DIVERSIDADE LINGUÍSTICA E INCLUSÃO DIGITAL: DESAFIOS PARA UMA IA BRASILEIRA

A PREPRINT

Raquel Meister Ko. Freitag

Departamento de Letras Vernáculas Universidade Federal de Sergipe - UFS rkofreitag@academico.ufs.br

ABSTRACT

Linguistic diversity is a human attribute which, with the advance of generative AIs, is coming under threat. This paper, based on the contributions of sociolinguistics, examines the consequences of the variety selection bias imposed by technological applications and the vicious circle of preserving a variety that becomes dominant and standardized because it has linguistic documentation to feed the large language models for machine learning.

Keywords Ethics · Sociolinguistics · Brazilian AI

1 Introdução

No documento do Ministério da Ciência, Tecnologia e Inovação *IA para o Bem de Todos*, em que é apresentada a proposta de um *Plano Brasileiro de Inteligência Artificial 2024-2028*, um dos cinco objetivos listados é "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].

A Sociolinguística é o campo da ciência que estuda as relações entre língua e sociedade, e o conjunto de trabalhos neste campo desenvolvidos no Brasil nos últimos 50 anos tem contribuições diretas para a consecução deste objetivo Freitag [2016]. E é sob esta perspectiva que este objetivo é discutido neste texto. Pensando em uma IA ética e socialmente sensível, a diversidade das comunidades na sociedade se reflete também (ou, pelo menos, deveria se refletir) na diversidade das comunidades em amostras linguísticas para treinar modelos de língua em larga escala (LLMS).

Uma IA ética precisa atender aos princípios de justiça, equidade, diversidade e inclusão, e no domínio linguístico, por meio da seleção das amostras de línguas e variedades de línguas que vão compor o corpus de treno dos modelos, assimetrias se acentuam, desde a exclusão ou apagamento de línguas, até a priorização de uma variedade – dita de prestígio – face às variedades consideradas não-padrão ou estigmatizadas. Os preconceitos decorrentes dessa hieraquização de variedades são reproduzidos em LLMs e geram respostas Shrawgi et al. [2024], Fleisig et al. [2024], Freitag and de Gois [2024], como já constatado no inglês afro-americano Mengesha et al. [2021], Dacon and Tang [2021], Dacon et al. [2022].

Considerando o objetivo do *Plano Brasileiro de Inteligência Artificial 2024-2028* que trata de diversidade linguística, primeiramente, 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 *Plano Brasileiro de Inteligência Artificial*.

2 No Brasil, não se fala só português

Sobre língua, 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, bem mais distante, 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].

E, pela vertente da patrimonialização, reconhecimento e valorização da diversidade linguística brasileira, o Instituto do Patrimônio Histórico e Artístico Nacional (Iphan), por meio do 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.

Modelos de língua em larga escala 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 especificadades 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 Roncarati et al. [2003], Abraçado and Martins [2015].

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".

Diversidade linguística e inclusão digital: desafios para uma IA brasileira

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 *Plano Brasileiro de Inteligência Artificial 2024-2028*, é necessário não só a intensificação de ações de documentação linguística de variedades subrepresentadas, mas a conscientização dos desenvolvedores de que as aplicações das tecnologias precisam refletir valores linguísticos e a representatividade para o grupo, sob pena de reforçar ainda mais o preconceito que já existe em relação às variedades linguísticas subrepresentadas.

Nos seus 50 anos de trajetória, além das descrições linguísticas que caracterizam cientificamente a diferença entre a variedade brasileira e europeia do português, os estudos sociolinguísticos brasileiros vêm acumulando como produto primário um expressivo acerco de documentação linguística Freitag et al. [2012, 2021,?], Machado Vieira et al. [2021a,b], Sousa and Freitag [2024]. Resultado de pesquisa de campo para subsidiar teses e dissertações, estas ações de documentação linguística resultam em produtos, e como tais, têm custo e valor. Dados linguísticos autênticos, especialmente dados transcritos e anotados, têm aplicação em diversas áreas das tecnologias de linguagem e inteligência artificial. Atualmente estes acervos são armazenados de maneira assistemática e provisória, sem protocololos específicos para compartilhamento e reuso. Embora as instituições de pesquisa tenham repositórios para compartilhamento de produção científica (teses, dissertações, etc.), estes não são apropriados para compartilhar coleções de dados linguísticos. A solução para este problema é a construção de um repositório próprio específico para este tipo de acervo, em uma iniciativa denominada **Plataforma da Diversidade Linguística Brasileira** Machado Vieira et al. [2021c].

Com a articulação da Comissão de Sociolinguística da Associação Brasileira de Linguística (ABRALIN) e do GT de Sociolinguística da Associação Nacional de Pós-Graduação em Letras e Linguística (ANPOLL), a **Plataforma da Diversidade Linguística Brasileira** configura-se como um projeto nacional alinhado à visão estratégica para o desenvolvimento sustentável da área de IA, como destadado no objetivo da proposta do *Plano Brasileiro de Inteligência Artificial 2024-2028*, e que demanda financiamento e institucionalização de um repositório nacional que oportunizem à sociedade a diversidade do patrimônio linguístico que vem sendo registrado e mapeado. A plataforma visa a catalogação nacional (salvaguarda e difusão) e a constituição de um repositório comum, com padrões de metadados e diretrizes de armazenamento de de coleções de dados sociolinguísticos Sousa and Freitag [2024], de modo a atender necessidades tanto do público amplo, para que possibilite a qualquer interessado ver, ouvir, repetir diferentes manifestações de uso linguístico no país, como para público especializado, tal como a alimentação 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

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

Raquel Meister Ko Freitag. Sociolinguística no/do brasil. Cadernos de Estudos Linguísticos, 58(3):445-460, 2016.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1841–1857, 2024.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. Linguistic bias in chatgpt: Language models reinforce dialect discrimination. *arXiv preprint arXiv:2406.08818*, 2024.

Raquel Meister Ko Freitag and Túlio Sousa de Gois. Performance in a dialectal profiling task of llms for varieties of brazilian portuguese. *arXiv preprint arXiv:2410.10991*, 2024.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. "i don't think these devices are very culturally sensitive." — impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence*, 4:725911, 2021.

Jamell Dacon and Jiliang Tang. What truly matters? using linguistic cues for analyzing the# blacklivesmatter movement and its counter protests: 2013 to 2020. *arXiv preprint arXiv:2109.12192*, 2021.

Diversidade linguística e inclusão digital: desafios para uma IA brasileira

- Jamell Dacon, Haochen Liu, and Jiliang Tang. Evaluating and mitigating inherent linguistic bias of african american english through inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1442–1454, 2022.
- Constituição. Constituição da República Federativa do Brasil de 1988. Assembleia Nacional Constituinte, 1988. URL https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm.
- Brasil. Lei nº 10.436, de 24 de abril de 2002. dispõe sobre a língua brasileira de sinais libras e dá outras providências. Diário Oficial [da] República Federativa do Brasil, 2002. ISSN 1677-7042. URL http://www.planalto.gov.br/ccivil_03/leis/2002/110436.htm.
- Raquel Meister Ko Freitag and Mônica Maria Guimarães Savedra. Contatos, mobilidades e línguas no brasil. In *Mobilidades e Contatos Linguísticos no Brasil*, pages 13–26. Blucher Open Access, 2023.
- 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] República Federativa do Brasil*, 2010. ISSN 1677-7042. URL https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2010/decreto/d7387.htm.
- Raquel Meister Ko. Freitag. Sociolinguistic repositories as asset: challenges and difficulties in brazil. *The Electronic Library*, 40(5):607–622, 2022.
- Isabel Cristina Michelan Azevedo, Ricardo Nascimento Abreu, and Raquel Meister Ko Freitag. Desafios do português brasileiro como língua adicional para a cidadania global. *Revista Linguagem & Ensino*, 24(2):263–288, 2021.
- Cláudia Roncarati, Jussara Abraçado, and Jürgen B Heye. *Português brasileiro: contato lingüístico, heterogeneidade e história*, volume 2. 7 Letras, 2003.
- Jussara Abraçado and Marco Antonio Martins. *Mapeamento sociolinguístico do português brasileiro*. Editora Contexto, 2015.
- Brasil. Base Nacional Comum Curricular. Ministério da Educação, 2018.
- Raquel Meister Ko Freitag, Marco Antonio Martins, and Maria Alice Tavares. Bancos de dados sociolinguísticos do português brasileiro e os estudos de terceira onda: potencialidades e limitações. *Alfa: Revista de Linguística (São José do Rio Preto)*, 56:917–944, 2012.
- Raquel Meister Ko Freitag, Marco Antonio Rocha Martins, Aluiza Araújo, Elisa Battisti, Iandra Maria Weirich da Silva Coelho, Marta Deysiane Alves Faria Sousa, Raimundo Gouveia da Silva, and Rodrigo Esteves de Lima Lopes. Desafios da gestão de dados linguísticos e a ciência aberta. *Cadernos de linguística. Campinas, SP. Vol. 2, n. 1 (jan. 2021), p. 1-19*, 2021.
- M dos S Machado Vieira, JB Barbosa, RMK Freitag, MM Borges, and ALS Medeiros. Collections of data open to society: linguistic and sociocultural memory and potential for (re) use. *Cadernos de Linguística*, 2(1):e607, 2021a.
- Marcia dos Santos Machado Vieira, Marcos Luiz Wiedemer, Raquel Meister Ko Frreitag, and Juliana Bertucci Barbosa. 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*, 2021b.
- Marta Deysiane Alves Faria Sousa and Raquel Meister Ko Freitag. Bancos de dados sociolinguísticos e a ciência aberta: compartilhamento de dados e conhecimentos. *Revista Diálogos*, 12(1):165–187, 2024.
- Marcia dos Santos Machado Vieira, Marcos Luiz Wiedemer, Raquel Meister Ko Frreitag, Juliana Bertucci Barbosa, Edenize Ponzo Peres, and Maria Cecília de Magalhães Mollica Mollica. 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, 2021c.