CADERNOS
de_LINGUÍSTICA

cadernos.abralin.org

RELATO DE PESQUISA

DESAFIOS DA **GESTÃO DE DADOS LINGUÍSTICOS** E A CIÊNCIA ABERTA

Raquel Meister Ko. FREITAG
Universidade Federal de Sergipe (UFS)

Marco Antonio Rocha MARTINS
Universidade Federal de Santa Catarina (UFSC)

Aluiza ARAÚJO
Universidade Federal do Ceará (UFC)

Elisa BATTISTI
Universidade Federal do Rio Grande do Sul (UFRS)

Iandra Maria Weirich da Silva COELHO
Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM)

Marta Deysiane Alves Faria SOUSA
Universidade Federal de Sergipe (UFS)

Raimundo Gouveia da SILVA
Instituto Federal de Educação, Ciências e Tecnologia do Acre (IFAC)

Rodrigo Esteves de LIMA-LOPES
Universidade Estadual de Campinas (UNICAMP)

OPEN ACCESS

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

AVALIADORES
– Sebastião Gonçalves (UNESP)
– Jacyra Mota (UFBA)

SOBRE OS AUTORES
– Raquel Meister Ko. Freitag
Conceptualização; Escrita – Rascunho Original, Análise e Edição.
– Marco Antonio Rocha Martins
Conceptualização; Escrita – Rascunho Original, Análise e Edição.
– Aluiza Araújo
Conceptualização; Escrita – Rascunho Original; Recursos.
– Elisa Battisti
Conceptualização; Escrita – Rascunho Original; Recursos.
– Iandra Maria Weirich da Silva Coelho
Conceptualização; Escrita – Rascunho Original; Recursos.
– Marta Deysiane Alves Faria Sousa
Conceptualização; Escrita – Rascunho Original; Escrita – Análise e Edição.
– Raimundo Gouveia da Silva
Escrita – Rascunho Original.
– Rodrigo Esteves de Lima Lopes
Conceptualização; Escrita – Rascunho Original; Escrita – Análise e Edição.

DATAS
– Recebido: 24/09/2020
– Aceito: 05/11/2020
– Publicado: 26/04/2021

COMO CITAR
FREITAG, R. M. K.; MARTINS, M. A. R.; ARAÚJO, A.; BATTISTI, E.; COELHO, I. M. W. DA S.; SOUSA, M. D. A. F.; SILVA, R. G. DA; LIMA-LOPES, R. E. Desafios da gestão de dados linguísticos e a Ciência Aberta. *Cadernos de Linguística*, v. 2, n. 1, p. 01-19.

RESUMO

O simpósio *Descrição linguística: gestão de dados linguísticos* teve como proposta retomar questões específicas ao gerenciamento de dados linguísticos, quer de fala quer de textos escritos, da atualidade ou históricos, em função das demandas latentes, especialmente face a

exigências como as da Ciência Aberta: i) Como atender aos princípios de ciência aberta quanto ao armazenamento, reuso e autoria de conjuntos de dados linguísticos? ii) Como lidar com a tensão entre a transparência e o sigilo de dados de fala? iii) Quais os formatos e as ferramentas mais adequados para a vitalidade dos conjuntos de dados linguísticos? iv) Quais ferramentas permitem o melhor armazenamento e sistemas de interface para consulta e pesquisa? Neste artigo, respondemos a estas a questões com o objetivo de motivar a discussão e o compartilhamento de boas práticas com a comunidade científica e sinalizamos as ações propositivas de natureza coletiva: i) a criação de políticas específicas da área para a replicabilidade dos estudos; ii) a adoção dessas políticas por programas de pós-graduação e periódicos; e iii) a criação e manutenção de repositórios de dados.

ABSTRACT

The symposium *Linguistic description: management of linguistic data* aimed to discuss issues regarding the management of linguistic data in its many formats: oral or written, contemporaneous or historical. Such symposium had its central motivation in the latent demands, especially those imposed by the Open Science movement. The participants' plenaries centred on four main questions: i) How to fulfil the principles of open science concerning the storage, reuse and authorship of linguistic data sets? ii) How to deal with the tension between transparency and confidentiality of language data? iii) Which formats and tools are best suited to make linguistic data sets viable? iv) Which would be best systems and interfaces for storing and accessing language data? In this article, we answered these questions with the aim of motivating the discussion and good practices with the scientific community. As an outcome, the participants proposed that some collective measures to ways to promote good practices regarding data use management: i) developing specific protocols in the area, encouraging studies' replicability; ii) engaging graduate programs and scientific journals in such protocols; and iii) creating and maintaining data repositories.

PALAVRAS-CHAVE

Dados Linguísticos; Ciência Aberta; Banco de Dados.

KEYWORDS

Linguistic Data; Open Science; Data Set.

INTRODUÇÃO

A descrição linguística requer um grande volume de dados para caracterizar fenômenos e dar suporte a generalizações, para validar teorias e definir normas de referência. A já clássica afirmação de Labov (1994, p. 11) no labor da linguística histórica sobre "a arte de fazer bom uso de dados ruins" pode ser estendida a todos os conjuntos de dados de quem trabalha com descrição e análise linguística: nem sempre dispomos de dados de qualidade (dados linguísticos coletados em laboratório vs. dados de campo), e nem sempre os dados são facilmente acessíveis (organização em repositórios assistemáticos, sem ferramentas de busca consolidada vs. anotação e alinhamento dos dados e busca automática). Ademais, se no início das pesquisas na área de linguística até final do século passado, as limitações para constituir amostras de dados, especialmente de fala espontânea, eram decorrentes da tecnologia ainda incipiente, hoje, as limitações são decorrentes de restrições impostas por comitês de ética em pesquisa. E, ainda, a política de transparência com dados abertos requer condições de armazenamento e disponibilização, e novos desafios precisam ser superados, como os custos de manutenção, o sigilo e direito de uso dos dados, dentre outros. Dados linguísticos são base para desenvolvimento de tecnologias assistivas, o que coloca a constituição e a gestão de dados linguísticos em alinhamento com as áreas estratégicas para a ciência no Brasil.

Diferentes perspectivas de análise linguística beneficiam-se de dados armazenados e disponibilizados em *corpora*. A coleta de dados linguísticos provenientes de situações autênticas de uso tem sido basilar em pesquisas em abordagens baseadas no uso, como a Sociolinguística (LABOV, 1972), a Linguística de Corpus (MCENERY; HARDIE, 2012; SINCLAIR, 1991), as diferentes correntes do Funcionalismo, dentre outras. Mais recentemente, tal relação se ampliou para a análise de interações em espaços tecnológicos, especialmente em mídias sociais no âmbito da Comunicação Mediada por Computadores (CAMERON; PANOVIĆ, 2014; EMIGH; HERRING, 2005) e da Ciência das Redes (BARABÁSI, 2002; SCOTT, 2000; WATTS, 2004). Em tais perspectivas, a pesquisa baseada em dados empíricos é uma forma de levantar padrões linguísticos e interacionais, que podem ser utilizados como indicativos de questões e representações dos diferentes atores sociais (VAN LEEUWEN, 2008). A utilização de dados é determinante nesse tipo de pesquisa, que se caracteriza por análises exploratórias e empiricamente fundamentadas.

Diferentemente de abordagens que se valem da intuição ou da introspecção, os resultados de descrições empíricas de padrões linguísticos emergem dos levantamentos realizados em *corpora*, que podem tanto ocorrer a partir de categorias pré-estabelecidas (como constituintes fonológicos, morfossintáticos, classes gramaticais ou palavras de valor semântico etc.), como também pela exploração das diferentes relações evidenciadas pelos

próprios dados. As unidades de análise podem ser, por exemplo, as da léxico-gramática (STUBBS, 2001), como a instanciação de categorias em um conjunto de textos (BIBER, 2015); ou, como no estudo de questões interacionais, fundamentadas em uma análise que parte do estabelecimento de redes de interação, tomando-se como base o papel que cada indivíduo pode assumir nas diferentes redes de que participa.[1]

Na programação do evento *Abralin ao Vivo*, diversos pesquisadores trataram da gestão de dados linguísticos, como nas atividades *Grandes Projetos da Linguística em rede: Tycho Brahe, PROHPOR e PHPB* e *Archiving and Language Documentation*, com relatos de experiências e compartilhamento de práticas. No campo da sociolinguística brasileira, a discussão também tem sido travada no âmbito do GT de Sociolinguística, com um simpósio específico sobre o tema no 1º Fórum Internacional de Sociolinguística, e publicações voltadas para a prática (FREITAG; MARTINS; TAVARES, 2012; FREITAG, 2014a, 2015, 2017a,b).

O simpósio *Descrição linguística: gestão de dados linguísticos* no *Abralin Ao Vivo*, que reuniu os autores do presente artigo, teve como proposta retomar questões específicas ao gerenciamento de dados linguísticos, quer de fala quer de textos escritos, da atualidade ou históricos, em função das demandas latentes, especialmente face a exigências como as da Ciência Aberta (seção 2). Com base na experiência dos participantes do simpósio em composição e gestão de bancos de dados, discutiram-se estratégias para superar os seguintes desafios:

- Como atender aos princípios de ciência aberta quanto ao armazenamento, reuso e autoria de conjuntos de dados linguísticos?
- Como lidar com a tensão entre a transparência e o sigilo de dados de fala?
- Quais os formatos e as ferramentas mais adequados para a vitalidade dos conjuntos de dados linguísticos?
- Quais ferramentas permitem o melhor armazenamento e sistemas de interface para consulta e pesquisa?

As reflexões do simpósio foram sistematizadas e repercutidas (CARDOSO, 2020), e são apresentadas a seguir, no intuito de motivar a discussão e o compartilhamento de boas práticas com a comunidade científica.

1 Aqui, importantes métricas podem ser estabelecidas a partir das conexões e relações topográficas apresentadas pela análise de redes (SCOTT, 2013), as quais podem representar nossa assinatura social (WATTS, 2003).

1. MOVIMENTO CIÊNCIA ABERTA: USO E REUSO DE DADOS

Ciência Aberta é um movimento de ruptura com um aspecto da ciência tradicional, a circulação das ações e resultados da pesquisa, estrita ao ambiente acadêmico. De acordo com Silva e Silveira (2019), essa ruptura incentiva a transparência, desde a concepção da pesquisa, com a publicação aberta do projeto de pesquisa, por exemplo, até seu produto, como o acesso a dissertações e teses, bem como à publicação dos resultados em periódicos científicos de acesso gratuito. Busca-se também por meio da Ciência Aberta um maior detalhamento de metodologias e gerenciamento de dados, de forma que eles possam ser acessados por toda a sociedade.

No escopo da Ciência Aberta, o gerenciamento dos dados deve seguir os princípios nomeados pelo acrônimo *FAIR* (WILKINSON *et al.,* 2016), no qual *F* significa *findable* (passível de ser encontrado), *A* significa *accessible* (acessível), I, *interoperable* (interoperável) e *R*, *reusable* (reutilizável). Para a discussão dos desafios que identificamos para a gestão de dados, nos detemos em dois desses princípios: ser acessível e ser reutilizável. 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 de os dados não estarem mais disponíveis, seus metadados sejam ainda acessíveis (WILKINSON *et al.,* 2016). Já para que um conjunto de dados seja considerado reutilizável, 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.

Ressaltamos, contudo, que os princípios de acessibilidade e reutilização devem ser considerados com cautela, observando sempre os aspectos éticos preconizados por documentos regulatórios no âmbito de cada país. No Brasil, por exemplo, os padrões no que tange a aspectos éticos em pesquisa com seres humanos nas ciências sociais atualmente são estabelecidos pela Resolução N° 510, de 07 de abril 2016, do Conselho Nacional de Saúde. No artigo 3°, inciso VII, deste documento, é estabelecido que deve haver "garantia da confidencialidade das informações, da privacidade dos participantes e da proteção de sua identidade, inclusive do uso de sua imagem e voz" (BRASIL, 2016, p. 45). Além disso, em seu artigo 9°, inciso V, a resolução (BRASIL, 2016) deixa claro que é direito do participante decidir se sua identidade poderá ser divulgada, e também quais das informações coletadas pelos pesquisadores poderão estar disponíveis ao público. Percebemos por esses artigos a existência de regras que estabelecem como os dados devem ser geridos pelos pesquisadores, assegurando ao participante total controle sobre os dados por ele fornecidos para a condução da pesquisa.

A preocupação com a confidencialidade dos participantes nas pesquisas envolvendo dados linguísticos tem sido abordada em diversos estudos tanto no Brasil (PAIVA, 2005; ABREU, 2014; FREITAG, 2017) quanto no exterior (CALAMAI; FRONTINI, 2018; CASILLAS; CRISTIA, 2019; CHILDS; VAN HERK; THORBURN, 2011). Em comum, todos eles observam as formas de obter os dados de fala e reportá-los em suas pesquisas, mantendo o anonimato dos participantes. O que tem ganhado mais notabilidade é a disponibilização desses dados, visto que, embora possa ser possível apagar dados sensíveis, como nome do entrevistado, lugares e nomes de pessoas de fácil identificação, existe ainda a possibilidade de identificação do participante pela voz. Ao discutir a aplicação dos padrões *FAIR*, Calamai e Frontini (2018), por exemplo, como uma forma de minimizar os danos na disponibilização de dados de fala e considerando o reconhecimento dos participantes pela voz, apontam as licenças de uso e o termo de consentimento como possíveis soluções. Num repositório como o *CLARIN* (*Common Language Resource Infrastructure for Social Sciences and Humanities*), os dados são disponibilizados conforme licenças de uso. No *CHILDES* (*Child Language Data Exchange System*) *corpora*, houve um planejamento para se chegar a um termo de consentimento e se escolherem licenças de uso adequadas, possibilitando inclusive a disponibilização dos dados para a comunidade em geral, mesmo em se tratando de dados obtidos de crianças, que normalmente, como ressaltam as autoras, são mais delicados no que tange a questões legais e éticas.

No Brasil, Freitag (2017a) apresenta detalhadamente os procedimentos de coleta de dados em entrevistas sociolinguísticas, bem como modelos de termo de consentimento e assentimento livres e esclarecidos. Na redação de ambos, consta o fato de que as interações gravadas serão disponibilizadas para a comunidade acadêmica e também armazenadas em uma base de dados. Essa observação, contida nos termos de consentimento e assentimento, evidencia indícios de uma boa prática a ser padronizada para todos aqueles que se interessarem em participar de uma gestão de dados de fala consorciada, visto que, em conjunto com licenças apropriadas a serem utilizadas, fazem uma combinação semelhante às apresentadas por Calamai e Frontini (2018) acerca da disponibilização de dados de fala em conformidade com os padrões *FAIR* (WILKINSON *et al.*, 2016).

Apesar da crescente visibilidade que áreas como a Sociolinguística e a Linguística de *Corpus* dão ao trabalho com dados empíricos e, consequentemente, ao uso de ferramentas específicas para pesquisa – como pacotes estatísticos ou programas de análise de *corpora* – (FINARDI *et al.*, 2020), a reflexão aprofundada sobre metodologias de coleta, processamentos e arquivamento de dados são raras. De maneira geral, há a ideia de que estudos qualitativos de dados empíricos não seriam reproduzíveis, dada sua ancoragem nos processos subjetivos de construção da análise e do objeto.

Por outro lado, Biber *et al.* (1998) defendem que a interpretação dos dados sempre ocorre a partir de uma perspectiva qualitativa, porque o significado das diferentes categorias poderia apenas ser definido pelo pesquisador, que parte do seu contato com o contexto (e com o texto) como parâmetro analítico. De fato, observamos que tal perspectiva acaba por se inserir naquilo que Johnson, Onwuegbuzie e Turner (2007) definem como métodos mistos, ou seja interpretações subjetivas que partem de dados verificáveis e replicáveis por outros pesquisadores.

No Brasil, o grupo de pesquisa *Mídia, Discurso, Tecnologia e Sociedade* (*MiDiTeS*) é um dos poucos que promove essa reflexão acerca de metodologias, processamento e arquivamento de dados linguísticos. As constantes discussões giram em torno de temáticas que podem ser relacionadas aos princípios *FAIR* da Ciência Aberta e se situam em dois eixos inter-relacionados: 1) Letramentos de Dados e 2) Ativismos. No caso do primeiro, destaca-se a preocupação em refletir sobre processos de interação que revelem questões estruturais em nossa sociedade por meio da compreensão, compilação e estrutura de dados. Letrar o indivíduo nesses termos é propor uma abordagem crítica aos diferentes sistemas de dados abertos de forma a possibilitar a criação de novas realidades epistemológicas e sociais distintas (GUTIÉRREZ, 2019; WOLFF *et al.*, 2016). No caso do segundo, o enfoque está em entender como a sociedade datificada leva à construção de diferentes formas de interação e propagação ideológica. Isso pode contemplar desde a construção de identidades de resistência (GABARDO; LIMA-LOPES, 2018; LIMA-LOPES; GABARDO, 2019; LIMA-LOPES; PIMENTA, 2017), passando pela propagação de ideologias conservadoras (LIMA-LOPES, 2018) ou mesmo pelo uso dos processos de Comunicação Mediada por Computadores como forma de controle social (MERCURI; LIMA-LOPES, 2020).

A ausência de discussão aprofundada sobre coleta, processamentos e arquivamento de dados, como a promovida pelo *MiDiTes*, tem duas grandes consequências. A primeira atinge os processos de replicabilidade de pesquisa. King (1995), por exemplo, evidencia que a possibilidade de replicar um estudo tem impacto direto em, pelo menos, três níveis: 1) proteção contra erros honestos e falsificação de resultados; 2) facilitação de processos pedagógicos e 3) acúmulo substantivo de conhecimento. Apesar das diferenças entre as áreas, King (1995) trabalha no espectro das ciências humanas, com o qual as ciências da linguagem dividem um lastro comum. A segunda grande consequência da ausência frequente de reflexões sobre a coleta e análise de dados pode ser observada no nível do ensino, nos cursos de graduação e pós-graduação na área: mesmo quando tal discussão está presente, ela é ainda pouco valorizada, e, muitas vezes, tratada por alunos e por parte do corpo docente como menos importante ou relevante.

Orientar-se pelos princípios *FAIR* da Ciência Aberta e promover a discussão sobre a coleta e manuseio dos dados traz benefícios para pesquisadores e sociedade, dos quais

destacamos três: i) a conscientização da necessidade de discussões metodológicas claras faz com que o pesquisador cresça em segurança ao afirmar seus resultados e suas assunções sobre os dados; ii) a possibilidade de análise por diferentes perspectivas de um mesmo conjunto de dados pode fazer crescer substancialmente nosso conhecimento sobre um determinado registro, variedade linguística, contexto interacional, entre muitos outros; iii) a possibilidade de refazimentos dos experimentos em contextos como o da sala de aula contribui não apenas para a formação científica dos nossos alunos, mas também para a sedimentação de conhecimentos em nível teórico e prático. Afinal, conforme Meie (1995), a replicabilidade é algo importante tanto para trabalhos quantitativos, como também qualitativos, por fazer crescer o cruzamento entre as diversas intersubjetividades.

2. COLETA DE DADOS LINGUÍSTICOS EM SITUAÇÕES DE USO: PRÁTICAS E DESAFIOS

Apresentamos a seguir duas experiências acerca da coleta de dados aos moldes da Sociolinguística Variacionista, abordagem que tradicionalmente se ampara em dados de fala produzidos em situação de entrevista, conforme o protocolo de entrevista sociolinguística (FREITAG, 2014b, 2016). Uma das coletas segue a rotina já tradicionalmente consolidada para a área, com estratificação homogênea (amostra *PORCUFORT* fase 2), e outra parte para uma estratificação que amplia as categorias sociodemográficas incluídas na documentação (*LínguaPOA*).

Com base na metodologia já tradicionalmente consolidada, especialmente pelo projeto *Norma Urbana Culta* (*NURC*), que envolveu UFRGS, USP, UFRJ, UFBA, UFPE nos anos 1970, a fase II do projeto *Descrição do Português Oral Culto de Fortaleza* (*PORCUFORT*) compreende a constituição de uma amostra linguística no período de 2018-2021 (a fase I compreende o período de 1993 a 1995). A fim de garantir a comparabilidade das amostras e possibilitar estudos de mudança linguística em tempo real, a nova fase segue a mesma estratificação do *PORCUFORT* – fase I: sexo (masculino e feminino), faixa etária (faixa I - 22 a 35 anos, faixa II - 36 a 55 anos e faixa III - a partir dos 56 anos) e tipo de registro (Diálogo entre Informante e Documentador: DID, Diálogo entre Dois Documentadores: D2 e Elocução Formal: EF). Os participantes da amostra do *PORCUFORT* - fase II, a exemplo do que ocorreu na fase I, apresentam as seguintes características: i) por se tratar de um banco de dados de fala culta, todos os falantes devem ter o nível superior completo em regime presencial; ii) além disso, os falantes devem ser nascidos em Fortaleza-CE; iii) filhos de pais fortalezenses, ou que seus pais, sendo cearenses, tenham vindo morar na cidade com até cinco anos de idade; iv) todos devem residir na capital cearense; além de não terem se afastado da capital por tempo superior a três meses, ou seja, são admitidos falantes que tenham viajado para outras

localidades apenas a passeio, pois um período superior a esse pode interferir na fala original do indivíduo. A amostra está estratificada homogeneamente de acordo com as variáveis sociodemográficas sexo e faixa etária, e o tipo de registro (DID, D2 e EF).

Como outros projetos de variação linguística baseados em amostras de fala socialmente estratificadas, o *LínguaPOA* (UFRGS, https://www.ufrgs.br/linguapoa/) resultou de um projeto de pesquisa, intitulado *Variação fonético-fonológica e classe social na comunidade de fala de Porto Alegre* (2015-2019), que buscou: i) investigar a configuração da comunidade de fala de Porto Alegre em termos de normas e características linguísticas partilhadas nas diferentes áreas do espaço geográfico e em seus estratos sociais; ii) esclarecer a estruturação de Porto Alegre em classes sociais, relacionando distinção social à padronização e distinção linguística; iii) verificar as variáveis linguísticas e extralinguísticas condicionadoras da aplicação de processos fonético-fonológicos variáveis no português brasileiro falado em Porto Alegre.

As 103 entrevistas, realizadas em português entre agosto de 2015 e setembro de 2019, contemplam narrativas de experiência pessoal, descrições e apreciações de lugares de Porto Alegre e da vida na cidade, feitas por sujeitos que nasceram em Porto Alegre ou se mudaram para a cidade até 5 anos de idade, tendo nela vivido desde então. Os critérios de estratificação do LínguaPOA são: i) quatro zonas: Centro (Central), Norte, Sul, Leste; ii) dois bairros por zona: renda alta e renda baixa (por renda domiciliar média mensal em salários mínimos); iii) três grupos etários: 20-39 anos, 40-59 anos, 60 ou mais anos; iv) três níveis de escolaridade: fundamental, médio, superior; v) dois gêneros: masculino e feminino. As 103 entrevistas atendem a todos os critérios de estratificação nos níveis médio e superior de escolaridade, mas não no fundamental.

A constituição do *LínguaPOA* conciliou técnicas da Dialetologia Perceptual de Preston (1989) às habitualmente seguidas na constituição de amostras para estudos sociolinguísticos variacionistas, conforme Labov (1972, 1994), e descritas em publicações brasileiras como as de Tarallo (2006), Guy e Zilles (2007), Freitag (2014). Além disso, assumiu o compromisso de comparabilidade, desde que possível, com amostras anteriores da fala de Porto Alegre, como as do *NURC*, dos anos 1970, e do projeto *VARSUL* (Variação Linguística na Região Sul do Brasil), que envolveu UFRGS, UFSC, UFTPR, PUCRS, nos anos 1990. A peculiaridade do *LínguaPOA* está na possibilidade de viabilizar estudos sobre os efeitos de classe social na variação linguística no português falado em Porto Alegre: contempla não só a variável escolaridade, mas também zona de residência e bairro por zona, este selecionado conforme a renda média domiciliar mensal (alta, baixa) das famílias residentes no bairro. Ademais, o *LínguaPOA* dispõe de informações complementares, registradas na Ficha de Entrevista (profissão e ocupação do participante, a renda mensal de sua família) e no Questionário Econômico, conforme o Critério Brasil de classificação econômica (http://www.abep.org/criterio-brasil).

A constituição de ambas as amostras de dados linguísticos apresentadas foi permeada por desafios. Na fase de realização das entrevistas para a amostra da fase 2 do *PORCUFORT*, em vários momentos, a equipe defrontou-se com o seguinte problema: como lidar com a tensão entre a transparência e o sigilo? Partindo da experiência com o *PORCUFORT*, algumas estratégias para minimizar este problema, já previstas na literatura sociolinguística, foram adotadas, tais como:

- não revelar, de imediato, a real motivação das entrevistas;
- mencionar que a pesquisa tem o objetivo de documentar a memória do fortalezense acerca de sua cidade, desfocando a atenção do entrevistado acerca de sua própria fala;
- saber conduzir a entrevista (de forma descontraída, perguntas nunca formuladas antecipadamente, mas elaboradas de forma simples, clara e precisa);
- instigar os entrevistados a falar de fatos de suas vidas que, de algum modo, os marcaram de forma positiva ou negativa;
- usar gravadores que sejam não só potentes, mas também pequenos, portáteis, discretos e de fácil manuseio;
- fornecer ao documentador não só o Termo de Consentimento Livre e Esclarecido (TCLE), mas também uma carta de apresentação com assinatura e carimbo da pesquisadora responsável, além da assinatura do coordenador do curso de graduação do aluno voluntário ou bolsista;
- solicitar do participante, no próprio TCLE, liberação do uso da entrevista para fins científicos e de estudos (livros, artigos e *slides*), em favor dos pesquisadores, obedecendo ao que está previsto na Resolução 510/2016/CNS;
- pedir ao entrevistado autorização, no TCLE, para que a interação fique disponível no banco de dados acima referido para ser utilizada em pesquisas futuras;
- realizar a gravação no dia, local e hora da conveniência do participante;
- apresentar-se como aluno perante o entrevistado, gerando empatia com o participante;
- apresentar-se para o entrevistado como amigo ou conhecido de um familiar ou amigo/colega seu, promovendo aproximação entre o pesquisador e o participante;
- garantir que o entrevistado pode inutilizar a gravação ao seu término ou durante a entrevista; - não permanecer no recinto onde a gravação está sendo realizada, o que dá maior liberdade de fala aos participantes, no caso das entrevistas do tipo D2.

No caso do *LínguaPOA*, observar todos os critérios de amostragem revelou-se uma tarefa complexa, especialmente no que diz respeito à escolaridade, bairro por zona e idade: por exemplo, em centros urbanos como Porto Alegre, jovens apenas com nível fundamental de escolaridade tendem a habitar bairros periféricos e de baixa renda. Além disso, e provavelmente porque a equipe de pesquisadores acionou sua rede de contatos para recrutar participantes, apenas nos níveis médio e superior de escolaridade localizaram-se porto-alegrenses para preencher todas as células por grupo etário e gênero nos dois tipos de bairro em cada uma das quatro zonas. Esses fatos revelam, portanto, i) a não ortogonalidade entre escolaridade, zona e bairro por zona e ii) os limites impostos à amostra pelo recrutamento via rede de contatos dos pesquisadores. Mesmo assim, reforça-se a ideia de que o valor do *LínguaPOA* esteja em sua comparabilidade com bancos de dados anteriores (*NURC*, *VARSUL*) e em seu potencial para viabilizar a realização de análises de variação linguística e classe social.

Em relação à tensão entre a transparência e o sigilo, apontamos aqui algumas estratégias tomadas frente a esse desafio: as transcrições das entrevistas do *LínguaPOA* (com o *software* ELAN) e os respectivos áudios são anonimizados, ou seja, nomes de pessoas e de certos lugares que identificariam os participantes são trocados nas transcrições e silenciados nos arquivos de áudio, para atender a demandas éticas. Além das transcrições e áudios das entrevistas, compõem o acervo do *LínguaPOA* os documentos referentes a cada participante – Termo de Consentimento Livre e Esclarecido, Ficha de Entrevista, Questionário Econômico, estes dois últimos disponibilizados em versão anonimizada.

3. ARMAZENAMENTO, USO E REUSO

As amostras referidas na seção anterior foram resultado de desenvolvimento de projetos de descrição linguística. Outras amostras, também constituídas pensando na replicabilidade, são geradas em projetos de outra natureza. É o caso do *Banco de Dados Linguísticos Reci*, compilado durante os anos de 2016 e 2017.

O *Banco Reci* é composto por entrevistas com 36 falantes do núcleo da Reserva Extrativista Cazumbá-Iracema, localizado no Acre, em uma amostra estratificada quanto às variáveis sociodemográficas sexo, escolaridade e idade. Além de ser uma das poucas amostras linguísticas da região Norte do país, ainda subdocumentada no que diz respeito às variedades do português, este banco de dados tem a peculiaridade de ter sido gerado como produto educacional, oriundo de uma pesquisa realizada no Mestrado Profissional em Ensino Tecnológico, do Instituto Federal do Amazonas (SILVA, 2017). Sua organização serviu como material de referência para a elaboração de uma proposta de ensino e

aprendizagem dos paradigmas verbais, indicativo e subjuntivo da língua portuguesa, com ênfase no trabalho da variação linguística. Por tratar-se de uma modalidade profissional, a produção e aplicação dos conhecimentos são orientados para a prática e resolução de problemas, por meio do planejamento, implementação e avaliação de processos e produtos educacionais (BARROS; VALENTIM; MELO, 2005). Assim, o produto educativo deve ser utilizado "em condições reais de sala de aula ou de espaços não formais ou informais de ensino, em formato artesanal ou em protótipo" (BRASIL, 2013, p. 24-25).

Levando em conta a necessidade de uso e reuso do *Banco Reci*, ampliando assim o acesso aberto, esse produto foi organizado em dois formatos: um *corpus* oral e um escrito. A amostra oral possui cerca de 3h de gravações, digitalizadas e armazenadas eletronicamente, e conta com 36 falantes do núcleo da Reserva. O *corpus* escrito possui um acervo de 30.798 palavras, constituído por amostras de fala que foram gravadas, transcritas e, posteriormente, armazenadas eletronicamente. Posteriormente, a amostra foi compilada como uma versão do produto educacional em duas modalidades, restrita e irrestrita. Na modalidade restrita, em formato de DVD, há fichas técnicas específicas, contendo informações relevantes sobre o produto e permitindo consultas na biblioteca da instituição.

A outra versão do produto, de modalidade irrestrita, teve como objetivo principal ampliar a reutilização dos dados linguísticos em contextos de ensino e aprendizagem de Língua Portuguesa, de forma que o reuso desse produto educacional possa suscitar novas práticas pedagógicas, baseadas no fomento à utilização de dados reais. Para tanto, essa base foi transcrita, mantendo-se o anonimato e substituição dos nomes dos entrevistados e então disponibilizada no repositório institucional em formato de um produto educacional, com reconhecimento de produção técnica e de autoria pelos organizadores[2] (SILVA, R; COELHO, 2018).

Para ampliar a visibilidade, o impacto e o reuso, por meio do acesso aberto, esse e outros produtos que foram gerados a partir da pesquisa encontram-se disponíveis no *site:* http://ppget.ifam.edu.br/dissertacoes-defendidas/. Essa prática de compartilhamento público e aberto da amostra escrita compilada, desenvolvida e validada em formato de produto educacional, teve como objetivos: i) servir como fonte para a elaboração de processos e produtos didáticos relacionados ao ensino e aprendizagem de fenômenos da Língua Portuguesa; ii) contribuir para estudos descritivos da variação linguística, no âmbito das línguas naturais; iii) fomentar propostas diferenciadas de ensino e aprendizagem da Língua Portuguesa, inserindo no contexto de sala de aula novas possibilidades educativas com o uso de bancos de dados; iv) potencializar a preservação e a salvaguarda da

2 O corpus escrito pode ser acessado por meio do link: https://drive.google.com/file/d/1YekKfpwDUZ23VMZG-mOqE1PWZmUmOQ_w/view

realidade linguística brasileira e os aspectos sociais que a circunscrevem e v) constituir e ampliar os bancos de dados linguísticos com populações tradicionais.

Entre esses objetivos, leva-se em conta o efetivo trabalho com dados autênticos, ressaltando não apenas o potencial uso de dados linguísticos nas práticas de ensino e aprendizagem, mas a importância das variedades linguísticas e de aspectos como a valorização da região amazônica, com destaque especial para a relação estreita que há entre os povos tradicionais e o ecossistema Amazônico e a vinculação da população que habita tradicionalmente esse diverso ecossistema, evidenciando a representatividade de grupos sociais específicos.

Apesar de não terem sido concebidos como produtos, como a amostra *Banco Reci*, cuidados com reuso e autoria são presentes na gestão de dados decorrentes de coletas aos moldes sociolinguísticos, como o *PORCUFORT* e o *LínguaPOA*. Mesmo que via e-mail, é regra solicitar a autorização.

Diferentemente do *Banco Reci*, cujos produtos estão disponíveis no *site* da instituição de origem, o armazenamento de amostras como a do *PORCUFORT* e a do *LínguaPOA* é um desafio. Todo o material do *LínguaPOA*, assim como as demais amostras de dados linguísticos que conhecemos, vem sendo armazenado em *drives* custeados e mantidos pelos próprios pesquisadores, já que suas instituições não possuem espaço, quer nas máquinas físicas, quer nas virtuais, para armazenar e disponibilizar a um grande público essas amostras. Vêm daí boa parte dos desafios enfrentados na gestão de acervos como o *LínguaPOA* e por bancos de dados linguísticos em geral, o de manter e gerir os dados após a coleta e a realização dos primeiros estudos, diretamente vinculados à pesquisa dos responsáveis pelos projetos.

Há questões relativas à salvaguarda (armazenamento e proteção/preservação) das amostras e ao acesso aos dados que requerem solução. Sem amparo institucional, acervos como o *LínguaPOA* e o *PORCUFORT* podem ter seu uso restrito aos responsáveis e membros da equipe de pesquisa; sua preservação pode estar condicionada a possibilidades e interesses individuais. Isso faz com que os dados, que deveriam ser públicos, tornem-se exclusivos a um certo grupo de usuários; e cria o risco de "desaparecimento" dos acervos em função, por exemplo, de alterações nas iniciativas dos pesquisadores responsáveis ou de fatos prosaicos como sua aposentadoria, por exemplo. Desvirtua-se, desse modo, o propósito original dos bancos: servir de fonte permanente de dados para pesquisa linguística ou afins, disponíveis a toda a comunidade de pesquisadores.

Os mesmos problemas e questões encontrados na constituição de amostras e bancos de dados de fala são uma constante em *corpora* do português escrito em sincronias passadas: dificuldades de armazenamento, uso e reuso e livre acesso via rede mundial. Algumas amostras estão disponibilizadas na rede mundial, como as dos *Projeto para a*

História do Português Brasileiro (*PHPB*) (https://sites.google.com/site/corporaphpb/)[3] e os projetos estaduais vinculados ao *PHPB* nacional, incluindo o *Corpus do Laboratório de História do Português* (https://laborhistorico.letras.ufrj.br/) coordenado por Célia Lopes na UFRJ; *Programa para a História da Língua Portuguesa* (*PROHPOR*) (https://www.prohpor.org/bit-prohpor); *Corpus Eletrônico de Documentos Históricos do Sertão (CE-DOHS)* (http://www.tycho.iel.unicamp.br/cedohs/corpora.html); *Corpus Histórico do português Tycho Brahe* (http://www.tycho.iel.unicamp.br/corpus/) (e essa não é uma lista exaustiva, naturalmente).

Desses *corpora* históricos, é importante destacar a relevância e pioneirismo do *Programa para a História da Língua Portuguesa (PROHPOR)*, atualmente sob a coordenação de Tania Lobo, que foi fundado em 1990 e permaneceu por muitos anos sob a coordenação de Rosa Virgínia Mattos e Silva, na Universidade Federal da Bahia. Além do *PROHPOR*, outro importante projeto desenvolvido em universidades na Bahia, para documentação de textos históricos do português escrito no Brasil, é o *Corpus Eletrônico de Documentos Históricos do Sertão (CE-DOHS)*, em Feira de Santana/BA, coordenado por Zenaide Carneiro e Mariana Oliveira.

O *Corpus histórico Tycho Brahe,* fruto de um importante projeto coordenado por Charlotte Galves, está anotado sintaticamente com textos em português escritos por autores diversos nascidos entre 1380 e 1881 (http://www.tycho.iel.unicamp.br/corpus/). A plataforma desse *corpus* permite realizar buscas de estruturas sintáticas automáticas em grande quantidade de dados nos textos que estão anotados via programa *E-Dictor*, desenvolvido pelos pesquisadores do projeto Pablo Faria, Fábio Kepler e Maria Clara Paixão de Sousa.

O *corpus* em elaboração pelo projeto *PHPB* disponibiliza textos impressos e manuscritos escritos no Brasil dos séculos XIX e XX com livre acesso. De acordo com Barbosa (2019, p. 18), a plataforma *PHPB* apresenta "uma sistematização de todos os materiais editados pelos membros do Projeto até o Censo de *corpora do PHPB* em junho de 2010".

Outro *corpus* de língua escrita é o *PROCORP* (*Corpus de Redações do ProFIS*), ainda em processo de compilação. Iniciativa de Ines Signorini e Rodrigo Esteves de Lima-Lopes, tem por objetivo a categorização e disponibilização das redações realizadas por alunos do ProFIS, um programa de integração universitária de alunos oriundos das escolas de Campinas e mantido pela universidade desde 2011. Quando compilado, o corpus oferecerá a possibilidade de estudar o trajeto de formação e letramento acadêmico, com ênfase na aquisição escrita, de centenas de alunos durante os dois anos que participam do programa.

De um modo geral, esses *corpora*, com livre acesso, como dito, disponibilizam textos em português escritos em diferentes períodos por autores portugueses e brasileiros (quando

3 É importante dizer que os *corpora* do PHPB não estão mais disponíveis na Plataforma, acesso em 25 de nov. de 2020.

possível identificar os dados sociais) e estão organizados em gêneros textuais/discursivos vários. Mas um grande impasse para a pesquisa e coleta de dados linguísticos em número expressivo, na maioria desses *corpora*, é que eles armazenam textos em arquivos *.doc ou *.pdf que podem ser baixados, mas não permitem uma busca rápida de dados específicos, sistematizada por gênero textual/discursivo ou período. A exceção desse modo de armazenamento dos textos são os *Corpus Tycho Brahe* e o *DEDOHS*, que estão sintaticamente anotados e são os únicos, dos aqui citados, que disponibilizam uma versão dos documentos editados em *.xml. A amostra do *PHPB* é disponibilizada em arquivos, organizados em gêneros e regiões, que podem ser baixados e as buscas de dados devem ser manuais.

Com o objetivo de elaborar e disponibilizar uma plataforma interativa, com programas de interface e ferramentas para busca e tratamento de dados de textos escritos em português no Brasil, em diferentes sincronias, o projeto *PB-Corpus Histórico* está sendo desenvolvido na Universidade Federal de Santa Catarina, em parceria com a Universidade de Colônia/Alemanha. O objetivo dessa plataforma é armazenar e disponibilizar textos escritos no Brasil no curso dos séculos XVIII a XXI de diferentes corpora, de modo que interfaces interativas e ferramentas operacionais permitam buscas de dados para estudos de diferentes fenômenos linguísticos.

Na organização da plataforma, buscam-se reunir materiais de diferentes gêneros textuais/discursivos: da esfera dos textos impressos de jornais: anúncios, cartas de leitores e cartas de redatores/editoriais; da esfera dos textos manuscritos: cartas privadas; e da esfera dos textos literários: peças de teatro. Os textos estão organizados por coleções que preservam as informações das fontes de onde foram extraídos os textos e estão dispostos em: (a) séculos, separados por intervalo de tempo de 50 anos (1701 a 1750; 1751 a 1800; 1801 a 1850; 1851 a 1900; 1901 a 1950; 1951 a 2000; e a partir de 2001); e isso permite uma pesquisa de dados que considere a data de publicação de cada texto individual ou a data de nascimento dos autores (quando for o caso de essa informação estar disponível); (b) regiões e Estados, e, até o momento, conta com textos do Sul – Rio Grande do Sul, Santa Catarina e Paraná; Sudeste – São Paulo, Rio de Janeiro e Minas Gerais; e Nordeste – Bahia, Pernambuco, Rio Grande do Norte e Ceará; Norte – Amazonas e Pará; e Centro-Oeste – Mato Grosso do Sul.

O diferencial do *PB-Corpus Histórico* em relação às amostras citadas reside no fato de constituir uma plataforma de armazenamento que permite a interatividade no manuseio dos textos e buscas de dados linguísticos por sistemas de interfaces. Neste momento, em fase de testes, foi elaborado e implementado um programa de interface na Plataforma que permite a coleta de sentenças com pronomes clíticos, o que possibilita a coleta e categorização para análise de um número robusto de dados (MARTINS, 2020).

4. PERSPECTIVAS

A gestão dos dados linguísticos é uma questão em aberto. O simpósio *Descrição linguística: gestão de dados linguísticos* discutiu práticas e experiências em nível individual ou de grupos de pesquisa que, reunidas, indicam ações propositivas de natureza coletiva:

1) a criação de políticas específicas da área para a replicabilidade dos estudos;
2) a adoção dessas políticas por programas de pós-graduação e periódicos;
3) a criação e manutenção de repositórios de dados.

A primeira ação passaria por discussões de caráter abrangente entre as diferentes associações de pesquisa na área, de forma a construir padrões a serem seguidos pelos diferentes tipos de estudo, tanto qualitativo como quantitativo. A segunda refere-se à efetiva adoção dessas políticas pelos atores formadores de pesquisadores. A terceira trata da construção de repositórios e modelos de licença que permitam o compartilhamento dos dados. As associações de pesquisa teriam um papel seminal na execução dessas ações, por permitirem visualizar a abrangência e qualidade na distribuição dos dados.

Acervos como o *LínguaPOA*, *PORCURFORT* e outros tantos precisam ser articulados, e organizados em uma forma de gestão consorciada. A um só tempo, essa articulação daria visibilidade aos bancos de dados linguísticos ora existentes no Brasil, salvaguardaria os dados e viabilizaria formas de acesso a eles, garantindo seu adequado uso por um número amplo de pesquisadores. Sabemos, no entanto, que iniciativas como essas certamente teriam implicações de ordem operacional com repercussão financeira (profissionais dedicados, equipamentos etc.).

Um passo inicial rumo à gestão consorciada poderia ser providenciar o mapeamento dos bancos brasileiros de dados linguísticos, para traçar um panorama de quantos e quais são os bancos de dados existentes, e qual a natureza dos dados que os compõem. O resultado do mapeamento, por seu turno, seria divulgado nos sites das associações nacionais, como a ABRALIN e a ANPOLL, e internacionais, como a ALFAL, informando, de forma padronizada, as características gerais das amostras, os pesquisadores responsáveis e uma forma de contato.

Com esse primeiro passo, seria possível sensibilizar os gestores de amostras linguísticas quanto às vantagens de ações articuladas. Além disso, essa ação sensibilizaria todos os envolvidos – gestores dos bancos de dados e pesquisadores – para o conjunto de práticas esperadas pela Ciência Aberta, como a disponibilização de dados com licenças de uso claras e acessíveis. Apoiado nesse primeiro passo, viria um segundo, necessário para viabilizar o armazenamento dos dados e sua disponibilização eletrônica: a elaboração de um projeto coletivo para obtenção de apoio financeiro junto a órgãos de fomento.

Sabemos que a gestão consorciada é complexa e que as soluções para desafios pontuais de uso e manutenção dos dados poderiam ser mais facilmente encontradas pelas equipes locais. Mas a ação articulada dos responsáveis torna-se vantajosa para todos os bancos de dados linguísticos se considerada a circulação mais ampla do conhecimento. Para isso, precisamos dar um primeiro passo, aquele proposto aqui ou outro. Um passo que desencadeie acreditar na articulação dos bancos de dados como algo possível e positivo, tanto para os pesquisadores quanto para a sociedade.

REFERÊNCIAS

ABREU, R. N. Aspectos legais envolvidos na coleta de dados linguísticos. *In*: Raquel Meister Ko. Freitag. (Org.). *Metodologia de Coleta e Manipulação de Dados em Sociolinguística*. São Paulo: Editora Edgard Blücher, 2014.

BARABÁSI, A.-L. *Linked:* The New Science of Networks. Cambridge: Perseus Pub, 2002.

BARBOSA, A. G. A Plataforma de *corpora* do PHPB: uma apresentação ad infinitum. In: CASTILHO, A, T de. (Org.) *História do Português Brasileiro* – corpus diacrônico do português brasileiro. São Paulo: Contexto, 2019, p. 16-67.

BARROS, E. C.; VALENTIM, M. C.; MELO, M. A. A. O debate sobre o mestrado profissional na Capes: trajetória e definições. *Revista Brasileira de Pós-graduação*, Brasília, v. 2, n. 4, p. 124-138, 2005. Disponível em: http://ojs.rbpg.capes.gov.br/index.php/rbpg/article/view/84/80. Acesso em: 7 mar. 2020.

BIBER, D. Corpus-Based and Corpus-Driven Analyses of Language Variation and Use. *In:* HEINE, B.; NARROG, H. (Eds.). *The Oxford Handbook of Linguistic Analysis*. Oxford University Press, 2015, p.-193-224.

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus Linguistics:* Investigating Language Structure and Use. Cambridge; New York: Cambridge University Press, 1998.

BRASIL. Ministério da Educação. *Documento de área 2013*. Brasília: Fundação CAPES, 2013. Avaliação trienal 2013. Disponível em http://capes.gov.br/component/content/article/44-avaliacao/4670-ensino

BRASIL. Resolução do Conselho Nacional da Saúde nº 510, de 07 de abril de 2016. *Diário Oficial [da] República Federativa do Brasil*, Poder Executivo, Brasília, DF, 24 maio 2016. Seção 1, p. 44-46.

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. https://doi.org/10.1080/09298215.2018.1473449

CAMERON, D.; PANOVIĆ, I. Computer-Mediated Discourse Analysis. *In*: CAMERON, D.; PANOVIĆ, I. (Eds.). *Working with Written Discourse*. London: SAGE, 2014. p. 112–129.

CARDOSO, P. B. O paradoxo entre a transparência dos dados e a privacidade dos informantes na gestão de dados linguísticos. *Revista da ABRALIN*, v. 19, n. 2, p. 1-9, 24 ago. 2020.

CASILLAS, M.; CRISTIA, A. A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra: Psychology*, v. 5, n,1, p.1-21. DOI http://doi.org/10.1525/collabra.209

CHILDS, B.; VAN HERK, G.; THORBURN, J. Safe harbour: Ethics and accessibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory*, v. 7, n., p. 163–180, 2011. DOI http://doi. org/ 10.1515/CLLT.2011.008.

EMIGH, W.; HERRING, S. C. Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, v. 4, 2005, Big Island, HI, USA, 2005. *Proceedings [...].* Big Island, HI, USA: IEEE, 200.5p. 99a-99a. DOI: 10.1109/HICSS.2005.149.

FINARDI, K. R. et al. uma rodada de perguntas com os membros do grupo de trabalho Linguagem e Tecnologias (ANPOLL). *Revista Linguagem em Foco*, v. 12, n. 2, p. 1-31, 28 ago. 2020. DOI: 10.46230/2674-8266-12-3859.

FREITAG, R. M. K. (Org.). *Metodologia de coleta e manipulação de dados em sociolinguística*. São Paulo: Editora Edgard Blücher, 2014a.

FREITAG, Raquel Meister Ko. Dissecando a entrevista sociolinguística: estilo, sequência discursiva e tópico. In: GORSKI, Edair; COELHO, Izete Lehmkuhl; DE SOUZA, Christiane Maria Nunes (Ed.). *Variação estilística:* reflexões teórico-metodológicas e propostas de análise. Florianópolis: Insular, 2014b, p. 125-141.

FREITAG, Raquel Meister Ko. Sociolinguística no/do Brasil. *Cadernos de Estudos Linguísticos*, v. 58, n. 3, p. 445-460, 2016.

FREITAG, R. M. K. *Documentação sociolinguística:* coleta de dados e ética em pesquisa. São Cristóvão: Editora UFS, 2017a.

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

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*, 56(3), p. 917-944, 2012.

GABARDO, M.; LIMA-LOPES, R. E. DE. Ni una menos: ciência das redes e análise de um coletivo feminista. *Humanidades & Inovação*, v. 5, n. 3, p. 44-58, 2018.

GUTIÉRREZ, M. Participation in a Datafied Environment: Questions about Data Literacy. *Comunicação e sociedade*, v. 36, p. 37-55, 2019.

GUY, G. R.; ZILLES, A. *Sociolinguística quantitativa* – instrumental de análise. São Paulo: Parábola Editorial, 2007.

JOHNSON, R. B.; ONWUEGBUZIE, A. J.; TURNER, L. A. Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research*, v. 1, n. 2, p. 112–133, abr. 2007.

KING, G. Replication, Replication. *PS: Political Science and Politics*, v. 28, n. 3, p. 444-452, 1995.

LABOV, W. *Principles of linguistic change* – Volume 1: Internal factors. Malden/Oxford: Blackwell, 1994.

LABOV, W. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press, 1972.

LIMA-LOPES, R. E. de; GABARDO, M. Ni una menos: a luta pelos direitos das mulheres na argentina e suas representações no Facebook. *Revista Brasileira de Linguística Aplicada*, v. 19, n. 4, p. 801–824, 2019.

LIMA-LOPES, R. E. de. O Conservadorismo Como Ideologia: Contribuições Da Ciência Das Redes Para a Linguística Sistêmico Funcional. *Letras*, v. 28, n. 56, p. 43–69, 2018.

LIMA-LOPES, Rodrigo Esteves de; PIMENTA, Izadora Silva. #MULHERESNOFUTEBOL: transitividade e avaliatividade na identificação de padrões sexistas. *Revista Humanidades & Inovação*, v. 4, n. 6, p. 116-131, 2017.

MARTINS, M. A. R. A plataforma PB-Corpus Histórico e uma investigação da ordem de clíticos e de sujeitos em jornais brasileiros oitocentistas. *Revista Letras*, v. 60, p.179-200, 2020.

MCENERY, T.; HARDIE, A. *Corpus Linguistics:* Method, Theory and Practice. Cambridge: Cambridge University Press, 2012.

MEIE, K. J. Replication: A View From the Streets. *PS: Political Science and Politics*, v. 28, n. 3, p. 456–459, 1995.

MERCURI, K. T.; LIMA-LOPES, R. E. de. Discurso de Ódio Em Mídias Sociais Como Estratégia de Persuasão Popular. *Trabalhos em Linguística Aplicada*, v. 59, n. 2, p. 1216-1238, 2020.

PAIVA, V. L. M. O. Reflexões sobre ética e pesquisa. *Revista Brasileira de Linguística Aplicada*, v. 5, n. 1, p.43-61, 2005.

PRESTON, D. *Perceptual Dialectology:* nonlinguists' views of Areal Linguistics. Dordrecht – Holanda /Providence: Foris Publications, 1989.

SCOTT, J. *Social Network Analysis:* A Handbook. 2. ed. London/Thousand Oaks/Calif: SAGE Publications, 2000.

SCOTT, J. *What is social network analysis?* London; New York: Bloomsbury Academic, 2013.

SILVA, F. C. C.; SILVEIRA, L. O ecossistema da Ciência Aberta. *Transinformação*, v. 31, e190001, 2019. http://dx.doi.org/10.1590/2318-889201931e190001.

SILVA, R. G. da; COELHO, I. M. W da S. *Tutorial para o uso do banco de dados linguísticos RECI*. 2018. 28 f. Disponível em: http://ppget.ifam.edu.br/dissertacoes-defendidas/. Acesso em: 15 set. 2020.

SILVA, R. G. da; COELHO, I. M. W. da S. *Banco de Dados Línguísticos RECI:* corpus Oral, 2017, 1 DVD (218 min.43s) WAV. Disponível em: http://ppget.ifam.edu.br/dissertacoes-defendidas/. Acesso em: 15 set. 2020.

SILVA, R. G. da. *Variação linguística na língua portuguesa:* uma proposta de ensino dos modos verbais com uso de banco de dados linguísticos. Dissertação (Mestrado Profissional em Ensino Tecnológico) – Instituto Federal de Educação, Ciências e Tecnologia do Amazonas, Campus Manaus Centro, 2017. Disponível em: http://ppget.ifam.edu.br/dissertacoes-defendidas/. Acesso em 15 set. 2020.

SILVA, R. G.; COELHO, I. M. W. da S. Banco de Dados Linguísticos Reci – *Corpus escrito:* 30.798 palavras, 2017. DVD (PDF) Digital. Disponível em: https://drive.google.com/file/d/1YekKfpwDUZ23VMZG-mOqE1PWZmUmOQ_w/view. Acesso em: 15 set. 2020.

SINCLAIR, J. *Corpus, Concordance, Colocation*. Oxford: Oxford University Press, 1991.

STUBBS, M. *Words and Phrases:* Corpus Studies of Lexical Semantics. Oxford; Malden: Blackwell Publishers, 2001.

TARALLO, F. *A pesquisa sociolinguística*. 7.ed. São Paulo: Ática, 2006.

VAN LEEUWEN, T. *Discourse and Practice:* New Tools for Critical Discourse Analysis. Oxford: Oxford University Press, 2008.

WATTS, D. J. *Six Degrees:* The Science of a Connected Age. New York: Norton, 2003.

WATTS, D. J. The "New" Science of Networks. *Annual Review of Sociology*, v. 30, n. 1, p. 243–270, 2004.

WILKINSON, M., et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, v.3, n. 160018, 2016. https://doi.org/10.1038/sdata.2016.18.

WOLFF, A. *et al*. Creating an Understanding of Data Literacy for a Data-driven Society. *The Journal of Community Informatics*. v. 12, n. 3, p. 9–26, 2016.