DOI: https://doi.org/10.14393/DLv19a2025-16
e019016

Artigo

Possibilidades de uso de acervos de dados do português brasileiro em Fonética Forense

Potential uses of Brazilian Portuguese data repositories in Forensic Phonetics

Daniel Fonseca VIEIRA[*]
Renata Regina PASSETTI[**]
Pablo ARANTES[***]

RESUMO: Este trabalho tem como objetivo realizar um levantamento dos acervos de dados de língua e fala do português do Brasil e analisar seu potencial de aplicação em tarefas de fonética forense, em especial a comparação de locutor. Foram identificados 45 acervos de dados de fala de diferentes regiões brasileiras, obtidos por meio de consulta a especialistas em sociolinguistas e dialetologia, além de pesquisa em fontes bibliográficas. A avaliação do potencial de uso destes acervos para os propósitos da fonética forense baseou-se em um questionário que abordou, entre outros aspectos, as características gerais de cada acervo, os tipos de materiais fornecidos, características dos participantes e a área geográfica coberta e as condições de acesso e uso do material coletado. Embora os acervos de língua e fala identificados atendam aos propósitos para os quais foram originalmente pensados, como o registro de fenômenos de variações nos níveis lexical, sintático e conversacional, quando se consideram as necessidades específicas da tarefa de comparação de locutor no campo da fonética forense, conclui-se que as opções são relativamente mais limitadas. Dos 45 acervos levantados, apenas sete servem de maneira mais adequada ao propósito de geração de estatísticas de distribuição de características linguísticas e fonéticas relevantes em tarefa de comparação de locutor, em especial quando baseada no uso de razão de verossimilhança (*likelihood ratio*). Os resultados evidenciam, ainda, a necessidade de ampliar a disponibilidade de acervos que sejam contemporâneos e que disponibilizem materiais de áudio para as regiões do país atualmente pouco representadas, uma vez que não foram encontrados acervos que disponibilizassem amostras de áudio para as regiões Centro-Oeste e Norte, e apenas um acervo com tais características para as regiões Sul e Nordeste. Diante desse cenário, é importante que projetos em andamento ou futuros considerem a disponibilização de materiais de áudio e informações complementares sobre os falantes e as gravações, de forma a ampliar a base de dados fonético-linguísticos representativa das diferentes regiões do país. As informações geradas nesta pesquisa são úteis tanto para especialistas em linguística e fonética forense quanto para pesquisadores de outras áreas de estudo da linguagem.
PALAVRAS-CHAVE: Fonética. Fonética Forense. Criminalística. Acervos de dados.

* Mestrando em Linguística (UFSCAR). Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos (UFSCar). São Carlos, SP – Brasil. danielvieira@estudante.ufscar.br
**Doutorado em Linguística (UNICAMP). Pós-doutoranda da Universidade Federal de São Carlos (UFSCar). São Carlos, SP – Brasil. re.passetti@gmail.com
***Doutorado em Linguística (UNICAMP). Professor Associado do Departamento de Letras da Universidade Federal de São Carlos (UFSCar). São Carlos, SP – Brasil. pabloarantes@ufscar.br

ABSTRACT: This study aims to survey existing data repositories of Brazilian Portuguese language and speech and to evaluate their potential applications in forensic phonetics, with a particular focus on speaker comparison tasks. A total of 45 speech corpora from various Brazilian regions were identified through consultations with specialists in sociolinguistics and dialectology, as well as through bibliographic research. To assess the potential use of these corpora for forensic purposes, a questionnaire was developed, addressing a range of aspects including general corpus characteristics, types of materials available, participant profiles, geographic coverage, and conditions for access and use of the collected data. Although the identified repositories fulfill their primary goal of documenting linguistic variation—especially at the lexical, syntactic, and conversational levels—they are not always suitable for the specific demands of speaker comparison within forensic phonetics. This task requires repositories that include high-quality audio recordings, metadata on speakers and recording conditions, and broad regional and demographic coverage to support the estimation of relevant linguistic and phonetic feature distributions. Among the 45 corpora surveyed, only seven were found to be adequately suited for use in speaker comparison analyses, particularly when employing likelihood ratio-based approaches, which require representative reference data. The results underscore a significant gap in the availability of appropriate and accessible resources for forensic phonetic applications in Brazil. In particular, there is a notable lack of corpora providing audio materials from the Central-West and North regions, and only one corpus includes such materials for the South and Northeast regions. This regional imbalance limits the ability to conduct robust speaker comparison analyses nationwide. Given these findings, it is essential that current and future corpus development initiatives consider including high-quality audio data, detailed metadata, and broad geographic representation. The outcomes of this research are relevant not only for linguists and forensic phonetics specialists but also for researchers working in related areas of language study, who may benefit from greater access to phonetically and regionally representative data.
KEYWORDS: Phonetics. Forensic Phonetics. Criminalistics. Data repositories.

Artigo recebido em: 02.10.2024
Artigo aprovado em: 25.03.2025

1 Introdução

A linguística forense é um campo de atuação e estudo científico interdisciplinar que aplica conhecimentos gerados pelos estudos linguísticos a contextos nos quais evidência de natureza linguística (registros de áudio, textos, entre outros objetos) são elementos relevantes em processos envolvendo a justiça. O tipo de análise linguística que pode ser mais útil em cada caso particular depende muito de sua natureza. Há, por exemplo, casos nos quais é preciso realizar análises de textos escritos para identificar padrões de escrita que podem sugerir a autoria de textos apócrifos ou cuja autoria é disputada. Em outros, o que está em jogo é a análise de registros de áudio,

seja para prover a transcrição do seu conteúdo, para traçar o perfil linguístico de um locutor ou, ainda, comparar uma voz cuja identidade é desconhecida com a voz de um ou mais suspeitos. Uma apresentação abrangente da linguística forense é apresentada em Olsson (2008).

Dentro da grande área da linguística forense, a fonética forense é o ramo responsável pela análise de amostras de fala em contextos legais, bem como do desenvolvimento de novos conhecimentos e métodos especificamente fonético-forenses (Jessen, 2008). Uma das tarefas mais comuns no contexto da fonética forense é a comparação de locutor (doravante CL), cujo objetivo é avaliar a probabilidade de uma mesma pessoa ter produzido (ou não) duas amostras de fala (Gold; French, 2011; 2019). Em um exame de CL são comparados os parâmetros linguísticos e fonéticos extraídos de amostras de fala de um locutor de identidade desconhecida, o que é chamado de amostra questionada (doravante AQ), com os parâmetros vindos de amostras de um ou mais falantes suspeitos de terem produzido a AQ, o que são chamadas de amostras padrão (doravante AP). O desafio na tarefa de CL é comparar o grau de convergência e de divergência entre AQ e AP, ou seja, o grau de similaridade entre elas, e identificar quão comuns são os parâmetros de fala examinados, tendo como referência uma determinada população, isto é, a tipicidade daquilo que é observado na AP.

Gold e French (2011; 2019) realizaram dois levantamentos sobre as práticas correntes entre profissionais do campo que mostram quais são os métodos de expressão de conclusão em laudos de CL mais utilizados. Os levantamentos constataram que uma das abordagens cuja adoção mais cresceu no intervalo entre os dois levantamentos é baseada no cálculo da chamada razão de verossimilhança, normalmente chamada de LR (abreviação do termo inglês *likelihood ratio*). O cálculo da LR estima o valor probatório da evidência avaliada no exame de CL, isto é, as semelhanças e diferenças encontradas entre AQ e AP, através da comparação da probabilidade de duas hipóteses com sentido contrário, uma que diz que AQ e AP têm

a mesma origem (foram produzidas pelo mesmo locutor) e outra que diz que a origem duas é diferente (Morrison, 2009; 2010). A probabilidade da hipótese de mesma origem é fundamentada na semelhança entre elas, enquanto a hipótese de origens diferentes considera o quão típicos são os padrões observados nas amostras em relação à distribuição desses padrões em uma população de referência.

A ideia que subjaz à aplicação da LR é que a similaridade entre as amostras em comparação não é suficiente para afirmar que ambas têm a mesma origem. Se apenas a semelhança é levada em conta, o exame de CL pode resultar em um falso positivo, uma vez que a semelhança encontrada entre AQ e AP poderia, em tese, ser também observada na comparação entre AQ e amostras produzidas por um número potencialmente elevado de outros locutores que partilham dos mesmos padrões linguísticos e fonéticos atestados no exame. Por isso, também é necessário considerar a tipicidade dos achados do exame de CL, isto é, quão comuns são na população relevante para o caso os padrões linguísticos e fonéticos extraídos da AQ. Se os padrões forem muito comuns na população, então cai o peso relativo da evidência em favor da hipótese de mesma origem. Se, ao contrário, a tipicidade é baixa, então a evidência em favor da hipótese de mesma origem ganha mais peso.

Para determinar a tipicidade, no entanto, é preciso recorrer a fontes de dados, em geral externas ao caso em exame, que descrevam a distribuição, na população relevante para o caso, dos parâmetros observados no exame, tanto os de natureza linguística (uso de léxico regional ou idiossincrático, ou de padrões morfossintáticos, por exemplo) quanto fonético-acústica (frequência fundamental, valores de formantes vocálicos, por exemplo). Uma fonte comum de dados para estabelecer distribuições desse tipo são acervos de dados de língua e fala[1], tais como *corpora* de fala ou atlas

[1] Usamos a expressão "acervo de dados de língua e fala" de forma abrangente. Isso inclui tanto corpora de fala quanto atlas linguísticos, bem como publicações derivadas deles, ainda que os produtos gerados por essas fontes de dados sejam distintos. A justificativa é que tanto os dados brutos coletados e disponibilizados por um corpus quanto produtos derivados de sua análise, como as cartas linguísticas de um atlas, são potencialmente relevantes para o estabelecimento da tipicidade de padrões linguísticos e fonéticos.

linguísticos, cujo material linguístico, na forma de amostras de áudio ou texto e cartas linguísticas, por exemplo, pode servir para a extração de padrões linguísticos e fonéticos prevalentes em uma população de interesse. Para exemplificar mais concretamente como acervos de dados podem contribuir para a linguística e fonética forenses, apresentaremos a seguir dois exemplos.

Cartas de um atlas linguístico podem indicar variantes lexicais predominantes em determinadas áreas geográficas. Assim, se, em uma região, predomina o uso da variante lexical "mormaço" e tanto na AQ quanto na AP (cujos produtores, em função de evidências independentes, se pode assumir que sejam oriundos da mesma área) aparece uma das variantes "normaço" ou "bormaço"[2], por exemplo, essa similaridade dá mais peso à hipótese de mesma origem para AQ e AP. Caso se verifique nas duas amostras em análise a ocorrência da variante predominante, esse padrão é de alta tipicidade e daria mais força à hipótese de origens diferentes.

Amostras de áudio de boa qualidade provenientes de acervos de fala podem ser usadas para gerar sumários estatísticos de parâmetros fonéticos que indicam a distribuição desse parâmetro na população de falantes. Distribuições populacionais permitem determinar qual é o grau de tipicidade do valor de um parâmetro acústico medido na AP. Suponhamos que o interesse em um determinado exame de CL recaia sobre o valor médio da frequência fundamental (f0) das amostras em comparação. Sabendo, por exemplo, que a média da f0 na população masculina brasileira é 135 Hz[3] e a média apurada na AP é 132 Hz, esse resultado, que caracteriza uma situação de alto grau de tipicidade para o achado do exame, dá força para a hipótese de origens diferentes para AP e AQ se houver baixa similaridade entre os valores encontrados em

[2] O exemplo hipotético foi inspirado pelo Questionário Semântico-Lexical (QSL) número 64 no Atlas Linguístico-etnográfico da Região Sul do Brasil – ALERS (Altenhofen; Klassmann, 2002, p. 144)

[3] Esse dado foi retirado de Cunha (2023), que usou amostras de fala do acervo Corpus Forense do Português Brasileiro (Portaria n° 934-Ditec/PF de 30 de julho de 2020) para gerar estatísticas populacionais da distribuição de f0 de falantes brasileiros do sexo masculino. Este é um exemplo concreto do aproveitamento de um acervo de fala para a geração de um recurso de apoio à prática da fonética forense.

AP e AQ. Caso a média observada em AP divergisse mais substancialmente do que é típico na população, esse resultado pesaria a favor da hipótese de mesma origem para AP e AQ, supondo alta similaridade nos valores encontrados em ambas, dada a baixa probabilidade de coincidência entre o padrão encontrado na AP e amostras de fala retiradas ao acaso da população de referência.

2 Justificativa e objetivos

No contexto nacional, existem muitos acervos de dados de fala do português brasileiro que foram compilados e reunidos para apoiar pesquisas em sociolinguística, dialetologia, linguística de *corpus*, entre outros campos da linguística. Alguns dos mais notáveis incluem o Programa de Estudos sobre o Uso da Língua (PEUL), que foi pioneiro na criação de acervos de dados para fins sociolinguísticos (Freitag, 2016), o Projeto da Norma Urbana Oral Culta (NURC), e o Projeto Variação Linguística na Região Sul do Brasil (Varsul). Como visto anteriormente, o estabelecimento da tipicidade de padrões linguísticos e fonéticos em uma população é um componente necessário para a aplicação do cálculo da LR em exames de CL e esses acervos podem ser usados para esse fim.

Duas questões se colocam a respeito das possibilidades de aproveitamento dos acervos de língua e fala como recursos de apoio ao exame de CL e justificam o presente trabalho. A primeira é a possibilidade de descoberta ou grau de visibilidade dos acervos existentes, uma vez que não há, até o momento, um catálogo ou repositório cujo propósito primeiro seja listar e caracterizar esses acervos. A segunda questão é que a maioria dos acervos não foi criada com o propósito de servir como recurso de apoio à linguística forense, de forma geral, ou os exames de CL, em particular, de modo que seria importante avaliar se eles podem servir adequadamente a essa finalidade.

O presente trabalho tem dois objetivos principais, que se relacionam com as justificativas apresentadas anteriormente. Em primeiro lugar, fazer um recenseamento

e caracterização dos acervos de dados de língua e fala existentes no cenário brasileiro. Em segundo lugar, avaliar de forma sistemática e motivada o potencial de aproveitamento de cada acervo para a geração de distribuição populacional de padrões linguísticos e fonéticos para uso no exame de CL.

A seção 3 apresenta a metodologia empregada para realizar o levantamento e os critérios usados para avaliar o potencial de cada acervo. A seção 4 apresenta os resultados da avaliação. Com base nela, indicamos, na seção 5, os acervos com maior potencial para aproveitamento como fonte de dados que permitam a aplicação da metodologia baseada em LR no contexto de exames de CL.

3 Metodologia

Na etapa inicial do projeto realizamos um levantamento dos acervos de dados a serem avaliados nesta pesquisa. Para isso, contatamos especialistas em sociolinguística e dialetologia de todas as regiões do Brasil, para que indicassem acervos de dados de língua e fala que poderiam contribuir para os objetivos deste trabalho[4]. Nessa fase, 26 acervos foram sugeridos. Além da consulta a especialistas, recorremos também a ferramentas de busca, como o Google e o Google Acadêmico, bem como repositórios acadêmicos institucionais. Os termos de busca utilizados para as buscas foram: "banco de dados linguístico", "dados linguísticos", "amostras de fala", "amostras linguísticas", "corpus linguístico", "sociolinguística", "atlas linguístico", "atlas sociolinguístico", "atlas fonético", "atlas geolinguístico", "atlas linguístico regional" e "repositório linguístico". Nessa etapa, foram encontrados 19 acervos.

A lista completa de acervos, contendo informações como referências bibliográficas que os descrevem e URLs para acesso digital, quando elas existem,

[4] Para a escolha dos especialistas usamos como fontes a lista de participantes do GT de sociolinguística da ANPOLL (https://anpoll.org.br/gt/sociolinguistica) e também pesquisas em motores de busca usando como palavras-chave termos como "sociolinguística", "dialetologia" e "bancos de dados". Foram contactados no total 47 pesquisadores, atuantes em todas as regiões do Brasil.

encontra-se no documento "info-acervos.ods" disponibilizado no repositório https://osf.io/avu9e/files/osfstorage/.

Foram excluídos do total de acervos levantados inicialmente aqueles que não se enquadraram nos propósitos da pesquisa. Foram desconsiderados acervos que não dão acesso aos dados coletados, que não divulgam informações relevantes sobre os dados ou, ainda, que o acesso ao material é possível apenas por mídia física, como CD-ROM ou fitas magnéticas. Os acervos excluídos e a justificativa para a exclusão são listados no Quadro 1.

Quadro 1 – Acervos de dados excluídos e critérios de exclusão.

Acervos de dados excluídos	Critério de exclusão
Amostras de áudio do NURC-SP sob a guarda do Centro de Documentação Cultural "Alexandre Eulalio" (CEDAE), da Unicamp	Impossibilidade de acesso às amostras
Pirá: A Bilingual Portuguese-English Dataset for Question-Answering about the Ocean	Não foram encontradas informações fundamentais para a tarefa de CL como regiões geográficas abrangidas e se há classificação de sexo/gênero e faixa etária dos informantes
Corpus de Textos Orais do Português Santareno (CTOPS)	Não foram encontradas informações gerais do acervo como tipos de materiais coletados e características dos informantes, além do projeto ter sido publicado apenas em formato físico (CD-ROM)
Aspectos linguísticos da fala londrinense: esboço de um atlas linguístico de Londrina.	Não foram encontradas informações gerais do acervo como tipos de materiais coletados e características dos informantes

Fonte: elaborado pelos autores.

Após a exclusão dos acervos desconsiderados, obtivemos uma lista com 45 diferentes acervos de várias regiões do Brasil. O Quadro 2 apresenta as regiões abrangidas pelos acervos.

Quadro 2 – Lista de acervos levantados separados por região geográfica brasileira.

Abrangência regional	**Acervos de dados de língua e fala**
Região Sul	1. Atlas Linguístico-Etnográfico da Região Sul do Brasil (ALERS)
	2. LínguaPOA
	3. Projeto Variação Linguística na Região Sul do Brasil (VARSUL)
	4. Atlas Linguístico do Paraná (ALPR)
	5. Atlas Geossociolinguístico de Londrina (AGeLO)
Região Sudeste	1. ALIP - Iboruna (Amostra Censo)
	2. NURC RJ
	3. NURC SP
	4. Programa de Estudos sobre o Uso da Língua (PEUL)
	5. Projeto SP2010
	6. C-ORAL-BRASIL (I)
	7. CORAA NURC-SP Minimal Corpus
	8. Esboço de um Atlas Linguístico de Minas Gerais (EALMG)
	9. Atlas Semântico-Lexical da Região do Grande ABC
	10. Atlas Semântico-Lexical de Caraguatatuba, Ilhabela, São Sebastião e Ubatuba - municípios do Litoral Norte de São Paulo
	11. Atlas Linguístico Pluridimensional do Português Paulista níveis semântico-lexical e fonético-fonológico do vernáculo da região do Médio Tietê
Região Nordeste	1. NURC Digital/Recife
	2. Norma Oral do Português Popular de Fortaleza (NORPOFOR)
	3. PORTAL - Variação linguística no português alagoano
	4. Projeto Variação Linguística no Estado da Paraíba (VALPB)
	5. Banco de dados falares sergipanos
	6. A língua portuguesa do semiárido baiano
	7. Programa de Estudos sobre o Português Popular de Salvador (PEPP)
	8. Estudos da Língua Oral do Cariri
	9. Dialetos Sociais Cearenses
	10. Português Oral Culto de Fortaleza (PORCUFORT)

	11. Atlas Linguístico da Paraíba (ALPB)
	12. Atlas Linguístico de Sergipe (ALS)
	13. Atlas Linguístico de Sergipe II (ALS II)
	14. Atlas Linguístico da Mata Sul de Pernambuco (ALMASPE)
	15. Atlas Linguístico do Estado do Ceará (ALECE)
	16. Atlas Linguístico de Pernambuco (ALiPE)
Região Centro-Oeste	1. Atlas Linguístico de Mato Grosso do SUL (ALMS)
	2. Atlas Linguístico da Mesorregião Sudeste de Mato Grosso (ALMESEMT)
Região Norte	1. Atlas Linguístico Sonoro do Pará (ALISPA)
	2. Atlas Geolingüístico do Litoral Potiguar (ALiPTG)
	3. Atlas Linguístico do Amazonas (ALAM)
	4. Atlas linguístico do Amapá
Multirregional[5]	1. ALIP - Iboruna (Amostra de Interação)
	2. Projeto Atlas Linguístico do Brasil (Projeto ALiB)
	3. Discurso & Gramática
	4. BrasilData
	5. Corpus Forense do Português Brasileiro (CFPB)
	6. Atlas Prévio dos Falares Baianos (APFB)
	7. Mozilla Common Voice

Fonte: elaborado pelos autores.

Uma vez compilada a lista de acervos identificados e encontrados por meio de pesquisa independente, iniciamos a avaliação sobre a aplicabilidade de cada um deles em tarefas de fonética forense. A avaliação dos acervos foi realizada com base em um questionário, destinado a identificar as informações linguísticas que podem ser extraídas de cada acervo, e, especial aquelas que podem ser úteis como apoio a tarefas em linguística e fonética forenses, em especial a CL. O questionário encontra-se no

[5] A categoria "multirregional" diz respeito aos acervos que abrangem mais de uma região brasileira, como o Atlas Prévio dos Falares Baianos (APFB), por exemplo, que abrange localidades adjacentes ao território da Bahia, incluindo cidades dos estados de Sergipe, do norte de Minas Gerais, do leste de Goiás e do atual Tocantins, abrangendo, assim, as regiões Norte, Nordeste, Sudeste e Centro-Oeste do Brasil.

documento "questionario.pdf" disponibilizado no repositório https://osf.io/avu9e/files/osfstorage/. As perguntas do questionário estão organizadas em seis seções principais, conforme listado abaixo.

Características gerais do acervo, como nome do projeto, tipos de materiais coletados e disponibilizados, período de coleta das amostras, número de amostras e condições para o acesso e utilização do material coletado.

Características sociodemográficas dos participantes.

Região geográfica abrangida pela coleta do acervo.

Características da coleta, como as condições em que as amostras foram coletadas e se existe um registro de diário de campo, por exemplo.

Informações a respeito de amostras de áudio (consideradas apenas em acervos que concedem acesso a materiais de áudio), como a qualidade perceptiva dos áudios, o formato dos arquivos, o tipo de situação comunicativa/interacional (por exemplo, se é um diálogo, monólogo, conteúdo dirigido, entre outros), como a duração das amostras foi registrada e as características vocais dos participantes (como uso de aparelhos ortodônticos, condições de saúde, histórico cirúrgico etc.).

Detalhes sobre as transcrições das amostras coletadas, se houver: informações sobre o material transcrito disponível, como o tipo de transcrição fornecido (fonética, ortográfica, ortográfica com marcas de conversação, entre outros) e o sistema de transcrição empregado.

4 Resultados e discussão

Os resultados da análise dos acervos serão apresentados aqui seguindo as seções do questionário usado para a avaliação dos acervos. É relevante mencionar que a informação apresentada no eixo vertical (eixo y) dos gráficos de barra nesta seção é o número bruto de acervos em cada categoria de resposta e o número no interior de cada barra é a porcentagem que o valor bruto representa em relação ao total de acervos avaliados naquela análise. Apresentaremos aqui, por limitação de espaço, uma seleção

dos resultados da análise dos acervos levantados. A totalidade das análises está disponível no documento "info-acervos.ods" do repositório https://osf.io/avu9e/files/osfstorage/. Ali é possível encontrar as respostas para todos os itens presentes no questionário referentes a todos os acervos analisados.

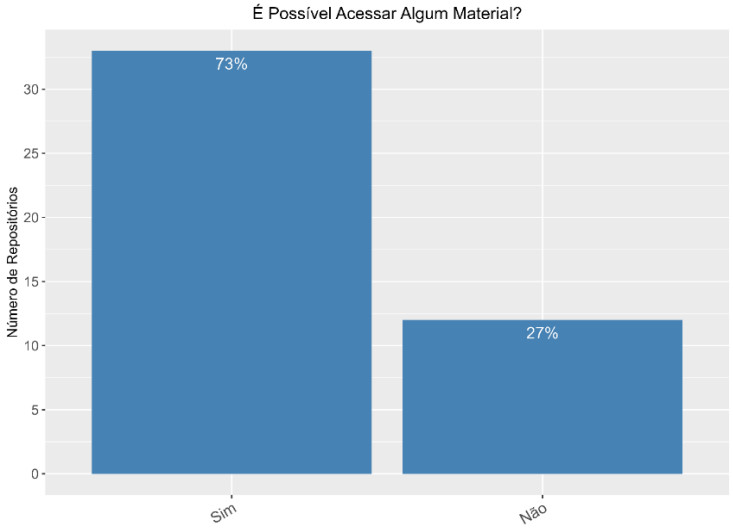
4.1 Características gerais

Esta seção apresenta informações gerais sobre os acervos, enfocando dados essenciais que seriam relevantes para o reaproveitamento desses acervos como subsídio para tarefas de CL. Em seguida, serão apresentados na forma de gráficos os resultados da análise dos acervos tendo em vista os aspectos gerais dos acervos, as características da coleta e disponibilização dos materiais, assim como o período durante o qual foram coletados e acessados.

4.1.1 Coleta e disponibilização dos materiais

Todos os acervos analisados coletaram de materiais de língua ou fala, porém nem todos os disponibilizam ao público. A diferença entre material coletado e material disponibilizado será abordada a seguir, com exemplos. A disponibilidade de acesso aos materiais coletados é mostrada no Gráfico 1.

Gráfico 1 – Possibilidade de acesso a materiais dos acervos.

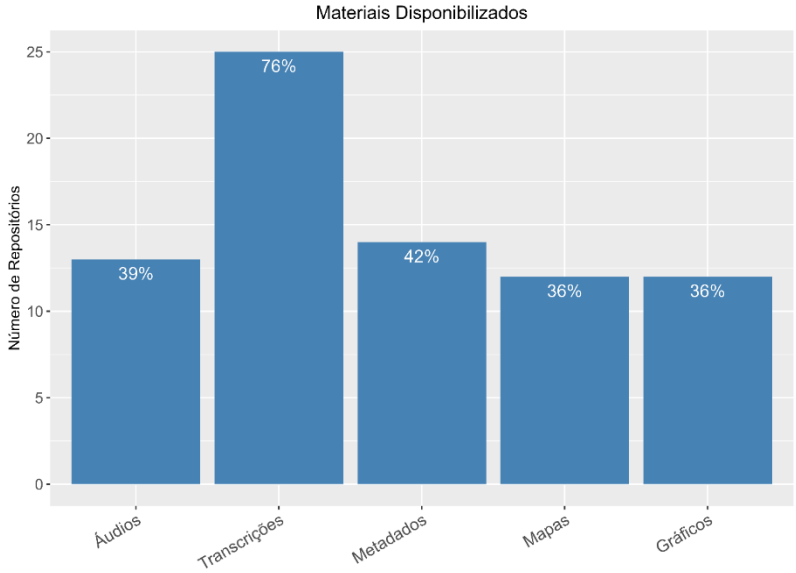


Fonte: elaborado pelos autores.

Dos 45 acervos analisados, 73% (que correspondem a 33 deles) dão acesso a algum material coletado. Nos 12 casos em que o material não está disponível, alguns oferecem apenas acesso a descrições dos acervos em forma de artigo ou outro tipo de publicação acadêmica em que não há informações disponíveis sobre a possibilidade de acesso aos materiais coletados. Em outros casos, os sites referentes aos acervos encontram-se em manutenção e as amostras estão temporariamente indisponíveis.

Após visualizar quantos dos acervos oferecem acesso aos materiais, é necessário identificar quais são esses materiais e como estão distribuídos entre os acervos levantados, assim como está representado no Gráfico 2. Os valores brutos e as porcentagens apresentados no Gráfico 2 têm como base os 33 acervos que disponibilizam materiais.

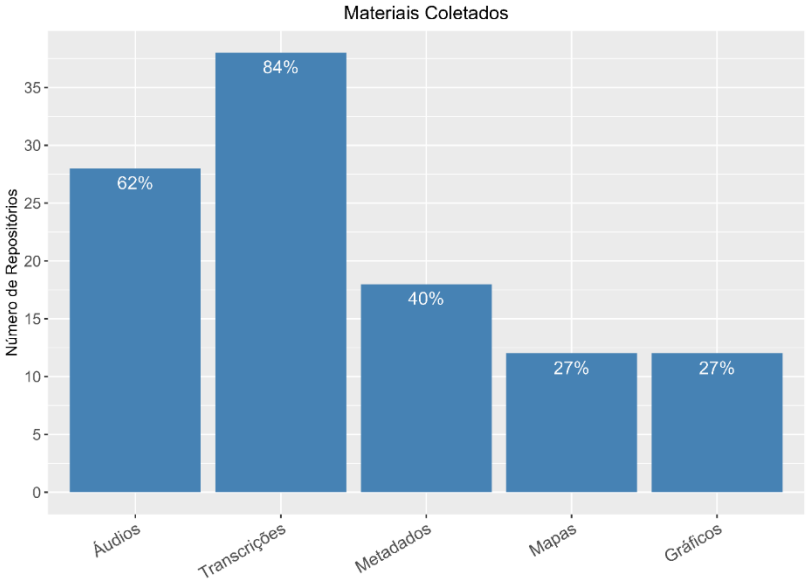
Gráfico 2 – Tipos de materiais disponibilizados.



Fonte: elaborado pelos autores.

Dessa forma, é possível observar no Gráfico 2 quais foram os tipos de materiais disponibilizados pelos acervos, para, no Gráfico 3, poder constatar a discrepância em relação aos números totais de bancos que coletaram cada tipo de material mencionado. Os valores brutos e as porcentagens mostrados no Gráfico 3 são baseados na análise dos 45 acervos avaliados neste trabalho.

Gráfico 3 – Tipos de materiais coletados.



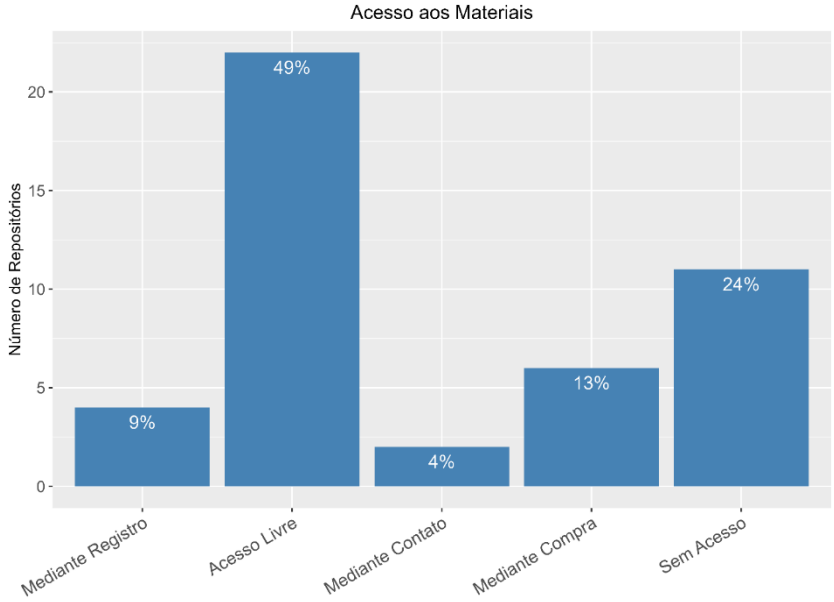
Fonte: elaborado pelos autores.

Observa-se que, exceto pelas amostras de mapas e gráficos, o número de materiais disponibilizados diminuiu em comparação com os materiais coletados. Isso indica que muitos dos acervos identificados coletaram materiais de língua e fala, mas não tinham como objetivo a disponibilização pública dos dados coletados. Em muitos casos, os organizadores dos acervos que não disponibilizam o material coletado não são explícitos em relação às razões para a não disponibilização. Podemos supor que isso possa se dever a questões de ética de pesquisa, que impedem a liberação pública do material coletado. Em outros casos, a questão pode ser que o acervo em questão tenha sido coletado visando servir de base para um estudo particular sem a proposta de posterior liberação dos dados coletados para o público. Finalmente, pode haver casos de acervos que coletaram os materiais de fala apenas como etapa anterior à transcrição do material, sem a intenção de disponibilização dos arquivos de áudio. Para os propósitos da fonética forense, é preferível que o acervo dê acesso tanto aos materiais de áudio quanto às transcrições, pois, em tarefas de CL, as amostras de áudio são usadas para análises acústicas, enquanto as transcrições permitem a análise de outros fenômenos linguísticos com potencial valor indiciário.

4.1.2 Acesso aos materiais e termos de uso

As condições para acesso aos materiais dos acervos são variadas. O Gráfico 4 permite visualizar os diferentes processos para o acesso desses materiais.

Gráfico 4 – Procedimento para acesso aos materiais.



Fonte: elaborado pelos autores.

O Gráfico 4 mostra que a maioria dos acervos de dados levantados disponibilizam suas amostras de forma livre, seja nos próprios *sites* dos acervos, ou por meio de *downloads* de arquivos com as amostras coletadas. Nos casos dos acervos que não dão acesso aos seus materiais, as razões para a indisponibilidade variam, como discutimos a seguir.

Alguns acervos realizaram suas coletas e armazenam os dados exclusivamente em formato físico, como CDs, disquetes e fitas magnéticas e não procederam às suas digitalizações, o que impede o compartilhamento em formato digital. Em outros casos, os acervos coletaram seus materiais com o propósito exclusivo de usar o conteúdo como subsídio para projetos específicos do grupo de pesquisa. Exemplos de acervos compilados sem o objetivo de tornar público o material reunido são o *Projeto Descrição do Português Oral Culto de Fortaleza (PORCUFORT)* e o *Projeto A língua portuguesa falada no semiárido baiano*.

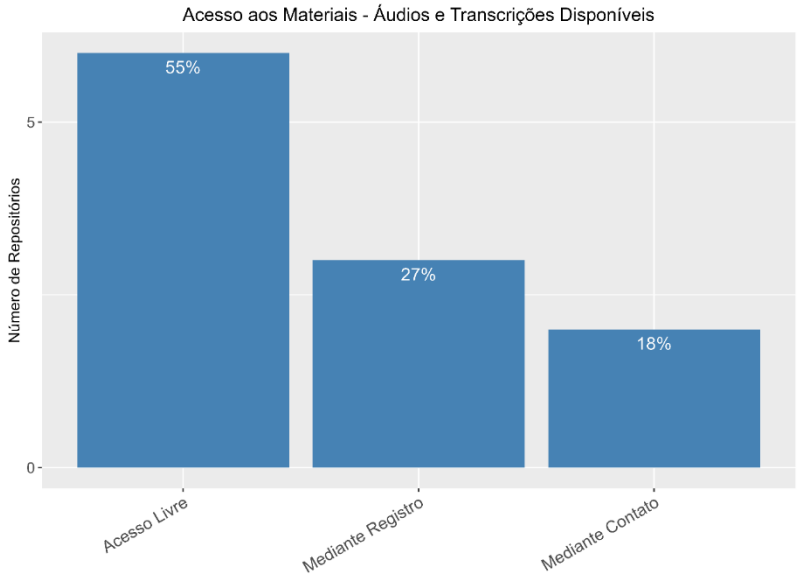
Os acervos que disponibilizam materiais mediante contato com os responsáveis são o *Projeto SP2010* e o *LínguaPOA*. É preciso pagar para ter acesso a materiais dos seguintes acervos: *Projeto Atlas Linguístico do Brasil (Projeto ALiB)*, *Atlas Prévio dos Falares Baianos (APFB)*, *Esboço de um Atlas Linguístico de Minas Gerais (EALMG)*, *Atlas Linguístico de Mato Grosso do Sul (ALMS)*, *Atlas Linguístico do Estado do Ceará (Alece)* e o *Atlas linguístico do Amapá*.

Há também casos de acesso unicamente por meio de cadastro, seja por meio de inscrição por e-mail, como nos dois acervos do *Projeto Amostra Linguística do Interior Paulista (Alip)*, ou com acesso imediato após completar o cadastro, como o *C-ORAL-BRASIL*.

Outro caso particular e relevante a ser destacado é o *Corpus Forense do Português Brasileiro (CFPB)*, elaborado especificamente para fins forenses. Os materiais deste acervo podem ser fornecidos, total ou parcialmente, a pesquisadores de outras instituições, dado que estejam envolvidos em projetos específicos alinhados aos interesses do Instituto Nacional de Criminalística, ligado à Polícia Federal. A descrição do acervo, de sua finalidade e das condições para seu compartilhamento constam da Portaria Nº 934-DITEC/PF, de 30 de julho de 2020.

Dada a relevância de ter acesso às amostras de fala dos acervos para sua utilização na fonética forense, o Gráfico 5 apresenta informações sobre o procedimento de acesso aos dados exclusivamente para os 11 acervos que disponibilizam tanto amostras de fala quanto transcrição.

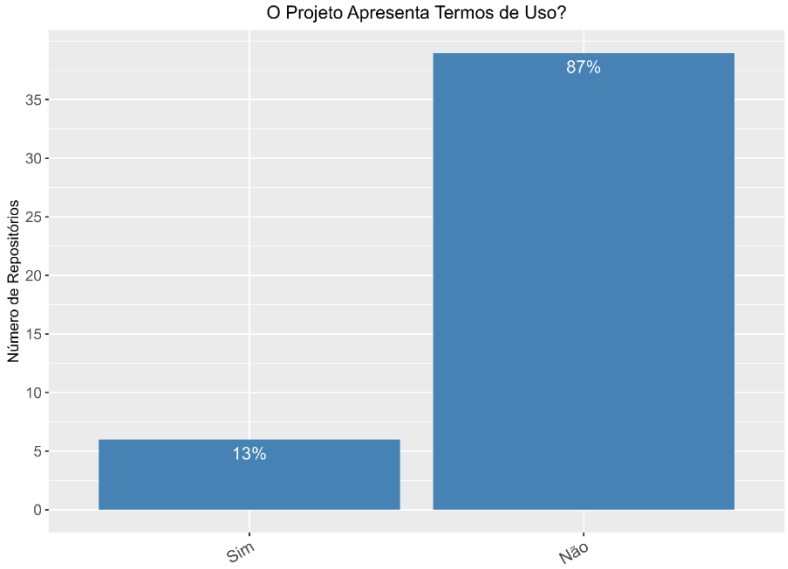
Gráfico 5 – Procedimento para acesso de acervos que disponibilizam áudio e transcrição.



Fonte: elaborado pelos autores.

Por fim, o Gráfico 6 mostra a informação sobre a presença ou ausência de um termo de uso para os acervos, para aqueles que oferecem acesso aos materiais de alguma forma.

Gráfico 6 – Exposição de termo de uso.



Fonte: elaborado pelos autores.

É notável o baixo número de acervos brasileiros de dados e amostras de língua e fala que incluem um termo de uso. Os acervos com termo de uso identificados nesta

pesquisa são de projetos realizados a partir da década de 2000, uma época marcada por uma crescente preocupação com os aspectos éticos relacionados à coleta e divulgação de dados de pesquisa envolvendo seres humanos.

Essa mudança pode ser em parte explicada pela edição, em 2016, de legislação que exige que os pesquisadores forneçam os termos de uso para os dados coletados em contextos de pesquisa em ciências humanas e sociais, conforme a Resolução nº 510 (Brasil, 2016) do Conselho Nacional de Saúde, bem como pela promulgação da Lei Geral de Proteção de Dados Pessoais (LGPD), em 2018, que determina que a coleta de dados nas ciências humanas siga princípios legais essenciais, como propósito específico, transparência, necessidade, segurança, anonimização e consentimento.

Por fim, é relevante citar que todos os acervos mencionados na coluna "Sim" do Gráfico 6 apresentam disponibilidade de acesso a materiais de áudio.

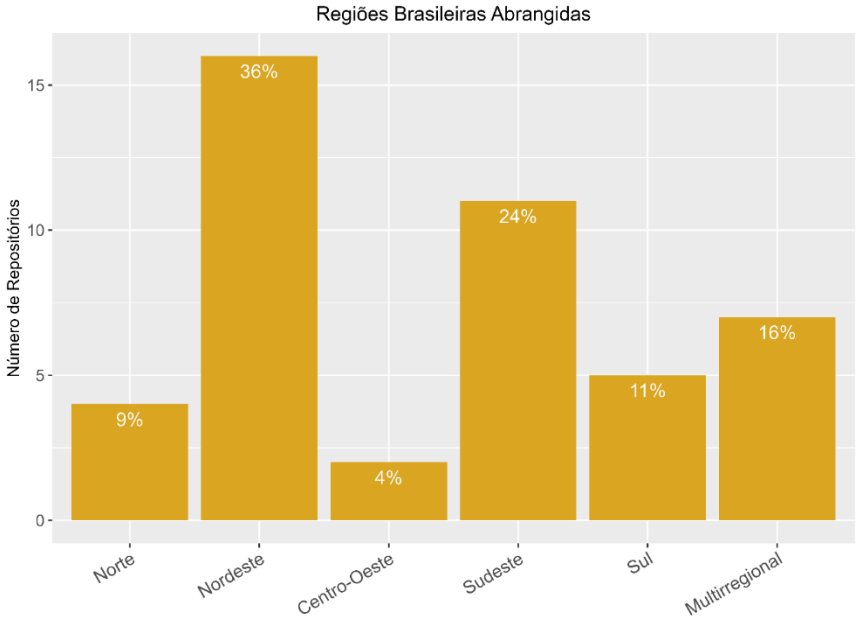
4.2. Abrangência e características sociolinguísticas

Esta seção aborda aspectos relacionados aos participantes cujos dados foram coletados pelos acervos de fala e língua selecionados. Entre esses aspectos estão informações sobre as regiões de coleta, estados abrangidos e outros detalhes das amostras de língua e fala.

4.2.1. Regiões brasileiras abrangidas

Todas as regiões do Brasil (segundo a divisão regional proposta pelo IBGE, que classifica o país nas regiões Norte, Nordeste, Centro-Oeste, Sudeste e Sul) são abrangidas pelos acervos levantados pela pesquisa, embora a concentração de acervos por região não seja uniforme. Existem casos em que a cobertura abrange várias regiões, ou seja, acervos que reuniram dados de locais em mais de uma região do país. O Gráfico 7 representa a abrangência regional dos acervos de dados levantados.

Gráfico 7 – Abrangência regional.



Fonte: elaborado pelos autores.

O Gráfico 7 permite ver que as regiões Nordeste e Sudeste são as que têm a maior cobertura em termos de acervos. No caso do Nordeste, é relevante mencionar que a maioria dos acervos são atlas linguísticos. Nesses casos, o uso forense desse tipo de acervo, especialmente no caso da geração de estatísticas populacionais de traços fonético-acústicos, é mais limitado, visto que os atlas que entraram em nosso levantamento não disponibilizam diretamente materiais de áudio, já que o propósito dos atlas, no caso geral, é a geração de mapas e gráficos apresentando a variação espacial de fenômenos linguísticos, sem o propósito de fornecer materiais de transcrição completa e/ou gravações de áudio. Contudo, mesmo sem o acesso a materiais de áudio, os atlas linguísticos podem ter utilidade em tarefa de CL no mapeamento de evidências sociofonéticas relevantes no processo de determinação do grau de tipicidade, como proposto por Gonçalves e Brescancini (2020). Propostas como essa justificam a inclusão dos atlas neste trabalho.

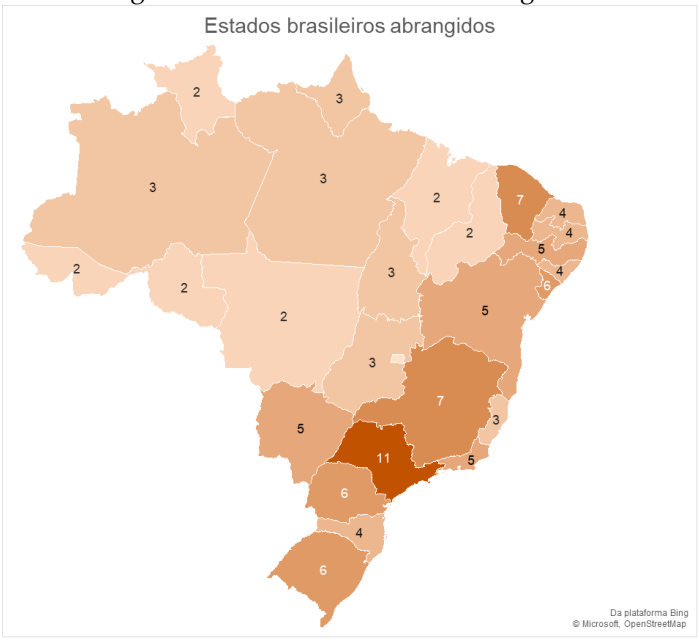
Dos acervos multirregionais mencionados, poucos deles tiveram a intenção de coletar amostras em todo o país: são esses o *Projeto Atlas Linguístico do Brasil (ALiB)*, o *BrasilData* e o *Corpus Forense do Português Brasileiro (CFPB)*. Nos demais casos, são

acervos que coletaram materiais em áreas de fronteira entre regiões, mas sem a pretensão de cobrir todo o território brasileiro, como o *Atlas Prévio dos Falares Baianos (APFB)* e o projeto *ALIP-Iboruna*.

4.2.2. Estados brasileiros abrangidos

O mapa do Brasil apresentado na Figura 1 ilustra a distribuição dos acervos de dados de língua e fala para cada estado do país.

Figura 1 –Estados brasileiros abrangidos.



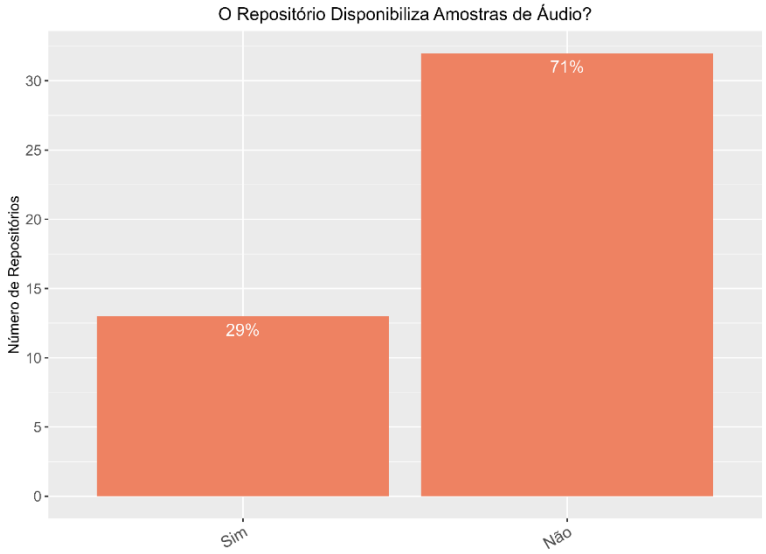
Fonte: elaborado pelos autores.

Na Figura 1, é possível observar a distribuição do número de acervos que coletaram material em cada um dos estados brasileiros. Há seis estados com as maiores concentrações de acervos: São Paulo, Minas Gerais, Ceará, Sergipe, Paraná e Rio Grande do Sul. A pesquisa revelou a ausência de acervos relevantes que representem o Distrito Federal, uma região metropolitana que atualmente conta com cerca de 3 milhões de habitantes ao redor da capital federal.

4.3. Áudio

Na seção *Áudio* do questionário, são avaliadas características gerais das amostras de áudio disponibilizadas pelos acervos. Primeiramente, é importante visualizarmos quantos dos acervos levantados disponibilizam materiais de áudio, como mostra o Gráfico 8.

Gráfico 8 – Disponibilidade das amostras de áudio.



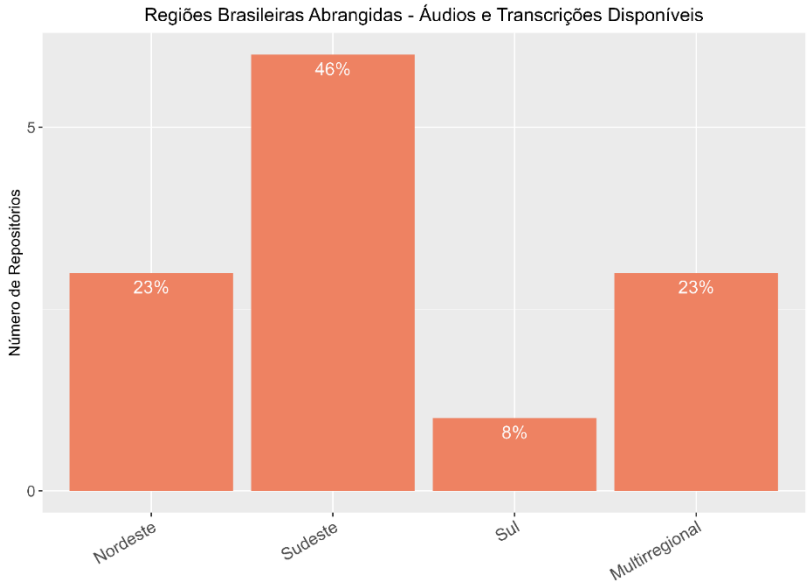
Fonte: elaborado pelos autores.

Agora sabemos que 13 dos acervos, ou seja, 29% do total, fornecem amostras de áudio. Essa classificação é aplicada a amostras de áudio que podem ser acessadas, ou seja, que estão disponíveis no momento da realização deste estudo. Assim, a resposta "Sim" não inclui os acervos que estão em manutenção ou temporariamente sem meios para a disponibilização dos dados.

O fato de apenas 29% dos acervos disponibilizarem amostras de áudio expõe uma limitação no seu uso para a geração de estatísticas populacionais de parâmetros fonético-acústicos, considerando que as análises acústicas e sociofonéticas são muito importantes para a produção de estatísticas sobre características linguísticas de interesse para determinar o grau de tipicidade dos padrões observados nas amostras

analisadas. Sabendo quantos acervos dão acesso a amostras de áudio, é relevante entender como eles estão distribuídos pelas regiões do país, como mostra o Gráfico 9.

Gráfico 9 – Regiões abrangidas por acervos que disponibilizam áudio.



Fonte: elaborado pelos autores.

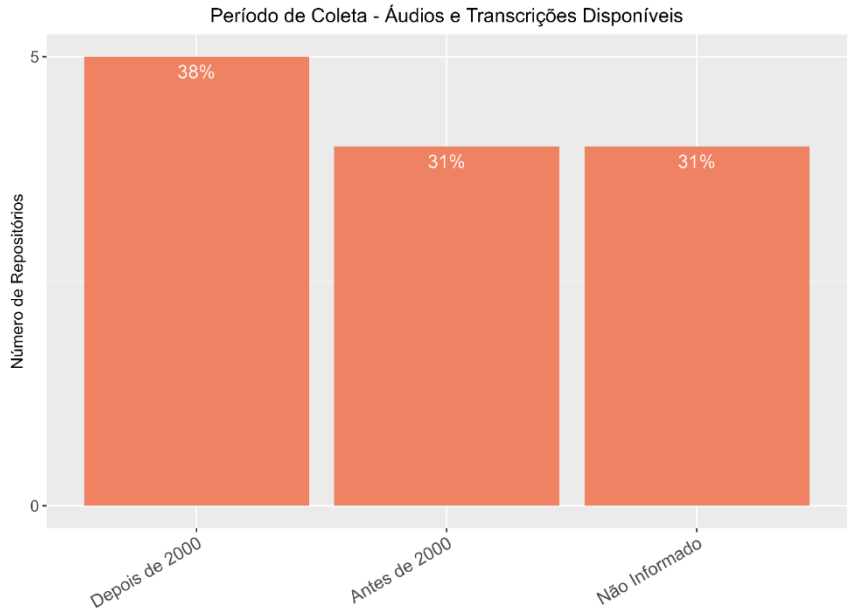
É significativa a comparação dos resultados do Gráfico 9 com o Gráfico 7, uma vez que ambos tratam de informações sobre abrangência regional de acervos que disponibilizam suas amostras. É notável que a maior parte dos acervos que disponibilizam amostras de áudio se concentra na região sudeste, ao passo que, considerando o total de acervos de língua e fala levantados neste trabalho, há uma abundância de acervos na região nordeste, fato explicado pelo grande número de atlas linguísticos regionais produzidos na região.

Os acervos na categoria "multirregional" são o *ALIP - Iboruna (Amostra de Interação)*, que abrange cidades da região noroeste do Estado de São Paulo e uma pequena parte do Estado de Mato Grosso, o *Corpus Forense do Português Brasileiro (CFPB)*, com abrangência nacional, e o *Mozilla Common Voice*, que coleta dados de todo o Brasil. Portanto, não foi possível identificar acervos que ofereçam amostras de áudio provenientes de falantes das regiões Centro-Oeste e Norte, o que limita a cobertura

regional dos dados disponíveis que podem ser aproveitados como subsídio para a compilação de distribuições populacionais de parâmetros linguísticos.

Quanto à época em que os materiais foram coletados, o Gráfico 10 mostra o período de coleta dos acervos, separando-os entre acervos coletados antes de 2000 e depois de 2000.

Gráfico 10 – Período de coleta de acervos que disponibilizam áudio.



Fonte: elaborado pelos autores.

No Gráfico 10 vemos que 5 dos acervos tiveram um período de coleta a partir do ano 2000, ou seja, são acervos mais recentes, o que indica que a qualidade acústica das amostras de fala não foi prejudicada pela degradação do áudio registrado em mídias não digitais, como fitas cassete, por exemplo.

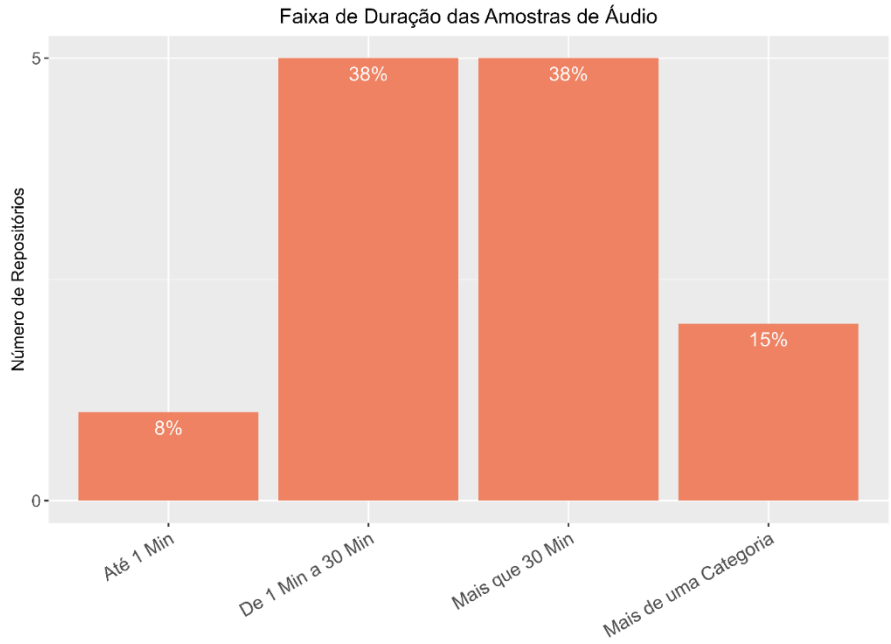
Também é importante mencionar que a preferência, na tarefa de CL, pelos acervos que coletaram suas amostras acústicas a partir dos anos 2000 se dá por fatores relacionados à variação geracional (Eckert, 1997), visto que o uso da língua é influenciado pelo contexto histórico e sociocultural da época em que as amostras foram coletadas.

Os 5 acervos que fizeram coletas depois de 2000 são o *Alip - Iboruna (Amostra Censo)*, o *Alip - Iboruna (Amostra de Interação)*, o *Projeto SP2010*, o *C-ORAL-BRASIL (I)* e o *LínguaPOA*.

4.3.1. Faixa de duração das amostras de áudio

A gama de duração das amostras disponibilizadas pelos acervos levantados pode ser observada no Gráfico 11.

Gráfico 11 – Faixa de duração das amostras.



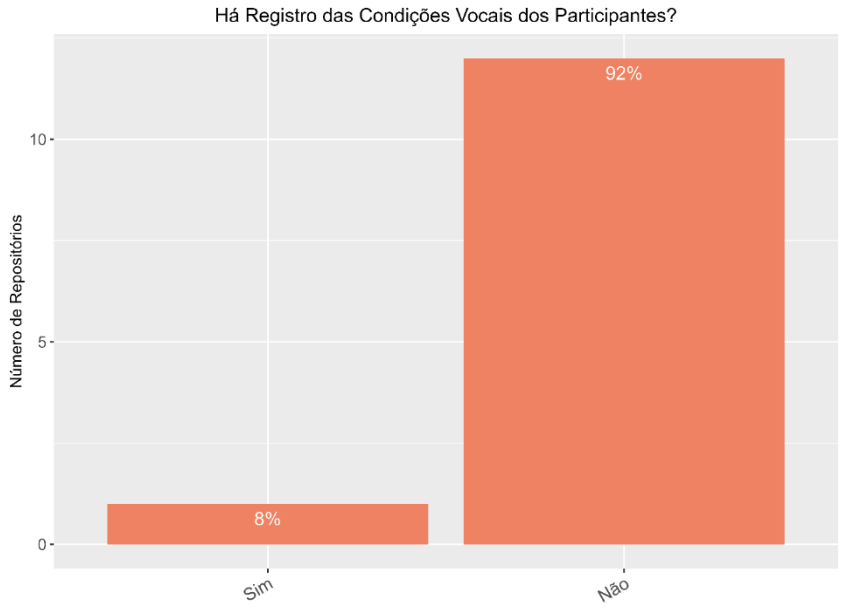
Fonte: elaborado pelos autores.

O Gráfico 11 mostra que a maioria dos acervos fornece amostras de áudio com mais de 1 minuto de duração ou de mais de 30 minutos. Essa informação é crucial para a tarefa de CL, já que amostras muito curtas dificultam a extração de parâmetros acústicos que representem de forma adequada os padrões de fala a longo prazo do participante. Portanto, é preferível utilizar áudios de maior duração.

4.3.2. Registro de condições vocais

Em análise acústica em contexto forense, pode ser relevante ter informação sobre fatores que podem afetar de maneira sistemática diferentes componentes do sistema de produção da fala e alterar a produção vocal dos participantes, tais como histórico de doenças respiratórias, de cirurgia no trato vocal, de tabagismo, bem como o uso de aparelho ortodôntico, entre outros. O Gráfico 12 mostra que, dos 13 acervos que dão acesso material de áudio, apenas um documenta as condições vocais dos participantes, indicando que essa não é uma preocupação tipicamente levada em conta no desenho dos acervos presentes em nossa amostra. O acervo em questão é o *Corpus Forense do Português Brasileiro (CFPB)*, que foi pensado especificamente para dar apoio à área da fonética forense.

Gráfico 12 – Quantidade de acervos que documentam condições que afetem a produção vocal dos participantes.



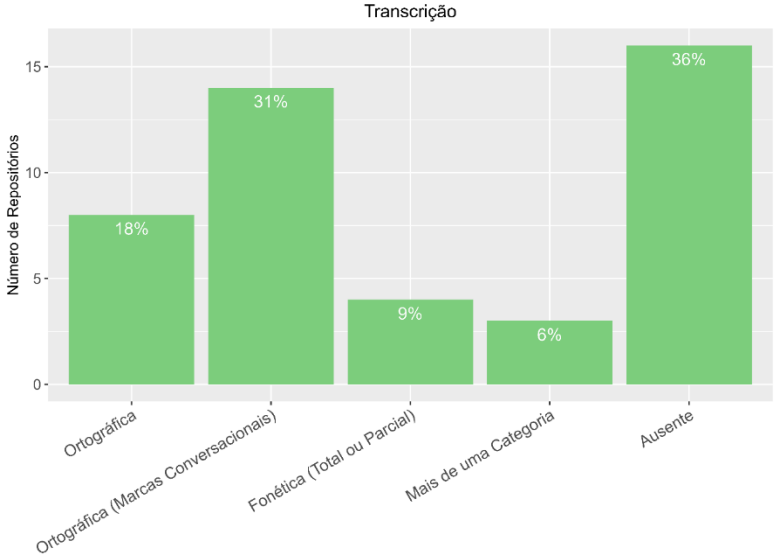
Fonte: elaborado pelos autores.

4.4. Transcrição

Nesta última seção dos resultados, analisamos os tipos de transcrição coletados e/ou disponibilizados pelos 45 acervos levantados. Sendo assim, o Gráfico 13 apresenta

a distribuição dessas informações. Explicamos a seguir a classificação dos tipos de transcrição que adotamos.

Gráfico 13 – Tipo de transcrição.



Fonte: elaborado pelos autores.

Nas transcrições ortográficas o conteúdo linguístico de uma amostra de fala é representado seguindo a norma ortográfica do português brasileiro. Isto é, as palavras presentes nas amostras coletadas são registradas na transcrição da mesma maneira como aparecem nos dicionários, de modo que não se registram possíveis variações sistemáticas ou idiossincráticas que pode haver na sua pronúncia pelo falante da amostra transcrita.

Em transcrições ortográficas com marcas conversacionais, além de registrar o conteúdo linguístico propriamente dito, convenções gráficas especiais são usadas para indicar elementos da organização da fala, como pausas, ênfases, palavras interrompidas ou não pronunciadas por completo, e outros aspectos característicos da conversação.

Em transcrições fonéticas, os sons específicos usados pelo produtor da amostra são registrados por meio de símbolos convencionais, tais como aqueles que formam o Alfabeto Fonético Internacional. As transcrições fonéticas podem ser completas,

quando todas as palavras do texto são transcritas dessa maneira, ou parciais, quando apenas algumas palavras (ou mesmo sons isolados dentro delas) são transcritas foneticamente. Encontramos transcrições parciais principalmente nos atlas linguísticos, nos quais se deseja apresentar variações fonéticas na pronúncia de palavras específicas e para isso usam símbolos fonéticos que registram os sons efetivamente produzidos pelos falantes entrevistados. Por exemplo, temos algumas variações fonéticas possíveis da palavra "sete", que podem ser transcritas foneticamente como [ˈsɛtɪ] ou [ˈsɛt͡ʃɪ̃]. Se o objetivo for destacar um som específico dentro da palavra, a transcrição fonética pode se concentrar apenas no fonema em questão, como em "se[t]e" e "se[t͡ʃ]e". Vale destacar que, apesar da transcrição fonética ser o tipo menos comum entre os acervos levantados, é a modalidade de transcrição mais relevante para a fonética forense. Isso se deve ao fato de que a transcrição fonética possibilita a identificação precisa das variações alofônicas sistemáticas na pronúncia de determinados fonemas da língua. Esse nível de detalhamento permite analisar minuciosamente as características acústicas e articulatórias que podem ser essenciais em contextos forenses.

5 Principais acervos de interesse para a tarefa de Comparação de Locutor

Com base nos resultados apresentados nas seções anteriores, estão compilados no Quadro 3 os acervos que, a nosso juízo, melhor se prestam ao aproveitamento para a fonética forense, sobretudo para a geração de estatísticas populacionais de parâmetros fonéticos e fonético-acústicos no âmbito de exames de CL. Tendo isso em mente, foram priorizados nessa lista os acervos que fizeram suas coletas a partir da década de 2000 e/ou que disponibilizam amostras de áudio com boa qualidade acústica. O critério de relativa contemporaneidade das amostras contidas nos acervos é relevante porque acervos cujos dados tenham mais de três décadas podem não refletir mais o estado sincrônico da língua em função de fatores diversos que podem causar mudança linguística, tanto internos quanto externos ao sistema (Passetti *et al.*,

2024). Boa qualidade acústica das amostras é importante para a extração de parâmetros acústicos, muitos dos quais são sensíveis às condições de registro do sinal acústico, como o cálculo de formantes e a observação de classes de fones específicos, como fricativas (Barbosa *et al*., 2020).

6 Conclusão

O presente trabalho apresenta o resultado de um levantamento de acervos de dados de língua e fala do português brasileiro e avalia sua aproveitabilidade na tarefa de CL em fonética forense, em especial como fonte de material para a geração de estatísticas populacionais de parâmetros linguísticos e fonéticos para uso em exames de comparação de locutor baseados na aplicação do arcabouço bayesiano.

Os resultados deste trabalho sugerem que o Brasil conta com um número relativamente pequeno de acervos de dados que podem ser úteis como fonte geradora de estatísticas populacionais para dar subsídio à tarefa de CL. Dos 45 acervos levantados, apenas sete servem, de maneira mais adequada, ao propósito de geração de estatísticas populacionais de traços linguísticos importantes para a determinação do grau de tipicidade de padrões linguísticos e fonéticos observados em amostras de fala.

No que se refere à variação linguística regional, nossos resultados mostram que seria importante dispor de mais acervos que disponibilizassem material de áudio para as regiões do país atualmente pouco cobertas. Não encontramos tais acervos nas regiões Centro-Oeste e Norte e, quanto às regiões Sul e Nordeste, identificamos, por fim, apenas um acervo que abranja cada uma delas. Isso mostra que é pertinente haver um esforço por parte de projetos em andamento ou já concluídos no sentido de viabilizar a disponibilização de amostras de áudios e outros metadados. Em relação a acervos a serem coletados futuramente, a sugestão é que os responsáveis prevejam desde o momento do planejamento a adoção de termos de uso que incluam a previsão

de disponibilização e utilização das amostras de áudio coletadas para outros fins além daqueles previstos pelo projeto.

Os acervos de língua e fala que incluímos em nossa amostra cumprem o papel para o qual foram pensados, que são análises linguísticas mais típicas, principalmente o registro de fenômenos de variações nos níveis lexical, sintático, conversacional, entre outros fins. No entanto, quando se leva em consideração as necessidades específicas da tarefa de CL no campo da fonética forense, a conclusão é que as opções são relativamente mais limitadas. Pretendemos, em trabalhos futuros, avaliar de forma mais detalhada os parâmetros fonético-acústicos que podem ser analisados em cada um dos acervos que disponibilizam amostras de áudio.

Espera-se que os resultados do presente trabalho sejam de utilidade tanto para especialistas nas áreas de linguística e fonética forenses quanto de outras áreas da linguagem, e ajudem a promover, a médio e longo prazo, a criação de novos acervos de dados voltados para estabelecer estatísticas de distribuição populacional de traços linguísticos do português brasileiro.

Agradecimentos

O primeiro autor agradece o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (bolsa PIBIC, processo n.º 142760/2023-4). A segunda autora agradece o apoio da Capes (bolsa de Pós-Doutorado, Processo Capes 88887.804443/2023-00). Os autores em conjunto agradecem também à professora Lívia Oushiro (Unicamp) pelos valiosos comentários e sugestões feitos a uma versão preliminar do trabalho e aos pareceristas anônimos da revista por sua leitura e sugestões.

Quadro 3 – Principais acervos de interesse para a tarefa de Comparação de Locutor em fonética forense.

Acervos	ALIP - Iboruna (Amostra Censo)	ALIP - Iboruna (Amostra de Interação)	Projeto SP2010	C-ORAL-BRASIL (I)	PORTAL - Variação linguística no português alagoano	LínguaPOA	Corpus Forense do Português Brasileiro (CFPB)
Período de coleta das amostras	2004 a 2007	2005 a 2006	2011 a 2013	2006 a 2011	Não informado	2015 a 2019	Não informado
Procedimento para acesso aos materiais	Mediante cadastro	Mediante cadastro	Contato com responsáveis	Mediante cadastro	Acesso livre	Contato com responsáveis	Por meio de convênio por projetos alinhados aos interesses da Polícia Federal
Há termo de uso?	Sim	Sim	Não	Sim	Não	Não	Sim
O termo de uso permite aproveitamento para contextos forenses?	Sim	Sim	Não informado	Sim	Não informado	Não informado	Sim
Regiões do Brasil abrangidas	Sudeste	Multirregional	Sudeste	Sudeste	Nordeste	Sul	Multirregional
Há informações sobre a coleta?	Sim	Sim	Sim	Sim	Sim	Não	Não
Qual o escopo das informações?	Por amostras	Por amostras	Por amostras	Por amostras	Informações gerais	Não informado	Não informado
Formato dos arquivos de áudio	.mp3	.mp3	.wav e .mp3	.wav	.wav	.mp3	.wav

Estilo de elocução	Conteúdo guiado	Conteúdo livre	Conteúdo guiado	Conteúdo guiado	Conteúdo guiado e conteúdo livre	Conteúdo guiado	Leitura de palavras/frases e conteúdo guiado
Faixa de duração das amostras	De 1 a 30 minutos	De 1 a 30 minutos	Mais do que 30 minutos	De 1 a 30 minutos	De 1 a 30 minutos	Mais do que 30 minutos	De 1 a 30 minutos
Há registros de condições vocais?	Não	Não	Não	Não	Não	Não	Sim
Há transcrições das amostras?	Sim	Sim	Sim	Sim	Sim	Sim	Não
Tipo de transcrição	Ortográfica com marcas conversacionais	Ortográfica com marcas conversacionais	Ortográfica com marcas conversacionais	Ortográfica com marcas conversacionais	Ortográfica	Ortográfica	Ausente

Fonte: elaborado pelos autores.

Referências

BRASIL. **Conselho Nacional de Saúde**. Resolução nº 510, de 7 de abril de 2016. Diário Oficial da União: seção 1, Brasília, DF, 24 maio 2016. Disponível em: http://conselho.saude.gov.br/resolucoes/2016/Reso510.pdf.

BRASIL. Lei n.º 13.709, de 14 de agosto de 2018. **Lei Geral de Proteção de Dados Pessoais**. Diário Oficial da União: Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm.

BRASIL. Portaria n° 934-Ditec/PF, de 30 de julho de 2020. **Institui o Corpus Forense do Português Brasileiro (CFPB) no âmbito do Sistema Nacional de Criminalística e estabelece regras para seu funcionamento, manutenção e compartilhamento**. Brasília, 2020.

BALDWIN, J.; FRENCH, P. **Forensic phonetics**. Londres: Pinter, 1990.

BARBOSA, P. A. *et al*. **Análise Fonético-Forense:** em tarefa de Comparação de Locutor. Campinas: Millenium Editora, 2020.

BRESCANCINI, C. R; GONÇALVES, C. S. O peso da evidência sociofonética na perícia de Comparação de Locutor. *In*: BARBOSA, P. A. *et al*. (ed.). **Análise fonético-forense em tarefa de comparação de locutor.** 1. ed. Campinas: Millenium Editora, 2020, p. 67-87.

CUNHA, M. S. da. **Estatísticas populacionais da frequência fundamental do português brasileiro para uso em fonética forense**. 2023. Dissertação de Mestrado - Universidade Federal de São Carlos, São Carlos, 2023.

ECKERT, P. Age as a sociolinguistic variable. *In*: COULMAS, F. (ed.). **Handbook of Sociolinguistics.** Oxford: Blackwell, 1997, p. 151-67. DOI https://doi.org/10.1002/9781405166256.ch9

FREITAG, R. M. Ko. Sociolinguística no/do Brasil. **Cadernos de Estudos Linguísticos**, Campinas, SP, v. 58, n. 3, p. 445–460, 2016. DOI https://doi.org/10.20396/cel.v58i3.8647170

GOLD, E.; FRENCH, P. International practices in forensic speaker comparison. **The International Journal of Speech, Language and the Law**, v. 18, n. 2, p. 293–307, 2011. DOI https://doi.org/10.1558/ijsll.v18i2.293

GOLD, E.; FRENCH, P. International practices in forensic speaker comparisons: second survey. **International Journal of Speech Language and the Law**, v. 26, n. 1, p. 1– 20, 2019. DOI https://doi.org/10.1558/ijsll.38028

JESSEN, M. Forensic Phonetics. **Language and Linguistics Compass**, v. 2, n. 4, p. 671–711, 2008. DOI https://doi.org/10.1111/j.1749-818X.2008.00066.x

ALTENHOFEN, C. V.; KLASSMANN, M. S. **Atlas lingüístico-etnográfico da região Sul do Brasil**: cartas semântico-lexicais. Porto Alegre: Editora da UFRGS, 2011.

MORRISON, G. S. Forensic voice comparison and the paradigm shift. **Science and Justice**, v. 49, n. 4, p. 298–308, 2009. DOI https://doi.org/10.1016/j.scijus.2009.09.002

MORRISON, G. S. Forensic voice comparison. *In*: FRECKELTON, I.; SELBY, H. (org.). **Expert Evidence**. Sydney: Thomson Reuters, 2010.

OLSSON, J. **Forensic Linguistics**: An Introduction to Language, Crime and the Law. 2. ed. Londres: Continuum, 2008.

PASSETTI, R. R. *et al*. Tipicidade e qualidade de voz: considerações metodológicas sobre o controle de critérios sociolinguísticos, fonéticos e de voz. **Cadernos de Estudos Linguísticos**, [s. l.], v. 66, p. e024020, 2024. DOI https://doi.org/10.20396/cel.v66i00.8675468