Diversidade linguística e inclusão digital: desafios para uma IA brasileira

Raquel Meister Ko. Freitag [1]

[1]Departamento de Letras Vernáculas (DLEV)
Laboratório Multiusuário de Informática e Documentação Linguística (LAMID)
Universidade Federal de Sergipe (UFS)
Avenida Marcelo Deda Chagas, s/n – 49107-230 – São Cristóvão – SE – Brazil

rkofreitag@academico.ufs.br

Abstract. *The paper explores how linguistic diversity, a fundamental human attribute, is threatened by generative AI. It highlights how technological applications often favor certain language varieties due to selection biases, creating a cycle where dominant, well-documented languages are preserved and standardized, as they provide the data needed for training large language models.*

Resumo. *O artigo explora como a diversidade linguística, uma característica humana fundamental, está sendo ameaçada pela IA generativa. É destaca do como as aplicações tecnológicas favorecem certas variedades linguísticas devido a vieses de seleção, criando um ciclo em que línguas dominantes e bem documentadas são preservadas e padronizadas, pois fornecem os dados necessários para o treinamento de modelos de língua em larga escala.*

1. Introdução

O Plano Brasileiro de Inteligência Artificial 2024-2028 inclui como um dos objetivos ”desenvolver modelos avançados de linguagem em português, com dados nacionais que abarcam nossa diversidade cultural, social e linguística, para fortalecer a soberania em IA.” [MCTI 2024].

Para cumprir este objetivo, o mito do monolinguismo do português precisa ser desfeito. Em seguida, o português falado no Brasil é apresentado sob a perspectiva da diversidade e a tensão entre variedades de prestígio e variedades ditas ”não-padrão” que divide a sociedade. Após essa contextualização sociolinguística, são apresentadas recomendações para a constituição de amostras linguísticas brasileiras para treinar LLMs, de modo a garantir a diversidade cultural, social e linguística prevista na proposta do PBIA.

2. No Brasil, não se fala só português

A Constituição de 1988 reconhece, no Art. 13., que ”A língua portuguesa é o idioma oficial da República Federativa do Brasil.” [Constituição 1988]. O objetivo do *Plano Brasileiro de Inteligência Artificial 2024-2028* reflete o dispositivo legal. No entanto, não é apenas português que se fala no Brasil. A existência de outras línguas, embora empírica e legalmente reconhecidas, não faz parte do imaginário da nação, que se molda por uma ideologia monolíngue – a de que aqui todos falamos português – que se reproduz nos LLMs, na medida que somente o português é reconhecido como língua de soberania nacional no documento norterador.

Na própria constituição, há pistas da diversidade linguística, como no § 2º do Art. 210, que garante que ”O ensino fundamental regular será ministrado em língua portuguesa, assegurada às comunidades indígenas também a utilização de suas línguas maternas e processos próprios de aprendizagem.”, ou, ainda mais longe, no Art. 231., que diz que ”São reconhecidos aos índios sua organização social, costumes, línguas, crenças e tradições, e os direitos originários sobre as terras que tradicionalmente ocupam, competindo à União demarcá-las, proteger e fazer respeitar todos os seus bens.” Mesmo status de reconhecimento tem a Libras. O art. 1º da Lei 10.436/2002 diz que “É reconhecida como meio legal de comunicação e expressão a Língua Brasileira de Sinais - Libras e outros recursos de expressão a ela associados.” [Brasil 2002].

A co-oficialização é outro processo que reconhece legalmente as línguas. As primeiras línguas co-oficializadas foram três línguas indígenas faladas no município de São Gabriel da Cachoeira, estado do Amapá: Tukano, Baniwa e Nheengatu. Desde então, já são 23 línguas cooficializadas no país, sendo 13 línguas indígenas e 9 e imigração [Freitag and Savedra 2023].

O Inventário Nacional da Diversidade Linguística (INDL) tem atuado na “identificação, documentação, reconhecimento e valorização das línguas portadoras de referência à identidade, à ação e à memória dos diferentes grupos formadores da sociedade brasileira” [Brasil 2010]. As línguas do Brasil, no escopo do INDL, são de seis grupos: indígenas, comunidades afro-brasileiras, imigração, sinais, crioulas e a Língua Portuguesa e suas variações dialetais. Já foram reconhecidas como *Referência Cultural* cinco línguas de base indígena (duas línguas do tronco Tupi, Asurini e Guarani M’bya, três línguas da família Karib (Nahukuá, Matipu e Kuikuro Kalapalo), duas línguas de contato (Talian e Portunhol) e uma língua geral Nheengatu [Freitag and Savedra 2023].

Além da informação de base legal sobre a existência de línguas, estudos linguísticos identificam e documentam outras tantas, de modo que não há consenso sobre quantas línguas são faladas no Brasil, nem quantas pessoas falam cada uma dessas línguas. Há, no entanto, consenso de que no Brasil não se fala apenas português, e uma política para a soberania nacional não deve ignorar a diversidade linguística, sob pena não só de excluir os povos originários, como também de excluir a identidade de uma população socialmente diversa.

LLMs para uma IA de soberania nacional precisam considerar a diversidade de línguas do Brasil, e não apenas eleger o português como língua de treino. E, mesmo dentro do português, há diversidade que reflete padrões sociais e culturais da realidade brasileira, que, como veremos na sequência, precisam ser considerados.

3. O português falado no Brasil é diverso

Seja como uma das línguas com o maior número de falantes ou como uma língua com o maior número de países onde é falado, o português aparece nos ranqueamentos de línguas do mundo. O português não é apenas falado em Portugal e no Brasil [Freitag 2022]. Não há um Português, há variedades de português, e cada uma destas variedades é polarizada em um centro, o que o configura o português como uma língua pluricêntrica.

O pluricentrismo do português é reconhecido nas ações de inclusão digital: é frequente encontrar documentação de *software* nas duas variedades hegemônicas do português (Português Europeu e Português Brasileiro) [Azevedo et al. 2021]. E, mesmo no

Brasil, as especifidades de cada uma das comunidades que têm o Português como sua língua refletem seus valores socioculturais e diferenciam as variedades, o que tem sido amplamente demostrado pela sociolinguística brasileira.

A diversidade do português brasileiro é reconhecida no INDL – Língua Portuguesa e suas variações dialetais – e também é alçada a direito de aprendizagem na Base Nacional Comum Curricular [Brasil 2018]: "Compreender as línguas como fenômeno (geo)político, histórico, cultural, social, variável, heterogêneo e sensível aos contextos de uso, reconhecendo suas variedades e vivenciando-as como formas de expressões identitárias, pessoais e coletivas, bem como agindo no enfrentamento de preconceitos de qualquer natureza".

Assim, para a soberania nacional, uma IA brasileira não pode se limitar a uma única língua, o português, nem a uma única variedade do português. O viés de seleção de uma única língua/variedade reforça e acentua ainda mais os preconceitos, em especial contra às variedades linguísticas subrepresentadas.

4. Recomendações para o desenvolvimento de uma IA brasileira linguisticamente diversificada

Para cumprir o objetivo do PBIA, é necessário não só a intensificação de ações de documentação linguística, mas também a conscientização de desenvolvedores de que as aplicações das tecnologias precisam refletir valores linguísticos do grupo, sob pena de reforçar ainda mais o preconceito que já existe em relação às variedades linguísticas subrepresentadas.

Ao longo de 50 anos, a sociolinguística brasileira acumulou vasto acervo de documentação linguística, oriundo de pesquisas de campo que subsidiam teses e dissertações. Esses dados autênticos, especialmente os transcritos e anotados, são valiosos para tecnologias de linguagem e inteligência artificial [Freitag et al. 2012, Freitag et al. 2021, Freitag et al. 2021, Machado Vieira et al. 2021a, Machado Vieira et al. 2021b, Sousa and Freitag 2024]. Contudo, esses acervos são armazenados de forma assistemática, sem protocolos específicos para compartilhamento e reuso. Como solução, propõe-se a criação da **Plataforma da Diversidade Linguística Brasileira** [Machado Vieira et al. 2021c], um repositório especializado para preservar e compartilhar essas coleções.

A Plataforma da Diversidade Linguística Brasileira, articulada pela Comissão de Sociolinguística da ABRALIN e o GT de Sociolinguística da ANPOLL, é um projeto nacional alinhado à estratégia do PBIA, ao catalogar e armazenar coleções de dados sociolinguísticos que podem subsidiar o treino de LLMs para uma IA brasileira.

Uma IA eticamente sensível para a soberania nacional requer que a diversidade linguística seja considerada de maneira plena equinâme, 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.

References

Azevedo, I. C. M., Abreu, R. N., and Freitag, R. M. K. (2021). Desafios do português brasileiro como língua adicional para a cidadania global. *Linguagem & Ensino*, 24(2):263–288.

Brasil (2002). Lei nº 10.436, de 24 de abril de 2002. dispõe sobre a língua brasileira de sinais - libras e dá outras providências.

Brasil (2010). Decreto nº 7.387, de 9 de dezembro de 2010. institui o inventário nacional da diversidade linguística e dá outras providências.

Brasil (2018). *Base Nacional Comum Curricular*. Ministério da Educação.

Constituição (1988). *Constituição da República Federativa do Brasil de 1988*. Assembleia Nacional Constituinte.

Freitag, R. M. K. (2022). Sociolinguistic repositories as asset: challenges and difficulties in brazil. *The Electronic Library*, 40(5):607–622.

Freitag, R. M. K., Martins, M. A., and Tavares, M. A. (2012). Bancos de dados sociolinguísticos do português brasileiro e os estudos de terceira onda: potencialidades e limitações. *Alfa: Revista de Linguística*, 56:917–944.

Freitag, R. M. K., Martins, M. A. R., Araújo, A., Battisti, E., Coelho, I. M. W. d. S., Sousa, M. D. A. F., Silva, R. G. d., and Lopes, R. E. d. L. (2021). Desafios da gestão de dados linguísticos e a ciência aberta. *Cadernos de linguística*, 2(1):1–19.

Freitag, R. M. K. and Savedra, M. M. G. (2023). Contatos, mobilidades e línguas no brasil. In *Mobilidades e Contatos Linguísticos no Brasil*, pages 13–26. Blucher Open Access.

Machado Vieira, M. d. S., Barbosa, J., Freitag, R., Borges, M., and Medeiros, A. (2021a). Collections of data open to society: linguistic and sociocultural memory and potential for (re) use. *Cadernos de Linguística*, 2(1):e607.

Machado Vieira, M. d. S., Wiedemer, M. L., Frreitag, R. M. K., and Barbosa, J. B. (2021b). Mapeamento de bancos de dados (socio) linguísticos no brasil. *Projeto desenvolvido pelo GT de Sociolinguística da ANPOLL e pela Comissão da Área de Sociolinguística da ABRALIN*.

Machado Vieira, M. d. S., Wiedemer, M. L., Frreitag, R. M. K., Barbosa, J. B., Peres, E. P., and Mollica, M. C. d. M. M. (2021c). Plataforma da diversidade linguística brasileira. *Projeto apresentado à Pró-Reitoria de Pós-Graduação e Pesquisa da UFRJ e à Fundação Universitária José Bonifácio, em razão do Edital BNDES-Chamada Pública para seleção de propostas no âmbito da iniciativa Resgatando a História*.

MCTI (2024). *IA para o Bem de Todos*. Ministério da Ciência, Tecnologia e Inovação.

Sousa, M. D. A. F. and Freitag, R. M. K. (2024). Bancos de dados sociolinguísticos e a ciência aberta: compartilhamento de dados e conhecimentos. *Revista Diálogos*, 12(1):165–187.