Caderno Estudos Linguísticos

Revista Diálogos (RevDia)

Bancos de dados sociolinguísticos e a Ciência Aberta: compartilhamento de dados e conhecimentos

Marta Deysiane Alves Faria Sousa[1]
Raquel Meister Ko. Freitag[2]

R E S U M O:
Bancos de dados sociolinguísticos têm sido fonte de referência para a descrição das diferentes variedades do português brasileiro e oferecido subsídios para outras funções sociais. Contudo, o conhecimento produzido pela Sociolinguística Variacionista não tem recebido a visibilidade no cenário brasileiro, assim como argumentado por Labov (2020) acerca das contribuições da sociolinguística nos Estados Unidos. Constatamos isso por meio de exemplos mecanicistas em livros didáticos da variação de sentido no ensino da língua portuguesa, bem como a visibilidade de "linguistas da internet" que propagam o preconceito linguístico. Assim, apresentamos, neste texto, uma reflexão sobre como a adoção de alguns preceitos de Ciência Aberta podem contribuir para o fortalecimento da área frente a sociedade brasileira.

PALAVRAS-CHAVE:

Sociolinguística;
Ciência Aberta;
Bancos de dados sociolinguísticos;

[1] Possui graduação em Letras pela Universidade Federal de Viçosa (2010) e mestrado em Letras pelo Programa de Pós-Graduação em Letras pela mesma universidade, sendo bolsista CAPES (2013). É doutora em Letras da Universidade Federal de Sergipe (2023). ORCID: https://orcid.org/0000-0002-0480-0422

[2] Professora do Departamento de Letras Vernáculas, do Programa de Pós-Graduação em Letras e do Programa de Pós-Graduação em Psicologia da Universidade Federal de Sergipe. Graduada em Letras, mestre e doutora em Linguística pela Universidade Federal de Santa Catarina. https://rkofreitag.github.io

1 Introdução

Bancos de dados sociolinguísticos têm sido fonte de referência para a descrição das diferentes variedades do português brasileiro e oferecido subsídios para outras funções sociais. Contudo, as contribuições da Sociolinguística Variacionista para a promoção de justiça social não recebem grade visibilidade, como argumentado por Labov (2020). O autor cita como consequências de suas pesquisas a criação do material didático "The Reading Road", que foi elaborado a partir do seu estudo no Harlem observando a estrutura do inglês vernacular afro-americano (*African American Vernacular English - AAVE*). Outro fato que o pesquisador narra é a criação de softwares de análise acústica e a descrição de fenômenos fonético-fonológicos variáveis que caracterizavam as diferentes variedades do inglês norte-americano a partir da coleta e análise dos dados do "*Atlas of North-American English*". Esse projeto também auxiliou na sentença absolvição de um homem injustamente acusado de ameaçar uma companhia aérea com bombas[3].

No Brasil, o cenário não é muito diferente. Apesar do fato de a variação e o reconhecimento da diversidade linguística serem questões que passaram a ser abordadas em documentos oficiais como nos Parâmetros Curriculares Nacionais (BRASIL, 1998), no Programa Nacional do Livro Didático (BRASIL, 2021), na Base Nacional Comum Curricular (BRASIL, 2017) e no Programa de Mestrado Profissional em Letras[4] (notadamente nas ementas das disciplinas "Fonologia, variação e ensino" e "Gramática, variação e ensino"), devido aos estudos variacionistas, a variação linguística contemplada nos livros didáticos ainda é incipiente. Na coleção Trilhas e Tramas do PNLD de 2018, há atividades relacionadas à variação nos níveis fonético, fonológico e morfológico, mas pouco espaço para a variação nos níveis sintático, semântico e pragmático. Como exemplo, citamos o lugar ocupado pelo aspecto verbal. Nesta coleção, são explicitadas as diferentes nuances semânticas dos usos dos aspectos perfectivo e imperfectivo por meio de exemplos mecanicistas e criados com o propósito de se focar na estrutura morfológica, mas não mostram como eles são empregados em situações de comunicação diária e como o aspecto exerce função importante na configuração de figura e fundo de textos narrativos.

---

[3] Em 1984, Paul Prinzivalli, havia sido detido sob acusação de ameaças de bomba à *Pan American*, uma companhia aérea. No entanto, ao se comparar a voz que fazia as ameaças e a de Prinzivalli, foi percebido que a pessoa dos áudios possuía traços fonéticos característicos do sudeste da Nova Inglaterra, enquanto Prinzivalli era um falante com traços fonéticos de Nova Iorque. Labov foi até o tribunal para apontar essas diferenças e devido a essa compilação do "*Atlas of North American English*" e o desenvolvimento de técnicas de análise acústica, o juiz se convenceu de que o réu era inocente (LABOV, 2020).

[4] Esse programa é realizado em âmbito nacional em rede, ou seja, em qualquer uma de suas unidades o regime didático e a matriz curricular é a mesma. Mais informações sobre esse programa estão disponíveis em: https://profletras.ufrn.br/. Acesso em: 20 nov 2021.

A profusão de "linguistas da internet" que propagam o preconceito linguístico e de debates em torno das questões do uso do gênero neutro para se fazer referência a pessoas que se identificam como não-binárias também demonstram a falta de conhecimento até mesmo da história linguística do Português Brasileiro (PB) ou da história das línguas de maneira mais geral. Por exemplo, uma professora de Língua Portuguesa se referiu aos usos inovadores das formas não-binárias como "imposição ideológica", que "coloca a ideologia antes da linguística, antes da gramática" e "artificial" de usos[5]. Essa fala desvela a crença de que a Língua Portuguesa Europeia, por exemplo, foi se infiltrando no Brasil de maneira natural, pacífica e sem algum tipo de ideologia subjacente.

Embora esses fatos revelem que o conhecimento produzido pelos estudos sociolinguísticos ainda não chega às massas, ressaltamos, neste texto, a importância da compilação de dados de fala em diversos aspectos. Bancos de dados de fala oferecem recursos linguísticos que podem ser operados por computadores e também ser utilizados como fonte para construção de materiais didáticos, gramáticas, dicionários, desenvolvimento de teorias linguísticas e desenvolvimento de recursos humanos (GONÇALVES, 2019; FREITAG, 2021). Como exemplos de desenvolvimento de recursos educacionais a partir de bancos de dados de fala, mencionamos: a "Gramática do Português Culto Falado no Brasil", produzida com dados de fala do projeto Norma Linguística Urbana Culta (NURC) com análise linguística em diferentes níveis e sob diferentes perspectivas teóricas (CASTILHO, 2021) e o Banco de Dados RECI, que forneceu dados para a construção de uma proposta de ensino dos modos indicativo e subjuntivo da Língua Portuguesa levando em consideração contextos mais formais e informais de comunicação (SILVA; COELHO, 2020).

A relevância dos bancos de dados linguísticos não se encerra nas áreas de descrição e análise linguística ou educacional. Bancos de dados linguísticos funcionam como uma fonte de dados de comunidades de fala, de suas histórias e costumes, sendo, portanto, um recurso de pesquisa para diferentes áreas do conhecimento, como antropologia, agricultura e história da arte (BURKE; CHELLIAH, 2021).

O desenvolvimento de ferramentas para o Processamento de Linguagem Natural (PLN) também ocorre por meio da utilização de bancos de dados linguísticos. Especificamente para o processamento da língua portuguesa, citamos algumas ferramentas de análise de corpora que foram feitas com base em bancos de dados linguísticos disponíveis: o *parser* PALAVRAS (BICK, 2000), um analisador sintático automático que se beneficiou da disponibilidade de corpora do NURC[6] (CASTILHO,

[5] Debate disponível em: https://www.youtube.com/watch?v=xBb1E5CNegU. Acesso em: 20 mai. 2023.

[6] Por se tratar de um projeto interinstitucional, nem todos os dados do NURC encontram-se digitalizados. Neste site: https://fale.ufal.br/projeto/nurcdigital/, encontram-se os dados referentes à coleta realizada em Recife. Acesso em: 10 jan. 2023.

2021), Tycho Brahe[7] (GALVES, ANDRADE, FARIA, 2017), NILC[8] (KUHN; ABARCA; NUNES, 2000) entre outros, para treinamento; o AELIUS, outro analisador sintático automático, que também utilizou o *corpus* Tycho Brahe (GALVES; ANDRADE; FARIA, 2017) e o Mac-Morpho[9] (ALUÍSIO et al., 2003) para treinamento, e, no caso do Tycho Brahe, utilizou também o conjunto de etiquetas; o spaCy 3.5 que usa o modelo Universal Dependencies do *corpus* Bosque[10] (RADEMAKER et al., 2017); e o etiquetador TreeTagger, que também criou seu modelo baseado no Bosque e no CETEMPúblico[11] (SANTOS; ROCHA, 2001).

Os bancos citados como fonte para o desenvolvimento das ferramentas de PLN encontram-se disponíveis online e estão sistematizados em conformidade com as intenções de pesquisa que culminaram em sua produção. Essa sistematização e disponibilização ainda é incipiente no Brasil na área da Sociolinguística Variacionista, com poucos bancos de dados online com amostras disponíveis, sejam em livre acesso, ou acesso restrito (sob demanda). Isso significa um entrave para reprodução de estudos para estudos contrastivos entre variedades do PB, generalizações mais robustas acerca da do PB e disseminação do conhecimento produzido para que se chegasse a tal compilação. Neste texto, argumentamos, então, em favor da pesquisa em Sociolinguística alinhada com os preceitos de Ciência Aberta, isto é, pesquisas com metodologias mais transparentes e que permitam o acesso aos dados, mesmo que sob demanda, sendo consideradas, portanto, mais democráticas.

Este texto está dividido em cinco partes, contando esta introdução. Na seção 2, discutimos a desestabilização da confiança nos conhecimentos científicos e evidenciamos as contribuições da Ciência Aberta para contribuir na resolução deste problema, apresentando os princípios FAIR de gerenciamento de dados. Na terceira seção, apresentamos as demandas por compartilhamento de dados e inciativas de disponibilização de dados no escopo da Sociolinguística Variacionista. Na quarta seção, apresentamos reflexões sobre a metodologia de coleta da pesquisa em sociolinguística e apresentamos pequenas ações para tornar a pesquisa mais transparente nessa área. Por fim, tecemos nossas considerações finais.

[7] Disponível em: https://www.tycho.iel.unicamp.br/corpus/. Acesso em 10 jan. 2023.
[8] Disponível em: http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm. Acesso em 10 jan. 2023.
[9] Disponível em: http://nilc.icmc.usp.br/macmorpho/. Acesso em: 10 jan. 2023.
[10] Disponível em: https://github.com/UniversalDependencies/UD_Portuguese-Bosque. Acesso em 10 jan. 2023.
[11] Disponível em: https://www.linguateca.pt/CETEMPublico/. Acesso em 10 jan. 2023.

2 Ciência Aberta: gerenciamento de dados e valorização da pesquisa científica

Práticas questionáveis de pesquisa, desestabilização do sistema de peritos. Essas são algumas das questões que tornam desejáveis a adoção do paradigma da Ciência Aberta para se conduzir pesquisas científicas. Fidler e Wilcox (2021) abordam alguns fatores acerca da crise de reprodutibilidade e replicabilidade na ciência: a) ausência de replicação de estudos em várias áreas; b) falha na reprodução de estudos; c) enviesamento das publicações; d) práticas questionáveis na pesquisa; e e) falta de transparência e completude seja nos métodos, dados e/ou análises nas publicações. Cesarino (2021), por sua vez, retrata como o populismo digital e o neoliberalismo vem ocasionando um esvaziamento dos sentidos produzidos na academia, estremecendo o sistema de peritos e fazendo novos conhecimentos baseados em experiências idiossincráticas emergirem e se propagarem. Como consequência, surge o que é chamado de regime de pós-verdade, no qual a validação científica perde espaço para a disseminação de informações falsas, mas validadas pelo enviesamento, muitas vezes político e ideológico, das massas populares.

O paradigma da Ciência Aberta, então, procura estabelecer critérios para tornar a ciência mais transparente e acessível, resultando em uma ciência mais replicável, robusta e democrática. Em outras palavras, há um maior detalhamento de metodologias científicas e do gerenciamento de dados coletados (SILVA; SILVEIRA, 2019), facilitando o acesso a esses dados e, por consequência, a reprodução e a replicação de estudos.

Os princípios que guiam o gerenciamento de dados, dentro do paradigma da Ciência Aberta, podem ser resumidos pelo acrônimo FAIR, segundo o qual F significa *findable* (ser localizável), A significa *accessible* (ser acessível), I, *interoperable* (ser *interoperável*) e R, *reusable* (ser reutilizável). No quadro 1 abaixo, reproduzimos os descritores de cada um dos princípios FAIR.

Quadro 1 – Descrição dos princípios FAIR.

Os Princípios Norteadores FAIR
Ser localizável: F1 Atribuir um identificador exclusivo e persistente aos dados e metadados. F2 Descrever dados com metadados ricos (definidos pelo R1 abaixo). F3 Incluir o identificador dos dados descritos nos metadados de maneira clara e explícita. F4 Indexar e registrar (meta) dados em um repositório pesquisável **Ser acessível:** A1. Usar protocolos padronizados para recuperar dados e metadados pelo seu identificador. A1.1 Os protocolos devem ser abertos, gratuitos e universalmente aplicáveis. A1.2 Os protocolos devem permitir um procedimento de autenticação e autorização, se necessário. A2. Garantir a acessibilidade aos metadados, mesmo que os dados não estejam mais disponíveis. **Ser interoperável**: I1. Usar linguagens formais (acessíveis, compartilhadas e padronizadas) para representar o conhecimento nos (meta)dados. I2. Usar um vocabulário que também segue os princípios do FAIR nos meta(dados). I3. Incluir referências a outros (meta)dados. **Ser reutilizável**: R1. Descrever os (meta)dados com uma pluralidade de atributos precisos e relevantes. R1.1. Publicar (meta)dados com uma licença de uso clara e acessível. R1.2. Usar detalhes de proveniência na descrição dos (meta)dados R1.3. Seguir padrões comuns da área de conhecimento a que os (meta)dados se referem.

Fonte: Adaptado de Wilkinson *et al.* (2016) para a língua portuguesa.

De maneira geral, os descritores apontam para ações que devem ser feitas para que um conjunto de dados e metadados estejam organizados e gerenciados para que este conjunto possa ser compartilhado de maneira segura e padronizada na comunidade científica onde se insere a pesquisa. Por exemplo, os princípios descrevem que que o depósito dos (meta)dados deve ser feito em repositórios pesquisáveis, para facilitar a sua localização por outros pesquisadores. Apontam também que, para um conjunto de dados ser considerado acessível, seus (meta)dados devem ser recuperáveis por meio de um identificador padronizado por um protocolo de comunicação aberto, gratuito, que permita autenticação e autorização, e também que, mesmo em caso desse conjunto não estar mais disponível, seus metadados sejam ainda acessíveis. Em outras palavras, espera-se que qualquer usuário consiga o acesso ao conjunto de dados sem a necessidade de ferramentas ou conhecimentos

especializados, e que, mesmo na indisponibilidade dos dados por algum motivo, seus metadados ainda sejam recuperáveis.

Os princípios apontam também medidas para tornar um conjunto de dados reutilizável. Nesse sentido, os meta(dados) devem ser liberados por meio de uma licença de uso clara e acessível, ter sua procedência detalhada e estar de acordo com domínios relevantes para a comunidade. Assim, para a verificação da utilidade dos dados para uma pesquisa, é necessário que os metadados estabeleçam uma descrição pormenorizada dos dados em diferentes camadas, uma mais técnica que permita interoperabilidade entre os sistemas digitais e outra com informações mais específicas dos dados. Na Sociolinguística Variacionista, informações mais específicas, por exemplo, seria o detalhamento das informações sociodemográficas da comunidade e dos informantes pesquisados, bem como dos critérios de seleção desses informantes, dos modelos estatísticos empregados, das ferramentas computacionais utlizadas, entre outros.

Seguir os preceitos de gerenciamento de dados da Ciência Aberta traz diferentes benefícios. Garellek et al. (2020) apontam, por exemplo, citações e por consequência oportunidades de financiamento (uma vez que o sistema de financiamento preza por beneficiar pesquisadores que possuam mais publicações e relevância em termos de citação), trabalho, colaboração, prevenção de perda dos dados e problemas com link, além de reprodução, replicação favorecendo um avanço mais rápido do conhecimento. Diante disso, percebemos a importância de se buscar formas para gerenciar os dados de pesquisas na Sociolinguística Variacionista seguindo preceitos de Ciência Aberta.

3 O cenário dos bancos de dados sociolinguísticos brasileiros frente ao compartilhamento de dados

O cenário da Sociolinguística Variacionista brasileira ainda se encontra distante do paradigma da Ciência Aberta, principalmente no que concerne ao compartilhamento, ao planejamento da gestão dos dados e autoria. Em relação ao compartilhamento, Freitag (2017a) constatou que o acesso a bancos de dados sociolinguísticos ainda é restrito, sendo que grande parte deles está armazenada em dispositivos de armazenamento móveis dos próprios pesquisadores. Situações prosaicas como problemas com HDs, queima de computadores e mofo são fatos que tornam desejável a adoção de um plano de preservação de dados para que eles não sejam perdidos (TAGLIAMONTE, 2006), isto é, a adoção de um planejamento de gestão, especificamente no que concerne ao armazenamento de dados.

Avanços tecnológicos, por outro lado, têm favorecido o armazenamento, a anotação e digitalização de amostras sociolinguísticas e a disponibilização dessas

amostras de forma online. Esses avanços, que têm permitido o compartilhamento de áudios, vídeos e transcrições digitais, levaram a uma nova forma de produção do registro linguístico, de forma que a transcrição impressa já é considerada obsoleta (VANN, 2021). Assim, conforme Mello (2021), corpora orais de terceira geração são aqueles em que a transcrição e o áudio estão alinhados.

Ademais, devido ao movimento da Ciência Aberta, os periódicos têm incentivado o acesso aos dados, evidenciando a necessidade de se ter um plano de gerenciamento para torná-los totalmente digitais e sistematizados. Como exemplo de incentivo à disponibilização de dados, o periódico *Open Linguistics* encoraja os autores a fornecer acesso irrestrito a tudo aquilo que foi necessário para gerar os resultados das pesquisas a serem publicadas pelo periódico, incluindo os dados crus, códigos, tabulações entre outros. Conforme a política editorial desse periódico, a acessibilidade aos dados favorece a replicação, a robustez e a solidez do trabalho, aumentando a qualidade das pesquisas da área e, consequentemente, o aumento quantitativo do fator de impacto e citações, métricas importantes para a concessão de fundos para pesquisa, como também apontado por Garellek et al. (2020).

Seguindo a mesma política editorial em consonância com o movimento Ciência Aberta, a Revista da Abralin também incentiva o compartilhamento dos dados, scripts para análises estatísticas e quaisquer outros materiais dessa natureza. Esses exemplos de ações demonstram o crescente interesse em sistematizar os dados advindos de pesquisas sociolinguísticas em bancos de dados alocados, preferencialmente, em sites ou repositórios online.

Em 2020, diferentes simpósios realizados no evento ABRALIN Ao Vivo, notadamente "Grandes Projetos da Linguística em rede: "Tycho Brahe, PROHPOR e PHPB"[12], "Archiving and Language Documentation"[13] e "Descrição linguística: gestão de dados linguísticos"[14] , discutiram ações sobre o gerenciamento de dados linguísticos frente às demandas da Ciência Aberta. O painel temático "Futuros Possíveis para Dados Sociolinguísticos"[15], no Festival do Conhecimento da Universidade Federal do Rio de Janeiro em 2021, também teve como foco a gestão de dados sociolinguísticos evidenciando o interesse por projetos de sistematização, gerenciamento e preservação de dados linguísticos.

Conforme Kendall (2013), nos Estados Unidos, vários pesquisadores têm discutido a necessidade de ações conjuntas para o desenvolvimento de corpora sociolinguísticos, no qual *corpus* é considerado em um sentido menos prototípico[16].

[12] Disponível em: https://www.youtube.com/watch?v=gJgrqArfDIw. Acesso em 10 nov. 2020.
[13] Disponível em: https://www.youtube.com/watch?v=uQY4dZnQKds&t=269s. Acesso em 15 ago. 2020.
[14] Disponível em: https://www.youtube.com/watch?v=S7YS57i7ogs&t=7348s. Acesso em: 15 ago. 2020.
[15] Disponível em: https://www.youtube.com/watch?v=ZrZxsd5QQns. Acesso em: 12 jul. 2021
[16] São considerados *corpora* menos prototípicos os conjuntos de dados linguísticos que não atendem a um ou mais critérios que caracterizam um *corpus* tradicionalmente descrito na Linguística de *Corpus*, como

Apesar desse interesse, por ações conjuntas, o autor relata que a sistematização dos dados linguísticos nesse país tem se dado no nível individual, ou seja, corpora sociolinguísticos são construídos, mas padronizados, armazenados e compartilhados em sites individuais dos próprios pesquisadores. Situação semelhante tem acontecido no Brasil. Há pesquisadores que já sistematizam e disponibilizam algumas de suas amostras em site próprio, porém, diversas publicações sobre o tema (FREITAG; MARTINS; TAVARES, 2012; FREITAG, 2016; 2017a; 2017b) consideram necessária a padronização tanto em termos de coleta quanto de disponibilização.

Em 2019, o Grupo de Trabalho de Sociolinguística da ANPOLL decidiu prospectar e começar a realizar ações para que um repositório consorciado da área seja criado. Em setembro de 2021, durante o XII Congresso Internacional da Abralin, foi apresentada ao público a proposta construída pelo grupo intitulada "Plataforma Digital da Diversidade Linguística Brasileira".

Dentro da idealização da proposta constava o mapeamento de amostras sociolinguísticas existentes no Brasil. Por meio de um questionário, até setembro de 2021, o grupo identificou 24 bancos de dados, em termos de 24 respondentes ao formulário. Por meio das respostas, Machado Vieira et. al (2021) relataram que a maioria dessas amostras está vinculada a programas de pós-graduação e concentrada na faixa litorânea do Brasil, o que pode estar relacionado à densidade populacional nessas áreas, conforme os pesquisadores. Na tabela 1, sistematizamos os dados de projetos que possuem site.

representatividade, balanceamento, coleta naturalística, ou a possibilidade de ser lido por máquinas (GRIES; BEREZ, 2017).

Tabela 1 – Bancos de dados sociolinguísticos com site.

Nome do Projeto	Sigla	Página na internet	Indica como citar	Local	Amostras no site
Variação Linguística no Português Alagoano	PORTAL	https://www.portuguesalagoano.com.br	Não.	AL/BR	Sim.
Corpora de Variedades do Português em Análise	COPORAPORT	https://corporaport.letras.ufrj.br/	Sim.	RJ/BR, PT e MOÇ	Sim.
Análise Contrastiva de Variedades do Português	VARPORT	https://varport.letras.ufrj.br/	Não.	BR e PT	Sim.[1]
Vertentes do Português Popular do Estado da Bahia	VERTENTES	http://www.vertentes.ufba.br/	Não.	BA/BR	Não.
Variação Linguística na Região Sul do Brasil	VARSUL	http://www.varsul.org.br/	Não.	RS, SC e PR/BR	Sim.
Amostra Linguística no Interior Paulista	ALIP	https://www.alip.ibilce.unesp.br/	Sim.	SP/BR	Sim.
Núcleo de Estudos sobre Interlínguas	NEIS	https://corpusneis.wixsite.com/home	Não.	RJ/BR	Não.
Corpus Histórico da Língua Portuguesa	HISTLING	https://histling.letras.ufrj.br/index.php	Não.	Variado	Sim.
Projeto Variação Linguística no Estado da Paraíba	VALPB	http://projetovalpb.com.br/index.html	Não.	PB/BR	Não.[2]
Programa de Estudos Sobre o Uso da Língua	PEUL	https://peul.letras.ufrj.br/	Não.	RJ/BR	Sim
Grupo de Estudos Variacionistas	GEVAR	https://sites.google.com/site/uftmgevar/	Não.	MG/BR	Não.
Núcleo de Estudos do Português em Uso	PORUS	http://porus.sites.uff.br/	Não.	RJ e MG/BR	Sim.
Não possui nome por extenso.	LínguaPOA	https://www.ufrgs.br/linguapoa/	Sim.	RS/BR	Não.

1.O acesso aos dados se faz por meio de *links*, mas alguns não funcionam. 2. Os *links* de acesso aos dados direcionam para pastas vazias, ou informam que os *links* não estão disponíveis. **Fonte:** Elaboração própria a partir do levantamento do GT de Sociolinguística em 2021 e apresentado em Machado Vieira *et. al* (2021).

Os dados da tabela 1 estão organizados pelo nome do projeto, sigla, página na internet, se indica como citar, o local que significa a comunidade pesquisada e se as amostras estão disponíveis em site próprio. Pode ser observado pela tabela 1 que: sete, dos 13 projetos que possuem site na internet, disponibilizam amostras em seus sites; três possuem amostras coletadas na região nordeste do Brasil (PORTAL, VERTENTES e VALPB); dois, na região sul (VARSUL e LÍNGUAPOA); cinco, na região sudeste (ALIP, NEIS, PEUL, GEVAR, PORUS); dois (CORPORAPORT e VARPORT), no Brasil (Rio de Janeiro) e em outros países; um com coleta de cartas e documentos produzidos por informantes com origem variável (HISTLING). Sobre referência aos dados, apenas três indicam as formas de citá-los (CORPORAPORT, ALIP e LÍNGUAPOA), evidenciando a necessidade de se adotar medidas mais claras em relação aos colaboradores envolvidos na compilação desses bancos.

O número maior de sites com dados compartilháveis confirma as experiências de Kendall (2008) e Tagliamonte (2006) sobre ter amostras sistematizadas digitalmente. Os autores perceberam que com a sistematização das amostras, houve maior integração entre os dados, melhora nas análises devido ao fácil acesso aos dados, maior facilidade em colaborar com pesquisas e compartilhamento de dados e resultados.

No entanto, no cenário brasileiro, ainda existem alguns desafios na constituição desses repositórios que precisam ser superados em relação aos seguintes aspectos: depreciação, autoria, compartilhamento e financiamento (FREITAG, 2021). A depreciação e o financiamento são aspectos que estão interligados. Nesse sentido, toda a constituição de uma estrutura para se produzir uma pesquisa com dados linguísticos bem como a sua salvaguarda demanda financiamento. A formação de pesquisadores em diferentes níveis (graduação e pós-graduação), a digitalização de amostras antigas, procedimentos de backup para amostras já digitalizadas são limitadas pela escassez de recursos para tal fim, sendo que muitas vezes os pesquisadores usam recursos próprios para manter uma estrutura mínima de salvaguarda.

Sobre a autoria, como visto na tabela 1, dos bancos de dados que possuem site, apenas três indicam a forma de como citar os dados. O CORPORAPORT e o ALIP indicam os coordenadores dos projetos que deram origem aos bancos de dados como seus respectivos autores, já o LINGUAPOA indica a citação ao site, seguindo as recomendações *Tromsø* para citações de dados em Linguística. Essa diferença na forma de citação aos bancos de dados confirma a necessidade de se refletir sobre questões relacionadas à autoria, como aponta Freitag (2021). A pesquisadora recomenda a descrição dos papéis desempenhados por cada pesquisador na compilação e gerenciamento dos dados utilizando a taxonomia *CRediT*, o que assegura que o trabalho de cada um dos envolvidos na pesquisa seja reconhecido, e também

recomenda o uso de copyright, uma vez que a constituição das amostras é um produto intelectual.

No que diz respeito ao compartilhamento, Freitag (2021) afirma que essa é uma das ações importantes na constituição de repositórios, visto que o acesso aos dados vem sendo motivado por periódicos, como já mencionamos, e também para conferir maior transparência e confiabilidade às pesquisas. A autora salienta que como pesquisadores, instituições e controladores envolvidos na pesquisa assumem a responsabilidade ética e legal pelos dados coletados, estes não podem ser disponibilizados à revelia apenas por se tratarem de produtos gerados a partir de recursos públicos. Além disso, dados linguísticos são importantes para o PLN, sendo considerados de grande valor comercial de forma que empresas privadas não devem obter acesso a eles sem a garantia de uma contrapartida (cf. FREITAG, 2021; GARELLEK et al., 2020).

Como observado, apesar de, no contexto brasileiro, haver iniciativas de gerenciamento e compartilhamento de dados, a aderência da área da Sociolinguística Variacionista ao paradigma da Ciência Aberta ainda é pequena. Há ainda falta de padronização no que diz respeito ao reconhecimento da autoria das amostras e também no compartilhamento dos dados.

4 Pequenos passos metodológicos para aderência à Ciência Aberta

Na Sociolinguística Variacionista, devido à preocupação em se entender como a mudança ocorre, a língua é estudada dentro de uma comunidade, e, por isso, consegue-se explicar a variação em situações de contato, a coexistência de diferentes formas linguísticas para indicar mesmo valor referencial dentro de uma mesma comunidade, e como os usos da língua são sistemáticos e estruturados conforme a sociedade (WEINREICH; LABOV; HERZOG, 1968; LABOV, 2006 [1966], 2008 [1972]). Para se fazer análise linguística dentro dessa perspectiva, dados são coletados, principalmente por meio de entrevistas sociolinguísticas. As análises recorrem a modelos estatísticos para explicar quais fatores (de natureza linguística, social ou cognitiva) interferem no uso de determinada variante em relação a outra (LABOV, 1969; TAGLIAMONTE, 2012; WEINREICH; LABOV; HERZOG, 1968).

No modelo tradicional de coleta de dados, parte-se da premissa de que a fala dos indivíduos representa as normas da comunidade de fala (TAGLIAMONTE, 2012), logo, para se captar os padrões linguísticos de uma dada comunidade, é preciso ter uma amostra que reflita esses padrões. A escolha por entrevistas, então, não é feita ao acaso. Apesar dos efeitos do paradoxo do observador, conforme Labov (1981), a entrevista sociolinguística representa o melhor meio para se capturar o vernáculo em

grandes volumes (entre uma e duas horas de duração) e com melhor qualidade em termos de gravação.

Tradicionalmente a escolha dos indivíduos para compor a amostra a ser estudada pode ser feita de duas maneiras: amostra aleatória simples e amostra aleatória estratificada (SILVA, 2004). De acordo com Silva (2004), na amostra aleatória simples, busca-se pela quantidade de informantes proporcionalmente à sua representação na população. Já a amostra aleatória estratificada é dividida em estratos (ou células sociais), os quais são compostos por indivíduos aleatoriamente selecionados, mas que compartilham das mesmas características sociais. No estudo empreendido por Labov em Nova Iorque, por exemplo, foi empregada a amostragem aleatória estratificada, pois esta pode oferecer uma representação adequada do grupo-alvo a ser estudado, no caso, falantes de inglês considerados nativos de Nova Iorque (LABOV, 2006[1966]).

Para uma amostragem ser considerada aleatória, a seleção dos informantes deve ser feita ao acaso, ou seja, qualquer indivíduo na população teria chances iguais de ser selecionado. Esse procedimento pode ser feito utilizando informações obtidas por agentes de saúde do Programa Saúde da Família, como feito na constituição da amostra do Povoado Açuzinho em Sergipe (FREITAG; SANTANA; ANDRADE, 2014), ou por meio de dados do censo como levantado pelo projeto NORPORFOR (Norma Popular Oral de Fortaleza) (ARAÚJO, 2011).

Em relação à quantidade de informantes, Labov (2006 [1966]) argumenta que, embora importante, apenas volume não indica a eficácia do método, sendo o detalhamento da metodologia de coleta (seleção da área, detalhes geográficos, composição da survey, detalhamento dos critérios de amostragem, quantidade de respondentes e fontes de erro) também significativo. Ademais, fazendo um retrospecto acerca da composição da amostragem empregada no estudo de Nova Iorque, Labov (2001, p.80) argumenta que:

> embora haja uma variação individual considerável dentro de cada grupo, ela não é normalmente grande o suficiente para perturbar a regularidade do padrão quando entre 5 e 10 falantes são incluídos em cada grupo. Indivíduos cujo desvio da média é suficientemente grande para perturbar o padrão são marcados por histórias sociais irregulares.

Dessa forma, para preencher as células sociais adequadamente, elas devem ter o mínimo de 5 informantes para que não haja discrepâncias no padrão linguístico apresentado. Feagin (2013) argumenta que, na prática, a Sociolinguística Variacionista faz sua análise baseada na quantidade de tokens por informante, ou seja, a quantidade de dados do fenômeno por informante torna-se mais relevante do que a quantidade de informantes por si. Freitag (2018), também salienta a importância do fenômeno para a amostragem, argumentando, com base em Meyerhoff, Schleef e Mackenzie (2015), que resultados com boa confiabilidade e acurácia são atingidos quando cada

fator apresenta 30 ocorrências por célula, de forma que para fenômenos em níveis de análise mais altos e raros são necessários mais informantes e horas de gravação. Fenômenos fonológicos, por sua vez, conforme a autora, podem ser encontrados em um dimensionamento menor da amostra.

De acordo com Freitag (2017b; 2018a), a amostragem, na maioria das pesquisas brasileiras, tem sido feita: em termos de cotas, ou seja, são determinadas as quantidades dos participantes por células sociais a priori; conveniência, os documentadores buscam aqueles informantes que se disponibilizam; ou julgamento, os informantes são selecionados por sua adequação à geração de dados para pesquisa. Silva (2004), por exemplo, aponta que existem amostras no Brasil com células de até dois informantes, como no caso de Araújo e Almeida (2014).

Araújo e Almeida (2014, p. 44) ao descreverem a constituição da amostra de fase 3 do projeto "A Língua Portuguesa no Semiárido Baiano" relatam algumas razões para justificar um tamanho diferente daquele considerado ideal: "falta de financiamentos, dificuldades em se conseguir informantes com certos perfis, perda de entrevistas por problemas técnicos e tempo para conseguir fazer a coleta". Esses mesmos desafios também foram relatados na constituição do Banco de Dados Fala-Natal (TAVARES; MARTINS, 2014), do Projeto SP2010 (MENDES, 2011) e do Banco RECI (SILVA; COELHO, 2020).

Labov (2001) argumenta que estudos não aleatórios têm produzido resultados que oferecem informações relevantes sobre variáveis sociolinguísticas influenciadas por "graus de distância social, e iluminam os mecanismos sociais que levam à conformidade e diversidade linguística" (p. 39), ou seja, por meio desses estudos, é possível observar a atuação das variáveis sociais na diversidade linguística. Contudo, ainda segundo o mesmo autor, os resultados advindos desses estudos não possuem um caráter explanatório que permita generalizações acerca do comportamento linguístico da comunidade de fala.

Não estamos fazendo uma crítica aos trabalhos acima. Ao contrário, estamos relatando como o levantamento de dados no escopo da sociolinguística no Brasil é condicionado por questões contextuais que são discrepantes em relação aos estudos clássicos empreendidos por Labov na década de 60 nos Estados Unidos. Além disso, dificuldades em se replicar a metodologia de Nova Iorque também são encontradas por pesquisadores em outras partes do mundo, como em Toronto (TAGLIAMONTE, 2012), e outros estudos apontados pelo próprio Labov (2001), como o de Haeri (1996) no Cairo e Milroy e Milroy (1978) em Belfast.

O trabalho de constituição de amostras de fala, como visto acima, não é somente caro, mas também dispendioso em termos do tempo gasto para se realizá-lo, e esse fato, como apontam Freitag, Martins e Tavares (2012), não consegue ser refletido na metodologia das publicações na área de sociolinguística. Os autores

relatam que existe uma padronização na escrita, sempre em voz passiva e com construções como "como corpus foram selecionados X informantes do banco de dados Y, estratificados em Z células sociais" (FREITAG; MARTINS, TAVARES, 2012, p. 917), de forma que o delineamento da pesquisa, os desafios em se fazer a amostragem, a entrada em campo, detalhes de transcrição, tempo dispendido, scripts e softwares utilizados, ficam em segundo plano, como também reportado por Freitag (2017b).

Salientamos, conforme Freitag e Rost-Snichelotto (2015) e Freitag (2018), que a forma como é feita a amostragem traz implicações metodológicas, pois ao se manter cotas fixas, a comparabilidade com outras amostras torna-se possível, mas uma amostragem proporcional à estratificação consegue representar melhor a população em estudo, sendo que em qualquer uma das metodologias empregadas, para se manter um rigor estatístico, grandes quantidades de dados devem ser geradas. Devido a isso, ressaltamos a necessidade apontada em diferentes estudos (COLLISCHONN; MONARETTO, 2012; FREITAG, 2016, 2018; FREITAG; MARTINS; TAVARES, 2012; LABOV, 2006 [1966]) de que os critérios de estabelecimento das amostras estejam evidentes nas publicações da área para que a confiabilidade dos resultados possa ser atestada.

A falta de acessibilidade aos dados, ao delineamento metodológicos, aos códigos utilizados na análise estatística, e aos protocolos de transcrição, por exemplo, vai de encontro aos princípios *Accessible* e *Reusable* da Ciência Aberta, isto é, aos princípios de os dados serem acessíveis e reutilizáveis.

Advogamos, então, como um importante passo para a adoção de preceitos da Ciência Aberta na pesquisa sociolinguística, a adoção de uma descrição mais pormenorizada da metodologia empregada, bem como compartilhamento de dados, uma vez que o detalhamento da pesquisa como um todo facilita a reprodução dos resultados originais e também a identificação de fatores que podem levar a resultados discrepantes, como também argumentam Gilmore, Kennedy e Adolph (2018). Segundo esses pesquisadores, o compartilhamento leva ao reuso dos dados, o que pode favorecer novas descobertas científicas. Por fim concordamos com Zee e Reich (2018), pesquisadores da área de educação, que o julgamento da qualidade e relevância da pesquisa é consequência do modo como os pesquisadores compartilham questões metodológicas e processuais subjacentes ao fazer científico.

No contexto internacional, contudo, a adoção do alinhamento entre a constituição de bancos de dados sociolinguísticos às práticas da Ciência Aberta, considerando-se o que está preconizado pelos princípios FAIR tem sido uma prática recorrente (CALAMAI; FRONTINI, 2018). Pesquisadores têm buscado publicar protocolos de coleta, transcrição e etiquetagem, e, ainda, disponibilizam esses dados

em plataforma com níveis de acesso, principalmente em repositórios consorciados, como o *Linguistic Data Consortium*[17], *The Language Archive*[18], *Talk Bank*[19] entre outros.

Oez (2018), por exemplo, descreve como foi feita a documentação do dialeto Beth Qustan, que pertence a uma língua neo-aramaica chamada Turyo falada na região central da província Mardin no sudeste da Turquia, reportando tanto informações de procedimentos de amostragem e preceitos éticos como também a transcrição e o nível de anotação linguística do corpus. O autor menciona também as ferramentas *ELAN* (utilizada para transcrição e segmentação) e *FleX* para a tradução e anotação morfológica. Além disso, o pesquisador alocou seus dados em um repositório consorciado (*ELAR – Endengered Language Archive*) com acesso livre. Embora se trate da descrição de uma variedade ameaçada, uma vez disponibilizado pelo autor, o protocolo pode ser replicado em outras situações.

Outros bancos de dados sociolinguísticos disponíveis e de livre acesso no contexto internacional são: *Corpus del español en el sur de Arizona* (CESA) (CARVALHO, 2012) que contém um protocolo de coleta e transcrição dos dados; The Miami corpus que possui protocolo com detalhamento da transcrição e marcações morfológicas; e o *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América* (PRESEEA, 2008) que possui protocolo de transcrição e etiquetagem de aspectos contextuais e interacionais.

Kendall e French (2006) também reportam a constituição do banco de dados *North Carolina Sociolinguistic Archive and Analysis Project*, apresentando com ênfase como foram feitas as transcrições e anotações prosódicas no corpus. O acesso aos dados desse banco é restrito.

No contexto internacional, os bancos de dados sociolinguísticos acima citados evidenciam exemplos de práticas que se alinham com o princípio da acessibilidade e reutilização ao disponibilizarem os dados na internet seja com níveis de acesso ou totalmente abertos e ao reportarem os processos de transcrição e etiquetagem em diferentes níveis como o morfológico, fonológico e discursivo. Essas práticas promovem também maior transparência na pesquisa o que contribui para a replicabilidade e reprodutibilidade dos estudos, favorecendo o fortalecimento e a confiança da ciência.

Práticas de disponibilização de protocolos de constituição de amostras, como as citadas nos parágrafos anteriores acerca dos bancos de dados sociolinguísticos no exterior, são desejáveis para que sejam garantidas a confiabilidade e intersubjetividade das análises, ou seja, que ao seguir a mesma metodologia, a reprodução dos estudos deve encontrar os mesmos resultados em relação ao mesmo fenômeno linguístico (BAILEY; TILLERY, 2004). O contexto brasileiro, no entanto, ainda carece de estudos e

[17] https://www.ldc.upenn.edu/
[18] https://archive.mpi.nl/tla/
[19] https://archive.mpi.nl/tla/

de bancos de dados que sigam as práticas abertas, de forma a tentar atingir um padrão transparente na metodologia das pesquisas em Sociolinguística.

A exemplo dos bancos de dados de fala bem como da divulgação dos protocolos dos respectivos bancos no contexto internacional, defendemos uma publicação mais detalhada dos procedimentos metodológicos de coleta e sistematização dos dados de amostras sociolinguísticas. São pequenos passos rumo à aderência da área aos preceitos de Ciência Aberta.

5 Considerações finais

Apresentamos, no decorrer deste texto, um ponto de vista que defende a adoção de preceitos de Ciência Aberta para a Sociolinguística Variacionista. Nosso argumento partiu da constatação de que o conhecimento gerado por essa área do conhecimento não tem chegado à população geral. Portanto, apresentamos a importância dos bancos de dados linguísticos em diversas áreas, os preceitos de Ciência Aberta, um panorama acerca do compartilhamento e autoria de dados sociolinguísticos no Brasil e pequenas ações que podem ser feitas para um movimento de atender a alguns preceitos da Ciência Aberta. Enquanto ainda não há um repositório consorciado da área, acreditamos que a disseminação do conhecimento produzido pela Sociolinguística Variacionista pode se tornar mais forte por meio do compartilhamento de dados e de mais detalhes metodológicos sobre os processos de constituição de amostras e análises dos dados coletados.

Referências:

ARAÚJO, A. A. o projeto norma oral do português popular de fortaleza NORPOFOR. In: XV CONGRESSO NACIONAL DE LINGUÍSTICA E FILOLOGIA, v. 15, n. 5, 2011, Rio de Janeiro. **Anais [...]** Rio de Janeiro: CiFEFiL, 2011. p. 835-845. Disponível em: https://url.gratis/h5Lgdm. Acesso em: 23 ago. 2019.

ARAUJO, S. S. F.; ALMEIDA, N. L. F. O Projeto A língua portuguesa no semiárido baiano – Fase 3: critérios de constituição e da amostragem do banco de dados. In: FREITAG, R. M. KO. (Org.). **Metodologia de Coleta e Manipulação de Dados em Sociolinguística**. São Paulo: Blücher, 2014. p. 27-48. Disponível em: https://url.gratis/fUwn0w. Acesso em: 05 ago. 2020.

ALUÍSIO, S., PELIZZONI, J., MARCHI, A.R., DE OLIVEIRA, L., MANENTI, R., MARQUIAFÁVEL, V. 2003. An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese. In: Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language (PROPOR). **Proceedings [...]** PROPOR, 2003.

BAILEY, G.; TILLERY, J. Some Sources of Divergent Data in Sociolinguistics. In: FOUGHT, C. (Ed.) **Sociolinguistic Variation: Critical Reflections**. Nova Iorque: Oxford University Press, 2004. p. 11-30.

BICK. E. **The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. Tese (Doutorado) - Departamento de Linguística, Universidade de Aarhus, 2000.

BRASIL. Ministério da Educação. Edital de convocação Nº 01/2021 – CGPLI. [EDITAL DE CONVOCAÇÃO PARA O PROCESSO DE INSCRIÇÃO E AVALIAÇÃO DE OBRAS DIDÁTICAS, LITERÁRIAS E PEDAGÓGICAS PARA O PROGRAMA NACIONAL DO LIVRO E DO MATERIAL DIDÁTICO - PNLD 2023. Diário Oficial da União, seção 3, Brasília, DF, n. 30, p. 47, 12 fev 2021.

BRASIL. **Base Nacional Comum Curricular**. Brasília: MEC, 2017.

BRASIL. **Parâmetros Curriculares Nacionais: Terceiro e quarto ciclos do ensino fundamental**. Língua Portuguesa. Brasília: MEC, 1998.

BURKE, M.; CHELLIAH, S. L. Challenges to Representing Personal Names and Language Names in Language Archives: Examples from Northeast India. In: ZAVALINA, O. L.; CHELLIAH, S. L. (Eds.) International Workshop on Digital Language Archives, LanArc2021, 2021, Barcelona. **Proceedings [...]** [s. l.] University of Texas, 2021. p. 44-46. Disponível em: http://hdl.handle.net/2142/111675. Acesso em: 04 out. 2021.

CALAMAI, S. FRONTINI, F. FAIR data principles and their application to speech and oral archives. **Journal of New Music Research**, v. 47, n. 4, 339–354, 2018. Disponível em: https://www.tandfonline.com/doi/full/10.1080/09298215.2018.1473449?scroll=top&needAccess=true. Acesso em: 30 jun. 2020.

CARVALHO, A. M. Corpus del Español en el Sur de Arizona (CESA). University of Arizona. 2012. Disponível em: cesa.arizona.edu. Acesso em: 12 set. 2020.

CASTILHO, A. T. **Gramática do português brasileiro: fundamentos, perspectivas. Cadernos de Linguística**, v. 2, n. 1, p. 01-17, 2021. Disponível em: https://cadernos.abralin.org/index.php/cadernos/article/view/252. Acesso em 08 out. 2021.

CESARINO, L. Pós-verdade e a crise do sistema de peritos: uma explicação cibernética. **Ilha Revista de Antropologia**, v. 23, n. 1, p. 73-96, 2021. Disponível em: https://periodicos.ufsc.br/index.php/ilha/article/view/75630. Acesso em: 16 mar. 2023.

COLLISCHONN, G.; MONARETTO, V. O. Banco de dados V ARSUL: a relevância de suas características e a abrangência de seus resultados. **ALFA: Revista de Linguística**, v.56, n.3, p. 835-853, 2012. Disponível em: https://periodicos.fclar.unesp.br/alfa/article/view/4953. Acesso em: 20 abr. 2019.

FIDLER, F.; WILCOX, J. Reproducibility of Scientific Results. In: ZALTA, E. N. (Ed.) **The Stanford Encyclopedia of Philosophy**. Disponível em: https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/. Acesso em: 08 jul. 2019.

FREITAG, R. M. K., Linguistic Repositories as Asset: Challenges for Sociolinguistic Approach in Brazil. In: In: ZAVALINA, O. L.; CHELLIAH, S. L. (Eds.) International Workshop on Digital Language

Archives, LanArc2021, 2021, Barcelona. **Proceedings [...]** [s. l.] University of Texas, 2021b. p. 33-35. Disponível em: http://hdl.handle.net/2142/111675. Acesso em: 04 out. 2021.

FREITAG, R. M. K. Amostras sociolinguísticas: probabilísticas ou por conveniência? **Revista de Estudos da Linguagem**, v. 26, n. 2, p. 667-686, 2018. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/12412/0. Acesso em: 10 set. 2021.

FREITAG, R. M. K. **Documentação sociolinguística: coleta de dados e ética em pesquisa**. São Cristóvão: Editora UFS, 2017a Disponível em: https://www.livraria.ufs.br/produto/documentacao-sociolinguistica-coleta-de-dados-e-etica-em-pesquisa/. Acesso em: 10 jun. 2020.

FREITAG, R. M.K. A dadidade (ou dadidão) do dado, **Linguística Rio**, v.3, n.1, p.1-10, 2017b

FREITAG, R. M. K. Sociolinguística no/do Brasil. **Cadernos de Estudos Linguísticos**, v. 58, n. 3, p. 445-460, 2016.

FREITAG, R. M. K.; MARTINS, M. A.; TAVARES, M. A. FREITAG, R. M. K.; MARTINS, M. A.; TAVARES, M. A. Banco de dados sociolinguísticos do português brasileiro e os estudos de terceira onda: Potencialidades e limites. **ALFA: Revista de Linguística**, v.56, n.3, p. 917-944, 2012. Disponível em: https://www.scielo.br/pdf/alfa/v56n3/a09v56n3. Acesso em: 16 set. 2020.

FREITAG, R. M. K.; ROST SNICHELOTTO, C. A. Análises contrastivas: estabilidade, variedade ou metodologia?. **Working Papers em Linguística**, v. 16, n. 1, p. 157-167, 2015. Disponível em: https://periodicos.ufsc.br/index.php/workingpapers/article/view/1984-8420.2015v16n1p157. Acesso em 01 dez 2021.

FREITAG, R. M. K.; SANTANA, C. C.; ANDRADE, T. R. C. Práticas constitutivas do povoado Açuzinho. **Revista Ambivalências**, v. 2, n. 3, p.194-217, 2014. Disponível em: https://seer.ufs.br/index.php/Ambivalencias/article/view/3129. Acesso 12 set. 2021.

GALVES, C.; ANDRADE, A. L.; FARIA, P. **Tycho Brahe Parsed Corpus of Historical Portuguese**. 2017.

GARELLEK, M. et al. Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls. **Journal of Speech Science**, v. 9, n. 1, p.3-16, 2020. Disponível em: https://halshs.archives-ouvertes.fr/halshs-02894375/. Acesso em: 01 jun. 2021.

GONÇALVES, S. C. L. Projeto ALIP (Amostra Linguística do Interior Paulista) e banco de dados Iboruna: 10 anos de contribuição com a descrição do português brasileiro. **Estudos Linguísticos**, v. 48, n. 1, p. 276-29, abr. 2019. Disponível em: https://revistas.gel.org.br/estudos-linguisticos/article/view/2430/1503. Aceso em: 25 abr. 2019.

GILMORE, R.; KENNEDY, J. L.; ADOLPH, K. E. Practical solutions for sharing data and materials from psychological research. **Adv Methods Pract Psychol Sci**, v. 1, n. 1, p. 121-130, 2018. Disponível em: https://journals.sagepub.com/doi/full/10.1177/2515245917746500. Acesso em: 01 jun. 2021.

JUSTICE as a linguistic matter. Conferência apresentada por William Labov. [s.l., s.n], 2020. 1 vídeo (1h 06min 33s). Publicado pelo canal da Associação Brasileira de Linguística. Disponível em: https://www.youtube.com/watch?v=cr5tyw8_gT0&t=2231s. Acesso em: 23 mai. 2020.

KENDALL, T. Data in the Study of Variation and Change. CHAMBERS, J. K.; SCHILLING, N. (Eds.). **The handbook of language variation and change**. 2ª Ed. Malden: Wiley-Blackwell, 2013. p. 38-56.

KENDALL, T. On the History and Future of Sociolinguistic Data. **Language and Linguistics Compass**, v. 2, n. 2, p. 332-351, 2008. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/j.1749-818X.2008.00051.xAcesso em 20 mai. 2019.

KENDALL, T.; FRENCH, A. Digital Audio Archives, Computer-Enhanced Transcripts, and New Methods in Sociolinguistic Analysis. **Digital Humanities**, Paris, Jul., p. 110-112, 2006. Disponível em: https://slaap.chass.ncsu.edu/pdfs/DH2006_pp110-112.pdf. Acesso em: 20 mai. 2019.

LABOV, W. **Padrões sociolinguísticos**. Trad. Marcos Bagno, Maria Marta Pereira Scherre, Caroline Rodrigues Cardoso. São Paulo: Parábola Editorial, 2008. [1972]

LABOV, W. **The social stratification of English in New York city**. Cambridge University Press, 2006. [1966]

LABOV, W. **Principles of linguistic change: Social factors**. Malden: Blackwell, 2001. v.2

LABOV, W. Contraction, deletion, and inherent variability of the English copula. **Language**, v. 45, n. 4, p. 715-762, 1969. Disponível em: https://www.jstor.org/stable/412333. Acesso em: 13 jul. 2021.

LABOV, W. Field methods of the project on linguistic change and variation. **Sociolinguistic working paper**, n.81, p. 1- 43, 1981. Disponível em: https://eric.ed.gov/?id=ED250938. Acesso em: 13 jul. 2021.

MACHADO VIEIRA, M. S. et. al. **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 No. 01/2021, agosto de 2021.

MELLO, H. Trabalhando com dados de fala: a experiência do Projeto C-Oral-Brasil. In: BRESCANCINI, C. R. **Projeto VARSUL: Variação Linguística no Sul do País 36 anos**. Porto Alegre: Zouk Editora. p. 41-69.

OEZ, M. A guide to the documentation of the Beth Qustan dialect of the Central Neo-Aramaic Language Turoyo. **Language Documentation and Conservation**, v. 12, p. 339-358, 2018. Disponível em: https://scholarspace.manoa.hawaii.edu/handle/10125/24773. Acesso em: 15 set 2020.

PRESEEA: Corpus del Proyecto para el estudio sociolingüístico del español de España y de América. (2014) Alcalá de Henares: Universidad de Alcalá. Disponível em: http://preseea.linguas.net. Acesso em: 15 set. 2020.

RADEMAKER, A. et al. Universal Dependencies for Portuguese. In: Proceedings of the Fourth International Conference on Dependency (Depling). 2017. Pisa. **Proceedings [...]**. Pisa, 2017. p. 197-206.

SANTOS, D.; ROCHA, P. Evaluating CETEMPúblico, a free resource for Portuguese. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. 2001. Toulouse. **Proceedings [...]**. Toulouse, 2001. p. 442-449.

SETTE, G. et al. **Português: trilhas e tramas**. Vol 2. 2 ed. São Paulo: Leya, 2016.

SILVA, F.C.C.; SILVEIRA, L. O ecossistema da Ciência Aberta. **Transinformação**, v.31, e190001, 2019. http://dx.doi.org/10.1590/2318-889201931e190001. Disponível em: https://www.scielo.br/pdf/tinf/v31/2318-0889-tinf-31-e190001.pdf. Acesso em: 03 jul. 2020.

SILVA, G. M. O. Coleta de dados. In: MOLLICA, M. C.; BRAGA, M. L. (Orgs). **Introdução à Sociolinguística**. São Paulo: Contexto, 2004. p. 117-134.

SILVA, R. G.; COELHO, I. M. W. S. Metodologia de criação de um banco de dados linguísticos: desafios e contribuições para o processo de ensino-aprendizagem. **Educitec-Revista de Estudos e Pesquisas sobre Ensino Tecnológico**, v. 6, p. 1-15, 2020. Disponível em: http://200.129.168.14:9000/educitec/index.php/educitec/article/view/905. 10 out. 2021.

TAGLIAMONTE, S. A. **Variationist sociolinguistics: Change, observation, interpretation**. Malden: Wiley-Blackwell, 2012.

TAGLIAMONTE, S. A. Data, data and more data. In: TAGLIAMONTE, S. A. **Analysing sociolinguistic variation**. Cambridge: Cambridge University Press; 2006. p.50-69.

VANN, R. E. Best Practices for Information Architecture, Organization, and Retrieval in Digital Language Archives within University Institutional Repositories. In: ZAVALINA, O. L.; CHELLIAH, S. L. (Eds.) International Workshop on Digital Language Archives, LanArc2021, 2021, Barcelona. **Proceedings [...]** [s. l.] University of Texas, 2021. p. 36-39. Disponível em: http://hdl.handle.net/2142/111675. Acesso em: 04 out. 2021.

WEINREICH, U.; LABOV, W.; HERZOG, M. Empirical Foundations for a Theory of Language Change. In: LEHMANN, W. P.; MALKIEL, Y. (EDS.) **Directions for Historical Linguistics: A Symposium**. Austin: The University of Texas Printing Division, 1968. p. 95-188. Disponível em: https://liberalarts.utexas.edu/lrc/resources/books/directions/index.php. Acesso em: 02 jan 2021.

WILKINSON, M., et al. The FAIR Guiding Principles for scientific data management and stewardship. **Sci Data**, v.3, n. 160018, 2016. DOI https://doi.org/10.1038/sdata.2016.18. Disponível em: https://www.nature.com/articles/sdata201618. Aceso em: 30 jun. 2020.

ZEE, V. D. T; REICH, J. Open Education Science. **AERA Open**, v. 4, n.3, p. 1-15, 2018. DOI: 10.1177/2332858418787466.

TEMÁTICA LIVRE

Revista Diálogos (RevDia)

Sociolinguistic databases and Open Science: data and knowledge sharing

A B S T R A C T:

Sociolinguistic databases have been a referential source for the description of the different varieties of Brazilian Portuguese and have been affording other social functions. However, the knowledge produced by Variationist Sociolinguistics has not been recognized in Brazilian scenario, as Labov (2020) argued about the contributions of Sociolinguistics in the United States. This fact can be observed through the mechanist examples of semantic variation in the teaching of Portuguese language, as well as the visibility received by "internet linguists" that disseminate linguistic prejudice. Therefore, in this text, we present a reflection about the adoption of a few principles of Open Science may contribute to the strengthening of the area in Brazilian society.

KEYWORDS:

Sociolinguistics;
Open Science;
Sociolinguistic databases;