Pernambuco, v. 8, n.3, ago/dez. 2023
Revista Falange Miúda – Revista de Estudos da Linguagem
Mestrado Profissional em Letras – UPE – Garanhuns
ISSN 2525-5169 (on-line) - Site da Revista

Banco de dados orais: importância e funcionalidades para os estudos do português amazônico

Celeste Maria da Costa Ribeiro[1]
Geisy Rodrigues Ferreira[2]

Resumo

O propósito deste trabalho é demonstrar a relevância da constituição de um banco de dados orais dos falares amapaenses, com vistas a contribuir para a ampliação de pesquisas linguísticas, dadas as possibilidades de estudos baseados em corpora organizados e informatizados. Parte-se dos pressupostos teóricos da Linguística de Corpus e dos estudos de Sardinha (2004), Oliveira (2009) e Cruz, Bulhões e Fernandes (2004). Assim, este estudo busca apresentar e descrever a organização e constituição do banco de dados dos falares amapaenses, a partir dos registros de inquéritos provenientes de pesquisas realizadas pelo Grupo Atlas linguístico do Amapá, com o fim de disponibilizá-los eletronicamente à comunidade científica e acadêmica. O percurso metodológico segue os parâmetros de bancos orais já consolidados no Brasil, como o variação linguística no estado da Paraíba - VALPB, o Banco Programa de estudos sobre o uso da língua - PEUL, entre outros; o corpus que constituirá o referido banco advém da aplicação de questionários, relatos pessoais e de narrativas, com amostras representativas de três grupos de falantes do estado: amapaenses falantes do português brasileiro L1; indígenas usuários do português brasileiro L2; e franceses usuários do português brasileiro L2; os dados possibilitarão estudos de natureza fonético-fonológico, semântico-lexical, morfossintática e discursivo-pragmática. Esses dados serão disponibilizados em arquivos de áudio digitalizados, por localidade e perfil do falante. Os resultados iniciais apontam as contribuições que o banco de dados dos falares amapaenses trará para o desenvolvimento e o fomento da pesquisa linguística no Amapá, assim como para o ensino e a aprendizagem da língua portuguesa no estado.

Palavras-chave:

Banco de dados orais; pesquisa linguística; Português amapaense

Sobre os autores:

[1] Doutora em Linguística pela Universidade Federal do Rio de Janeiro.
EMAIL: celribeiro042002@gmail.com | **ORCID:** https://orcid.org/0000-0003-4934-515X
[2] Graduanda em Letras pela Universidade Federal do Amapá.
EMAIL: geisyrios52@gmail.com | **ORCID:** https://orcid.org/0009-0009-2605-6698

CONSIDERAÇÕES PRELIMINARES

As variedades do português brasileiro faladas no Amapá ainda são pouco estudadas, um maior conhecimento sobre elas pode ser alcançado somente a partir das pesquisas de Razky, Ribeiro e Sanches (2014; 2015; 2016) que resultaram na publicação do Atlas Linguístico do Amapá (ALAP) em 2017, evidenciando algumas das principais variantes linguísticas, sobretudo as de natureza fonético-fonológica e semântico-lexical empregadas neste estado. Assim, o referido atlas não só proporcionou um aumento na realização de estudos linguísticos, como também serviu de gatilho para o desenvolvimento de pesquisas acerca dos falares amapaenses.

Com base nos seus dados, o seu conjunto de dados foi sendo organizado e armazenado o que aponta para a constituição de um banco de dados sociolinguísticos do português amapaense. Objetivando subsidiar investigações linguísticas dessa variedade do português, em diferentes níveis e com diferentes finalidades, estamos propondo o projeto Banco de Dados dos Falares Amapaenses cujo objetivo central é a criação de um banco de dados linguísticos desses falares, numa perspectiva sincrônica.

Este estudo parte dos pressupostos teórico de Cruz, Bulhões e Fernandes (2004), Sardinha (2004) e Oliveira (2009) no tocante ao papel e à significância dos bancos de dados orais, sobretudo considerando que esses bancos se constituem solução para muitas lacunas e carências de pesquisas nas áreas de Dialetologia, Geolinguística e Sociolinguística, surgidas, geralmente, pelo custeio e tempo que requer uma pesquisa de campo. Ao passo que com o banco constituído, os dados linguísticos coletados serão disponibilizados à comunidade científica e educacional, fomentando o incentivo às pesquisas e às contribuições para o ensino de língua.

Assim, com vistas a trazer contribuições para investigações linguísticas do português amapaense, estamos propondo o desenvolvimento e a constituição do banco de dados orais amapaense, a fim de disponibilizar à comunidade acadêmico-científica insumos para a realização de estudos voltados ao falar amapaense em seus diferentes níveis, do morfofonológico ao discursivo.

Este texto está estruturado da seguinte forma: inicialmente, são apresentadas algumas concepções teóricas sobre a Linguística de Corpus, sobre o uso de bancos de dados linguísticos, destacando-se os principais bancos de dados do português brasileiro; em seguida, serão apresentados os procedimentos metodológicos que estão sendo utilizados na organização do banco dos falares amapaenses e, por fim, são explicitados alguns exemplos de como esse banco está sendo constituído.

A LINGUÍSTICA DE CORPUS

Esta é uma área que vem se desenvolvendo há décadas, graças aos estudos pioneiros de linguistas britânicos e americanos, os quais não se pode deixar de citar os nomes de John Sinclair, Geoffrey Leech, Douglas Biber e Jan Svartvik, que ao realizarem suas pesquisas descobriram novas metodologias de estudo da língua, partindo do uso real para formular teorias linguísticas, tal como a iniciativa de compilação de textos. Desta feita, surgiu o *Survey of English Usage* (SEU), o primeiro corpus não computadorizado considerado representativo do inglês compilado por Randolf Quirk e sua equipe na década de 1953.

Segundo Sardinha (2004) já havia organizações de corpora antes da invenção do computador, como ilustrações dessas organizações destacam-se Alexandre, O Grande, com o Corpus Helenístico na Grécia Antiga e corpora de citações da Bíblia na Antiguidade e Idade Média. No entanto, foi graças aos avanços tecnológicos que essa forma de fazer pesquisa se desenvolveu significativamente, haja vista que os equipamentos computacionais passaram a oferecer a possibilidade de construção de corpus eletrônicos e, assim, reduzir o tempo de organização e diminuir a exaustão do trabalho manual.

Um dos primeiros corpora representativos nos estudos da Linguística de Corpus é o SEU - *Survey of English Usage*, cujo *corpus* do inglês foi planejado para alcançar a marca de 1 milhão de palavras e foi organizado manualmente com fichas feitas de papel contendo os dados a serem analisados gramaticalmente, servindo de base para organização de corpora posteriores, como o Brown Corpus, lançado em 1964 pela Universidade de Brown, composto por um milhão de palavras do inglês americano, sendo considerado o primeiro *corpus* do inglês americano; posteriormente foram lançados banco de *corpora* como o *American Heritage Intermediate Corpus (AHI)* e o *Corpus of Spoken American Learner English.*

No que se refere ao inglês britânico, o primeiro corpus desenvolvido na Inglaterra, em parceria com a Noruega, foi denominado *Lancaster-Oslo/Bergen Corpus (LOB).* O corpus britânico-norueguês é caracterizado por ter o tamanho e formato muito próximos do *Brown Corpus*, que serviram de modelo para os demais corpora surgidos como o *British National Corpus* e *o Band of English*, todos de importância inquestionável para o estabelecimento da Linguística de Corpus como linha de pesquisa.

O desenvolvimento desse campo se deu desde os primeiros esforços dos linguistas na construção de *corpus* manualmente até chegar na década de 80 com as invenções tecnológicas que possibilitaram as primeiras compilações de *corpora* eletrônicos. A partir dos anos 80, "a área se expandiu devido a condições favoráveis em diferentes aspectos: sócio-históricos, acadêmicos, tecnológicos e pragmáticos" (OLIVEIRA, 2009, p. 56), isso porque as inovações tecnológicas, a divulgação e o incentivo da pesquisa no ramo, bem como otimização da

metodologia de pesquisa adotada possibilitaram o florescimento da Linguística de Corpus. Para Sardinha (2004) essa área:

> ocupa-se da coleta e exploração de corpora, ou conjuntos de dados lingüísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador. (SARDINHA, 2000, p.325)

Atualmente, a Linguística de Corpus está intimamente relacionada ao uso do computador, caracterizando-se por uma metodologia seguida de coleta, sistematização de dados e análises feitas com auxílio de ferramentas computacionais sob a abordagem de estudo empírico. Esse ramo parte do trabalho com dados da língua em uso orais e/ou escritos, organização de *corpus* e extração de informações analisadas a partir do método descritivo-interpretativo, fornecendo novas perspectivas para o estudo da língua. Conforme Oliveira (2009):

> os estudos de corpus caracterizam-se pela busca de tendências, probabilidades ou padrões de ocorrência ao lidarem com grande quantidade de dados. Nesses casos, os números servem de base para que estes padrões possam ser identificados e, então, interpretados pelos pesquisadores. Os resultados quantitativos produzidos com base no corpus são assim indicadores numéricos que devem ser discutidos à luz de diferentes posicionamentos teórico-metodológicos, para serem compreendidos. (OLIVEIRA, 2009, p.51)

Convém dizer que esse campo abre possibilidades para estudos de diferentes enfoques como fonético-fonológicos, tradutórios, discursivos, lexicográficos e pragmáticos, a serem desenvolvidos com base em *corpora* construídos a partir de textos escritos e/ou orais. Porém ressalta-se que uma pesquisa em corpus precisa ser norteada por objetivos bem delineados, dispor de dados linguísticos autênticos e de uma teoria-base alinhada ao que se propõe para, assim, realizar-se a identificação de tendências, observação de fenômenos linguísticos e constatação de variação da língua em uso.

Em relação às atividades que descrevem as ações que caracterizam a Linguística de Corpus, Kennedy (1998) destaca quatro principais: organização de corpora, invenção de ferramentas computacionais próprias, trabalhos de descrições linguísticas e aplicação dos estudos com base em *corpus* para as demais áreas de interesse linguístico. Por meio dessas ações, essa área vem trazendo importantes contribuições para os domínios da tradução, processamento de linguagem natural, linguagem artificial, mostrando-se ser vantajoso e profícuo no âmbito de fornecer as bases para a formulação de teorias linguísticas, para a construção de gramáticas e para o ensino de línguas.

BANCO DE DADOS ORAIS

O banco de dados é uma ferramenta desenvolvida na área de Tecnologia da Informação para comportar dados e informações com características definidas, podendo ser acessados, analisados e servirem de base para análises de diferentes finalidades para fins empresariais, institucionais, informacionais e científicos. É a partir desse recurso que os corpora eletrônicos têm ganhado forma "a fim de servir de suporte para toda e qualquer investigação linguística, assim como para aplicação tecnológica envolvendo engenharia de fala" (CRUZ, BULHÕES E FERNANDES, 2004, p. 194).

O uso desse recurso aproxima ainda mais a Linguística de Corpus da área da Informática, uma vez que os resultados de pesquisas ganham corpus a partir do fator eletrônico, pois com as invenções tecnológicas referentes à criação de softwares, modernização de computadores e organização de dados em maior dimensão, somados à possibilidade de divulgação e acesso online de informações, tem-se a possibilidade de construção de banco de dados cada vez mais robustos. Trata-se, então, de uma coletânea de dados linguísticos construída computacionalmente através da qual é possível registrar, armazenar e promover o acesso facilitado às informações por meio da consulta, podendo ser classificados em orais, escritos e orais-escritos de acordo com a modalidade da língua atendida.

Segundo Oliveira (2009), os corpora podem ser classificados em gerais e especializados. Os primeiros são caracterizados pela variedade do conteúdo, abarcando gêneros, discursos, temas e autores diversos que ensejam a realização de pesquisas variadas, enquanto os segundos são marcados pela especificidade dos dados selecionados com base em objetivos predeterminados. Ainda sobre a classificação de bancos de dados, Sardinha (2004) trabalha com a nomenclatura específica adotada em Linguística de Corpus baseada em conteúdo e propósito. Segundo o teórico, um corpus pode ser classificado em modo escrito ou falado pela modalidade da língua comportada; quanto ao tempo, pode ser sincrônico, diacrônico, contemporâneo ou histórico, a depender do contexto de produção dos dados; no que se refere ao conteúdo, o corpus pode ser especializado, regional ou multilíngue, estando diretamente relacionado à variabilidade e à especialidade apresentada.

Outros critérios de definição de corpora são a autoria e a finalidade. Quanto à primeira, há duas possibilidades de denominação, os corpora de aprendiz ou de língua nativa. Em relação à segunda, apresentam-se corpora de estudo, de referência ou de treinamento, construídos para subsidiar pesquisas linguísticas de viés descritivo, analítico e contrastivo. Há outros parâmetros de classificação de corpora em relação à pluralidade de autores, integralidade dos textos apresentados e quanto à renovação, no caso desse último, aponta-se para a possibilidade de atualização constante de dados. Desse modo, é possível compreender que a criação de corpora eletrônicos pressupõe a coleta, seleção e

sistematização de dados linguísticos, de acordo com propósitos somados à natureza dos dados que serão fatores caracterizadores do banco a ser construído.

Cruz, Bulhões e Fernandes (2004, p. 194) apontam para a "necessidade de documentar e disponibilizar eletronicamente dados de fala espontânea representativos do português brasileiro", cabendo ressaltar que os corpora de textos escritos são de extrema relevância para os estudos do uso da língua escrita; no entanto, em comparação com as pesquisas de língua falada, certamente os corpora de língua oral são mais abundantes e prestigiados. Desse modo, torna-se necessário conhecer mais a realidade oral das variedades do português brasileiro, oportunizando, assim, o desenvolvimento e o aumento de bancos de dados orais, para fins de pesquisas linguísticas.

BANCOS DE DADOS ORAIS BRASILEIROS

No Brasil, temos bancos de dados orais constituídos que se encontram disponíveis por meio de acesso online, podendo ter seu conteúdo não só consultado, como também utilizado para desenvolvimento de pesquisas na área. Alguns desses bancos são formados por *corpora* especializados, outros são *corpora* mais gerais marcados pela diversidade de textos orais e escritos com as variações do português brasileiro. Entre esses bancos destacam-se o Norma Urbana Oral Culta do Rio de Janeiro (NURC), o Programa de Estudos sobre o Uso da Língua (PEUL), o Variação Linguística na Região Sul do Brasil (VARSUL) e o Variação Linguística no Estado da Paraíba (VALPB), todos organizados por pesquisadores e grupos de pesquisas vinculados às universidades brasileiras. A seguir descrevemos sucintamente o perfil de cada um.

O NURC, projeto Norma Urbana Oral Culta do Rio de Janeiro, é um banco de dados orais da variedade culta da língua portuguesa falada; ele teve suas origens no final da década de 1960, focalizando cinco capitais brasileiras: Recife, Salvador, São Paulo, Rio de Janeiro e Porto Alegre; seu objetivo central é coletar sistematicamente material que permita a análise da linguagem oral culta do português brasileiro em seus diversos níveis. O banco comporta registros feitos nas décadas de 70 e 90 do século XX, abrangendo dados sincrônico do português contemporâneo com registros da oralidade em situações de uso da língua formal como em entrevistas, aulas, conferências, palestras além de outros contextos. Os dados que fazem parte do acervo do Projeto NURC têm sido utilizados para a elaboração de um grande número de trabalhos acadêmicos apresentados em encontros científicos ao redor do mundo.

O PEUL, Programa de Estudos sobre o Uso da Língua, volta-se para os estudos ligados às modalidades da língua escrita e falada com enfoque nas variações e na mudança linguística. Segue uma orientação essencialmente baseada na Sociolinguística Variacionista, os pesquisadores que integram o PEUL vêm se dedicando, ao longo de mais de quarenta anos, à análise da língua em uso, em

especial a fenômenos de variação e mudança na variedade carioca. O objetivo geral do projeto é a descrição da língua em uso e a sua interrelação com aspectos sociais, estruturais e funcionais; os dados podem ser acessados por estudantes e pesquisadores interessados em *corpora*, onde se encontram disponíveis amostras divididas em Censo de 1980, Censo 2000, Indivíduos Recontactados e mais amostras do discurso Jornalístico, da fala infantil, interacional e do projeto Mobral. Esse banco é de grande dimensão e comporta dados orais, escritos e semiortográficos.

Outro banco de destaque é o VARSUL, Variação Linguística na Região Sul do Brasil, desenvolvido por pesquisadores da Universidade Federal do Rio Grande do Sul, Pontifícia Universidade Católica do Rio Grande Sul, Universidade Federal de Santa Catarina e Universidade do Paraná com a finalidade de descrever o português escrito e falado nas cidades mais representativas da região. As áreas correspondem a quatro cidades dos três grandes estados da região: Rio Grande do Sul (Porto Alegre, Flores da Cunha, Panambi e São Borja), Santa Catarina (Florianópolis, Blumenau, Lages e Chapecó) e Paraná (Curitiba, Pato Branco, Londrina e Irati). O VARSUL disponibiliza o acesso a dados de 288 entrevistas realizadas em zonas urbanas, composta de áudios de tempo entre 5 e 15 minutos. O banco VARSUL se constitui como um banco de dados linguísticos e socioculturais.

Destaque ainda para o VALPB, Projeto Variação Linguística no Estado da Paraíba, fundado nos anos 90 pelo pesquisador Dermeval da Hora com o objetivo de caracterizar o português falado pelos paraibanos, a partir de aspectos fonético-fonológicos e gramaticais. A dimensão do banco é significativa e abrange três amostras, *Corpus 1993, Recontato 2015 e Corpus 2018* produzidas por informantes estratificados em sexo, faixa etária e escolaridade.

Somados a esses, juntam-se outros bancos de dados linguísticos de escopo menor, mas que compartilham dos pressupostos teóricos-metodológicos da Sociolinguística Variacionista, considerando, na seleção dos falantes, variáveis extralinguísticas como idade, sexo, escolaridade, origem do falante, entre outros fatores caracterizadores de comunidades de fala. Nesse viés citamos o projeto banco de dados dos falares amapaenses sobre o qual abordaremos no tópico seguinte.

PROJETO BANCO DE DADOS DOS FALARES

Este projeto tem suas origens ligadas ao Atlas Linguístico do Amapá – ALAP, publicado em 2017, que mapeou os principais usos linguísticos do estado amapaense, por meio de cartas fonéticas e lexicais, que possibilitaram um conhecimento melhor da realidade linguística do estado do Amapá. Desse modo, esse atlas deu origem a um *corpus* coletado a partir dos registros de fala dos moradores de 10 cidades amapaenses, "considerando a densidade demográfica e populacional, a priori, além de critérios históricos (tempo de origem), econômicos

e socioculturais" (RAZKY, RIBEIRO e SANCHES, 2017, p. 306).

Assim, esses registros passaram a constituir o principal acervo de dados orais para o desenvolvimento de estudos e pesquisas sobre o português amapaense, concretizadas em monografias de conclusão de curso, artigos científicos publicados e dissertação, além de apresentações e comunicações orais em eventos acadêmico-científicos pelo país.

Considerando o referido acervo somado a outros dados coletados para pesquisa linguística no estado, como o caso de registros de falantes que utilizam o português brasileiro como segunda língua (L2) surgiu a iniciativa de organizar todo esse material para a constituição e composição de um Banco de Dados Orais, relativo a registros de fala de usuários do português brasileiro, moradores do Amapá, tanto na condição de usuário do português como L1 quanto como L2. Acreditamos que a criação deste banco permitirá a realização de novos e diversificados estudos, buscando-se um maior conhecimento e entendimento da variedade do português falada no estado amapaense. Este projeto busca, sobretudo, apresentar, analisar e evidenciar a realidade linguística amapaense, por meio dos dados armazenados.

É sabido que o Norte é a maior região do Brasil em extensão territorial, embora não seja a região mais populosa, nas últimas décadas, tem crescido econômica e demograficamente atraindo imigrantes de vários estados brasileiros e, inclusive, de países fronteiriços, tais como venezuelanos e franceses. Por isso, acreditamos que em uma localidade onde transitam falantes de dialetos e línguas diferentes, haja maior abertura e possibilidade de investigação das consequências linguísticas, dada a convivência e o contato linguístico entre povos variados, pois não se pode deixar de reconhecer que analisar a língua é, antes de tudo, observar a sociedade na qual ela está inserida, considerando, nesses processos, os diversos fenômenos surgidos.

No caso do Amapá, o estado faz fronteira com dois países: Suriname e Guiana Francesa, porém a fronteira com o país francês é a mais expressiva, uma vez que é muito grande a interação entre estrangeiros que transitam na fronteira com o município amapaense de Oiapoque, onde há ainda a presença de várias aldeias de povos indígenas na região, cujos falantes fazem não somente uso de suas respectivas línguas maternas como também do português como L2. Diante desse cenário, observa-se o contexto diversificado no Amapá, no qual é notória a constituição de um campo linguístico que enseja uma gama de pesquisas.

Destaca-se ainda que os falantes desse estado empregam a língua portuguesa como língua materna (L1) e vivenciam uma situação de contato intensivo e extensivo com as línguas indígenas e a francesa no dia a dia. Conforme Sardinha (2004, p. 337) “um corpus deve ser planejado e concretizado seguindo critérios linguísticos de seleção”, em função disso, o objetivo dessa proposta é construir um Banco de Dados linguísticos que abranja tanto os usos do português L1, quanto os usos do português L2, falados nas áreas geossocioculturalmente representativas

do Amapá, destacando as suas funcionalidades, abrangência e importância para a representatividade do português falado no estado.

Assim, o referido banco está em fase de construção, cujo *corpus* comporta horas de gravações relativas a registros orais de amapaenses falantes de português brasileiro como primeira língua (PB L1) e também de franceses e indígenas falantes de português brasileiro como segunda língua (PB L2); todos esses falantes estão distribuídos em grupos sociais relativos a sexo, escolaridade, idade, etnia e língua materna e os dados de fala provenientes de cada um permitem o desenvolvimento de pesquisas diversas acerca dos principais usos feitos pelos falantes, no tocante a aspectos fônicos, morfossintáticos, semântico-lexicais e discursivo-pragmáticos da língua portuguesa. Dessa forma, salientamos que este banco de dados poderá fornecer bases para a realização de análise e descrição de fenômenos linguísticos que caracterizam o português amapaense, bem como poderá ser utilizado ainda como um recurso didático, no trabalho com a variação linguística nas aulas de língua portuguesa.

METODOLOGIA

O Banco de Dados dos Falares Amapaenses segue procedimentos metodológicos que orientam os estudos sobre montagem, instalação da ferramenta e sua organização. Os dados que farão a composição do banco de registros orais são provenientes de inquéritos realizados para fins de produção do Atlas Linguístico do Amapá (RAZKY, RIBEIRO E SANCHES, 2017) e de estudos/pesquisas de natureza sociolinguística na região; todos esses dados totalizam cerca de 150 horas de gravação e abarcam 90 informantes, todos moradores do estado, distribuídos entre falantes de PB L1 e falantes de PB L2.

ESPECIFICIDADE DA AMOSTRA

A coleta de dados caracterizadora de comunidades de fala, nos moldes labovianos (LABOV, 2008 [1972]), é predominante nos bancos de dados já constituídos; a sua reprodução viabiliza a comparabilidade de amostras de comunidades distintas. Para o banco com dados do PB L1 que estamos propondo, foram selecionadas, inicialmente, as 10 localidades representativas do estado do Amapá, pontos de inquérito do ALAP (RAZKY, RIBEIRO& SANCHES, 2017) a saber, a capital Macapá, Santana, Mazagão, Laranjal do Jari, Pedra Branca do Amapari, Porto Grande, Tartarugalzinho, Amapá, Calçoene e Oiapoque. Os dados do PB L2 concentram apenas falantes moradores de Oiapoque.

Todos os falantes estão organizados por escolaridade (ensino fundamental, o médio e o superior); por sexo (mulheres e homens); por faixa etária (18 a 35 anos e acima de 40 anos); acrescentam-se para os dados do PB L2, as categorias de

origem do falante (franceses e indígenas); grau de contato com o português brasileiro (baixo ou alto); e ainda língua materna (francesa e indígena). Até o momento o banco de dados contará com 60 entrevistas para a categoria de falantes de PB L1; 16 inquéritos com falantes indígenas de PB L2 e 14 com falantes franceses também de PB L2; totalizando, assim, 90 entrevistas sociolinguísticas, na amostra de comunidades de fala. Vale dizer que este banco, posteriormente, sofrerá ampliação em seu conjunto de dados, haja vista a perspectiva de serem incorporados registros de fala de outros grupos da região, como os quilombolas, afrodescendentes e ribeirinhos, além de dados mais atuais do PB L1 do estado, sobretudo da capital Macapá.

O perfil descrito acima foi identificado na fase de triagem e levantamento dos dados orais para que se pudesse pensar na organização do futuro banco, o qual deverá ser projetado através de estudos sobre a área que envolve a variação e heterogeneidade linguísticas, assim como sobre a criação da ferramenta. Posteriormente, passou-se à sistematização dos registros de fala por categoria: cidade, sexo, idade, escolaridade, grau de contato e língua materna de cada falante. Para organizar e sistematizar os muitos arquivos de áudios, foi elaborada uma codificação para cada falante, considerando o perfil descrito anteriormente.

Desse modo, essa codificação visa facilitar e organizar com clareza e ordenamento cada arquivo. A título de exemplificação para os falantes de PB L1, temos um arquivo denominado 01AMS, onde se lê: falante da cidade de Macapá (01), jovem (A), mulher (M), de ensino superior(S). Para os falantes de PB L2, temos o arquivo 10BHFEI, para o qual se lê: falante da cidade de Oiapoque (10), mais velho (B), homem (H), de ensino fundamental(F), grau de contato com o PB elevado (E) e língua materna indígena (I).

Esses arquivos que estão sendo codificados concretizarão o banco de dados dos falares amapaenses. Dessa maneira, esse Banco terá disponível acervo oral tanto de falantes de PB L1, como de PB L2 para posterior observação, identificação, análise e evidências dos principais usos linguísticos feitos pelos falantes amapaenses de português brasileiro, seja na condição de L1 ou de L2.

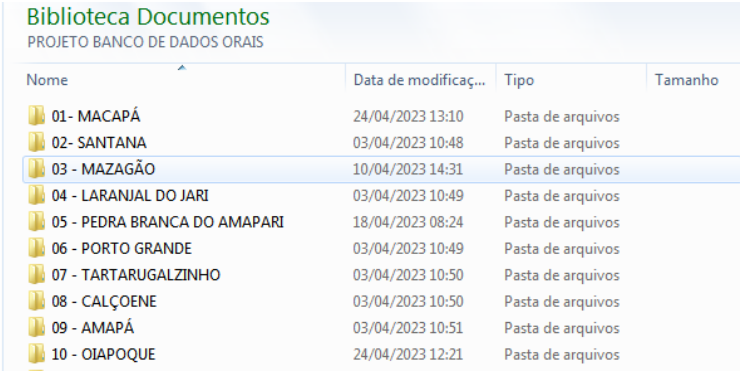
ANÁLISE DE DADOS

O projeto Banco de Dados Orais que se encontra em fase de execução, uma vez concluído, evidenciará informações relevantes para o conhecimento da realidade linguística que caracteriza a comunidade de fala amapaense. Ressalta-se que esse conhecimento se torna importante para a área variacionista, pois não apenas evidencia o fenômeno da variação linguística, como também reflete o perfil linguístico do estado amapaense e da própria língua portuguesa falada no Brasil.

Atualmente o *corpus* que constituirá o banco conta com registros orais de indivíduos amapaenses, falantes de PB L1 e PB L2 armazenados em arquivos MP3,

que estão sendo rigorosamente sistematizados e organizados, conforme a figura 1 seguinte.

Figura 1 – Organização dos dados de PB L1 por município amapaense



Fonte: ALAP (Razky, Ribeiro e Sanches, 2017)

Ao consultar os inquéritos por município que compõe o banco, pode-se ter acesso aos dados de cada falante de PB L1 e que se encontram organizados por questionários: o Fonético-fonológico (QFF), o Semântico-lexical (QSL) e as Narrativas. Convém ressaltar que esses questionários correspondem aos que foram utilizados na coleta de dados para a produção do atlas linguístico do Brasil (CARDOSO *et al.*, 2014). Os registros estão organizados conforme parâmetros da pesquisa dialetológica e geolingüística, utilizados na coleta de dados do Atlas Linguístico do Amapá (RAZKY, RIBEIRO & SANCHES, 2017) que, por sua vez, segue os parâmetros metodológicos do Atlas Linguístico do Brasil (CARDOSO *et al.*, 2014). Os dados dos falantes de PB L2 estão distribuídos apenas em relatos pessoais, visto que a coleta desses não seguiu o mesmo percurso metodológico do ALAP. A figura 2 seguinte evidencia a organização dos dados por falante e tipo de inquérito, se questionário ou narrativa.

Figura 2 – Organização dos dados de PB L1 por falante

Fonte: ALAP (Razky, Ribeiro e Sanches, 2017)

Os arquivos com os dados fonético-fonológicos apresentam, em média, 159 registros audíveis produzidos por cada falante de cada município, a partir do inquérito de questões fechadas. Esses dados estão diretamente relacionados à

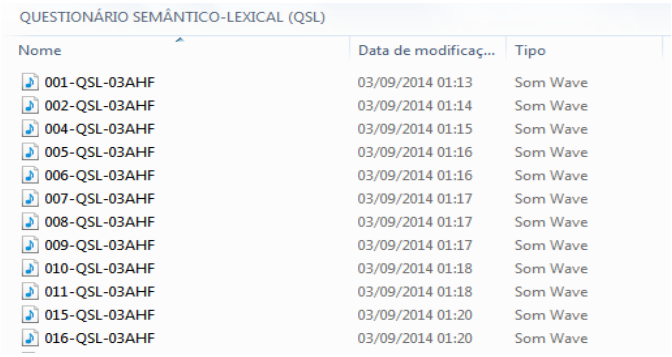
produção oral e evidenciam variações a nível da pronúncia de palavras, rendendo estudos sobre fenômenos linguísticos, como monotongação, rotacismo, metaplasmos, palatalização, assimilação, entre outros. Na figura 3 abaixo é possível observar alguns desses dados.

Figura 3 – Dados do questionário fonético-fonológico

Fonte: ALAP (Razky, Ribeiro e Sanches, 2017)

Com relação aos dados de natureza semântico-lexical, tem-se, aproximadamente, 202 registros orais produzidos por falantes a partir de perguntas semi-abertas. Esses registros apontam as variantes lexicais utilizadas pelos amapaenses para nomear fenômenos da natureza, partes do corpo humano, ações do dia a dia, brincadeiras, vestuário, alimentação, habitação, flora, vida urbana, hábitos, religião, crenças, entre outros temas ligados ao cotidiano das pessoas. A figura 4 seguinte ilustra um arquivo com dados do QSL.

Figura 4 – Dados do questionário semântico-lexical



Fonte: ALAP (Razky, Ribeiro e Sanches, 2017)

Por fim, convém dizer ainda que o banco de dados dos falares amapaenses disponibilizará também registros de narrativas curtas e pessoais, as quais versam sobre experiências, opinião, visão, relatos sobre festividades locais e acontecimentos mais marcantes na vida do falante, na cidade. Esses arquivos destacam dados de fala mais espontâneos, por meio dos quais é possível observar fenômenos fonético-fonológicos, morfossintáticos, semânticos, pragmáticos e discursivos, que evidenciam a correlação entre os mecanismos linguísticos

utilizados para a comunicação e os aspectos do contexto de inserção de quem fala, ensejando também estudos culturais.

Além desse formato, há ainda os dados de fala relativos aos falantes indígenas e franceses, usuários do PB L2, os quais, por meio de relatos pessoais com duração em média de 5 a 30 minutos cada um, possibilitam desenvolver estudos, que se voltam para esse perfil do PB, acerca de determinadas realizações, sobretudo, no campo fonético-fonológico. Acrescente-se ainda que o banco de dados orais caracterizadores dos falares dessa região, permitirá a difusão de novos estudos e pesquisas científicas, possibilitando um melhor conhecimento do perfil linguístico da língua portuguesa falada nesse local, a partir da descrição de fenômenos linguísticos, tanto na condição de PB L1 como de PB L2, contribuindo, assim, para o aumento do acervo de estudos variacionistas, sob um aspecto inovador em uma localidade do extremo norte do país.

Atualmente, o projeto banco de dados dos falares amapaenses encontra-se na fase de conclusão da codificação e organização dos dados, seguindo para a etapa de armazenamento em *software* adequado para posterior fase de testes e, finalmente, consolidação e conclusão do banco.

CONSIDERAÇÕES FINAIS

O desenvolvimento deste trabalho segue-se dos referenciais teóricos-metodológicos para a criação de banco de dados representativos das variedades linguísticas, nesse caso, do português falado no estado Amapá. Por se tratar de um quantitativo numeroso de registros a serem informatizados, exigiu-se, embora com a viabilidade proporcionada pelas tecnologias de informação, a realização de atividades graduais e concentradas em reconhecimento, transferência, download, distribuição e alocação de arquivos de acordo com critérios definidos na pesquisa.

Ressaltamos que a organização e a disponibilidade desse tipo de banco são de extrema relevância para as esferas científica e social, visto que os dados disponíveis podem ser tanto objeto de pesquisas, como suporte didático cujo enfoque se faz por meio de abordagens sobre a língua falada e as variações linguísticas no ensino de língua.

Por meio do banco de dados dos falares amapaenses pretende-se trazer contribuições para a identificação, levantamento e constituição dos perfis sociolinguísticos das variedades faladas do português, no estado amapaense e, desse modo, somá-lo aos outros bancos linguísticos já criados, com vistas a colaborar nos estudos da língua portuguesa falada no Brasil e, sobretudo, no estado nortista.

REFERÊNCIAS

CRUZ, R.; BULHÕES, J. U.; FERNANDES, L.S. Banco de Dados Orais: uma nova

perspectiva aos estudos do português brasileiro. *Rev. Est. Ling.*, Belo Horizonte, v. 12, n 2, p. 193-212, 2004.

KENNEDY, G. *An introduction to Corpus Linguistics*. New York: Longman, 1998.
LABOV, W. *Padrões Sociolinguísticos.* Tradução: Marcos Bagno et al. São Paulo: Parábola, 2008.

NURC-RJ Projeto Norma Linguística Urbana Culta. Rio de Janeiro: Universidade Federal do Rio de Janeiro, Faculdade de Letras, 2023. Disponível em: https://nurc.fflch.usp.br/o-nurc-brasil-origens. Acesso em: 20 marc. 2023.

OLIVEIRA, L.; Linguística de Corpus: teoria, interfaces e aplicações. *Revista Matraga*, Rio de Janeiro, v.16, n.24, p. 48-76, 2009.

PEUL. Programa de Estudos sobre o Uso da Língua. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2023. Disponível em: https://peul.letras.ufrj.br/. Acesso em: 22 mar. 2023.

PROJETO VARSUL. Variação Linguística no Rio Grande do Sul. Rio Grande do Sul: Universidade Federal do Rio Grande do Sul, 2023. Disponível em: https://www.varsul.org.br/. Acesso em: 22 mar. 2023.

PROJETO VALPB. Variação Linguística no Estado da Paraíba. Paraíba: Universidade Federal da Paraíba, 2023. Disponível em: http://projetovalpb.com.br/. Acesso em: 20 mar. 2023.

RAZKY, A.; RIBEIRO, C.M.R.R.; SANCHES, R. D. *Atlas Linguístico do Amapá*. São Paulo: Labrador, 2017.

RAZKY, A.; RIBEIRO, C.M.R.; SANCHES, R. D. O projeto atlas linguístico do amapá (ALAP): caminhos percorridos e estágio atual. *Alfa*, São Paulo, v.61, n.2, p.303-317, 2017.

SARDINHA, T. B.; Linguística de Corpus: histórico e problemática. *D.E.L.T.A*, São Paulo, v.16, n. 02, p. 323-367, 2000.

SARDINHA, T. B. *Linguística de corpus***.** São Paulo: Manole, 2004.

Pernambuco, v. 8, n.3, ago/dez. 2023
Revista Falange Miúda – Revista de Estudos da Linguagem
Mestrado Profissional em Letras – UPE – Garanhuns
ISSN 2525-5169 (on-line) - Site da Revista

Oral databases: importance and functionalities for the studies of the amazon portuguese

Celeste Maria da Costa Ribeiro[1]
Geisy Rodrigues Ferreira[2]

Abstract

The purpose of this work is to demonstrate the relevance of the constitution of oral databases of Amapaense dialects, with views to contribute to the amplification of linguistics research, given the possibilities of studies based on organized and computerized corpora. It starts from the theoretical postulates of Corpus Linguistics and studies of Sardinha (2004), Oliveira (2009) e Cruz, Bulhões e Fernandes (2004). Thus, this study looks to present and describe about the constitution of the database of Amapaense dialects, going from survey records from research conducted by the Linguistic Atlas of Amapá Group, with the finality of making it available electronically to the scientific and academic community. The methodological approach follows the oral databases parameters already consolidated in Brazil, such as the Linguistic Variation in the State of Paraiba Project – VALPB, the Studies Program about the Usage of Language – PEUL, among others; the corpus that will constitute the database comes from the application of questionnaires, personal accounts and narratives, with representative samples from three groups of speakers of the state: Amapaenses as speakers of Brazilian Portuguese as language L1; Indigenous users of Brazilian Portuguese L2; and Frenches users of Brazilian Portuguese L2, the data will possibility studies of phonetic-phonology nature, semantical-lexical, morphosyntactic and discursive-pragmatic. Those data will be made available in digitalized audio files, by localization and profile of the speaker. The initial results point to the contributions that the database of Amapá speeches tend to bring to the development and promotion of linguistic research in Amapá, as well as to the teaching and learning of the Portuguese language in the state.

Keywords:

Oral databases; Linguistic Research; Amapaense Portuguese