variedade dominante e a tome como referência.

Esse viés produz um efeito de padronização. Ao gerar linguagem, o modelo tende a reproduzir a variedade sobre a qual foi majoritariamente treinado, apresentando-a como a forma natural e correta de se expressar, e corrigindo ou desconsiderando as formas que dela se afastam (BAGNO, 2015). A tecnologia converte-se, assim, em agente de normatização: ao sugerir, corrigir e gerar texto na variedade dominante, as aplicações reforçam, em cada interação, a ideia de que essa é a língua correta, ampliando o alcance do preconceito linguístico para muito além da escola e da gramática tradicional.

O alcance dessa padronização é sem precedentes. Diferentemente das instâncias tradicionais de normatização, restritas a determinados espaços, as aplicações de inteligência artificial generativa difundem-se por incontáveis interações cotidianas, mediando a escrita de milhões de pessoas (FINGER, 2021). Essa ubiquidade confere ao viés de seleção de variedade uma capacidade de influência sobre os usos linguísticos que nenhuma instituição normatizadora anterior possuiu, o que amplia a preocupação com seus efeitos sobre a diversidade.

3.4 O CICLO VICIOSO: DOCUMENTAÇÃO, DADOS E INVISIBILIZAÇÃO

O aspecto mais preocupante do fenômeno é seu caráter autorreforçador, que configura um verdadeiro ciclo vicioso. As línguas e variedades mais documentadas dispõem de mais dados para alimentar os modelos, o que lhes garante melhor desempenho tecnológico; esse melhor desempenho estimula seu uso e sua documentação, ampliando ainda mais os dados disponíveis (JOSHI *et al.*, 2020). As línguas e variedades menos documentadas, ao contrário, dispõem de menos dados, obtêm pior

desempenho e são menos utilizadas nas tecnologias, o que reduz ainda mais sua documentação, num círculo que aprofunda progressivamente a desigualdade.

Esse ciclo tende a converter a desigualdade de recursos em desaparecimento tecnológico. As variedades sub-representadas, ao não figurarem nas aplicações que mediam cada vez mais a comunicação, tornam-se invisíveis no ambiente digital, o que compromete não apenas seu processamento pela tecnologia, mas sua própria presença no espaço público (FINGER, 2021). A ausência de dados, que começa como limitação técnica, converte-se em invisibilização cultural, à medida que as variedades ausentes das tecnologias deixam de ser reconhecidas como legítimas por gerações crescentemente imersas no ambiente digital.

O ciclo vicioso opera, assim, uma seleção que não é linguística, mas tecnológica e social. Não é a qualidade ou a expressividade de uma variedade que determina sua permanência, mas sua documentação e sua disponibilidade como dado — critério alheio ao valor linguístico e que reproduz, no plano tecnológico, as hierarquias sociais que a sociolinguística denunciou (BAGNO, 2015). A variedade dominante prevalece não por ser melhor, mas por estar mais documentada, o que revela a natureza social, e não natural, da seleção que a tecnologia opera e naturaliza.

3.5 IMPLICAÇÕES E CAMINHOS DE MITIGAÇÃO

O enfrentamento dessa ameaça exige, antes de tudo, o reconhecimento de que a inclusão linguística nas tecnologias não é automática, mas depende de esforço deliberado. Investigações sobre a diversidade linguística no processamento de linguagem natural apontam a necessidade de investir na documentação e no desenvolvimento de recursos para as línguas e variedades sub-representadas, sem o que a disparidade tende a se perpetuar (JOSHI *et al.*, 2020). A inclusão, nesse sentido, não resulta do simples avanço tecnológico, que tende a reforçar as desigualdades existentes, mas de uma escolha consciente de documentar e representar a diversidade.

No caso do português, esse esforço já se manifesta em iniciativas de construção de corpora e de recursos que buscam superar a condição de língua de poucos recursos, empreendidas por centros de pesquisa dedicados ao processamento do português brasileiro (FINGER, 2021). A ampliação e a diversificação desses recursos, de modo a contemplar não apenas a variedade de prestígio, mas as variedades populares e regionais, é condição para que a tecnologia represente a riqueza real da língua, e não apenas sua face dominante.

A contribuição da sociolinguística é, nesse ponto, decisiva. O conhecimento acumulado sobre a variação e sobre o preconceito linguístico pode orientar o desenvolvimento de tecnologias mais inclusivas, ao alertar para os vieses e ao fornecer os fundamentos para reconhecer e valorizar a

diversidade, evitando que a tecnologia reproduza acriticamente as hierarquias sociais (BORTONI-RICARDO, 2004). A articulação entre a sociolinguística e a ciência da computação desponta, assim, como caminho para conciliar o avanço tecnológico com a preservação da diversidade linguística, colocando o conhecimento sobre a língua a serviço de uma tecnologia que a respeite em sua pluralidade.

4 DISCUSSÃO

A análise desenvolvida permite articular convergências, tensões e implicações. A convergência fundamental é o reconhecimento de que a tecnologia não é neutra em relação à língua: ao aprender a partir de dados socialmente marcados, ela reproduz e amplia as hierarquias linguísticas, o que aproxima a crítica sociolinguística da crítica aos vieses da inteligência artificial.

A principal tensão opõe a promessa de acesso à ameaça de homogeneização. As aplicações de inteligência artificial generativa ampliam o acesso à informação e à produção de texto, mas ao fazê-lo majoritariamente na variedade dominante, arriscam homogeneizar os usos linguísticos e reduzir a diversidade que constitui a riqueza das línguas (BENDER *et al.*, 2021). Superar essa tensão exige que o desenvolvimento tecnológico incorpore, desde a concepção, a preocupação com a diversidade, em vez de tratá-la como correção posterior de um viés já instalado.

As lacunas identificadas são de duas ordens. A primeira é a escassez de estudos que documentem e quantifiquem o viés de seleção de variedade especificamente no português brasileiro e em suas variedades populares e regionais (FINGER, 2021). A segunda é a insuficiente integração entre a pesquisa sociolinguística e o desenvolvimento tecnológico, que ainda operam, em larga medida, em campos separados, quando o enfrentamento do problema exige precisamente sua articulação.

Do ponto de vista das implicações, a análise sugere que a preservação da diversidade linguística na era da inteligência artificial não será resultado espontâneo do avanço técnico, mas dependerá de escolhas deliberadas de documentação, de representação e de desenho das aplicações. Reconhecer o viés de seleção de variedade e o ciclo vicioso que ele alimenta é o primeiro passo para interrompê-los, substituindo a padronização automática pela valorização consciente da pluralidade linguística.

5 CONCLUSÃO

A análise realizada permite responder ao objetivo central: as aplicações de inteligência artificial generativa impõem um viés de seleção de variedade que favorece a variedade dominante e mais documentada, alimentando um ciclo vicioso que padroniza essa variedade e tende a invisibilizar as demais. Esse mecanismo amplia, em escala tecnológica e com alcance inédito, o preconceito

linguístico que a sociolinguística havia denunciado, convertendo a tecnologia em agente de normatização.

Retomando os objetivos específicos, recuperaram-se os pressupostos sociolinguísticos sobre a variação como fenômeno estruturado e legítimo; examinou-se como os modelos de linguagem aprendem e reproduzem os vieses presentes nos dados; caracterizou-se o viés de seleção de variedade e seu efeito de padronização; e analisou-se o ciclo vicioso, no qual a documentação determina a permanência tecnológica das variedades. Ao longo do percurso, sustentou-se que a seleção operada pela tecnologia é social, e não linguística, reproduzindo as hierarquias que a sociolinguística questiona.

Reconhecem-se as limitações de um trabalho teórico, cuja discussão se beneficiaria de estudos empíricos que documentem o viés de seleção de variedade no português e suas consequências. Delineiam-se, assim, perspectivas para investigações futuras: a documentação e a quantificação do viés nas variedades do português brasileiro; o desenvolvimento de corpora que contemplem a diversidade; e a construção de metodologias de avaliação da inclusão linguística nas aplicações. Em síntese, a diversidade linguística encontra-se sob ameaça da inteligência artificial generativa, e sua preservação dependerá da articulação entre a sociolinguística e a tecnologia, orientada a fazer das aplicações instrumentos de valorização, e não de apagamento, da pluralidade das línguas.

REFERÊNCIAS

BAGNO, Marcos. Preconceito linguístico: o que é, como se faz. São Paulo: Parábola, 2015.

BENDER, E. M.; GEBRU, T.; McMILLAN-MAJOR, A.; SHMITCHELL, S. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Nova York: ACM, 2021. p. 610-623. DOI: 10.1145/3442188.3445922.

BORTONI-RICARDO, Stella Maris. Educação em língua materna: a sociolinguística na sala de aula. São Paulo: Parábola, 2004.

FINGER, Marcelo. Inteligência artificial e os rumos do processamento do português brasileiro. Estudos Avançados, v. 35, n. 101, p. 51-64, 2021. DOI: 10.1590/s0103-4014.2021.35101.005.

JOSHI, P.; SANTY, S.; BUDHIRAJA, A.; BALI, K.; CHOUDHURY, M. The state and fate of linguistic diversity and inclusion in the NLP world. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020. p. 6282-6293. DOI: 10.18653/v1/2020.acl-main.560.

LABOV, William. Padrões sociolinguísticos. Tradução de Marcos Bagno, Maria Marta Pereira Scherre e Caroline Rodrigues Cardoso. São Paulo: Parábola, 2008.

SOUZA, M. T.; SILVA, M. D.; CARVALHO, R. Revisão integrativa: o que é e como fazer. Einstein (São Paulo), v. 8, n. 1, p. 102-106, 2010. DOI: 10.1590/S1679-45082010RW1134.