APRESENTAÇÃO

COMPARTILHAMENTO DE DADOS LINGUÍSTICOS: DA PRÁTICA À INFRAESTRUTURA

Juliana Bertucci BARBOSA
Departamento de Linguística e Língua Portuguesa - Universidade Federal do Triângulo Mineiro (UFTM)
Uberaba, MG, Brasil

Marcia dos Santos MACHADO VIEIRA
Departamento de Letras Vernáculas - Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, RJ, Brasil

Raquel Meister Ko. FREITAG
Departamento de Letras Vernáculas - Universidade Federal de Sergipe (UFS)
São Cristóvão, SE, Brasil

RESUMO

Este dossiê reúne trabalhos originados do *Abralin em Cena 17 - Dados Linguísticos* (2023), evento online promovido pela ABRALIN, pelo GT de Sociolinguística da ANPOLL e pelo Projeto Plataforma da Diversidade Linguística Brasileira. Os textos abordam compartilhamento, preservação e reutilização de corpora linguísticos; metodologia de metadados e comparabilidade entre bancos de dados; e processamento automático de linguagem. Os artigos integram e exemplificam a cultura de Ciência Aberta que motivou o evento, e convergem para demonstrar a necessidade de uma infraestrutura nacional de dados linguísticos para a soberania digital do Brasil.

PALAVRAS-CHAVE

Dados Linguísticos; Ciência Aberta; Repositórios; Compartilhamento; Diversidade Linguística Brasileira.

EDITORES
– Miguel Oliveira, Jr. (UFAL)
– René Almeida (UFRB)

AVALIADORES
– Miguel Oliveira, Jr. (UFAL)
– René Almeida (UFRB)

COMO CITAR
BARBOSA, J. B.; MACHADO VIEIRA, M. S.; FREITAG, R. M. K. (2026). Compartilhamento de dados linguísticos: da prática à infraestrutura. *Cadernos de Linguística*, v. 7, n. 2, e1012.

TITLE

LINGUISTIC DATA SHARING: FROM PRACTICE TO INFRASTRUCTURE

ABSTRACT

This dossier brings together works originated from *Abralin em Cena 17 – Linguistic Data* (2023), an online event organized by ABRALIN, the Sociolinguistics Working Group of ANPOLL, and the Brazilian Linguistic Diversity Platform Project. The texts address sharing, preservation, and reuse of linguistic corpora; metadata methodology and comparability across databases; and natural language processing. The papers both integrate and exemplify the Open Science culture that motivated the event, converging to demonstrate the need for a national linguistic data infrastructure for Brazil's digital sovereignty.

KEYWORDS

Linguistic Data; Open Science; Repositories; Data Sharing;
Brazilian Linguistic Diversity.

APRESENTAÇÃO

O *Abralin em Cena* é um evento científico itinerante, organizado pela Associação Brasileira de Linguística (ABRALIN) desde 2008, cujo objetivo é promover o intercâmbio acadêmico, dialogar sobre linguística em diferentes regiões do Brasil e destacar pesquisas locais, conectando especialistas brasileiros e estrangeiros. Na sua 17ª edição, *Abralin em Cena 17 – Dados Linguísticos* (https://abralin.org/abralin-em-cena-17/), realizado entre os dias 26 e 28 de junho de 2023 na modalidade remota e gratuita, a temática transversal foi a abordagem de dados linguísticos em interface com Ciência Aberta, Humanidades Digitais, ética e conformidade legal, patrimônio cultural, processamento de linguagem natural e políticas públicas. O evento foi promovido pela Comissão de Sociolinguística da ABRALIN, pelo Grupo de Trabalho de Sociolinguística da ANPOLL, para dar suporte ao projeto Plataforma da Diversidade Linguística Brasileira.

O que fazer com os dados que a pesquisa linguística brasileira acumulou ao longo de décadas? Projetos como o NURC, o VARSUL e o ALiB constituíram, entre as décadas de 1960 e 1990, acervos que moldaram gerações de pesquisadores e subsidiaram descrições do português brasileiro em múltiplos níveis (Freitag, a sair). Entretanto, esses acervos permanecem, em sua maioria, fragmentados em repositórios institucionais isolados, sem protocolos comuns de documentação, sem metadados padronizados e sem interoperabilidade (Freitag, 2022). O esforço de curadoria que sua constituição demandou é, paradoxalmente, invisibilizado na literatura: a referência à metodologia de um banco de dados costuma ocupar "duas a três linhas, quando muito um parágrafo" nos artigos, sem que o real dispêndio de tempo e recursos seja registrado (Freitag, Martins, Tavares, 2012). O *Abralin em Cena 17* teve o papel de colocar esse diagnóstico na ordem do dia e de propor caminhos.

A programação do evento refletiu essa agenda de forma articulada. Na Mesa-redonda 1, Isabel Monguilhott, autora de um dos artigos deste dossiê, integrou o debate sobre planejamento, governança e curadoria das grandes tradições de documentação linguística brasileira. A Mesa-redonda 2, mediada por Juliana Bertucci Barbosa, aprofundou as boas práticas de organização e reuso de dados. A Mesa-redonda 3, mediada por Marcia dos Santos Machado Vieira, discutiu o conceito de ecossistema de dados segundo os princípios FAIR, com a participação de Raquel Meister Ko. Freitag, que apresentou evidências sobre a sub-representação do português brasileiro nas infraestruturas digitais (Freitag e Gois, 2024, Freitag 2024). A Mesa-redonda 4 confrontou os desafios éticos e legais decorrentes da disponibilização de dados de fala. A Mesa-redonda 5 projetou as interfaces transdisciplinares da sociolinguística com computação e ciência da informação. Além disso, a sessão de comunicações, também na modalidade remota, reuniu trabalhos e ações relacionados à coleções de dados linguísticos. As mesas foram transmitidas pelo canal do YouTube da ABRALIN:

- Sessão de Abertura e Mesa-redonda 1 – Bancos de dados linguísticos brasileiros: planejamento, construção, governança e curadoria de coleções de dados <https://www.youtube.com/live/YDeKCxw-p-A?si=ahtNXfG9lJvKpV0q>
- Mesa-redonda 2 – Boas práticas de documentação, organização, tratamento, gestão, exposição e (re)uso de dados linguísticos <https://www.youtube.com/live/KGsbta72bG4?si=Qz51M4ZHqxgVqN33>
- Mesa-redonda 3 – Ecossistema de dados sociolinguísticos: conceito, tipologia, segurança e impacto de banco de dados linguísticos em Ciência e Educação abertas e cidadãs guiados por princípios FAIR <https://www.youtube.com/live/_97maExD1vE?si=71nQz-CJOU5GSzTY>
- Mesa-redonda 4 – Direitos, Conformidade legal, Ética, Etnossensibilidade, Cidadania: Ciência e Educação <https://www.youtube.com/live/7kZ4TgUkcgY?si=2Sti6oveh7QGl5Q2>
- Mesa-redonda 5 – Saberes e fazeres transdisciplinares: Sociolinguística(s) em ação com (repercussão em) outras áreas do conhecimento <https://www.youtube.com/live/QdLBMYsC0KQ?si=L9vVKiE09qGqng9c>

Por ter sido realizado na modalidade online e ter ficado registrado em canal de acesso público, o evento teve alcance e capilaridade amplamente expandidos, permitindo que pesquisadores de diferentes regiões do Brasil e de outros países assistissem e interagissem com o conteúdo para além do período de realização. Esse alcance se refletiu na composição deste dossiê: além dos trabalhos apresentados no evento, outros foram incorporados com o objetivo de fortalecer a discussão, ampliando o escopo temático para envolver, de forma articulada, três grandes campos do conhecimento sobre a linguagem: a Sociolinguística, com sua tradição de documentação de variedades e construção de bancos de dados; a Psicolinguística, com seus estudos sobre aquisição e processamento de linguagem a partir de corpora naturalísticos; e o Processamento de Linguagem Natural (PLN), com suas ferramentas de análise automática de dados de fala e texto.

Os artigos deste dossiê mostram desde o problema concreto de abrir dados existentes até a infraestrutura que tornaria possível fazê-lo em escala nacional.

Em **Os desafios para disponibilização e compartilhamento de dados linguísticos da Amostra Base VARSUL,** Isabel de Oliveira e Silva Monguilhott, Izete Coelho e Claudia Brescancini relatam os desafios para a disponibilização da Amostra Base VARSUL, com 288 entrevistas sociolinguísticas coletadas entre 1989 e 1996, hoje em processo de anonimização para conformidade com a Lei Geral de Proteção aos Dados (LGPD) e com as diretrizes da Ciência Aberta. Os desafios relatados neste artigo ecoam o que outros acervos também constatam: temos dados valiosos, mas abri-los é difícil, exige curadoria ativa, protocolos éticos e infraestrutura tecnológica. O VARSUL é um caso emblemático, e sua trajetória ilustra o desafio que outros repositórios já constituídos de dados linguísticos precisarão enfrentar.

Em contraponto, em **Linha verbal infantil: 21 meses de expressão verbal de uma criança brasileira,** Pedro Perini-Santos apresenta vinte e um meses de registros longitudinais da expressão verbal de uma criança brasileira integrante do corpus CIL (Corpus Infantil Longitudinal). A disponibilização de dados naturalísticos de aquisição desde a concepção do projeto relatada no artigo mostra como um dado pode nascer já planejado para o compartilhamento, com contribuição à psicolinguística e aos estudos de aquisição de linguagem pela via do reuso.

Em **Construindo o Carolina: Metadados de Proveniência e Tipologia em um Corpus do Português Brasileiro Contemporâneo**, Marcelo Finger, Maria Clara Paixão de Sousa, Cristiane Namiuti, Vanessa Martins do Monte, Aline Silva Costa, Felipe Ribas Serras, Mariana Lourenço Sturzeneker, Miguel de Mello Carpi, Mayara Feliciano Palma e Gabriela Alves Lachi descrevem o processo de construção do Carolina, grande corpus aberto de português brasileiro em desenvolvimento desde 2020, pela via do PLN. O artigo detalha os desafios de proveniência, tipologia e metadados segundo padrões internacionais, uma lição metodológica que outros repositórios nacionais também precisam aprender, e posiciona o projeto como resposta direta ao problema da escassez de recursos textuais representativos do português em sistemas de IA.

A perspectiva internacional é trazida em **Sharing and Preserving Sociolinguistic Corpora on the U.S.-Mexico Border**, de Katherine Christoffersen, Isabella Calafate, Julio Ciller, Ana Carvalho, Ryan Bessett, Brandon Martinez, Hannia Rojas Barreda, William Flores e Richard Quiroz, que apresentam dois corpora do espanhol na fronteira EUA-México – o CESA e o CoBiVa – como estudos de caso de compartilhamento por colaboração. O artigo mostra que os desafios da pesquisa brasileira são globais, e que a abertura de dados sociolinguísticos impacta diretamente acessibilidade, reprodutibilidade e democratização do conhecimento. A solução colaborativa que os autores propõem ecoa o modelo de articulação interinstitucional que a Plataforma da Diversidade Linguística Brasileira busca implementar.

Do plano da disponibilização, o dossiê passa para o plano do reuso e da comparabilidade. Em **Meta-análise dos estudos de negação verbal nas Regiões Nordeste e Sudeste: investigação da relevância dos condicionamentos sociais**, Pedro Henrique Sousa dos Santos e Elyne Giselle de Santana Lima Aguiar Vitório apresentam um estudo de uma meta-análise de trabalhos sobre negação verbal nas regiões Nordeste e Sudeste do Brasil. O artigo expõe as dificuldades de realizar este tipo de análise em função da falta de padronização metodológica entre os bancos de dados sociolinguísticos, o que limita as comparações regionais. No escopo do dossiê, o artigo mostra que, sem interoperabilidade entre coleções, a sociolinguística brasileira continuará produzindo ilhas de conhecimento. Ainda sobre a negação, em **Negativas: um protótipo para a busca e classificação de negação sentencial em dados de fala,**

Túlio Sousa de Gois e Paloma Batista Cardoso, que descrevem o negativas, ferramenta para identificação automática das três posições da negação verbal no português brasileiro em dados de fala transcritos, aplicada ao banco Falares Sergipanos com taxa de acerto de 93%. Apesar da alta

taxa de acerto, o artigo evidencia que o avanço do PLN em língua portuguesa depende de dados linguísticos curados por especialistas; a contribuição da Sociolinguística ao PLN revela-se bidirecional e necessária.

O dossiê se encerra com a apresentação da **Plataforma da Diversidade Linguística Brasileira** como proposta submetida à chamada do CNPq para Institutos Nacionais de Ciência e Tecnologia (INCTs). A proposta assume a diversidade linguística do Brasil como patrimônio científico e cultural – mais de 250 línguas, além do português e de suas variedades – à necessidade de dados estruturados e etiquetados por linguistas para modelos de língua em larga escala (LLM) representativos, e o papel de uma infraestrutura nacional integrada para transformar o esforço disperso de documentação em recurso estratégico para a soberania digital. A Plataforma da Diversidade Linguística Brasileira é a resposta institucional a uma demanda formulada desde as décadas de 1960 a 1990 (Freitag, a sair) e construída coletivamente desde 2018 pelo GT de Sociolinguística da ANPOLL e pela Comissão de Sociolinguística da ABRALIN, com marcos no 1º Fórum Internacional em Sociolinguística (2019), no 12º InterAb (Machado-Vieira et al., 2022), no Festival de Conhecimento da UFRJ (2021) e no Abralin em Cena 17 (2023) que deu origem a este dossiê.

Tomados em conjunto, os artigos deste dossiê mostram que a abertura, o compartilhamento, a salvaguarda e a reutilização de dados linguísticos são condições para que a diversidade linguística brasileira não seja invisibilizada na era dos algoritmos. A relevância deste dossiê transcende os limites da comunidade linguística.

O Plano Brasileiro de Inteligência Artificial 2024-2028 define como uma de suas metas estratégicas o desenvolvimento de LLMs em português, "com dados nacionais que abarcam nossa diversidade cultural, social e linguística, para fortalecer a soberania em IA". Assim, os dados linguísticos produzidos pela pesquisa científica brasileira são insumos estratégicos para o desenvolvimento tecnológico nacional. A linguística, com sua longa tradição de documentação rigorosa da diversidade linguística, tem a responsabilidade de liderar a curadoria especializada sem a qual nenhum LLM representará adequadamente o Brasil.

O protagonismo das associações científicas é central nesse processo. A ABRALIN e o GT de Sociolinguística da ANPOLL se posicionam como agentes institucionais que articulam a comunidade, pautam a agenda científica e constroem pontes com órgãos governamentais, agências de fomento e redes de infraestrutura de pesquisa. A trajetória que levou ao *Abralin em Cena 17*, desde o 1º Fórum Internacional em Sociolinguística (2019) à proposta à chamada CNPq/SECTICS/CAPES/FAPs Nº 46/2024 – Programa Institutos Nacionais de Ciência e Tecnologia – INCT (aprovada no mérito, mas fora da faixa de financiamento), é o resultado da organização coletiva da comunidade científica para enfrentar um desafio que nenhuma instituição isolada poderia resolver. Sem iniciativas como essa, os dados linguísticos brasileiros continuarão dispersos; a Plataforma da Diversidade Linguística Brasileira estrutura um projeto nacional de

salvaguarda e compartilhamento de dados. É esse protagonismo associativo, ainda raro no cenário científico brasileiro, que este dossiê também registra e celebra.

Por fim, este dossiê cumpre também uma função formativa. Grande parte da comunidade linguística brasileira ainda não está familiarizada com os procedimentos e protocolos da Ciência Aberta, que envolve planos de gestão de dados, esquemas de metadados, licenças de uso, identificadores persistentes, pré-registro de análises e repositórios abertos. A publicação na revista *Cadernos de Linguística* também cumpre um papel educativo, contribuindo para a formação de uma cultura de Ciência Aberta na área de Linguística, e preparando pesquisadoras e pesquisadores para as exigências crescentes das agências de fomento, dos editores científicos e das políticas de dados de pesquisa. Assim, publicar sobre compartilhamento de dados linguísticos em um periódico que é, ele mesmo, um exemplo de prática aberta é uma escolha propositiva: os artigos reunidos neste dossiê não apenas discutem a Ciência Aberta, como também a colocam praticam. A Plataforma da Diversidade Linguística Brasileira é o projeto que propõe tornar esse caminho infraestrutura permanente.

REFERÊNCIAS

FREITAG, R. M. K. Sociolinguistic repositories as asset: challenges and difficulties in Brazil. *The Electronic Library*, v. 40, n. 5, p. 607-622, 2022.

FREITAG, R. M. K. (a sair). Sociolinguística no Brasil: marcos, contribuições e perspectivas.

FREITAG, R. M. K. Diversidade linguística e inclusão digital: desafios para uma IA brasileira. In: *Conferência Latino-Americana de Ética em Inteligência Artificial*. SBC, 2024. p. 157-160.

FREITAG, R. M. K.; GOIS, T. S. Performance in a dialectal profiling task of LLMs for varieties of Brazilian Portuguese. In: *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*. 2024. p. 317-326.

FREITAG, R. M. K.; MARTINS, M. A.; TAVARES, M. A. Bancos de dados sociolinguísticos do português brasileiro e os estudos de terceira onda: potencialidades e limitações. *Alfa: Revista de Linguística*, v. 56, p. 917-944, 2012.

MACHADO VIEIRA, M. DOS S.; BARBOSA, J. B.; FREITAG, R. M. K.; BORGES, M. M.; MEDEIROS, A. L. S. Collections of data open to society: linguistic and sociocultural memory and potential for (re)use. *Cadernos de Linguística*, v. 2, n. 1, p. e607, 24 Jan. 2022. Disponível em: https://cadernos.abralin.org/index.php/cadernos/article/view/607

MACHADO VIEIRA, M. dos S.; BARBOSA, J. B. Coleções de dados brasileiras para o ensino de Português. MEIRELES, V.; MACHADO VIEIRA, M. dos S. (ed.). *Variação e ensino de português no mundo*. São Paulo: Blucher Open Access, 2022. Disponível em: <https://openaccess.blucher.com.br/journal-list/linguistica-68>.

SOUSA, M. D. A. F.; FREITAG, R. M. K. Bancos de dados sociolinguísticos e a Ciência Aberta: compartilhamento de dados e conhecimentos. *Revista Diálogos*, v. 12, n. 1, p. 165-187, 2024.

Painel "Futuros possíveis para dados sociolinguísticos". Festival de Conhecimentos da UFR. 12 de junho de 2021. Disponível em: https://www.youtube.com/watch?v=ZrZxsd5QQns

Mesa Redonda "Acervos de dados abertos à sociedade: memória linguística e sociocultural e potencialidade de (re)uso". InterAb21. 24 set 2021, Disponível em: https://www.youtube.com/watch?v=BsCvqcTo-qc&feature=emb_imp_woyt