UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS

DANIEL FONSECA VIEIRA

POSSIBILIDADES DE USO EM LINGUÍSTICA FORENSE DE BANCOS DE DADOS BRASILEIROS DE LÍNGUA E FALA

SÃO CARLOS
2024

DANIEL FONSECA VIEIRA

POSSIBILIDADES DE USO EM LINGUÍSTICA FORENSE DE BANCOS DE DADOS BRASILEIROS DE LÍNGUA E FALA

Trabalho de conclusão de curso apresentado como requisito parcial para a obtenção do título de Bacharel em Linguística pela Universidade Federal de São Carlos.

Orientador: Prof. Dr. Pablo Arantes
Coorientadora: Profa. Dra. Renata Regina Passetti

SÃO CARLOS

2024

DANIEL FONSECA VIEIRA

POSSIBILIDADES DE USO EM LINGUÍSTICA FORENSE DE BANCOS DE DADOS BRASILEIROS DE LÍNGUA E FALA

Trabalho de conclusão de curso apresentado como requisito parcial para a obtenção do título de Bacharel em Linguística pela Universidade Federal de São Carlos.

Aprovado em: _____/_____/_____.

_____

Prof. Dr. Pablo Arantes

Universidade Federal de São Carlos

_____

Profa. Dra. Renata Regina Passetti

Universidade Federal de São Carlos

_____

Profa. Dra. Livia Oushiro

Universidade Estadual de Campinas

RESUMO

Este projeto tem como objetivo realizar um levantamento a respeito dos bancos de dados de língua e fala disponíveis a respeito do português brasileiro e avaliar seu potencial de uso para tarefas em linguística forense. Na etapa de levantamento dos bancos de dados existentes foram identificados 45 bancos de dados de fala de várias regiões do Brasil a partir de consulta a especialistas na área de sociolinguística e dialetologia e de pesquisa bibliográfica. A análise das possibilidades de uso desses bancos para os propósitos específicos da linguística e fonética forenses foi feita através da aplicação de um questionário que aborda as características gerais de cada projeto, os tipos de materiais disponibilizados, as características dos participantes, a região geográfica abrangida, as condições de acesso e uso do material coletado, entre outros aspectos relevantes. Como resultado, este trabalho apresenta de forma sistematizada o levantamento dos bancos de dados de língua e fala no português brasileiro, permitindo a identificação daqueles adequados para a geração de estatísticas de distribuição populacional de características linguísticas e fonéticas relevantes para a tarefa de Comparação de Locutor. As informações geradas serão úteis tanto para especialistas em linguística e fonética forense quanto de outras áreas da linguagem.

Palavras-chave: fonética, fonética forense, criminalística, banco de dados.

ABSTRACT

This project aims to survey language and speech databases in Brazil and assess their potential for use in forensic linguistics tasks. During the survey stage of the research, we identified 45 databases from various regions of Brazil. The survey comprised consultation with experts in sociolinguistics and dialectology and bibliographic research. The assessment of potential usefulness for forensic purposes consisted in the application of a questionnaire addressing general project characteristics, types of materials provided, participant characteristics, geographical coverage, access conditions, and other relevant aspects. As a result, this work systematically presents the survey of language and speech databases in Brazilian Portuguese, enabling the identification of those suitable for generating population distribution statistics of linguistic and phonetic features relevant to the Speaker Comparison task. The generated information will be valuable for experts in forensic linguistics, phonetics, as well as other language-related fields.

Keywords: phonetics, forensic phonetics, criminalistics, database.

LISTA DE ILUSTRAÇÕES

TABELAS

FIGURAS

GRÁFICOS

SUMÁRIO

1 INTRODUÇÃO E JUSTIFICATIVA

A linguística forense é uma área de estudo científico interdisciplinar que combina os estudos linguísticos a investigações policiais a fim de dar suporte a evidências que sejam relevantes para a resolução de casos judiciais. Dessa forma, suas tarefas podem auxiliar em decisões judiciais a partir de análises pautadas em métodos e protocolos específicos. Os variados contextos criminais determinam que tipo de análise linguística pode ser mais útil para cada caso.

As tarefas mais comuns em linguística forense envolvem, em geral, a identificação de padrões linguísticos ou a elaboração de perfis de fala para identificar características discriminatórias de indivíduos. Há casos em que são necessárias análises de amostras escritas para extrair padrões estilísticos de escrita, estratégias retóricas, padrões de formas sintáticas, características comunicativas socioculturais, entre outros. Outros casos demandam análises de áudios, ou seja, análises fonético-forenses, que abrangem uma série de métodos e diferentes tarefas que serão aprofundadas no presente projeto.

A fonética forense é, dentro da linguística forense, a área que se ocupa da análise de materiais de fala em contextos criminais. Dessa forma, pode ser definida como "a aplicação de conhecimentos, teorias e métodos da fonética geral em tarefas práticas que emergem da atuação policial ou da apresentação de evidências em tribunais, bem como o desenvolvimento de novos conhecimentos, teorias e métodos especificamente fonético-forenses" (Jessen, 2008)[1].

Entre as tarefas mais comuns no contexto da fonética forense, destacam-se: a identificação de falantes por vítimas e testemunhas (em situações em que não há amostras de áudio como vestígios dos crimes), a elaboração de perfis de fala, que consiste em traçar um perfil linguístico de um falante para casos em que há apenas a amostra de fala de um indivíduo desconhecido (amostra linguística questionada) mas ainda não há amostra de fala de suspeitos (amostra linguística padrão), bem como a comparação de locutor, em casos em que é possível fazer análises linguísticas e acústicas de amostras de fala e determinar a probabilidade delas terem a mesma origem ou origens diferentes. O presente projeto buscou contribuir para a tarefa de comparação de locutor, uma vez que se trata da tarefa desenvolvida com maior frequência na área de Fonética Forense (cf. Passetti, 2022).

A tarefa de Comparação de Locutor (doravante CL) busca determinar a probabilidade de duas amostras de fala terem sido, ou não, produzidas por um mesmo indivíduo. Nesse tipo

[1] Traduzido por Passetti (2022)

de tarefa, são comparados os parâmetros de fala de um locutor desconhecido (o locutor que produziu o que é chamado no campo de amostra *questionada*), com os do suspeito, acusado, indiciado ou réu (locutor cuja identidade é conhecida e produz o que são chamadas de amostras *padrão*). Assim, são determinadas a amostra questionada (AQ) e a amostra padrão (AP). O desafio na tarefa de CL está em definir o grau de convergência e de divergência entre as duas amostras, ou seja, a *similaridade*, bem como definir o quão típicos são os parâmetros de fala analisados em relação a uma população específica, ou seja, a *tipicidade* (Brescancini; Gonçalves, 2020).

Em relação a metodologias que estabelecem o valor de probabilidade das evidências para corroborar ou refutar as hipóteses de mesma origem ou origens distintas das amostras de fala, Gold e French (2011; 2019) realizaram dois levantamentos em um intervalo de oito anos, em que foi possível compreender quais os métodos de expressão de conclusões mais utilizados e quais estavam em maior ascensão de uso na tarefa de CL. Constatou-se que, nesse intervalo de tempo, uma das metodologias com maior adesão entre os especialistas consiste no uso do que é conhecido como arcabouço bayesiano (Morrison, 2009). No contexto do arcabouço bayesiano, o peso atribuído à evidência avaliada no exame pericial é determinado através do cálculo da razão de verossimilhança, comumente denominada LR (abreviação da expressão inglesa *Likelihood Ratio*), que avalia a chance de as falas de interesse terem a mesma origem ou terem origens diferentes. A hipótese de mesma origem é baseada no grau de *similaridade* entre as amostras e a de origens diferentes leva em conta o grau de *tipicidade* dos padrões observados nas amostras analisadas tendo em vista a distribuição desses padrões em uma população de referência.

Nesse método, além da comparação de similaridades entre a AQ e a AP, levam-se em consideração também dados de estatística de distribuição populacional de características da fala de uma população específica, o que permite que os especialistas comparem as características linguísticas encontradas nas amostras questionadas com as características típicas de uma determinada população, para então calcular a razão entre duas hipóteses: a primeira sendo de ambas as amostras terem uma mesma origem, e a segunda sendo de ambas as amostras terem origens diferentes. Portanto, para estabelecer o grau de tipicidade em tarefa de CL usando LRs é preciso recorrer a estatísticas populacionais, que podem ser geradas por bancos de dados que forneçam tais dados de distribuição populacional da fala em determinada língua ou dialeto regional.

No contexto brasileiro, existem muitos bancos de dados de fala do português brasileiro que foram elaborados e coletados como subsídio para estudos sociolinguísticos, dialetológicos,

de linguística de *corpus*, entre outros campos dentro da linguística. Entre alguns dos mais conhecidos estão o Programa de Estudos sobre o Uso da Língua (PEUL), pioneiro na construção de bancos de dados para fins sociolinguísticos (FREITAG, 2016), o Projeto da Norma Urbana Oral Culta (NURC), o qual começou as coletas na década de 1970, e o Projeto Variação Linguística na Região Sul do Brasil (VARSUL). Muito embora os projetos que deram origem a esses bancos de dados não tenham necessariamente sido pensados como recursos de apoio a práticas forenses, é relevante a avaliação de sua utilidade potencial para a fonética forense, especialmente no que toca a geração de estatísticas de distribuição populacional de traços fonéticos e linguísticos do PB para a tarefa de CL.

O presente trabalho se justifica tendo em vista que a maioria dos bancos de dados de língua e fala brasileiros não foram compilados com finalidade forense e sua aproveitabilidade para esse uso precisa ser avaliada de forma sistemática. Partindo dessa justificativa, propusemos fazer um recenseamento dos diversos bancos de dados de língua e fala[2] produzidos no Brasil e avaliar a possibilidade de seus aproveitamentos em tarefas de linguística forense. Para isso, o trabalho foi pensado para ser desenvolvido em etapas que contemplassem, primeiramente, o levantamento de uma lista de bancos de dados de fala do português brasileiro, para posteriormente, a partir de uma série de critérios, fossem feitas as avaliações de cada um dos bancos selecionados a fim de ponderar a viabilidade de aproveitamento de seus dados e, com isso, considerá-los como dados de referência na determinação do indicador de tipicidade referente a uma população relevante.

2 OBJETIVOS

O projeto teve dois objetivos principais que se articularam:

- realizar um levantamento dos bancos de dados de língua e fala do português brasileiro a partir da consulta a especialistas na área de sociolinguística e dialetologia e de pesquisa bibliográfica;
- avaliar os bancos levantados quanto ao seu potencial de uso para tarefas de linguística forense a partir de critérios relevantes para esse campo de saber.

Os critérios para a realização do levantamento dos bancos e para sua avaliação serão descritos na seção Metodologia.

[2] Usamos a expressão “bancos de dados de língua e fala” para abranger os bancos que coletam tanto informações no nível fonético quanto em outros níveis de análise linguística, como morfologia, sintaxe ou variação lexical, por exemplo.

3 METODOLOGIA

A primeira etapa do projeto consistiu em fazer um levantamento dos bancos de dados com potencial de uso na presente pesquisa. Para isso, entramos em contato com especialistas nas áreas de sociolinguística e dialetologia, de todas as regiões do Brasil, para colaborarem conosco indicando ou recomendando bancos de dados de língua e fala que pudessem ser úteis para os objetivos deste trabalho. Além do levantamento feito através de indicações e recomendações de especialistas, parte dos bancos também foram encontrados em pesquisa individual por meio do uso de motores de busca, como o Google e o Google Acadêmico.

Foram excluídos do levantamento inicial bancos que não poderiam ser avaliados pelo questionário elaborado para o projeto. Os critérios para a exclusão variam de impossibilidade de acesso a amostras ou informações básicas dos bancos à constatação de que o banco não cumpre o papel de ser um banco de dados de língua e fala do português brasileiro. Os bancos excluídos e os critérios de exclusão estão representados na Tabela 1.

Tabela 1 - Bancos de dados excluídos e critérios de exclusão

Bancos de dados excluídos	Critério de exclusão
Amostras de áudio do NURC-SP contidas no 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	Trata-se de um banco de dados bilíngue, o que foge do propósito deste trabalho
Corpus de Textos Orais do Português Santareno (CTOPS)	Impossibilidade de acesso a informações básicas
Aspectos linguísticos da fala londrinense: esboço de um atlas linguístico de Londrina.	Impossibilidade de acesso a informações básicas

Após o levantamento dos bancos de dados de interesse, obtivemos uma lista com 45 diferentes bancos de várias regiões do Brasil. A Tabela 2 apresenta as regiões abrangidas pelos bancos.

Tabela 2 - Lista de bancos levantados separados por região geográfica brasileira

Abrangência regional	Bancos 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
	1. ALIP - Iboruna (Amostra de Interação)

Multirregional[3]	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

Uma vez compilada a lista de bancos indicados e descobertos por meio de pesquisa independente, iniciou-se a etapa de análise acerca da possibilidade do uso de cada um deles para tarefas em linguística forense. A análise dos bancos foi feita a partir de um questionário, cuja função foi identificar as características linguísticas que podem ser extraídas de cada banco.

O questionário contempla:

- Características gerais do banco, como nome do projeto, tipos de materiais coletados e disponibilizados[4], período de coleta das amostras, quantidade das amostras e condições para acesso e uso do material coletado.
- Características dos participantes, como idade, sexo, proveniência regional, escolaridade, entre outras.
- Região geográfica abrangida pela coleta do banco.
- Características da coleta, isto é, circunstâncias em que as amostras foram obtidas e se há o registro de diário de campo, por exemplo.

[3] A categoria "multirregional" diz respeito aos bancos que abrangem mais de uma região brasileira, como o Atlas Prévio dos Falares Baianos (APFB), que abrange localidades adjacentes à 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.

[4] É importante esclarecer a diferença entre os materiais **coletados** e os **disponibilizados**. Todos os bancos avaliados realizaram coleta de materiais de língua ou fala, mas nem todos os bancos *disponibilizam* o material coletado. Por exemplo, alguns atlas linguísticos, como o "Atlas Semântico-Lexical da Região do Grande ABC", coletaram amostras de áudio, mas não as disponibilizaram devido à metodologia de utilizar os áudios apenas como etapa anterior à transcrição, que nesse caso é disponibilizada. Para os casos em que não há disponibilização de material, os motivos variam: em alguns casos tivemos acesso apenas a descrições dos bancos em forma de artigo ou outros tipos de publicação acadêmica, mas não foi possível encontrar informações sobre a possibilidade de acesso aos materiais coletados. Em outros casos, como no projeto "Variação Linguística no Estado da Paraíba (VALPB)" e no projeto "Variação Linguística na Região Sul do Brasil (VARSUL)", os dados encontram-se em manutenção ou estão temporariamente sem possibilidade de disponibilização.

- Informações a respeito de amostras de áudio, preenchidas apenas em bancos que concedem materiais de áudio, diz respeito a elementos dos materiais de oralidade e reúne informações gerais das amostras, como nível de qualidade perceptiva dos áudios, formato dos arquivos, situação comunicativa/interacional (se se trata de um diálogo, monólogo, conteúdo guiado, entre outros), modo de registro da duração das amostras e de características vocais dos participantes (presença de aparelhos ortodônticos, patologias, históricos de cirurgia, etc.).
- Informações sobre transcrições das amostras coletadas, caso estejam presentes: informações do material transcrito disponibilizado, como o tipo de transcrição apresentada (fonética, ortográfica, ortográfica com marcas conversacionais etc.) e o sistema de transcrição utilizado.

4 RESULTADOS

Os resultados deste trabalho serão apresentados a seguir conforme a divisão de seções feitas no questionário criado. É relevante mencionar que o número apresentado acima de cada barra nos gráficos é o número bruto de observações em cada categoria de resposta e o número abaixo de cada rótulo descritivo é a porcentagem que o número bruto representa em relação ao número total de bancos avaliados (45 bancos).

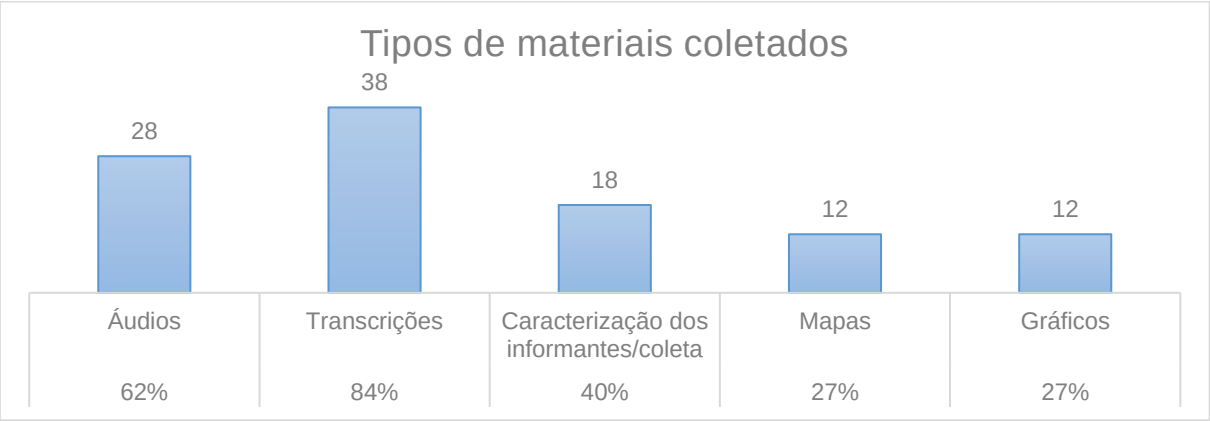
4.1 Características gerais

Esta primeira seção diz respeito a informações gerais dos bancos acerca do que seria relevante como dados básicos para os propósitos de reaproveitamento para análise em tarefa de CL. A seguir, serão apresentados gráficos acerca dos aspectos gerais dos bancos, como de coleta e disponibilização dos materiais, período da coleta e de acesso aos materiais.

4.1.1 Coleta e disponibilização dos materiais

Todos os bancos avaliados realizaram coleta de materiais de língua ou fala, mas nem todos os disponibilizaram. Dessa forma, é possível observar, no Gráfico 1, quais foram os tipos de materiais coletados, para, nos Gráficos 2 a 4, poder constatar a discrepância em relação ao número de bancos que disponibilizam, de alguma forma, os materiais de seus projetos.

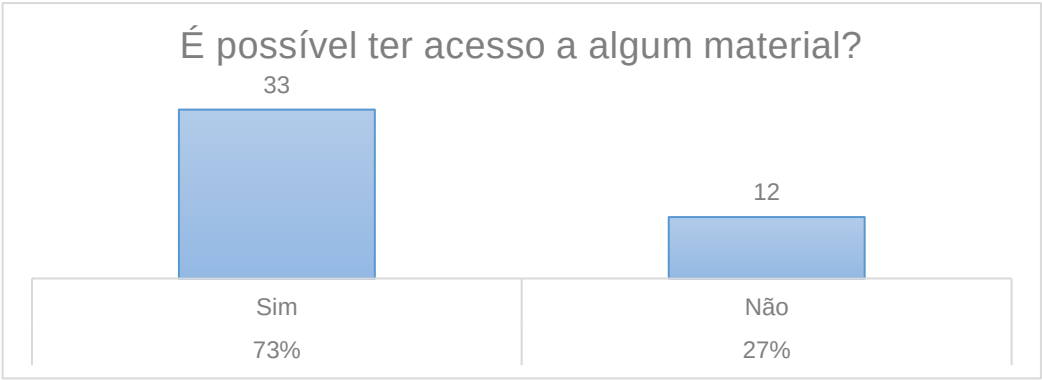
Gráfico 1 - Tipos de materiais coletados



Fonte: autoria própria.

Além dos dados expostos no Gráfico 1, é importante saber como se dá a disponibilização desses materiais, uma vez que se percebe uma predominância da coleta de materiais de áudio, transcrição e de caracterização dos informantes ou a coleta, os quais são relevantes para a tarefa de CL. A informação da possibilidade de acesso aos materiais coletados está ilustrada no Gráfico 2.

Gráfico 2 - Possibilidade de acesso a materiais dos bancos

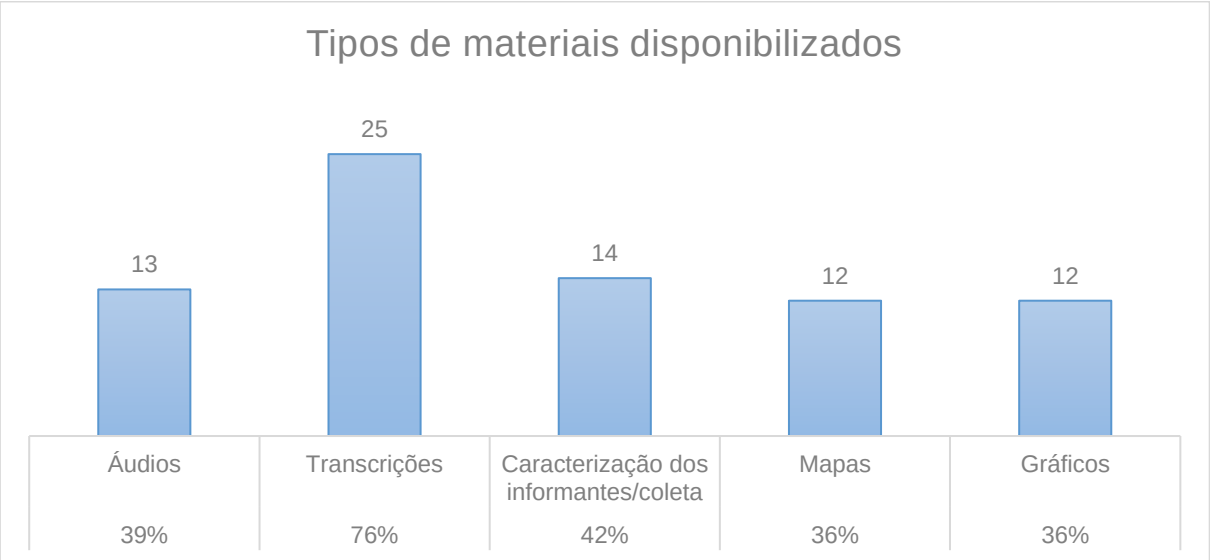


Fonte: autoria própria.

Dos 45 bancos analisados, 33 disponibilizam acesso a algum material. Para os 12 casos em que não há disponibilização de material, alguns possibilitam acesso apenas a descrições dos bancos em forma de artigo ou outros tipos de publicação acadêmica e não é possível encontrar informações sobre a possibilidade de acesso aos materiais coletados. Em outros desses casos, os dados encontram-se em manutenção ou estão temporariamente sem possibilidade de disponibilização.

Após visualizar quantos dos bancos possibilitam acesso a materiais, cabe saber quais são esses materiais e como eles se distribuem dentre os bancos levantados, assim como está representado no Gráfico 3. Os valores e as porcentagens apresentados no Gráfico 3 têm base nos 33 bancos que disponibilizam materiais.

Gráfico 3 - Tipos de materiais disponibilizados



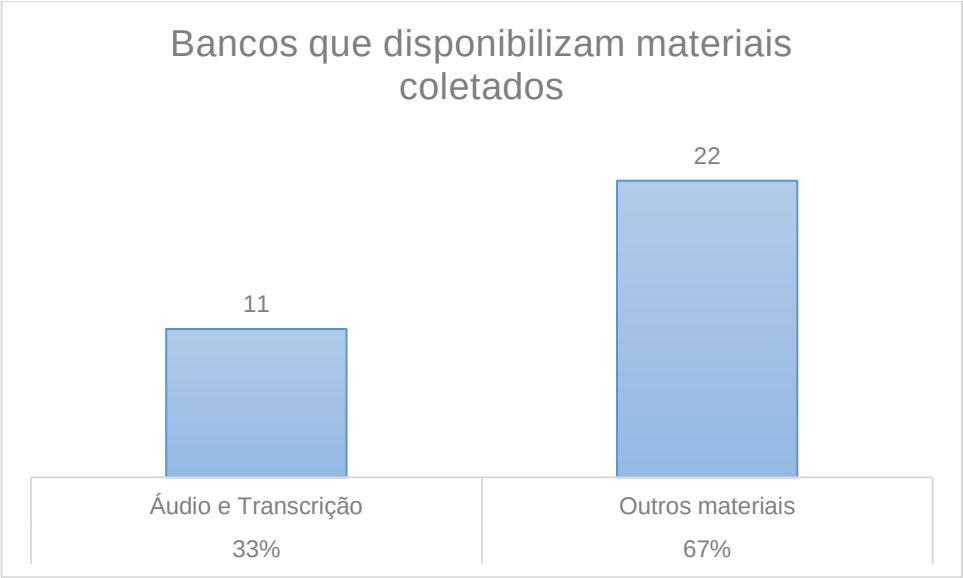
Fonte: autoria própria.

Nota-se que, com exceção das amostras de mapas e gráficos, houve uma redução no número de materiais disponibilizados em relação aos materiais coletados, o que significa que muitos dos bancos levantados fizeram a coleta dos materiais de língua e fala com propostas distintas à de disponibilização pública dos dados. Em alguns casos, pode ser que a proposta do banco em questão seja apresentar um estudo a partir de dados internos e não divulgados, bem como em casos de bancos coletarem os materiais de fala apenas como etapa anterior à transcrição, sem a intenção de disponibilização dos arquivos de áudio, ou por questões éticas. Para os propósitos da fonética forense, é preferível que o banco disponibilize os materiais de áudio e também as transcrições, uma vez que, em tarefa de CL, utilizam-se amostras de áudio para análises acústicas e as transcrições possibilitam análises de outros fenômenos linguísticos idiossincráticos.

O Gráfico 4 representa os bancos que disponibilizam materiais de áudio e transcrição em comparação aos bancos que não disponibilizam tais amostras simultaneamente (ou seja, bancos que disponibilizam apenas áudio ou apenas transcrições) ou bancos que disponibilizam

outros materiais. Os valores e porcentagens apresentados no Gráfico 4 têm base nos 33 bancos que disponibilizam materiais.

Gráfico 4 - Bancos que disponibilizam materiais coletados



Fonte: autoria própria.

Dessa forma, pode-se constatar que um terço dos bancos que permitem acesso a algum tipo de material disponibilizam materiais de áudio e transcrição, que são os de maior interesse para os propósitos linguístico-forenses.

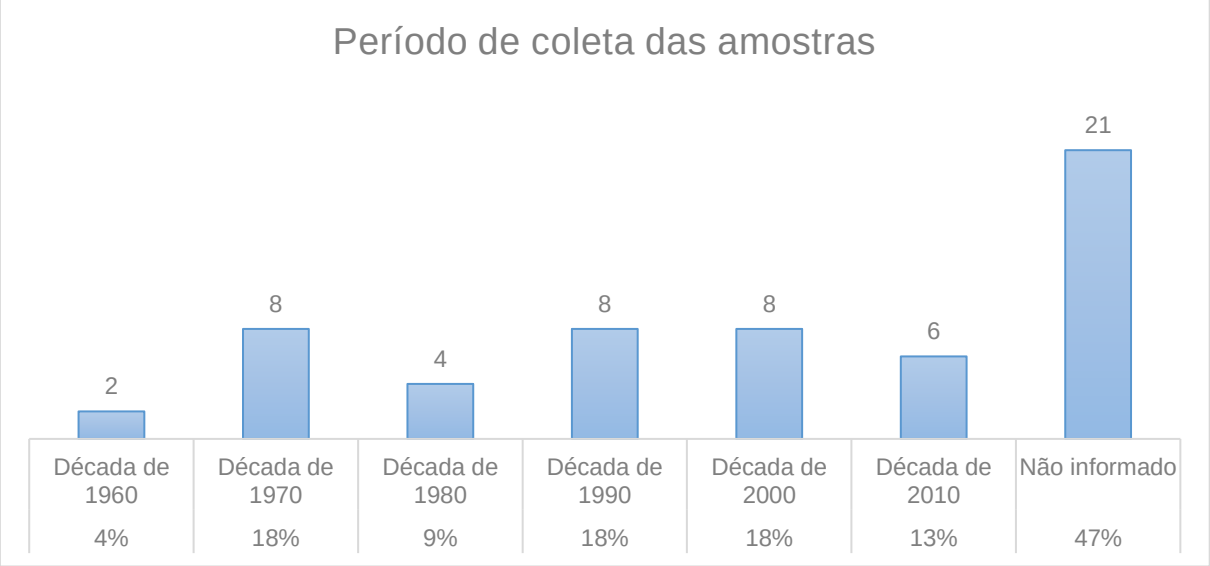
A partir desses primeiros gráficos é possível dizer que, de todos os bancos levantados, apenas 33% disponibilizam os materiais de maior interesse para tarefas em linguística forense. Entretanto, os outros materiais disponibilizados, uma vez que foram coletados para propósitos de análise específicos, podem ser úteis de outras formas dentro da área da linguística forense, como para descrever a variação espacial de fenômenos de natureza morfológica, lexical ou sintática, o que pode ser útil em situações nas quais por alguma razão não é possível fazer análises acústicas nas amostras de áudio, mas a extração de informações linguísticas a partir da oitiva é possível.

4.1.2 Período de coleta das amostras

É importante para a linguística e fonética forense o acesso a informações a respeito do período de coleta das amostras avaliadas. Bancos que disponibilizam amostras de falas mais recentes são preferíveis à medida que a qualidade acústica de um material de áudio pode ser

comprometida pela deterioração do áudio pelo tempo, em casos de mídias físicas, como fitas cassete. Além disso, fenômenos linguísticos socioculturais de épocas distintas podem impor vieses nas conclusões de análises de CL. Sendo assim, o Gráfico 5 representa, em décadas, a divisão dos períodos de coleta de material dos bancos. Há casos em que um mesmo banco de dados coletou dados em mais do que uma década, podendo fazer parte de mais de uma das colunas do gráfico.

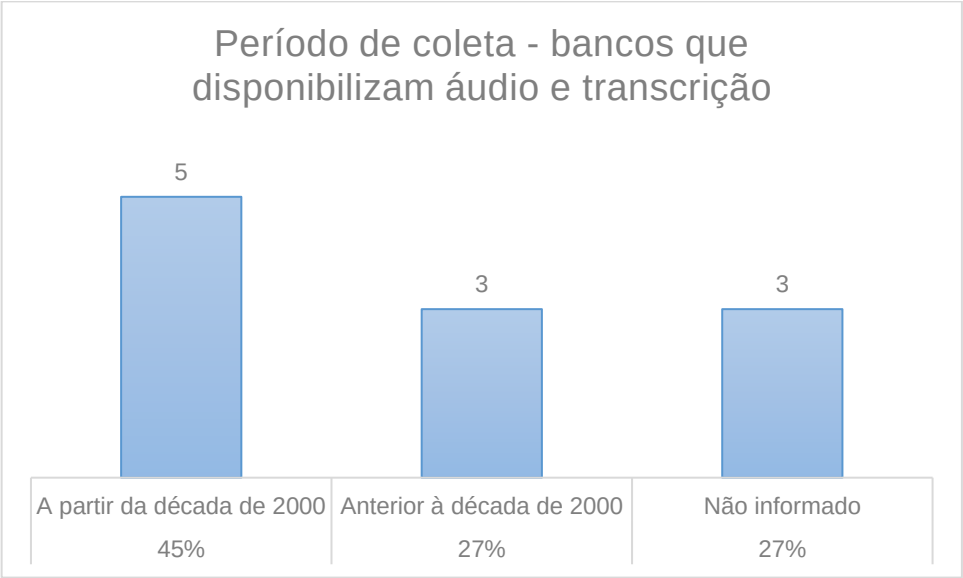
Gráfico 5 - Período de coleta das amostras



Fonte: autoria própria.

Pelos dados apresentados pelo gráfico, é notável que a maioria dos bancos não traz informação facilmente acessível a respeito do período de coleta das amostras, e, por conta disso, podem ter seu uso comprometido em tarefas forenses. Dos bancos que disponibilizam essa informação, vemos que os períodos de coleta se espalham desde a década de 1960 até a década de 2010. Considera-se mais apropriado ao uso forense amostras contemporâneas ao período em que as amostras periciadas foram coletadas e grande parte dos exames envolvem gravações coletadas no presente. O Gráfico 6 mostra a distribuição temporal dos períodos de coleta apenas para os bancos que disponibilizam áudios e transcrições a partir da década de 2000, que seriam os mais apropriados para aproveitamento em tarefas de CL.

Gráfico 6 - Período de coleta de bancos que disponibilizam áudio e transcrição



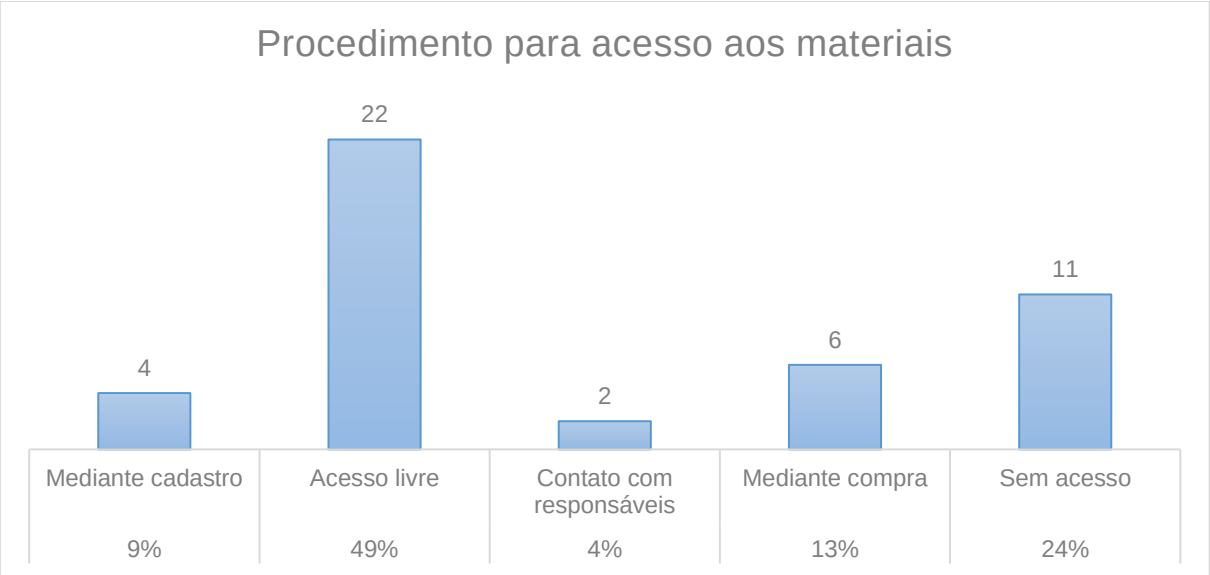
Fonte: autoria própria.

O gráfico mostra que 5 dos 11 bancos que disponibilizam áudios coletaram dados nas duas décadas desde o ano 2000. Do ponto de vista da importância da contemporaneidade dos materiais coletados, pode-se dizer que são poucos os bancos de dados de língua e fala que disponibilizam materiais mais recentes.

4.1.3 Acesso aos materiais e termos de uso

Os procedimentos para acesso aos materiais podem variar bastante de banco para banco. O Gráfico 7 permite visualizar os diferentes processos para o acesso desses materiais.

Gráfico 7 - Procedimento para acesso aos materiais[5]



Fonte: autoria própria.

O Gráfico 7 mostra que a maioria dos bancos de dados levantados disponibilizam suas amostras de forma livre, seja com amostras disponíveis nos próprios sites dos bancos, ou por *downloads* de arquivos contendo as amostras coletadas. Nos casos dos bancos sem possibilidade de acesso, o motivo para a indisponibilidade dos materiais varia, como exemplificamos a seguir.

Alguns bancos fizeram suas coletas e as armazenaram apenas em forma física, como em CDs, disquetes, fitas, entre outros, impossibilitando o compartilhamento por meios digitais. Outro caso se dá por conta de alguns bancos fazerem a coleta de seus materiais apenas com o objetivo de realizar trabalhos pontuais por parte do grupo que fez a coleta, como é o caso do *Projeto Descrição do Português Oral Culto de Fortaleza (PORCUFORT)* e do *Projeto A língua portuguesa falada no semiárido baiano*. Nesses casos, não houve a intenção de disponibilizar publicamente o material coletado.

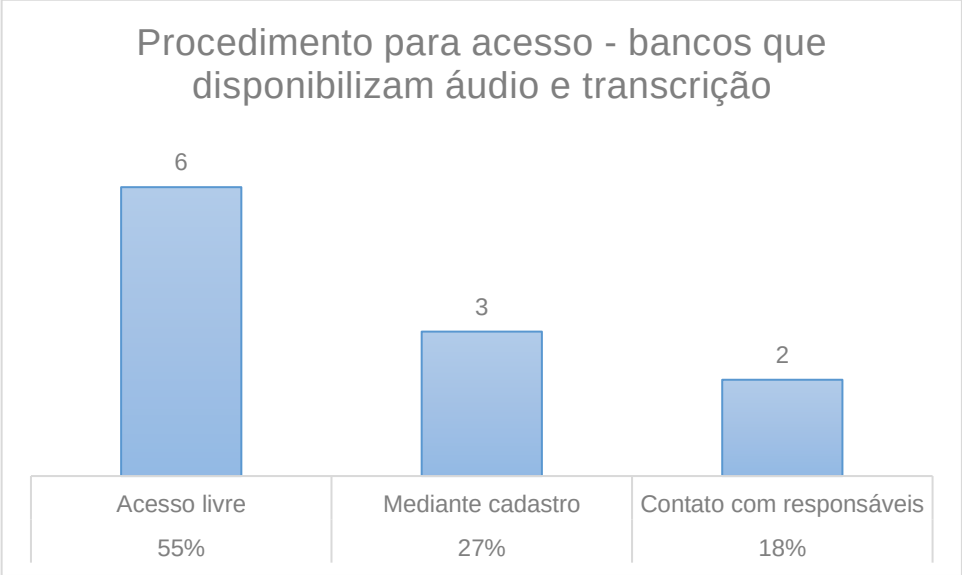
Há também o caso de acesso por meio de cadastro, seja por meio de inscrição por e-mail, como nos dois bancos do *Projeto Amostra Linguística do Interior Paulista (ALIP)*, ou com o acesso instantâneo após realizar o cadastro, como no caso do *C-ORAL-BRASIL*. O acesso mediante compra é o caso, geralmente, de atlas linguísticos publicados em livros e vendidos de forma física e/ou digital.

[5] Os bancos que disponibilizam materiais mediante contato com os responsáveis são: *Projeto SP2010* e o *LínguaPOA*; mediante compra são: *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*.

Outro caso específico e importante de ser mencionado diz respeito ao acesso a um dos bancos avaliados, *o Corpus Forense do Português Brasileiro (CFPB)*. Os materiais deste banco podem ser disponibilizados, total ou parcialmente, para pesquisadores de outras instituições, uma vez que estiverem envolvidos em projetos específicos alinhados aos interesses da criminalística da Polícia Federal, segundo a Portaria Nº 934-DITEC/PF, de 30 de julho de 2020, Art. 6º.

Devido à importância de ter à disposição as amostras de fala dos bancos para o aproveitamento destes no contexto da fonética forense, o Gráfico 8 mostra a informação a respeito do procedimento de acesso aos dados apenas para os bancos que disponibilizam amostras de fala e de transcrição.

Gráfico 8 - Procedimento para acesso de bancos que disponibilizam áudio e transcrição

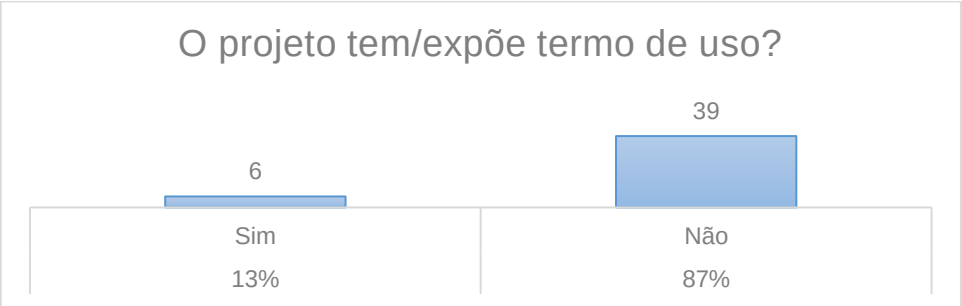


Fonte: autoria própria.

Dos bancos que disponibilizam áudio e transcrição, há casos em que existe a necessidade de um cadastro ou com contato com responsáveis dos projetos, enquanto na maioria dos casos o acesso aos materiais é livre.

Finalmente, no Gráfico 9 mostramos a informação sobre a existência ou não de termo de uso para os bancos, no caso daqueles que dão acesso aos materiais de alguma maneira.

Gráfico 9 - Exposição de termo de uso



Fonte: autoria própria.

É notável o baixo número de bancos de dados e língua e fala brasileiros que expõem termo de uso. Todos os bancos que apresentam termo de uso em nosso levantamento são trabalhos realizados a partir da década de 2000, período a partir do qual há maior preocupação com os aspectos éticos relativos à coleta e divulgação de dados de pesquisa que envolvem seres humanos. Em especial, a partir de 2016, segundo normas aplicáveis a ciências humanas e sociais (Resolução nº 510/2016), existe legislação que exige dos pesquisadores termos de uso para dados coletados em contexto de pesquisa, bem como a existência da Lei Geral de Proteção de Dados Pessoais (LGPD), que estabelece que a coleta de dados no contexto das ciências humanas deve ser realizada com base em princípios fundamentais legais, incluindo finalidade específica, transparência, necessidade, segurança, anonimização e consentimento.

Por fim, é relevante citar que todos os bancos mencionados na coluna “Sim” do Gráfico 9 apresentam disponibilidade de materiais de áudio.

4.2 Abrangência e características sociolinguísticas

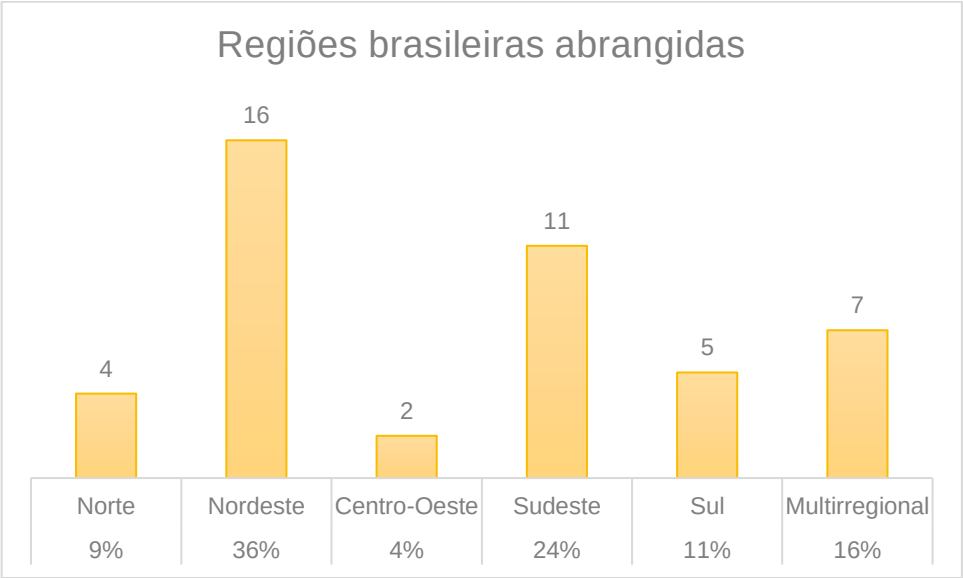
Esta seção trata de características relacionadas aos participantes dos bancos de fala e língua selecionados. Dizem respeito a informações de regiões de coleta, estados abrangidos, bem como de outras informações das amostras de língua e fala.

4.2.1 Regiões brasileiras abrangidas

Todas as regiões do Brasil (considerando aqui a divisão regional proposta pelo IBGE, que divide o Brasil nas regiões Norte, Nordeste, Centro-Oeste, Sudeste e Sul) estão cobertas pelos bancos levantados pela pesquisa, embora a concentração de bancos por região não seja homogênea. Há casos em que há abrangência multirregional, isto é, para bancos que coletaram

dados em localidades situadas em mais de uma região do país. O Gráfico 10 representa a abrangência regional dos bancos de dados levantados.

Gráfico 10 - Regiões brasileiras abrangidas



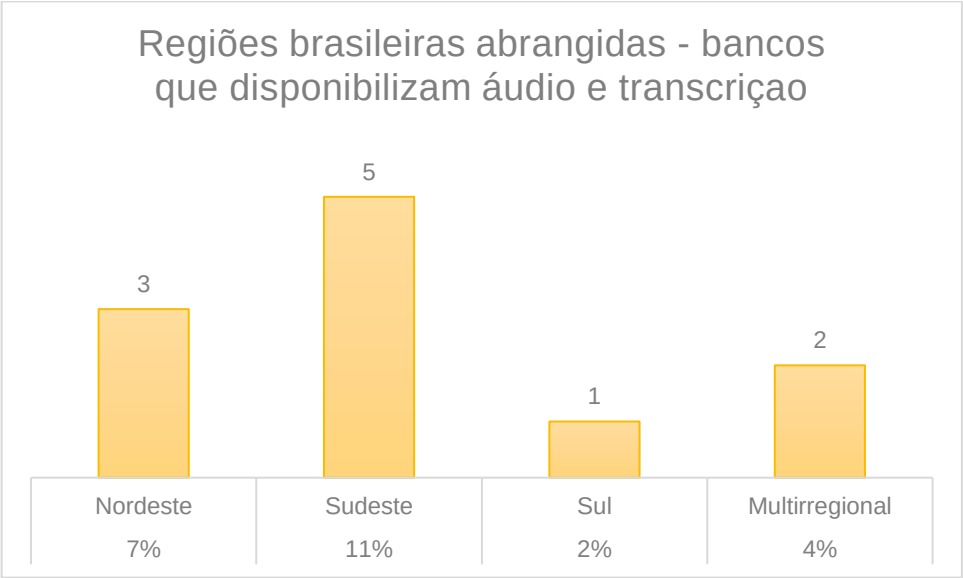
Fonte: autoria própria.

Percebe-se uma abundância de bancos de dados que cobrem as regiões Nordeste e Sudeste. No caso do Nordeste, é relevante mencionar que o número alto de bancos é resultado do fato da maioria dos bancos serem atlas linguísticos. Nesses casos, o uso forense é limitado, visto que os atlas dificilmente disponibilizam materiais de áudio, já que os propósitos, em geral, tratam da exposição de mapas e gráficos apresentando variação de fenômenos linguísticos, sem a intenção de disponibilizar materiais de transcrição bruta e/ou de áudios.

Dos bancos multirregionais mencionados, poucos deles tiveram a intenção de coleta de amostras a nível nacional: são esses o *Projeto Atlas Linguístico do Brasil (ALiB)*, o *BrasilData* e o *Corpus Forense do Português Brasileiro (CFPB)*. Nos demais casos, são bancos que coletaram materiais em regiões de mais de um Estado brasileiro, mas dentro de uma mesma região geográfica.

Para os propósitos forenses, é interessante relacionar as informações de abrangência regional aos bancos que disponibilizam materiais de áudio e transcrição, que pode ser visualizado no Gráfico 11.

Gráfico 11 - Regiões brasileiras abrangidas por bancos que disponibilizam áudio e transcrição



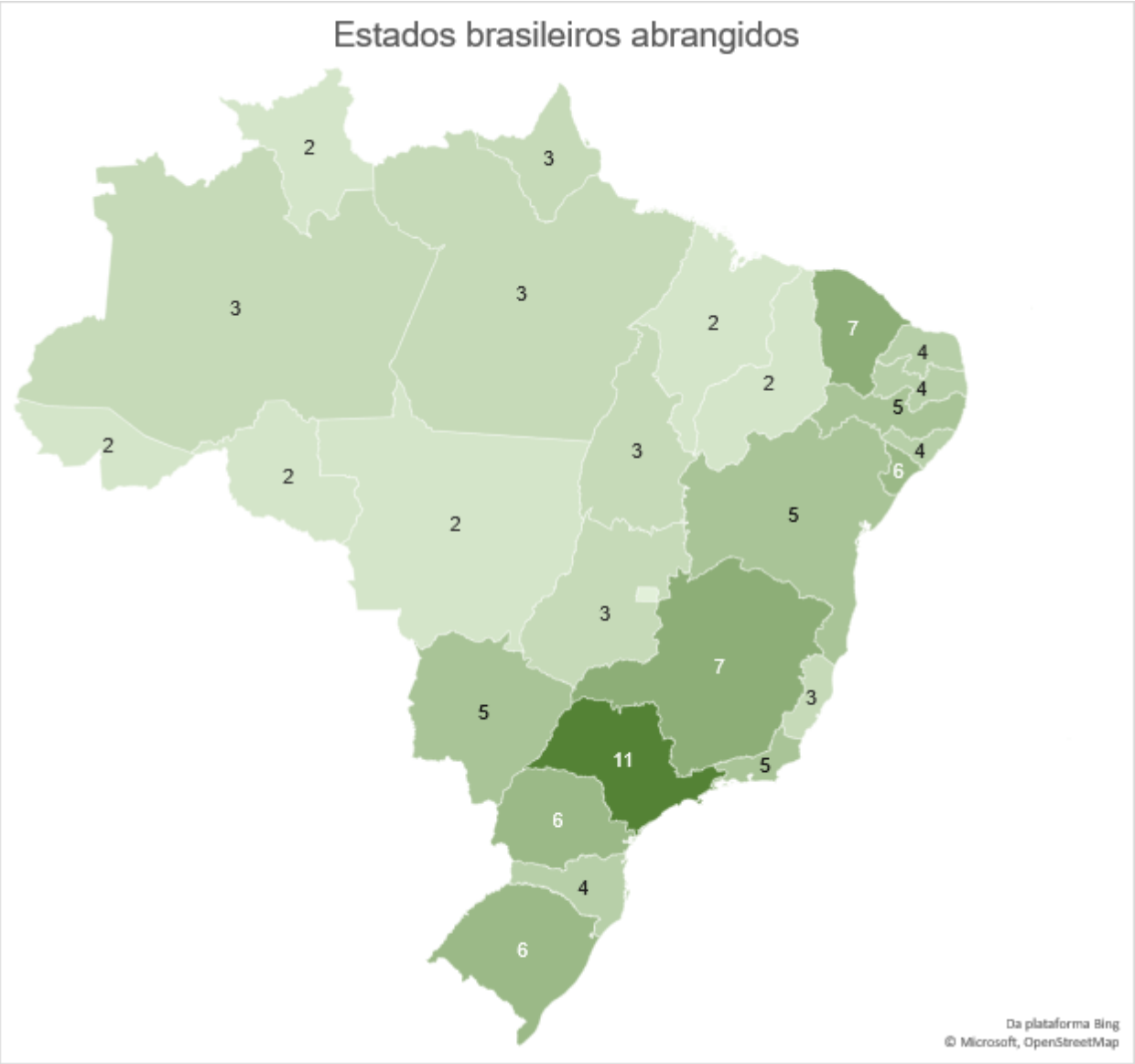
Fonte: autoria própria.

Com base no Gráfico 11 pode-se dizer que, do ponto de vista do aproveitamento dos dados para as finalidades da linguística e fonética forense, a distribuição geográfica dos bancos que disponibilizam amostras de áudio reflete a mesma concentração regional mostrada no gráfico 10: predominam bancos das regiões Nordeste e Sudeste, além dos dois casos em que houve coletas multirregionais: o *Mozilla Common Voice*, de abrangência nacional, e o *ALIP – Iboruna (Amostra de interação)*, que abrange o estado de São Paulo e um pouco do estado de Mato Grosso do Sul.

4.2.2 Estados brasileiros abrangidos

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

Figura 1 - Estados brasileiros abrangidos



Fonte: autoria própria.

Na Figura 1, é possível observar a distribuição do número de bancos que fizeram coleta de material em cada um dos estados brasileiros. Os 6 estados com as maiores concentrações de bancos são, respectivamente: São Paulo, Minas Gerais, Ceará, Sergipe, Paraná e Rio Grande do Sul. Destaca-se a inexistência de bancos na amostra levantada na pesquisa que retratem o Distrito Federal, que hoje abriga uma região metropolitana com cerca de 3 milhões de habitantes em torno da capital federal.

4.2.3 Informações de sexo/gênero e faixa etária

Consideradas fundamentais para os propósitos de linguística e fonética forense, as informações de sexo/gênero e de faixa etária dos participantes da coleta foram unanimidade no que diz respeito a coleta: todos os bancos de dados de língua e fala avaliados neste trabalho coletaram tais dados, como explicitam o Gráfico 12 e o Gráfico 13.

Gráfico 12 - Informação de sexo/gênero das amostras

Fonte: autoria própria.

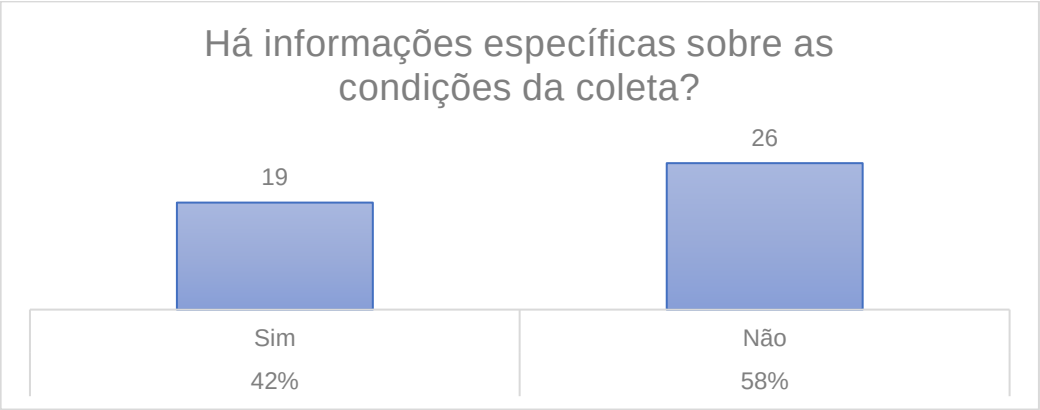
Gráfico 13 - Informação de faixa etária das amostras

Fonte: autoria própria.

4.3 Características da coleta

Nesta seção, tratou-se das informações a respeito da coleta dos materiais dos bancos. O Gráfico 14 explicita quantos dos bancos levantados coletam ou informam dados sobre as condições de coleta.

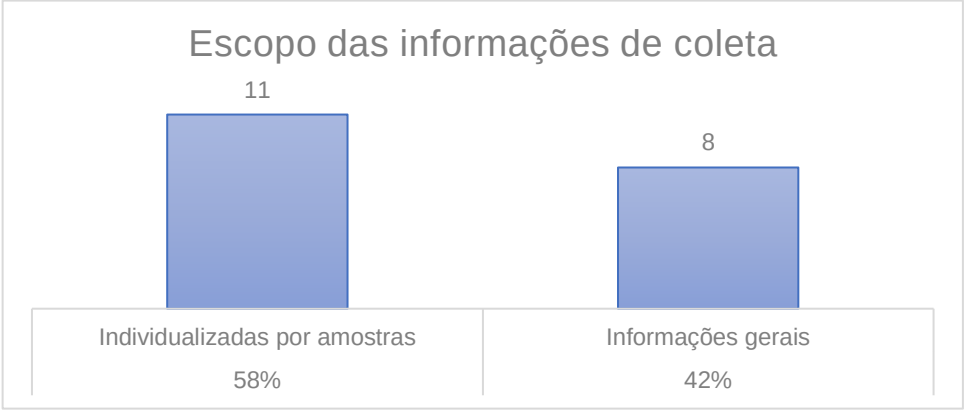
Gráfico 14 - Sobre informações específicas sobre as condições de coleta



Fonte: autoria própria.

Percebe-se que a maioria dos bancos não coletou ou não informou sobre as condições de coleta. Dos 19 bancos que coletaram/informam essas informações, cabe saber sobre o escopo delas: se são informações gerais, ou seja, apresentam de uma só vez os dados sobre a coleta de todas as amostras, ou se trata de informações sobre cada uma das amostras coletadas. O Gráfico 15 ilustra os escopos das informações de coleta.

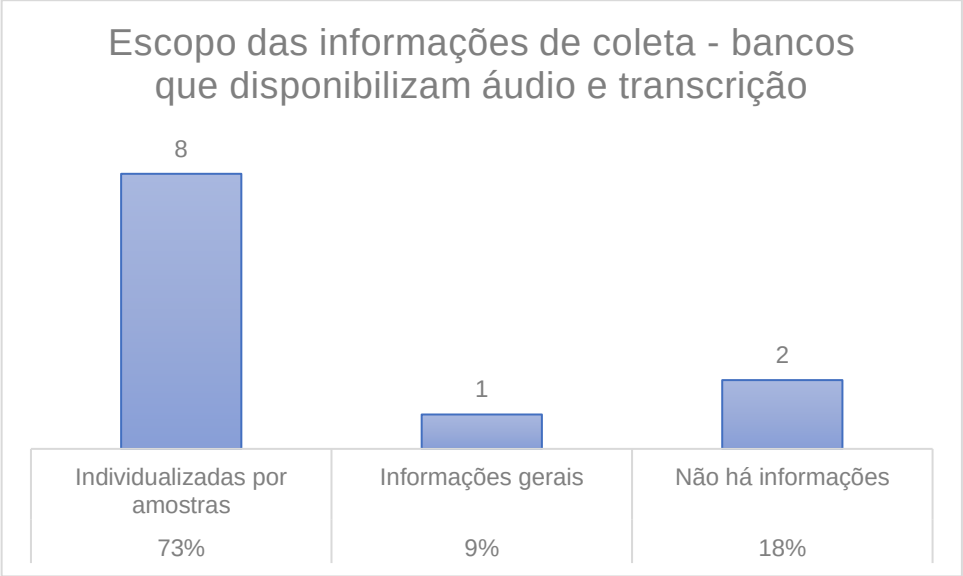
Gráfico 15 - Escopo das informações de coleta



Fonte: autoria própria.

Após constatar quantos bancos apresentam informações individualizadas ou gerais, é pertinente saber como essa distribuição acontece em relação aos bancos de dados que disponibilizam materiais de áudio e transcrição, conforme ilustrado no Gráfico 16.

Gráfico 16 - Escopo das informações de coleta para bancos que disponibilizam áudio e transcrição



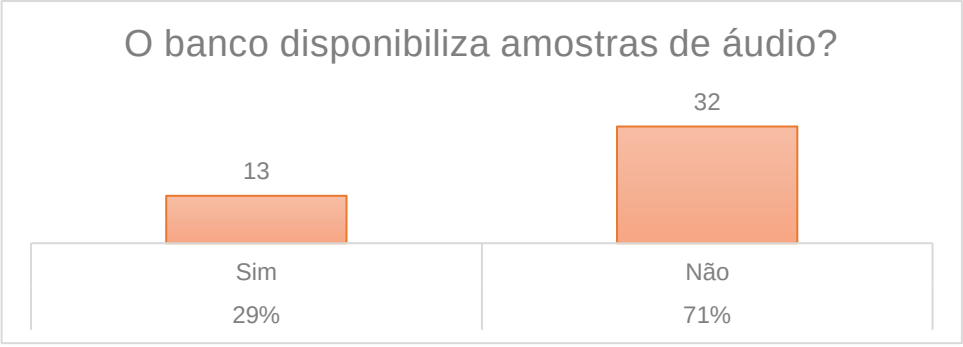
Fonte: autoria própria.

Dessa forma, temos que, para os propósitos forenses, ou seja, para bancos que disponibilizam materiais de áudio e transcrição, 8 dos 11 bancos têm informações de condições de coleta de forma individualizada, por amostras, o que é de utilidade em tarefas de linguística e fonética forense.

4.4 Áudio

Na seção *áudio*, são avaliadas características gerais das amostras de áudio disponibilizadas pelos bancos. Primeiramente, é importante visualizarmos quantos dos bancos levantados disponibilizam materiais de áudio, como mostra o Gráfico 17.

Gráfico 17 - Disponibilidade das amostras de áudio



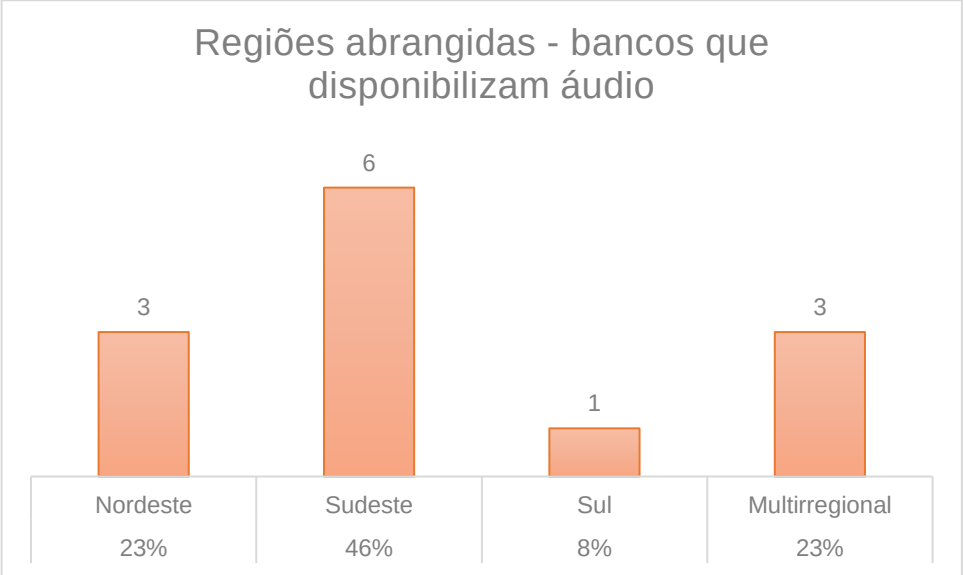
Fonte: autoria própria.

Agora sabemos que 13 dos bancos, ou seja, 29% dos bancos levantados, disponibilizam amostras de áudio. É importante esclarecer que essa categoria se refere a dados de áudio que podem ser acessados, isto é, que estão disponíveis no momento de realização deste estudo. Portanto, a resposta “Sim” não computa os bancos que estão em manutenção ou temporariamente sem possibilidade de disponibilização dos dados.

O fato de apenas 29% dos bancos disponibilizarem amostras de áudio expõe uma limitação no uso em fonética forense para a tarefa de CL, visto que as análises acústicas e sociofonéticas são de extrema relevância para a geração de estatísticas de traços linguísticos para determinar graus de similaridade e tipicidade de amostras questionada (AQ) e padrão (AP).

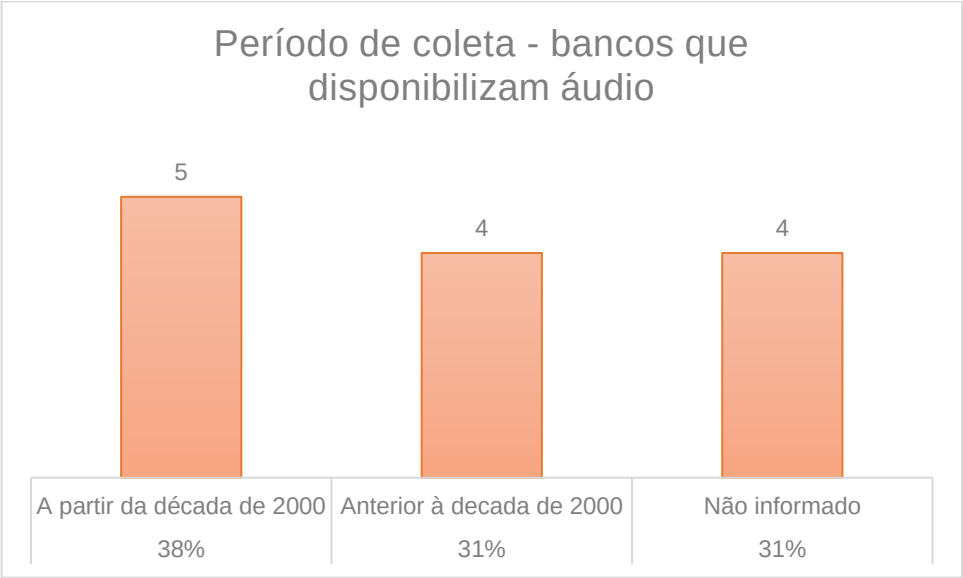
Sabendo quantos bancos disponibilizam áudio, é interessante compreender como eles se distribuem nas regiões do país, bem como a década em que os materiais foram coletados. O Gráfico 18 e o Gráfico 19 ilustram essa combinação de informações.

Gráfico 18 - Regiões abrangidas por bancos que disponibilizam áudio



Fonte: autoria própria.

Gráfico 19 - Período de coleta de bancos que disponibilizam áudio



Fonte: autoria própria.

Nos Gráficos 18 e 19, é importante citar os bancos classificados como “multirregional”. São eles 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)*, que tem abrangência nacional, e o *Mozilla Common Voice*, que coleta dados de todo o Brasil. Sendo assim, não foi possível encontrar bancos que disponibilizem materiais de áudio nas regiões Centro-Oeste (em grande escala) e Norte, o que dificulta a tarefa de CL no que diz respeito a estatísticas populacionais de traços linguísticos.

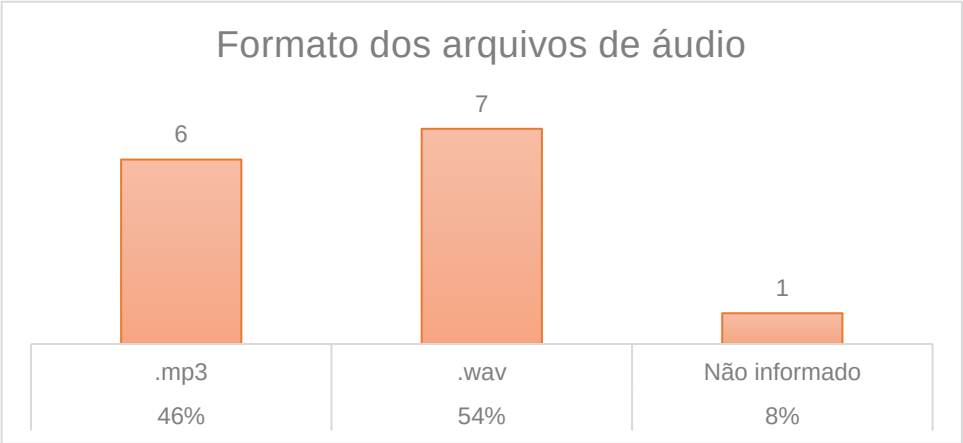
Além disso, vemos que 5 dos bancos tiveram um período de coleta a partir da década de 2000, ou seja, são bancos mais recentes, o que significa que a qualidade acústica das amostras de fala não foi comprometida pela deterioração do áudio pelo tempo. São esses o *ALIP - Iboruna (Amostra Censo)*, o *ALIP - Iboruna (Amostra de Interação)*, o *Projeto SP2010*, o *C-ORAL-BRASIL (I)* e o *LínguaPOA*.

4.4.1 Formato dos arquivos de áudio

Dentre os bancos que disponibilizam áudio, alguns bancos optaram pela utilização de arquivos .mp3 e outros optaram pelo formato .wav. Para os propósitos deste trabalho, é interessante observar a distribuição da quantidade de bancos para cada formato de arquivo de áudio, uma vez que .wav é um formato sem perdas por compressão, enquanto os arquivos .mp3 sofrem compressão e perda de informação acústica, fazendo com que o formato .wav permita maior

possibilidade de análises acústicas robustas em comparação ao formato .mp3. A distribuição dos formatos de arquivo de áudio está representada no Gráfico 20.

Gráfico 20 - Formato dos arquivos de áudio[6]



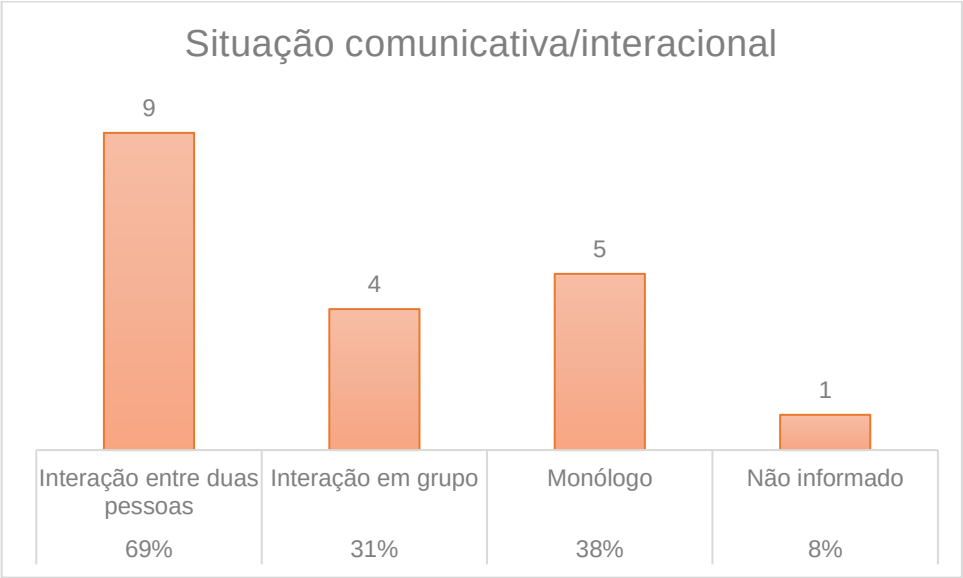
Fonte: autoria própria.

4.4.2 Situação comunicativa/interacional e estilo de elocução

É fundamental, na tarefa de CL, que se tenha informações sobre a situação comunicativa das amostras de áudio a serem analisadas. Um monólogo pode ter propriedades fonético-perceptivas, conversacionais e discursivas diferentes de uma conversa entre dois indivíduos, bem como de uma conversa em grupo. A importância de tais informações se dá pelo fato de que, para a tarefa de CL, é preferível que haja congruência entre os estilos de fala presente nas amostras questionadas e padrão, bem como nas amostras utilizadas para gerar as estatísticas populacionais. Pode-se visualizar no Gráfico 21 a distribuição das diferentes situações comunicativas encontradas nas amostras dos bancos mencionados nesta seção.

[6] Nesse caso, um dos bancos ocupa as duas das barras do gráfico, pois disponibiliza arquivos em ambos os formatos .wav e .mp3. Trata-se do *Projeto SP2010*.

Gráfico 21 - Situação comunicativa/interacional[7]

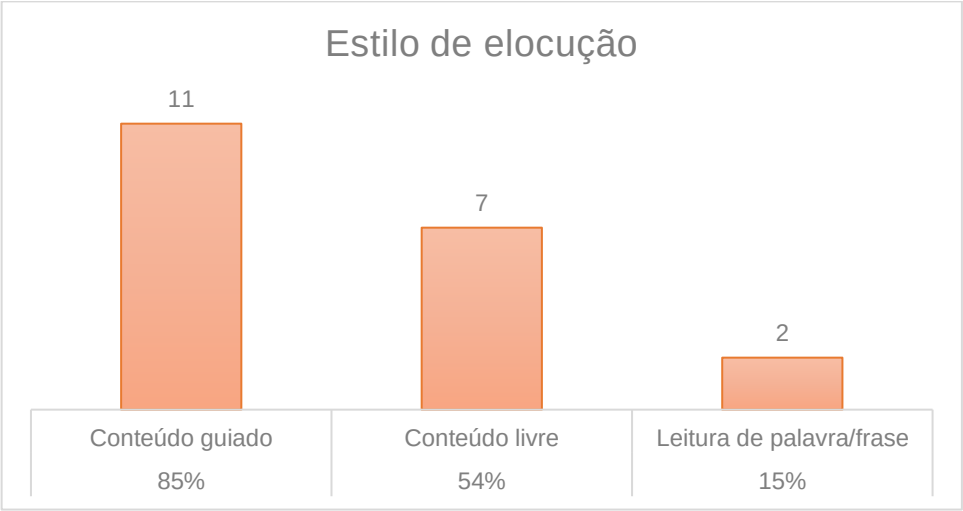


Fonte: autoria própria.

Além de compreender a natureza da situação comunicativa das amostras, o estilo de elocução é, também, peça fundamental na compreensão dos traços conversacionais que podem ser decisivos em análises fonético-forenses, e podem ser observadas no Gráfico 22.

[7] Aqui, há bancos que ocupam mais de uma barra do gráfico, visto que expõem amostras com mais de uma situação comunicativa: Os bancos *NURC SP* e *NURC Recife* têm amostras de monólogo e interação entre duas pessoas; o *Programa de Estudos sobre o Uso da Língua (PEUL)* tem amostras de interação entre duas pessoas e interação em grupo; os bancos *C-ORAL-BRASIL (I)* e o *CORAA NURC-SP Minimal Corpus* têm amostras de monólogo, interação entre duas pessoas e interação em grupo.

Gráfico 22 - Estilo de elocução[8]



Fonte: autoria própria.

A categoria "conteúdo guiado" se refere ao tipo de coleta de amostra em que o documentador guia a fala do informante por assuntos, temas conduzidos intencionalmente para obter amostras visando objetivos específicos. Conteúdos guiados podem ser úteis em análise forense, uma vez que podem ser classificados como fala semiespontânea.

Conteúdo livre diz respeito a amostras em que há fala de informante(s) sem nenhuma forma de condução específica. São, basicamente, conversas gravadas e, a depender do banco, transcritas com intuitos específicos, variando de banco para banco. Amostras por conteúdo livre são interessantes para a o contexto de fonética forense, visto que podem ser mais facilmente comparadas a uma amostra questionada. Como trata-se de situações conversacionais reais, elementos fonético-linguísticos, como os elementos prosódicos, podem ser mais fielmente analisados.

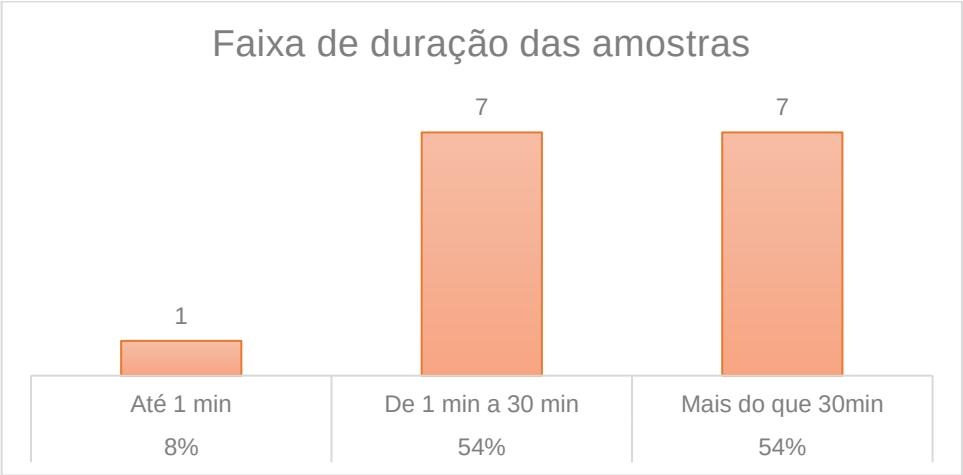
A categoria "Leitura de palavra/frase" não é a mais interessante no contexto forense, já que limita a análise a parâmetros unicamente acústicos e pode não haver congruência em relação ao estilo de elocução de uma eventual amostra questionada.

4.4.3 Faixa de duração das amostras de áudio

A faixa de duração das amostras dos bancos levantados pode ser observada no Gráfico 23.

[8] Nesse caso, a maioria dos bancos que disponibilizam áudio expõem amostras com mais de um tipo de elocução, conciliando conteúdo guiado com conteúdo livre.

Gráfico 23 - Faixa de duração das amostras[9]



Fonte: autoria própria.

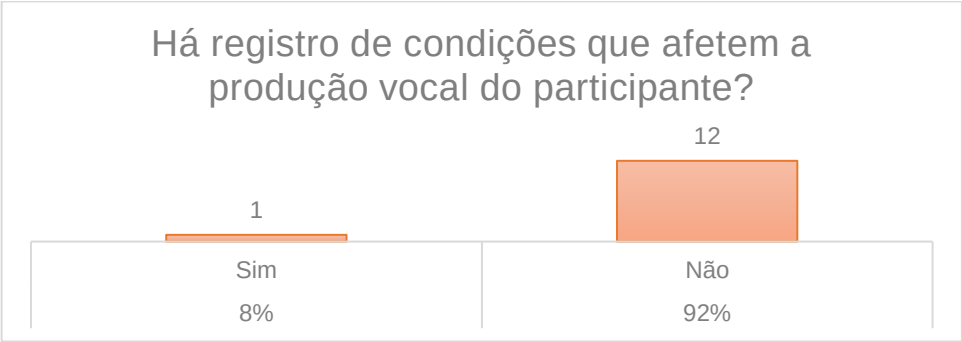
O Gráfico 23 mostra que a maioria dos bancos disponibiliza amostras de áudio com duração superior a 1 minuto e vários deles com duração superior a 30 minutos. Para a tarefa de CL, tal informação é extremamente importante, uma vez que amostras de fala muito curtas impossibilitam a extração de parâmetros acústicos que sejam representativos da fala do participante. Portanto, é preferível o uso de áudios de maior duração.

4.4.4 Registro de condições vocais

Em análise acústica em contexto forense, informações como doenças respiratórias, histórico de cirurgia no trato vocal, histórico de tabagismo, presença de aparelho ortodôntico, entre outras, determinam parâmetros acústicos específicos a respeito do trato vocal que, se ignorados, podem resultar em conclusões errôneas. Assim, o Gráfico 24 ilustra a importância da consideração de dados sobre as condições vocais dos informantes.

[9] Aqui, dois dos bancos ocupam duas barras do gráfico, visto que possuem amostras de 1 a 30 minutos e de mais de 30 minutos. São eles: *Estudos da Língua Oral do Cariri* e *CORAA NURC-SP Minimal Corpus*.

Gráfico 24 - Sobre registros de condições que afetem a produção vocal do participante



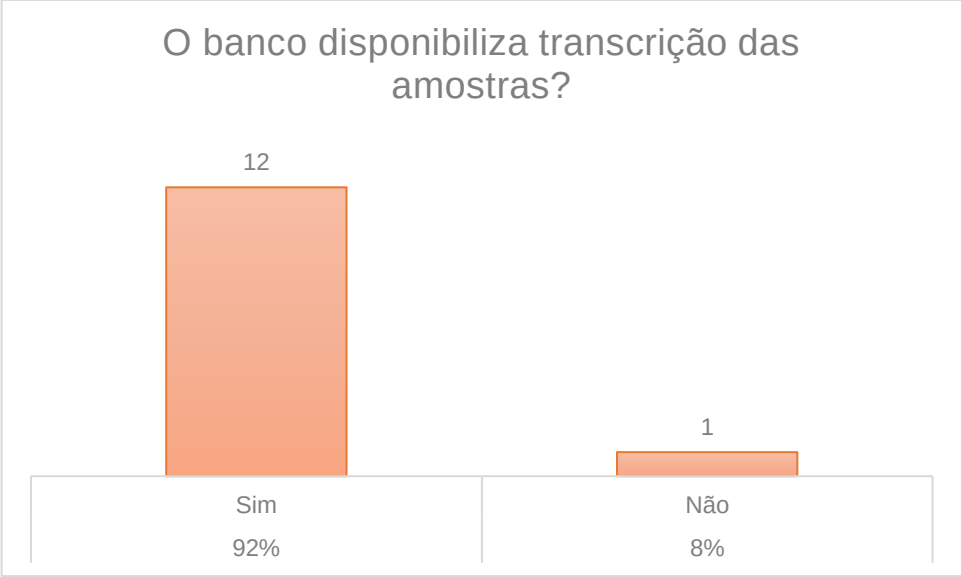
Fonte: autoria própria.

Vemos que, dos 13 bancos de dados de língua e fala que disponibilizam material de áudio, apenas um deles registra as condições vocais que podem afetar nas análises acústicas. Trata-se do *Corpus Forense do Português Brasileiro (CFPB)*, o único banco criado com propostas fundamentalmente forenses que se tem na atualidade.

4.4.5 Disponibilização de transcrição dos materiais de áudio

Dos bancos a serem avaliados nesta seção, o Gráfico 25 expõe que apenas um deles não disponibiliza transcrição das amostras de fala.

Gráfico 25 - Sobre disponibilização da transcrição das amostras pelos bancos



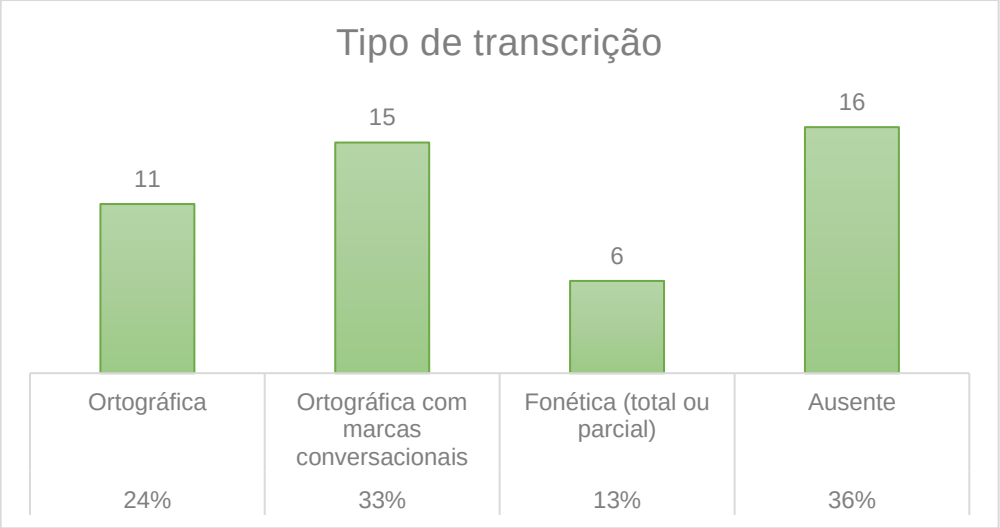
Fonte: autoria própria.

Trata-se, novamente, do *Corpus Forense do Português Brasileiro (CFPB)*, visto que o banco disponibiliza unicamente amostras de áudio.

4.5 Transcrição

Nesta última seção, avalia-se o tipo de transcrição que foi coletada e/ou disponibilizada pelos bancos levantados. Sendo assim, o Gráfico 26 apresenta a distribuição dessas informações.

Gráfico 26 - Tipo de transcrição



Fonte: autoria própria.

Transcrições ortográficas são aquelas em que o conteúdo linguístico de uma amostra de fala é registrado conforme as regras da ortografia padrão do português brasileiro. Ou seja, as palavras são escritas da mesma forma como são encontradas nos dicionários, sem considerar variações na pronúncia ou características específicas da fala.

Transcrições ortográficas com marcas conversacionais são aquelas em que além de registrar o conteúdo linguístico, são adicionadas notações para indicar elementos da organização do texto falado. Isso pode incluir pausas, ênfases, truncamentos (quando uma palavra é interrompida ou não é pronunciada completamente), e outros aspectos próprios da conversação.

Transcrições fonéticas representam a pronúncia fonética das palavras, ou seja, como elas são realmente faladas, independentemente da ortografia. Podem ser totais, cobrindo todas as palavras do texto, ou parciais, abordando apenas algumas palavras selecionadas. Geralmente são utilizadas em atlas linguísticos, onde se deseja apresentar variações fonéticas na pronúncia

de palavras específicas, acompanhadas por símbolos fonéticos que representam os sons da fala de forma mais precisa. Como exemplo, temos as variações fonéticas da palavra “sete”, podendo ser representadas foneticamente como [ˈsɛtɪ] ou [ˈsɛt͡ʃɪ]. Ou, se a intenção for enfatizar um som isolado na palavra, pode haver a transcrição fonética apenas do fone em questão, como na comparação entre “se[t]e” e “se[t͡ʃ]e”.

4.6 Principais bancos de interesse para a tarefa de Comparação de Locutor

Ao longo desta seção de resultados foi possível visualizar as informações que concernem ao contexto forense por meio dos gráficos apresentados. Em alguns dos casos, foram apresentados gráficos que fossem de maior interesse para a área de fonética forense, como a ênfase nos bancos de dados que disponibilizam materiais de áudio e transcrição. Nesta subseção, estão compilados os bancos que cumprem com determinadas características na tabela apresentada. Foram priorizados os bancos que fizeram suas coletas a partir da década de 2000, ou que disponibilizem amostras de áudio bem audíveis.

Tabela 3 - Principais bancos de interesse para a tarefa de Comparação de Locutor

	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	Sim	Sim	Não informa	Sim	Não informado	Não informado	Sim

aproveitamento para contextos forenses?			do				
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

5 CONSIDERAÇÕES ACERCA DOS RESULTADOS OBTIDOS

Como este é um trabalho de levantamento, descrição e sistematização dos bancos de dados, não houve hipóteses a serem testadas. O resultado do trabalho é a apresentação do levantamento dos bancos de dados de fala de forma sistematizada e organizada, de modo que um indivíduo interessado pelo tema possa identificar quais bancos são adequados para a geração de distribuição populacional de determinadas características linguísticas e fonéticas.

A seguir, portanto, serão feitas algumas considerações a partir dos resultados apresentados na seção anterior.

5.1 Das características gerais dos bancos

Há alguns pontos a serem mencionados em relação às características dos bancos levantados. Primeiramente, é vale esclarecer que há grande importância de disponibilização da caracterização dos informantes participantes das coletas de amostras de cada um dos bancos levantados neste trabalho. Tais características podem ser decisivas na análise de casos em contexto forense.

Além disso, pôde-se notar que há um grande número de bancos avaliados que, no momento de condução desta pesquisa, carecem de informações acessíveis sobre o período de coleta. Essa lacuna de informação prejudica o aproveitamento dos dados desses bancos no contexto forense, visto que dificulta o processo de análise acústica por conta da deterioração de áudios pelo tempo, por exemplo. Além do mais, pode-se dizer também que há relativamente pouco dado sobre o período de coleta desde a virada do século, visto que 18% dos bancos trazem a informação de coleta de suas amostras na década de 2000, e apenas 13% dos bancos trazem a informação de coleta de suas amostras na década de 2010.

Outro fator que pode ser considerado problemático para o contexto forense é a falta da presença de termos de uso explícitos nos bancos avaliados. Como a existência de termos de uso é comum atualmente, a inexistência desses termos pode desencorajar o seu uso por parte de peritos por medo de enfrentar consequências legais.

5.2 Das características de condições de coleta dos bancos

Poucos dos bancos de dados levantados trouxeram informações relevantes a respeito das condições de coleta. Desses, destacam-se os dois bancos do *ALIP - Iboruna* (Amostra Censo e

Amostra Interação), o *Projeto SP2010* e o *C-ORAL-BRASIL (I)*. Tais informações, principalmente aquelas que trazem considerações sobre as condições ambientais de coleta, como presença de ruído e sobreposição de vozes, bem como caracterização das condições de gravação, são extremamente bem-vindas para o uso forense.

5.3 Dos áudios disponibilizados pelos bancos

Dos bancos que disponibilizam amostras de áudio, há relativamente pouca disponibilidade de dados recentes com grande abrangência geográfica, o que foi constatado nos gráficos sobre regiões abrangidas e período de coleta desses bancos com materiais acústicos.

Em relação às faixas de duração das amostras, o resultado foi relativamente bom: a maioria dos bancos com materiais de áudio apresentaram amostras relativamente longas quando comparadas à situação forense típica. Quanto maior a duração do material a ser analisado, mais completa pode ser a análise feita pelo perito.

Além disso, espera-se que o presente projeto seja de utilidade tanto para especialistas na área de linguística e fonética forense quanto para especialistas em outras áreas da linguagem e possa incentivar a médio e longo prazo o desenvolvimento de novos bancos de dados para estatísticas de distribuição populacional de traços linguísticos brasileiros para fins fundamentalmente forenses.

6 CONCLUSÃO

O presente trabalho apresenta um levantamento de diversos bancos de dados de língua e fala do português brasileiro para avaliar a sua aproveitabilidade em tarefas de linguística forense e, mais especificamente, em fonética forense na tarefa de CL. Foram trazidos dados que podem auxiliar na geração de estatísticas de traços linguísticos de populações relevantes, além de contribuir para esta área do conhecimento tão rica e com tanto a ser explorada no contexto brasileiro.

A partir dos resultados obtidos com este trabalho, pode-se concluir que o Brasil dispõe de um bom número de bancos de dados que podem ser úteis para a tarefa de CL. Entretanto, em relação ao processo de geração de estatísticas de traços linguísticos importantes para a determinação do grau de similaridade e tipicidade de amostras, percebe-se que poderíamos dispor de muitos mais bancos que disponibilizassem material de áudio em muitas das regiões do país. Infelizmente, não encontramos tais bancos nas regiões Centro-Oeste e Norte e, em

relação às regiões Sul e Nordeste, selecionamos, no final, apenas um banco de cada uma delas. Isso mostra que é pertinente que haja trabalhos futuros de coleta de dados de língua e fala em todas as regiões do país, os quais disponibilizem suas amostras de áudio e exponham termos de uso que permitam a utilização em trabalhos periciais.

Muitos dos bancos apresentados podem ser reaproveitados para análises linguísticas, principalmente no que concerne a fenômenos de natureza lexical, sintática, conversacional, etc. Porém, no que diz respeito ao seu aproveitamento como recursos para auxiliar peritos forenses em casos de CL, verifica-se que as opções são relativamente limitadas. Contudo, os bancos que possibilitam o reaproveitamento desejado cumprem bem o papel esperado.

Este trabalho buscou trazer informações pertinentes à área de linguística e fonética forense de bancos de dados de língua e fala brasileiros. Entretanto, há ainda a intenção de detalhar os parâmetros acústicos que podem ser analisados em cada um dos bancos. Como trata-se de uma questão complexa, devido ao aprofundamento em fonética acústica, isso justificaria um trabalho separado.

Por fim, espera-se que o presente projeto seja de utilidade tanto para especialistas na área de linguística e fonética forense quanto para especialistas em outras áreas da linguagem e possa incentivar a médio e longo prazo o desenvolvimento de novos bancos de dados para estatísticas de distribuição populacional de traços linguísticos brasileiros para fins fundamentalmente forenses.

REFERÊNCIAS

BARBOSA, P. A. et al. (EDS.). **Análise fonético-forense em tarefa de comparação de locutor.** 1. ed. 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. (EDS.). **Análise fonético-forense em tarefa de comparação de locutor.** 1. ed. Campinas: Millenium Editora, 2020, p. 67-87.

FREITAG, R. M. K. Sociolinguística no/do Brasil. **Cadernos de Estudos Linguísticos**, Campinas, SP, v. 58, n. 3, p. 445–460, 2016. DOI: 10.20396/cel.v58i3.8647170. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/cel/article/view/8647170>. Acesso em: 20 jul. 2023.

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.

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.

JESSEN, M. Forensic Phonetics. **Language and Linguistics Compass**, v. 2, n. 4, p. 671–711, 2008.

MORRISON, G. S. Forensic voice comparison and the paradigm shift. **Science and Justice**, v. 49, n. 4, p. 298–308, 2009.

PASSETTI, R. R **Fonética Forense**. *In*: Verbetes LBASS. Disponível em: <http://www.letras.ufmg.br/padrao_cms/index.php?web=lbass&lang=1&page=&menu=&tipo=1>. Acesso em: 18 jul 2023

ROSE, P. **Forensic speaker identification**. London ; New York: Taylor & Francis, 2002.

APÊNDICES

APÊNDICE A - Questionário

Dados a serem levantados para cada banco de dados

1. **Características gerais do corpus**

Nome do banco de dados:

Site do banco:

Referência de descrição do banco:

Tipos de materiais coletados:

Tipos de materiais coletados: ● Áudios ● Transcrições ● Caracterização dos informantes/coleta ● Vídeos ● Mapas ● Gráficos ● Outro

É possível ter acesso a algum material? ● Sim ● Não

Tipos de materiais disponibilizados: ● Áudios ● Transcrições

- Caracterização dos informantes/coleta
- Vídeos
- Mapas
- Gráficos
- Outro

Período de coleta das amostras:

Quantidade disponível de amostras de áudio do banco:

Quantidade disponível de amostras transcritas do banco:

Procedimento de permissão para acesso aos materiais:

- Mediante cadastro
- Acesso livre
- Contato com responsáveis
- Mediante compra
- Outro

O projeto tem/expõe termo de uso?

- Sim
- Não

O termo de uso permite o aproveitamento de dados para os cenários típicos de fonética forense?

- Sim
- Não

2. Características dos participantes

Regiões geográficas do Brasil que o corpus abrange:

- Norte
- Nordeste
- Centro-Oeste
- Sudeste
- Sul

Estados brasileiros abrangidos:

- Acre (AC)
- Alagoas (AL)
- Amapá (AP)
- Amazonas (AM)
- Bahia (BA)
- Ceará (CE)
- Distrito Federal (DF)
- Espírito Santo (ES)
- Goiás (GO)
- Maranhão (MA)
- Mato Grosso (MT)
- Mato Grosso do Sul (MS)
- Minas Gerais (MG)
- Pará (PA)
- Paraíba (PB)
- Paraná (PR)
- Pernambuco (PE)
- Piauí (PI)
- Rio de Janeiro (RJ)
- Rio Grande do Norte (RN)
- Rio Grande do Sul (RS)
- Rondônia (RO)

- Roraima (RR)
- Santa Catarina (SC)
- São Paulo (SP)
- Sergipe (SE)
- Tocantins (TO)

Cidades/regiões abrangidas:

Há informação de sexo/gênero das amostras?

- Sim
- Não

Há informação de faixa-etária das amostras?

- Sim
- Não

Há outras características sociodemográficas? (escolaridade, profissão, etc.)

3. Características da coleta

Há informações específicas sobre as condições da coleta?

- Sim
- Não

Qual o escopo das informações?

- Individualizadas por amostras
- Informações gerais

Tipos de informações presentes:

- Nome do informante
- Nome do entrevistador/documentador
- Caracterização física do local da coleta
- Caracterização das condições de gravação
- Grau de familiaridade entre entrevistador e informante
- Grau de familiaridade entre interlocutores
- Informação sobre a situação comunicativa/interacional

Para a próxima seção:

O banco disponibiliza amostras de áudio?

- Sim
- Não

4. Áudio

Formato dos arquivos de áudio:

- .wav
- .mp3
- .ogg (opus)
- .aac
- .aiff
- Outro

Situação comunicativa/interacional:

- Monólogo
- Interação entre duas pessoas
- Interação em grupo
- Outro

Estilo de elocução:

- Leitura de lista de palavras, frases ou texto
- Conteúdo guiado (por tema, por perguntas do entrevistador)
- Conteúdo livre

Faixas de duração das amostras

- Até 1 minuto
- De 1 minuto a 30 minutos
- Mais do que 30 minutos

Há registro de condições que afetem a produção vocal do participante? (histórico de cirurgia oral, aparelho ortodôntico, queixas vocais, afecções do trato respiratório, estado de saúde, histórico de tabagismo)

- Sim
- Não

Para a próxima seção:

O banco disponibiliza transcrição das amostras?

- Sim
- Não

5. Transcrição

Tipo de transcrição:

- Ortográfica
- Ortográfica com marcas conversacionais (pausas, disfluências, sobreposições)
- Fonética: IPA
- Fonética (sistema próprio desenvolvido para o projeto)
- Outro

ANEXOS

ANEXO A - Referências bibliográficas dos bancos de dados levantados no trabalho

Nome do banco de dados:	Referência de descrição do banco:
ALIP - Iboruna (Amostra Censo)	GONÇALVES, Sebastião Carlos Leite; TENANI, Luciani Ester. Projeto ALiRP: constituição de um banco de dados para o estudo do português falado na região de São José do Rio Preto. Mosaico (São José do Rio Preto), São José do Rio Preto, v. 3, n.2, p. 13-37, 2004.
ALIP - Iboruna (Amostra de Interação)	GONÇALVES, Sebastião Carlos Leite; TENANI, Luciani Ester. Projeto ALiRP: constituição de um banco de dados para o estudo do português falado na região de São José do Rio Preto. Mosaico (São José do Rio Preto), São José do Rio Preto, v. 3, n.2, p. 13-37, 2004.
NURC RJ	CASTILHO, A. T. Informações sobre o Projeto de Estudo da Norma Urbana Lingüística Culta (Projeto NURC). Cadernos de Estudos Lingüísticos, v. 6, p. 187–190, 1984.
NURC Digital/Recife	Nurc Digital. Disponível em: <https://fale.ufal.br/projeto/nurcdigital/index.php?action=home>.
NURC SP	CASTILHO, A. T. Informações sobre o Projeto de Estudo da Norma Urbana Lingüística Culta (Projeto NURC). Cadernos de Estudos Lingüísticos, v. 6, p. 187–190, 1984.
Programa de Estudos sobre o Uso da Língua (PEUL)	SCHERRE, M. M. P.; RONCARATI, C. Programa de estudos sobre o uso da língua (PEUL): origens e trajetórias. In: VOTRE, S.; RONCARATI, C. Anthony Julius Naro e a linguística no Brasil: uma homenagem acadêmica. Rio de Janeiro: 7Letras, 2008. p.37-49. PAIVA, M. C.; SCHERRE, M. M. P. Retrospectiva sociolinguística: contribuições do PEUL. D.E.L.T.A, São Paulo, v.15, n. especial, p.201-222, 1999

Projeto Atlas Linguístico do Brasil (Projeto ALiB)	AGUILERA, V. de A.; MILANI, G. A. L.; MOTA, J. A. (Org.). Projeto Atlas Lingüístico do Brasil – Documentos I. Salvador: ILUFBA–EDUFBA, 2004.
Projeto SP2010	MENDES, R.B. (2013) Projeto SP2010: Amostra da fala paulistana
C-ORAL-BRASIL (I)	RASO, Tommaso; MELLO, Heliana (Org.). C-oral-Brasil I: corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012.
Discurso & Gramática	DA CUNHA, Maria Angélica Furtado; TAVARES, Maria Alice; COSTA, Marcos Antônio. Grupo de estudos Discurso & Gramática: pesquisa em desenvolvimento. Leitura, n. 35, p. 207-219, 2005
Atlas Linguístico-Etnográfico da Região Sul do Brasil (ALERS)	Koch, Walter; Altenhofen, Cléo V. & Klassmann, Mário (Orgs.). Atlas Lingüístico-Etnográfico da Região Sul do Brasil (ALERS): Introdução, Cartas fonéticas e morfossintáticas. 2a. ed. Porto Alegre: Editora da UFRGS; Florianópolis: Editora da UFSC, 2011. 512 p. Disponível em: https://lume.ufrgs.br/handle/10183/232185. Demais autores: Agostini, Basílio; Altenhofen, Cléo V.; Furlan, Oswaldo; Klassmann, Mário; Koch, Walter (†); Margotti, Felício Wessling; Mercer, José Luiz da Veiga; Vieira, Hilda Gomes (†)
Norma Oral do Português Popular de Fortaleza (NORPOFOR)	DE ARAÚJO, Aluiza Alves. O PROJETO NORMA ORAL DO PORTUGUÊS POPULAR DE FORTALEZA NORPOFOR. 2011.
PORTAL - Variação linguística no português alagoano	OLIVEIRA, Alan Jardel. Projeto PORTAL: variação linguística no português alagoano. http://www. portuguesalagoano. com. br/>.
BrasilData	DE PAULA MACHADO, Aline. Análise fonético-acústica para fins forenses em cinco localidades brasileiras a partir da base Brasildata. 2018. Doutora em Linguística – Universidade Estadual de Campinas, Campinas, 2018.

LínguaPOA	LÍNGUAPOA. Universidade Federal do Rio Grande do Sul. 2015-2019 (período de coleta). Disponível em: https://www.ufrgs.br/linguapoa/.
Projeto Variação Linguística no Estado da Paraíba (VALPB)	HORA, Dermeval da; PEDROSA, Juliene Lopes Ribeiro. Projeto variação lingüística no Estado da Paraíba (VALPB). João Pessoa: Idéia, v. 5, 2001.
BANCO DE DADOS FALARES SERGIPANOS	FREITAG, Raquel Meister Ko. Banco de dados falares sergipanos. Working Papers em Linguística, v. 14, n. 2, p. 156-164, 2013.
A língua portuguesa do semiárido baiano	DA AMOSTRAGEM, D. O. O PROJETO A LÍNGUA PORTUGUESA NO SEMIáRIDO BAIANO–FASE 3: CRITÉRIOS DE CONSTITUIÇÃO E. METODOLOGIA DE COLETA E MANIPULAÇÃO DE DADOS EM SOCIOLINGUÍSTICA, p. 26, 2014.
Programa de Estudos sobre o Português Popular de Salvador (PEPP)	DA SILVA LOPES, Norma. O PEPP e os estudos sobre o português de Salvador. A Cor das Letras, v. 19, p. 23-39, 2018.
Estudos da Língua Oral do Cariri	SOARES, Maria Elias (Org.). O português falado no Ceará: corpus do projeto. Fortaleza: Universidade Federal do Ceará – UFC. Disponível em: <www.profala.ufc.br>
Dialetos Sociais Cearenses	ARAGÃO, Maria do Socorro Silva de; SOARES, Maria Elias(Org.). A linguagem falada em Fortaleza: diálogos entre informantes e documentadores – materiais para estudo. Fortaleza: Universidade Federal do Ceará – UFC. Disponível em: <www.profala.ufc.br>
Projeto Variação Linguística na Região Sul do Brasil (VARSUL)	VARSUL, PROJETO. Variação Linguística na Região Sul do Brasil: banco de dados.
Português Oral Culto de Fortaleza (PORCUFORT)	MONTEIRO, José Lemos. O português oral culto de Fortaleza–PORCUFORT. Manuscrito inédito, 1993.

Corpus Forense do Português Brasileiro (CFPB)	Indisponível
CORAA NURC-SP Minimal Corpus	SANTOS, Vinıcius G. et al. CORAA NURCSP Minimal Corpus: a manually annotated corpus of Brazilian Portuguese spontaneous speech. In: 6th International Conference on Speech and Language Technologies on Iberian languages, IberSPEECH. 2022.
Atlas Prévio dos Falares Baianos (APFB)	ROSSI, Nelson. Atlas Prévio dos Falares Baianos. Rio de Janeiro: INL, 1963.
Esboço de um Atlas Linguístico de Minas Gerais (EALMG)	RIBEIRO, José. Esboço de um atlas lingüístico de Minas Gerais. Ministério da Educação e Cultura, Fundação Casa de Rui Barbosa, 1977.
Atlas Linguístico da Paraíba (ALPB)	ARAGÃO, Maria do Socorro Silva de.; BEZERRA DE MENEZES, Cleusa P. Atlas Lingüístico da Paraíba. Brasília: UFPB/CNPq, Coordenação Editorial, 1984; v. 1, 2
Atlas Linguístico de Sergipe (ALS)	FERREIRA, Carlota et al. Atlas Lingüístico de Sergipe. Salvador: UFBA - Instituto de Letras/Fundação Estadual de Cultura de Sergipe, 1987.
Atlas Linguístico do Paraná (ALPR)	AGUILERA, Vanderci de Andrade. Atlas Lingüístico do Paraná. Curitiba: Imprensa Oficial do Estado, 1994.
Atlas Linguístico de Sergipe II (ALS II)	CARDOSO, Suzana Alice Marcelino da Silva. Atlas Lingüístico de Sergipe II. Rio de Janeiro: S. A. M. da S. Cardoso, 2002. 2v.
Atlas Linguístico Sonoro do Pará (ALISPA)	RAZKY, Abdelhak. (Org.) Atlas lingüístico sonoro do Pará. Belém: PA/CAPES/UTM, 2004. CDRoom.
Atlas Geolingüístico do Litoral Potiguar (ALiPTG)	PEREIRA, Maria das Neves. Atlas geolinguístico do litoral potiguar. Rio de Janeiro: UFRJ, Faculdade de Letras, 2007. 2v. Vol I: 123p. mimeo. Vol II 189p. mimeo. Tese de Doutorado em Letras Vernáculas.

Atlas Linguístico de Mato Grosso do SUL (ALMS)	OLIVEIRA, Dercir. Pedro de (Org.). ALMS - Atlas Lingüístico de Mato Grosso do Sul. 1. ed. Campo Grande: Editora UFMS, 2007. 271 p.
Atlas Semântico-Lexical da Região do Grande ABC	CRISTIANINI, Adriana Cristina. Atlas Semântico-Lexical da Região do Grande ABC. 2007. 772f. Tese (Doutorado – Programa de Pós-Graduação em Lingüística. Área de concentração: Semiótica e Lingüística Geral) – Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, São Paulo, 2007.
Atlas Linguístico da Mata Sul de Pernambuco (ALMASPE)	ALMEIDA, Edilene Maria de Oliveira. Atlas Linguístico da Mata Sul de Pernambuco-Almaspe. 2009. 149f. Dissertação (Mestrado em Linguagens e Cultura) – Universidade Federal da Paraíba, João Pessoa, 2009.
Atlas Linguístico da Mesorregião Sudeste de Mato Grosso (ALMESEMT)	CUBA, M. A. Atlas Linguístico da Mesorregião Sudeste de Mato Grosso. Dissertação (Mestrado em Estudo de Linguagens) – Universidade Federal de Mato Grosso do Sul, Campo Grande-MS, 2009.
Atlas Linguístico do Estado do Ceará (ALECE)	BESSA, José Rogério Fontenele (coordenador). Atlas Linguístico do Ceará. Vol.I – Introdução, Vol.II – Cartogramas. Universidade Federal do Ceará. Fortaleza: Edições UFC, 2010.
Atlas Semântico-Lexical de Caraguatatuba, Ilhabela, São Sebastião e Ubatuba - municípios do Litoral Norte de São Paulo	ENCARNAÇÃO, Márcia Regina Teixeira da. Atlas semântico-lexical de Caraguatatuba, Ilhabela, São Sebastião e Ubatuba - municípios do Litoral Norte de São Paulo. 2010. 741f. Tese (Doutorado) – Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, São Paulo, 2010.
Atlas Geossociolinguístico de Londrina (AGeLO)	ROMANO, Valter Pereira. Atlas Geossolinguístico de Londrina: um estudo em tempo real e tempo aparente. 2012. 366f. Dissertação (Mestrado em Estudos da Linguagem) – Universidade Estadual de Londrina, Londrina, 2012.
Atlas Linguístico de Pernambuco (ALiPE)	SÁ, Edmilson José de. Atlas Linguístico de Pernambuco. Tese (Doutorado em Letras) – Universidade Federal da Paraíba. João Pessoa. 2013.

Atlas Linguístico Pluridimensional do Português Paulista níveis semântico-lexical e fonético-fonológico do vernáculo da região do Médio Tietê	FIGUEIREDO JUNIOR, Selmo Ribeiro. Atlas linguístico pluridimensional do português paulista: níveis semântico-lexical e fonético-fonológico do vernáculo da região do Médio Tietê. 2018. Tese de Doutorado. Universidade de São Paulo.
Atlas Linguístico do Amazonas (ALAM)	CRUZ-CARDOSO, Maria Luiza de C. Atlas Linguístico do Amazonas (ALAM). 2004. Tese de Doutorado. Tese de Doutorado em Letras Vernáculas da Faculdade de Letras Língua Portuguesa da Universidade Federal do Rio de Janeiro. Rio de Janeiro.
Mozilla Common Voice	ARDILA, Rosana et al. Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670, 2019.
Atlas linguístico do Amapá	RAZKY, Abdelhak et al. Atlas linguístico do Amapá. 1. ed. São Paulo: Labrador, 2019. E-book. Disponível em: https://plataforma.bvirtual.com.br.