ENSAIO TEÓRICO

OS DESAFIOS PARA DISPONIBILIZAÇÃO E COMPARTILHAMENTO DE DADOS LINGUÍSTICOS DA **AMOSTRA BASE VARSUL**

Isabel de Oliveira e Silva MONGUILHOTT
Centro de Ciências da Educação – Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Santa Catarina, Brasil

Izete Lehmkuhl COELHO
Programa de Pós-Graduação em Linguística – Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Santa Catarina, Brasil

Cláudia Regina BRESCANCINI
Escola de Humanidades – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Porto Alegre, Rio Grande do Sul, Brasil

RESUMO

O objetivo deste artigo é apresentar os desafios para disponibilização e compartilhamento de dados linguísticos da Amostra Base Varsul. O projeto Varsul foi constituído desde a década de 1980 para o estudo do português falado na Região Sul do Brasil, incluindo dados das capitais e dos centros urbanos histórica e socioculturalmente mais importantes. Atualmente é constituído por pesquisadores de quatro universidades da Região Sul do Brasil: Universidade Federal do Rio Grande do Sul (UFRGS), Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Universidade Federal de Santa Catarina (UFSC) e Universidade Tecnológica Federal do Paraná (UTFPR). Para compor a Amostra Base do banco de dados Varsul foram realizadas, entre 1989 e 1996, 288 entrevistas de experiência pessoal, nos moldes labovianos, sendo 96 por estado e 24 em cada uma das 12 cidades selecionadas, levando em conta as etnias

OPEN ACESS

SOBRE OS AUTORES
- Isabel de Oliveira e Silva Monguilhott
  Conceptualização; Curadoria de Dados; Metodologia; Administração do Projeto; Supervisão; Visualização; Escrita – Rascunho Original; Escrita – Análise e Edição.
- Izete Lehmkuhl Coelho
  Conceptualização; Curadoria de Dados; Metodologia; Visualização; Escrita – Rascunho Original; Escrita – Análise e Edição.
- Cláudia Regina Brescancini
  Conceptualização; Curadoria de Dados; Metodologia; Administração do Projeto; Supervisão; Visualização; Escrita – Rascunho Original; Escrita – Análise e Edição.

Recebido: 01/02/2025
Aceito: 17/07/2025
Publicado: 02/10/2025

COMO CITAR
MONGUILHOTT, I.O.S.; COELHO, I.L.; BRESCANCINI, C.R. (2025). Os desafios para disponibilização e compartilhamento de dados linguísticos da amostra base VARSUL. *Cadernos de Linguística*, v. 6, n. 4, e813.

formadoras das regiões. Além da etnia, a amostra foi estratificada em sexo (masculino e feminino), duas faixas de idade (de 25 a 49 anos; acima de 50 anos) e três escolaridades (4 a 5 anos, 8 a 9 anos e 10 a 11 anos). Todas as 288 entrevistas foram transcritas e armazenadas nas agências do projeto Varsul e têm servido, desde então, como rico material para a descrição e a análise das variedades do português do sul do Brasil. O projeto conta também com amostras complementares de ampliação da Amostra Base das três capitais, bem como com amostras diversas de bairros urbanos e não urbanos de Florianópolis e do Rio Grande do Sul. Para que possamos disponibilizar ao acesso público os dados linguísticos da Amostra Base, em conformidade com as diretrizes da ciência aberta, estamos, neste momento, em fase de desidentificação e anonimização dos áudios e transcrições das entrevistas, a fim de garantir o anonimato dos participantes em atendimento aos preceitos éticos e legais. Faz parte ainda da agenda atual do projeto Varsul a implementação do projeto 'Estudo da Mudança Linguística em Tempo Real: ampliação do Banco de Dados de Fala do Projeto VARSUL' cujo objetivo principal é a ampliação da Amostra Base, através de estudos de painel e estudos de tendência, com o intuito de investigar a mudança em tempo real.

PALAVRAS-CHAVE
Banco de dados Varsul; Português falado na Região Sul do Brasil; Ética no compartilhamento de dados linguísticos.

TITLE
THE CHALLENGES OF ACCESS AND SHARING OF LINGUISTIC DATA FROM THE VARSUL BASE SAMPLE

ABSTRACT
This article aims to present the challenges of making available and sharing linguistic data from the VARSUL Base Sample. The VARSUL project was set up in the 1980s to study the Portuguese spoken in the southern region of Brazil, including data from the capital cities and the most historically and socio-culturally important urban centres. It currently brings together linguists from four universities in the southern region: the Federal University of Rio Grande do Sul (UFRGS), the Pontifical Catholic University of Rio Grande do Sul (PUCRS), the Federal University of Santa Catarina (UFSC) and the Federal Technological University of Paraná (UTFPR). To make up the Basic Sample of the VARSUL database, 288 personal experience interviews were carried out between 1989

and 1996 - 96 per state and 24 in each of the 12 cities selected – taking into account the ethnic groups that make up the regions. In addition to ethnicity, the sample was stratified by gender (male and female), two age groups (25 to 49; over 50) and three levels of schooling (4 to 5 years; 8 to 9 years and 10 to 11 years). All 288 interviews were transcribed and stored at the VARSUL project branch offices and have served as rich material for describing and analysing the varieties of Southern Brazilian Portuguese. The database has complementary samples to expand the Base Sample, as well as samples from urban and non-urban neighbourhoods in Florianópolis-SC. In order to make the linguistic data of the Base Sample publicly accessible, in accordance with the guidelines of Open Science, we are currently in the process of de-identifying and anonymizing the audio recordings and transcripts of the interviews, in order to guarantee the confidentiality of the participants, in accordance with ethical and legal requirements. Also on the current agenda of the VARSUL project is the implementation of the project 'Study of Linguistic Change in Real Time: expansion of the VARSUL project speech database', whose main goal is to expand the base sample through panel studies and trend studies, with the aim of studying change in real time.

KEYWORDS
VARSUL Database; Brazilian Portuguese Spoken in the Southern; Ethics in Sharing Language Data.

INTRODUÇÃO

O objetivo deste artigo é apresentar os desafios para disponibilização e compartilhamento de dados linguísticos da Amostra Base Varsul. Inicialmente, trazemos um pouco da história da constituição do projeto Variação Linguística na Região Sul do Brasil (Varsul), um projeto que nasce em 1984 – há 41 anos, portanto – como um projeto em rede nacional, no âmbito do grupo de pesquisa Variação Linguística, com o objetivo geral de organizar um banco de dados linguísticos do português falado na Região Sul do Brasil.

Para compor a Amostra Base do banco de dados Varsul foram realizadas, entre os anos de 1989 e 1996, 288 entrevistas de experiência pessoal, nos moldes labovianos, em quatro cidades de cada estado da Região Sul (Rio Grande do Sul, Santa Catarina e Paraná): as capitais e mais três cidades histórica e sócio culturalmente representativas e relevantes.

A formação de um banco de dados de língua falada dessa região, fruto da imigração de etnias diversas, como açoriana, alemã, italiana, polonesa e ucraniana, além das migrações internas e dos contatos fronteiriços, visava cumprir alguns dos objetivos da área de Sociolinguística e Dialetologia dos Programas de Pós-Graduação das instituições federais de ensino superior (inicialmente UFRGS, UFSC e UFPR), atrelados à pesquisa e à formação de pessoal.

No decorrer da implementação e consolidação do banco de dados Varsul, a coordenação geral foi sendo sucessivamente exercida por diferentes universidades que sediam o projeto. Além do coordenador geral, cada agência conta com um coordenador regional, a quem compete, entre outras atribuições, zelar pela Amostra Base Varsul e ampliar o banco de dados.

A partir da constituição da Amostra Base, a ampliação do banco de dados Varsul vem ocorrendo em todas as agências do projeto. O banco Varsul conta atualmente com outras amostras, contemplando novas faixas etárias (jovens e adolescentes), mais um nível de escolaridade (universitários) e novas regiões (urbanas e não urbanas). Segundo Bisol, Menon e Tasca (2008, p. 52), "o Varsul vem se tornando um lugar privilegiado de formação de novos pesquisadores, abrindo portas a alunos de graduação, mestrandos e doutorandos".

No período em que a Amostra Base foi constituída, os protocolos éticos e legais para o compartilhamento de dados linguísticos restringiam-se, basicamente, à proteção do nome dos informantes, que passavam a ser identificados por meio de números, letras ou nomes fictícios. As amostras coletadas posteriormente, identificadas no projeto Varsul como amostras complementares, passaram a atender às resoluções que fundamentam a ação dos Comitês de Ética em Pesquisa com seres humanos das universidades onde se situam as agências do Varsul.

Neste momento, para que possamos disponibilizar e compartilhar os dados linguísticos da Amostra Base, em conformidade com as diretrizes da ciência aberta quanto à acessibilidade a dados em bases legais e éticas (*open access*) e à interoperabilidade, estamos em fase de desidentificação

e anonimização dos áudios e transcrições das entrevistas com o objetivo de garantir plenamente o anonimato dos entrevistados, em conformidade com as Resoluções no.466 e no.510[1] do Conselho Nacional de Saúde e a Lei Geral de Proteção de Dados Pessoais (LGPD).

Faz parte ainda da agenda atual do projeto Varsul a implementação do projeto 'Estudo da Mudança Linguística em Tempo Real: ampliação do Banco de Dados de Fala do Projeto VARSUL', cujo objetivo principal é, como o título sugere, a ampliação da Amostra Base, através de estudos de painel e estudos de tendência, com o intuito de investigar a mudança em tempo real[2].

O artigo em tela está organizado em três seções. Na primeira, apresentamos um pouco de história a respeito da constituição do banco de dados Varsul. Na segunda seção, detalhamos aspectos éticos e legais do compartilhamento de dados linguísticos e do processo de desidentificação e anonimização em andamento. Na terceira seção, descrevemos a agenda atual do projeto Varsul, que segue comprometido, após 40 anos de atividade, com a tarefa de documentar a fala do sul do Brasil, desta vez com a constituição de novas amostras para a condução de estudos em tempo real de curta duração.

1. A HISTÓRIA DO PROJETO VARSUL

Estudos realizados sob a perspectiva da Sociolinguística Variacionista surgiram no Brasil na década de 1970 com o Projeto Censo de Variação Linguística do Estado do Rio de Janeiro, hoje denominado de Programa de Estudos sobre o Uso da Língua (Peul), coordenado por Anthony Naro na Universidade Federal do Rio de Janeiro (UFRJ). Com o objetivo de estudar a variedade do português carioca, o projeto foi o pioneiro na operacionalização com bancos de dados para o estudo da variação linguística. A formação desse primeiro banco de dados de língua falada no Brasil seguiu a metodologia da Sociolinguística Variacionista veiculada na obra *The Social Stratification of English in New York City*[3], de William Labov, publicada em 1966. Nessa obra, foram consideradas amostras de fala coletadas no bairro *Lower East Side,* com base em um protocolo para realização de entrevista sociolinguística estratificada por critérios sociodemográficos amplos, como sexo, idade, escolaridade, classe social, etnia e região de origem.

A coleta de dados realizada no bairro *Lower East Side* de Nova York foi exemplar, considerada de alto rigor metodológico. Com o objetivo de estudar as variações linguísticas entre pessoas de

1 Discutiremos as Resoluções no.466 e no.510, na seção 2.

2 O projeto, em andamento, foi contemplado com recurso da Chamada CNPq/MCTI Nº 10/2023 – Universal, Processo 406239/2023-1, Faixa B, grupos consolidados.

3 Em português: "A Estratificação Social do Inglês na Cidade de Nova York".

diferentes classes sociais, etnias e locais de origem, de modo a considerar os efeitos sociais e estilísticos que poderiam condicionar o uso de cada um dos contextos em variação[4], foi escolhido o *Lower East Side* por ser, à época, representado por diferentes grupos étnicos (italianos, judeus, alemães, ucranianos, poloneses, afro-americanos, porto riquenhos, entre outros) e por ser uma região com larga escala de delinquência juvenil.

Definida a área geográfica em que os dados seriam coletados, as etapas de seleção dos informantes passaram a ser definidas, contando com critérios que envolveram desde o mapeamento de onde seria possível encontrar informantes falantes do inglês (com base no censo populacional da região) e a seleção aleatória de quase mil informantes que participaram de entrevistas assistemáticas com duração de 2 a 4 horas, à realização das entrevistas sociolinguísticas propriamente ditas com uma amostra mais reduzida. Com base em uma amostragem ampla, a amostra final foi, por fim, delimitada, com a exclusão dos informantes não nativos, dos que chegaram ao bairro depois dos oito anos de idade, dos que morreram e dos que se mudaram. As entrevistas sociolinguísticas foram conduzidas com 122 informantes. Esse trabalho de coleta foi realizado entre os anos de 1961 e 1963 e contou com a participação de uma equipe formada por 40 entrevistadores.

Seja pelas contingências brasileiras relacionadas à falta de recursos e à consequente falta de pessoal para a condução da constituição de bancos de dados numerosos, seja por conta dos diferentes objetivos de formação de bancos de dados, fato é que no Brasil uma outra Sociolinguística de orientação variacionista se constituiu, uma Sociolinguística estritamente brasileira, que foi delimitada a partir do Peul, segundo Freitag (2016).

Com o propósito de estudar a língua em uso e de pensar questões gerais de variação e mudança linguística, o Peul constituiu o primeiro banco de dados sobre a língua falada do Brasil composto por 64 entrevistas sociolinguísticas, estratificadas por critérios sociodemográficos amplos, como sexo, idade e escolaridade.

As diretrizes para a realização das entrevistas do banco de dados Peul são baseadas em seleção aleatória de informantes falantes de português a partir dos seguintes critérios: ser morador da cidade do Rio de Janeiro há pelo menos 2/3 de sua vida; não ter morado fora da região por mais de um ano no período da aquisição da língua; ser filho/a de pais com as mesmas características e ser reconhecido pelos pares como membros da mesma comunidade de fala. A amostra é estratificada em função de características sociodemográficas amplas, "formando, a partir da confluência das características sociodemográficas, células sociais, sempre ortogonais, isto é, correspondentes à mesma quantidade de informantes por célula" (Freitag, 2016, p. 453).

4 Segundo Freitag (2016, p. 446), "Labov faz o delineamento das variáveis, considerando que certos traços do sistema linguístico parecem dar pistas sobre a categorização social do falante: /r/, /aeh/, /oh/, /th/, /dh/."

A importância desse banco de dados e das descrições que se sucederam a respeito da variedade carioca levou muitos pesquisadores de diferentes instituições brasileiras a replicarem a metodologia de formação do banco de dados Peul. O projeto Varsul foi o primeiro a replicar o modelo do Peul.

Pesquisadores da UFRGS, da UFSC e da UFPR reuniram-se em 1982 em prol de um grande projeto regional do sul do Brasil constituído por três grupos de trabalho: Atlas Linguístico e Etnográfico, Bilinguismo e Variação Linguística. No âmbito do grupo de trabalho de Variação Linguística foi discutida a proposta de Leda Bisol de organizar um banco de dados linguísticos da Região Sul, de orientação laboviana, nos moldes do Peul.

Em 1984 nasce o projeto Varsul[5], abrangendo os estados do Rio Grande do Sul, Santa Catarina e Paraná com os objetivos de oferecer (i) subsídios para a descrição do português falado no País; (ii) condições para testar e desenvolver teorias linguísticas; (iii) condições para a formação de novos pesquisadores; e (iv) subsídios para programas educacionais, promovendo o conhecimento e o respeito às variedades linguísticas (Bisol, Menon e Tasca, 2008, p. 50-51). O projeto Varsul foi desenvolvido inicialmente em três agências (UFRGS, UFSC e UFPR). Em 1990, a PUCRS passou a integrar a equipe e, em 2015, a UTFPR assumiu o lugar da UFPR.

Para a constituição do banco de dados, o projeto Varsul seguiu critérios sociodemográficos amplos adotados pelo Peul, como sexo, idade e escolaridade. Além desses critérios amplos, a equipe levou em consideração a região de origem dos informantes para estabelecer as cidades que seriam investigadas. Esse critério é fundamental por ser a Região Sul rica em contatos linguísticos. No século XVIII, observou-se grande fluxo de imigração açoriana principalmente para as regiões litorâneas do Sul do Brasil. Ainda, nesse século, os estados contaram com migrações vicentinas, reconhecidamente chamadas de caminho dos tropeiros. No século XIX, a imigração de alemães, italianos, poloneses e ucranianos, especialmente para o interior dos estados, foi delimitando etnicamente as diferentes localidades, contribuindo para a sua expansão. A Região Sul contou também com o contato do português com os falantes de línguas indígenas, povos originários, e com o contato do português com o espanhol nas regiões de fronteira, em especial com Argentina, Paraguai e Uruguai.

Com base nesses critérios, foram selecionadas quatro cidades de cada um dos estados do sul do Brasil, incluindo sempre as capitais e mais três cidades etnicamente distintas, para serem investigadas. No Rio Grande do Sul: Porto Alegre (capital cuja colonização teve início no século XVII com a chegada de casais açorianos), Flores da Cunha (colonização italiana), Panambi (colonização alemã) e São Borja (fronteira - contato com o espanhol). Em Santa Catarina, Florianópolis (capital

5 Em 1989, finalmente o projeto Varsul foi aprovado por agência de fomento (FINEP), recebendo a partir de então recursos de agências diversas para a constituição do banco de dados de língua falada (FINEP, CNPq, CAPES, FAPERGS, FUNCITEC).

e de colonização açoriana), Blumenau (colonização alemã), Chapecó (colonização italiana) e Lages (colonização gaúcha-tropeiros vicentinos). No Paraná, Curitiba (capital que surge como povoado no século XVII, com colonizadores portugueses e espanhóis), Irati (colonização eslava), Londrina (colonização mineira e paulista) e Pato Branco (colonização gaúcha).

Para a seleção dos informantes foram estabelecidos critérios parecidos com os do Peul. O informante deveria ser falante de português, morador da cidade investigada há pelo menos 2/3 de sua vida, não ter morado fora da região por mais de um ano no período da aquisição da língua, não causar estranheza a outros falantes da região, ser filho/a de pais nascidos na cidade em questão.

Entre os anos de 1989 e 1996 foi realizada a coleta de 288 entrevistas sociolinguísticas para a constituição do banco de dados do projeto Varsul – conhecidas hoje por Amostra Base – assim distribuídas: 96 entrevistas em cada estado e 24 em cada uma das 12 cidades selecionadas. A amostra foi estratificada em sexo (masculino e feminino), idade (25 a 49 anos e acima de 50 anos) e escolaridade (primário, ginasial e segundo grau). As 96 entrevistas do Rio Grande do Sul foram realizadas pelas agências da UFRGS e da PUCRS, as 96 de Santa Catarina pela agência da UFSC e as 96 do Paraná pela agência da UFPR (conforme Knies, Costa, 1996).

Durante todo o período de coleta, transcrição e digitação das 288 entrevistas da Amostra Base, o projeto Varsul contou com a coordenação geral de Paulino Vandresen, com a Coordenação Científica de Leda Bisol e com coordenações regionais em cada uma das agências do projeto (UFRGS, PUC-RS, UFSC, UFPR). Além disso, contou com a participação efetiva de professores, bolsistas de Iniciação Científica e do Programa de Educação Tutorial (PET) e bolsistas de mestrado e doutorado da área de Sociolinguística e Dialetologia que atuaram em todas as etapas de constituição do banco de dados.

Assim que as entrevistas eram realizadas, seguiam-se as etapas de transcrição e de digitação de todo o material coletado em cada uma das agências, de acordo com o modelo de três linhas desenvolvido pelo Peul: na primeira linha, o transcritor anotava a sintaxe real do texto com comentários sobre ruídos, hesitações, gaguejos etc; na segunda linha, anotava as informações fonéticas relevantes que caracterizavam cada uma das variedades; na terceira linha, o transcritor fazia anotações morfossintáticas de cada palavra ou sintagma e registrava a velocidade da fala dos entrevistados. Transcritos, os dados foram digitados e as entrevistas foram eletronicamente armazenadas.

Durante todo o período de coleta, transcrição e digitação das entrevistas, foram realizadas reuniões anuais das quatro agências Varsul. Nessas reuniões, eram feitas discussões a respeito da montagem do banco de dados, de aspectos teóricos e metodológicos da pesquisa Sociolinguística Variacionista, da estratificação da amostra, de critérios para a seleção de informantes, dos procedimentos de coleta, de critérios de transcrição e de digitação das entrevistas, entre outros aspectos. Essas reuniões anuais contavam também com minicursos de professores convidados e com a apresentação de trabalhos em andamento com base na coleta que estava sendo realizada. Os primeiros resultados a respeito da descrição do português falado na Região Sul já se anunciavam.

Segundo Vandresen (2016, p. 17),

> O projeto VARSUL, desde seu início, colocou a formação de recursos humanos para a pesquisa sociolinguística como um de seus objetivos. (...) Ao se envolverem na pesquisa, passaram pela experiência do trabalho de campo, fazendo entrevistas, transcrição, digitação, envolvendo-se com a informática, numa época em que um computador ainda era de difícil acesso. Mas, o principal aspecto a salientar é o estímulo dado aos alunos bolsistas a prepararem trabalhos para apresentar em seminários/salões de iniciação científica, em nossos encontros regionais e mesmo em outros espaços acadêmicos como SBPC, CELSUL e CELIP.

Finalizadas as etapas de coleta, transcrição e digitação das 288 entrevistas da Amostra Base, o banco de dados Varsul foi apresentado às comunidades científicas nacional e internacional de 2 a 5 de setembro de 1996, na UFRGS, no *VI Encontro Regional do Projeto Varsul* e *I Encontro de Variação linguística do CONE SUL*. O Evento contou com convidados renomados estrangeiros (Gregory Guy, Adolfo Eliziancin, Elizabeth Rigatuso, Joaquim Born) e brasileiros (Anthony Naro, Rosa Virgínia Mattos e Silva, Milton do Nascimento).

Com base na metodologia de constituição dos bancos de dados Peul e Varsul, outros bancos de dados de língua falada se formaram no Brasil para descrever e mapear a diversidade linguística de cada uma das regiões, mantendo-se uma certa padronização. A esse respeito, afirma Freitag (2016, p. 453) que

> A abordagem a partir de bancos de dados sociolinguísticos trouxe subsídios para a descrição do português brasileiro, com a padronização da amostragem e coleta de dados, que permite, de certa forma, a comparação de resultados, e, assim, traz contribuições para uma norma brasileira, com descrições sociolinguísticas em interface teórica tanto com abordagens formais (...) como com abordagens funcionais.

Muitos pesquisadores formados pelo Projeto Varsul passaram a atuar em universidades de outras regiões do país e vêm contribuindo com a tarefa de descrição do português brasileiro e com a replicação do protocolo de coleta de dados definido no projeto.

2. ASPECTOS ÉTICOS E LEGAIS DO COMPARTILHAMENTO DE DADOS LINGUÍSTICOS: A EXPERIÊNCIA DO VARSUL

Considerando que os documentos norteadores da condução ética de pesquisa cientificam no Brasil – Resolução no. 466 (2013), Resolução no. 510 (2016), Lei Geral de Proteção de Dados Pessoais (LGPD) (2018) – foram publicados no início do século XXI e que, conforme apresentado na seção 1, a amostra base do banco de dados Varsul foi constituída no final do século XX, entre 1989 e 1996, autorizações formais dos entrevistados, hoje imprescindíveis para constituição de qualquer banco de dados que envolva seres humanos, inexistem nos registros do projeto Varsul. Tal desencontro temporal foi tomado como justificativa para que, em 02 de julho de 2013, o Comitê de Ética em

Pesquisa da PUCRS emitisse um documento reconhecendo o banco de dados de fala Varsul, a partir de sua constituição e objetivos, e declarando que

> O banco de dados Varsul poderá ser acessado por pesquisadores após seus projetos de pesquisa serem registados na Plataforma Brasil e aprovados pelo Comitê de Ética em pesquisa. A identidade dos entrevistados não será revelada em publicações e/ou em qualquer outro tipo de atividade de exposição de trabalho científico realizado a partir de dados obtidos por meio das gravações reunidas nesse banco de dados (Declaração, 2013).

Os termos da declaração referida são embasados pela Resolução 466, de 12 de dezembro de 2012, emitida pelo Conselho Nacional de Saúde, de acordo com a qual, especificamente com relação ao artigo III.1, i), a eticidade em pesquisa implica em

> i) prever procedimentos que assegurem a confidencialidade e a privacidade, a proteção da imagem e a não estigmatização dos participantes da pesquisa, garantindo a não utilização das informações em prejuízo das pessoas e/ou das comunidades, inclusive em termos de autoestima, de prestígio e/ou de aspectos econômico-financeiros (Artigo III.1, i).

Dado que as especificidades éticas das pesquisas em Ciências Humanas e Sociais e as particularidades de suas metodologias são reconhecidas pela Resolução 466 (artigo III.3), o Conselho Nacional de Saúde Pública publicou, em 07 de abril de 2016, a Resolução no. 510, que dispõe sobre as normas aplicáveis a pesquisas dessas áreas. Considerando aquilo que é aplicável à constituição de bancos de dados de fala, mantém-se, nessa Resolução, as orientações básicas da Resolução 466, a saber, "a garantia da confidencialidade das informações, da privacidade dos participantes e da proteção de sua identidade, inclusive no uso de sua imagem e voz" (Capítulo II, artigo 3º., item VII) e "garantia da não utilização, por parte do pesquisador, das informações obtidas em pesquisa em prejuízo dos seus participantes" (Capítulo II, artigo 3º., item VIII).

Dois anos depois, em 14 de agosto de 2018, é aprovada a Lei 13.709/2018, denominada Lei Geral de Proteção de Dados Pessoais (LGPD), que também dispõe sobre a questão da confidencialidade, mas com um enfoque particular, a partir da proteção do indivíduo e de sua personalidade por trás dos *dados pessoais*, ou seja, por trás das informações relacionadas à pessoa identificada ou identificável (artigo 5º., I). Sua relevância para a constituição dos bancos de dados de fala, nos moldes sociolinguísticos, está justamente no fato de que se debruça sobre dados sensíveis, entendidos como aqueles "[...] sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual [...] (artigo 5º., II)", frequentemente presentes nas entrevistas de experiência pessoal coletadas à luz dos pressupostos teórico-metodológicos variacionistas e que compõem as amostras desse tipo de banco.

A LGPD prevê a proteção dos dados pessoais em quaisquer das *formas de tratamento*, assim definidas:

> Toda operação realizada com dados pessoais, como as que se referem a coleta, produção, recepção, classificação, utilização, acesso, reprodução, transmissão, distribuição, processamento, arquivamento, armazenamento, eliminação, avaliação ou controle da informação, modificação, comunicação, transferência, difusão ou extração; (artigo 5º., X).

Dentre os dispositivos previstos na LGPD, o de interesse para a presente discussão é o da anonimização, conceituada como a "utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo" (artigo 5º., XI). Seu produto, o dado anonimizado, é relativo ao "titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento (artigo 5º., III). Dito de outro modo, de acordo com Carvalho (2021), a anonimização pode ser compreendida como a dissociação entre dados e seus titulares, independentemente do tipo de processo adotado.

Outra técnica utilizada para confidencialidade de dados pessoais é a de-identificação ou desidentificação, que permite a remoção das informações de identificação pessoal e as de quase-identificação, ou seja, aquelas que, se combinadas a outras, poderiam identificar o indivíduo (Carvalho, 2021). Nesse caso, mantém-se, geralmente, uma espécie de mapeamento que permite a re-identificação dos dados e, consequentemente, o acesso às informações originais, o que torna a técnica reversível de certo modo (Pinho, 2017)[6].

Para a Amostra Base do banco de dados Varsul, adotou-se o processo de anonimização do tipo *ex post*, "situação em que o dado é dissociado de seu titular após a sua captura" (Carvalho, 2021, p.46), pois as entrevistas de experiência pessoal, já gravadas e transcritas, caracterizam o material específico de análise para a identificação das informações a serem anonimizadas. Desse modo, para cada entrevista de experiência pessoal, informações referentes à identificação pessoal dos informantes, como nomes e endereços, assim como dados auxiliares, como nomes de pessoas próximas ao entrevistado, nomes de empresas em que os entrevistados trabalham/trabalharam ou identificação de locais citados como referência de endereços residenciais ou de trabalho foram alvo de desidentificação.

Nas transcrições das entrevistas, esse tipo de dado pessoal é localizado com o auxílio da ferramenta VarsulApp (versão 1.0) [7], um software de uso exclusivo do Varsul para edição das transcrições das 288 entrevistas da Amostra Base e busca de informações específicas nas três linhas de transcrição. Com esse aplicativo, dados pessoais são substituídos pela palavra NOME na primeira

6 A LGPD prevê ainda a chamada pseudonimização (art. 13, parágrafo 4), definida como "o tratamento por meio do qual um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo, senão pelo uso de informação adicional mantida separadamente pelo controlador em ambiente controlado e seguro." A criptografia , outra técnica, ainda é discutível como tipo de anonimização (Carvalho, 2021).

7 O VarsulApp foi desenvolvido pelo engenheiro Dr. Fábio Gonçalves Teixeira (UFRGS), em 2022, com apoio do Programa PROEX/CAPES do Programa de Pós-Graduação em Letras da PUCRS.

linha, a referente à transcrição ortográfica, sem que haja dessincronização com as outras duas linhas, respectivamente referentes à variação fonética e variação morfossintática. Nos arquivos de áudio equivalentes às entrevistas transcritas, o mesmo dado é apagado e, em seu lugar, é acrescentado um silêncio de aproximadamente um segundo de duração, através do recurso copiar e colar disponível no aplicativo de edição e gravação de áudio Audacity (versão 3.7.1)[8].

Considerando que os pesquisadores membros do projeto Varsul têm acesso às fichas sociais completas dos informantes, a reidentificação é uma possibilidade, já que cada informante é identificado por uma sequência de símbolos rastreável. Nesse caso, o processo aplicado restringe-se à desidentificação. Para o público externo, que obterá acesso às amostras de fala do banco de dados mediante cadastro prévio, serão fornecidas apenas informações referentes ao gênero, faixa etária e escolaridade, conforme descritas na seção 2 anterior, sendo a reidentificação altamente improvável, já que a possibilidade de associação direta ou indireta é perdida, caso a técnica para a omissão das informações relevantes seja adequadamente aplicada. Nesse caso, em que as informações únicas foram eliminadas ou agregadas, tem-se, de fato, a anonimização. Tanto em um caso quanto no outro, todos os usuários da Amostra Base devem assinar o Termo de Compromisso para Utilização de Dados de Fala do Banco de Dados Varsul, de acordo com o qual se comprometem a manter a confidencialidade sobre qualquer informação referente ao conteúdo das entrevistas e declaram ciência da impossibilidade de divulgação e/ou exposição da íntegra das entrevistas.

A LGPD não considera dados anonimizados como dados pessoais, pois entende que, uma vez anonimizados, os dados perdem a característica de dados pessoais. Desse modo, nos termos da lei:

> Os dados anonimizados não serão considerados dados pessoais para os fins desta Lei, salvo quando o processo de anonimização ao qual foram submetidos for revertido, utilizando exclusivamente meios próprios, ou quando, com esforços razoáveis, puder ser revertido (artigo 12).

No âmbito do banco de dados de fala VARSUL, a desidentificação e a anonimização garantem a confidencialidade dos entrevistados; no entanto, é a anonimização que permite, de fato, a disponibilidade das amostras para um público maior de interessados, dada a ausência das restrições de tratamento decorrentes das leis de proteção.

8 O software Audacity (versão 3.7.1) é uma ferramenta gratuita de edição e gravação de áudio.

3. AGENDA ATUAL DO PROJETO VARSUL

O projeto Varsul, como vimos, conta com quatro agências situadas em quatro instituições do sul do Brasil, que compartilham protocolos comuns para gerir os bancos de dados constituídos ao longo dessas quatro décadas de história do projeto (conforme Monguilhott *et al.*, 2023). No entanto, cada agência possui autonomia para sua organização interna, como periodicidade de reuniões e gerenciamento de tarefas de bolsistas vinculados ao projeto.

No que se refere aos encontros com pesquisadores de todas as agências, a coordenação geral, alternada a cada dois anos, organiza um evento em sua gestão para compartilhamento das pesquisas em andamento dos grupos das diferentes instituições, reservando um espaço para que os pesquisadores possam discutir questões comuns.

Na década de 1990, quando a Amostra Base foi constituída, havia outros protocolos éticos e legais para o compartilhamento de dados linguísticos, diferentemente das amostras complementares que já passaram a atender aos Comitês de Ética em Pesquisa com Seres Humanos das universidades. Para que possamos disponibilizar e compartilhar os dados linguísticos da Amostra Base, atendendo a demanda da ciência aberta de socializar de modo amplo o banco de dados, estamos, neste momento, enfrentando um desafio que é a desidentificação e anonimização dos áudios e transcrições das entrevistas coletadas na década de 90 para que possamos garantir o sigilo dos participantes, conforme discussão na seção anterior.

Na atual gestão do Projeto Varsul, submetemos, e fomos contemplados com recurso da Chamada CNPq/MCTI Nº 10/2023 – Universal, o projeto em rede 'Estudo da Mudança Linguística em Tempo Real: ampliação do Banco de Dados de Fala do Projeto VARSUL'. Trazemos aqui aspectos importantes desse projeto, em andamento, elaborado por algumas das pesquisadoras do grupo Varsul.

Iniciamos apontando seu objetivo principal que é ampliar a Amostra Base das capitais da Região Sul do país, com foco na realização de estudos de mudança em tempo real de curta duração, a partir de estudos de painel e estudos de tendência (Labov, 1994).

No que se refere aos estudos de painel, pretendemos obter regravações com os mesmos informantes da Amostra Base das capitais, visando a verificar o comportamento linguístico desses indivíduos cuja fala foi documentada cerca de trinta anos atrás. Já em relação aos estudos de tendência procederemos à constituição de uma nova amostra de fala das capitais, de acordo com os mesmos parâmetros estratificados da Amostra Base: etnia/colonização, sexo, faixa etária e escolaridade, visando a verificar o comportamento linguístico da comunidade. Serão considerados ainda aspectos estilísticos e identitários, ampliando, assim, as possibilidades de investigação desses aspectos relacionados às variáveis linguísticas.

A fim de garantir a comparabilidade das amostras de fala, todas as entrevistas obedecem aos mesmos princípios utilizados na constituição da Amostra Base fundamentados na metodologia da

sociolinguística variacionista (conforme Weinreich, Labov e Herzog, 2006 [1968]; Labov, 1994, 2008 [1972]).

A questão que pretendemos responder com a proposta é se o comportamento linguístico dos informantes da Região Sul do país mudou ou estabilizou-se no período entre as décadas de 1990 e 2020. Acreditamos que, ao comparar as amostras de fala, será possível captar mudança ou estabilidade no comportamento do indivíduo e da comunidade e reconhecer movimentos de mudança no curso de sua implementação.

A ampliação da amostra do banco de dados Varsul, seguida de armazenamento, transcrição e disponibilização das entrevistas, possibilitará aos pesquisadores descrever e revisitar fenômenos linguísticos em variação e/ou mudança com temas fonológicos, morfológicos, morfossintáticos, sintáticos, léxico-semânticos e discursivos, com base em estudos de painel e de tendência, comparando as amostras de fala de 1990 e 2020. Além disso, ressaltamos a importância do projeto no que se refere à formação de pesquisadores, refletida na alta capilaridade atingida em seus 40 anos de atividade, com pesquisadores formados no Varsul atuando em todas as regiões do país.

A proposta também disponibiliza o acesso aberto (Ciência Aberta - Open Science) e compartilhado para a comunidade científica da amostra de fala a ser coletada, o que significa a possibilidade de atualizar e comparar a descrição do português falado na Região Sul do Brasil em relação à amostra Varsul da década de 1990.

Essa nova amostra também poderá subsidiar estudos acústicos, considerando a melhora na qualidade do sinal a partir de equipamentos de captura de áudio avançados capazes de reconhecer com mais eficiência detalhes na produção de fala dos informantes, possibilitando assim estudos de cunho sociofonético, realizados a partir da interface entre Sociolinguística e Fonética Experimental, ainda pouco explorados no Brasil. Além disso, a documentação desses dados têm amplo potencial para subsidiar o ensino de língua materna, além de permitir análises contrastivas com caráter multidisciplinar e fornecer suporte ao processamento automático de linguagem natural.

Há a previsão de disponibilização no site do Projeto Varsul (www.varsul.org.br) das amostras de fala com alta qualidade de áudio, que proporcionarão o acesso a material de interesse para o desenvolvimento de pesquisas tanto para a comunidade acadêmica, composta por linguistas, como também a pesquisadores/profissionais de outras áreas do conhecimento, como a Fonoaudiologia, a Perícia Forense e a Ciência de Dados.

As amostras de recontato, obtidas a partir da realização de novas entrevistas de experiência pessoal com os mesmos informantes da Amostra Base coletada nos anos 1990, apresentam potencial de interesse para estudos sobre o envelhecimento da voz e suas causas, por exemplo. Adicionalmente,

o conjunto de amostras de fala constitui relevante banco de dados de referência de voz para a testagem de softwares de reconhecimento de locutor, como o Pacote Voice (Zabala, 2022)[9].

Também investimos, com este projeto, na formação de novos pesquisadores para atuação na área de Sociolinguística, assim como no fortalecimento dos Programas de Pós-Graduação aos quais vinculam-se os pesquisadores membros do Varsul, mediante apresentações de trabalhos sobre os resultados do projeto e das pesquisas dele decorrentes em eventos nacionais e internacionais, bem como a publicação de artigos em periódicos e revistas científicas da área, nacionais e internacionais. Pretende-se ainda divulgar os resultados das pesquisas em página no Instagram (@projetovarsul), para o público não acadêmico, contribuindo para o conhecimento das características do falar dos três Estados e para o olhar curioso, e despido de preconceitos, para o português brasileiro.

Prevê-se também a produção de material composto por fotos, painéis, gravações e vídeos para uma exposição itinerante "Falares do Sul" a ser instalada nas Câmaras de Vereadores, nas Assembleias Legislativas, museus e espaços públicos das três capitais, dando visibilidade à diversidade e à mudança linguística para o público não acadêmico que circula nesses espaços.

Nessa direção, pode-se mobilizar representantes de instâncias legislativas de âmbito municipal e estadual para o investimento em pesquisa com temáticas envolvendo o uso da língua para a construção de posicionamentos embasados cientificamente e isentos de preconceito. Pretende-se, ainda, em articulação com escolas e professores da rede básica de ensino das três capitais, criar espaço destinado à exposição itinerante e à organização de rodas de conversa com pesquisadores sobre as pesquisas desenvolvidas no projeto, desmistificando o ensino de gramática como conteúdo pronto e enfadonho e estimulando os estudantes a assumirem o papel agentivo de pesquisadores da língua.

A produção científica acerca da ampliação da amostra do banco de dados Varsul prevê a discussão dos aspectos metodológicos envolvidos, o que contribui para o conhecimento das dinâmicas de recontato de informantes, e a descrição e análise de fenômenos linguísticos em variação e/ou mudança, com a realização de pesquisas nos níveis fonético/fonológico, morfológico, morfossintático, sintático e discursivo.

Por fim, pretende-se divulgar a história do projeto Varsul e do projeto de recontato na grande mídia (jornais de ampla circulação, programas de rádio, programas de TV) para o público não acadêmico, visando socializar o conhecimento científico sobre a fala na Região Sul e estreitar os laços entre pesquisadores e jornalistas, parceiros no movimento de tradução dos conhecimentos científicos para o público mais amplo.

9 O Pacote Voice constitui-se em uma ferramenta computacional para análise de voz, reconhecimento de locutor e inferência de estado emocional.

4. CONSIDERAÇÕES FINAIS

O Varsul é um projeto consolidado e ativo há quatro décadas que vem contribuindo, ao longo desse período, com a comunidade acadêmica da área da Linguística, em especial, da Sociolinguística Variacionista, tanto por constituir-se em um banco de dados robusto, com inúmeras publicações científicas resultantes, quanto pela formação de pesquisadores que replicam seus conhecimentos em todas as regiões do Brasil.

Assim como todo projeto de pesquisa de tamanha complexidade, o projeto Varsul sempre lidou com desafios, seja pela dimensão das suas propostas, seja pela dimensão da rede de pesquisadores que o constitui. Neste momento, o projeto Varsul trabalha para enfrentar o desafio de disponibilizar e compartilhar os dados linguísticos da Amostra Base, em consonância com as diretrizes da ciência aberta e perfilado aos preceitos éticos e legais, preconizados pelas Resoluções do Conselho Nacional de Saúde no. 466 e no. 510 e pela Lei Geral de Proteção dos Dados. Apesar de sua complexidade, dado seus diversos agentes, o processo deve ser finalizado em breve.

A equipe de pesquisadores membros do Projeto Varsul empenha-se nesta nova fase de trabalho, que envolve a ampliação da Amostra Base para a investigação da mudança em tempo real, através de estudos de painel e estudos de tendência, na fala da Região Sul do Brasil.

AGRADECIMENTOS

Agradecemos aos bolsistas das agências UFSC, UFRGS e PUCRS pelo envolvimento no processo de anonimização e desidentificação dos áudios e transcrições das entrevistas da Amostra Base.

Agradecemos às pesquisadoras: Carla Regina Martins Paza, Cláudia Andrea Rost Snichelotto, Elisa Batistti e Loremi Loregian-Penkal pela contribuição na elaboração do projeto "Estudo da Mudança Linguística em Tempo Real: ampliação do Banco de Dados de Fala do Projeto VARSUL". Agradecemos, também, aos demais pesquisadores que compõem a equipe do Projeto: Ana Lívia dos Santos Agostinho, Christiane Maria Nunes de Souza, Edson Domingos Fagundes, Gustavo Nishida, Luiz Carlos da Silva Schwindt, Marco Antonio Rocha Martins, Marizete Bortolanza Spessatto e Odete Pereira da Silva Menon.

Agradecemos, também, às pareceristas da revista *Cadernos de Linguística* pelas sugestões para a versão final deste texto.

INFORMAÇÕES COMPLEMENTARES

CONFLITO DE INTERESSE

As autoras declaram que não possuem interesses financeiros ou relações pessoais que possam ter influenciado o trabalho relatado neste artigo.

DECLARAÇÃO DE DISPONIBILIDADE DE DADOS

O compartilhamento de dados não é aplicável a este artigo, pois nenhum dado novo foi criado ou analisado neste estudo.

DECLARAÇÃO DE USO DE IA

As autoras declaram que nenhuma ferramenta de IA foi utilizada na criação deste manuscrito nem em qualquer aspecto dos trabalhos realizados cujo resultado está reportado no manuscrito.

AVALIAÇÃO E RESPOSTA DOS AUTORES

Avaliação: https://doi.org/10.25189/2675-4916.2025.V6.N4.ID813.R
Resposta dos Autores: https://doi.org/10.25189/2675-4916.2025.V6.N4.ID813.A

REFERÊNCIAS

BISOL, Leda.; MENON, Odete Pereira da Silva; TASCA, Maria. VARSUL, um banco de dados. *In*: VOTRE, Sebastião; RONCARATI, Claudia (ed.*). Anthony Julius Naro e a linguística no Brasil*: uma homenagem acadêmica. Rio de Janeiro: 7Letras, 200., p. 50–58.

BRASIL. Ministério da Saúde. Conselho Nacional de Saúde. *Resolução nº 466, de 12 de dezembro de 2012.* Diretrizes e normas regulamentadoras de pesquisas envolvendo seres humanos. Diário Oficial da União, Brasília, DF, 13 jun. 2013.

BRASIL. Ministério da Saúde. Conselho Nacional de Saúde. *Resolução nº 510, de 07 de abril de 2016.* Diretrizes e normas regulamentadoras de pesquisas envolvendo seres humanos em Ciências Humanas e Sociais. Diário Oficial da União, Brasília, DF, 24 maio 2016.

BRASIL. *Lei nº 13.709, de 14 de agosto de 2018.* Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). Diário Oficial da União, Brasília, DF, 15 ago. 2018.

CARVALHO, Fernanda Potiguara. *O ser atrás do dado:* limites e desafios da anonimização e seus reflexos nos requisitos estabelecidos pela LGPD. 2021. Dissertação (Mestrado em Direito). Programa de Pós-Graduação em Direito, Universidade de Brasília, Brasília, Distrito Federal, 2021.

COMITÊ DE ÉTICA EM PESQUISA DA PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL. *Declaração.* Porto Alegre-RS, 02 jul. 2013.

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

KNIES, Clarice Bohn; COSTA, Iara Bemquerer. *Banco de dados linguísticos VARSUL:* manual do usuário. UFRGS/UFSC/UFPR, 1996.

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

LABOV, William. *Principles of linguistic change*: Internal factors. Cambridge: B. Blackwell, 1994.

MONGUILHOTT, Isabel de Oliveira e Silva.; CALLOU, Dinah; LOPES, Célia Regina dos Santos; MOTA, Jacyra Andrade. *Bancos de dados linguísticos brasileiros*: planejamento, construção, governança e curadoria de coleções de dados. Mesa-redonda: Abralin em Cena 17, 26/06/2023. Disponível em: https://www.youtube.com/watch?v=YDeKCxw-p-A&t=2240s.

PINHO, Frederico António Sá Oliveira. *Anonimização de bases de dados empresariais de acordo com a nova regulamentação europeia de produção de dados*. 2017. Dissertação. (Mestrado em Segurança Informática) - Faculdade de Ciências da Universidade do Porto, Porto, Portugal, 2017.

VANDRESEN, Paulino. A montagem do banco de dados VARSUL: 1990 a 1996. *ReVEL,* edição especial n. 13, 2016. [www.revel.inf.br].

WEINREICH, Uriel; LABOV, William; HERZOG, Marvin. *Fundamentos empíricos para uma Teoria da Mudança Linguística*. Tradução de Marcos Bagno. São Paulo: Parábola Editorial. 2006 [1968].

ZABALA, Filipe. *Voice*: Tools for Voice Analysis, Speaker Recognition and Mood Inference, R package version 0.4.14, 2022. Disponível em https://CRAN.R-project.org/package=voice. Acesso em: 30/09/2022.