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

Article

Potential uses of Brazilian Portuguese data repositories in Forensic Phonetics

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

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

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. 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.

[*] 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

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.

Article received: 10.02.2024
Article approved: 03.25.2025

1 Introduction

Forensic linguistics is an interdisciplinary field of scientific study and practice that applies linguistic knowledge to contexts where linguistic evidence (such as audio recordings, texts, and other materials) plays a relevant role in legal proceedings. The type of linguistic analysis most useful in each case depends on its nature. For example, some cases require the analysis of written texts to identify writing patterns that may suggest the authorship of disputed or apocryphal texts. In others, the focus is on analyzing audio recordings, whether to transcribe their content, outline the linguistic profile of a speaker, or compare an unknown voice with that of one or more suspects. A comprehensive overview of forensic linguistics is presented in Olsson (2008).

Within the broad field of forensic linguistics, forensic phonetics concerns itself with speech samples in legal contexts, as well as developing new knowledge and methods specifically for forensic phonetics (Jessen, 2008). One of the most common tasks in forensic phonetics is speaker comparison (henceforth SC), which aims to assess the probability that the same person produced (or did not produce) two speech samples (Gold; French, 2011, 2019). In an SC examination, linguistic and phonetic parameters extracted from speech samples of an unknown speaker (referred to as the questioned sample, or QS) are compared with parameters from samples of one or more suspects believed to have produced the QS (referred to as the reference samples, or RS). The challenge in SC tasks is to compare the degree of convergence and divergence between QS and RS—that is, their similarity—and to determine how common the observed speech parameters are within a given population, i.e., the typicality of what is observed in the RS.

Gold and French (2011, 2019) conducted two surveys on current practices among professionals in the field, revealing the most commonly used methods for expressing conclusions in SC reports. The authors report that one of the approaches that saw the greatest growth in adoption between the two surveys is based on calculating the so-called likelihood ratio (LR). The LR methodology estimates the probative value of the evidence evaluated in a SC examination—that is, the similarities and differences found between QS and RS—by comparing the probability of two competing hypotheses: one stating that QS and RS share the same origin (i.e., were produced by the same speaker) and the other stating that their origins are different (Morrison, 2009; 2010). The probability of the same-origin hypothesis is based on their similarity, while the different-origin hypothesis considers how typical the observed patterns in the samples are relative to their distribution in a reference population.

The underlying idea of applying the LR is that similarity between the compared samples is not sufficient to conclude that they share the same origin. If only similarity is considered, the SC examination could result in a false positive, since the similarity between QS and RS could, in theory, also be observed in comparisons between QS and samples produced by a potentially large number of other speakers who share the same linguistic and phonetic patterns identified in the examination. Therefore, it is also necessary to consider the typicality of the SC findings—that is, how common the linguistic and phonetic patterns extracted from the QS are in the relevant population for the case. If the patterns are very common in the population, the relative weight of the evidence supporting the same-origin hypothesis decreases. Conversely, if the typicality is low, the evidence supporting the same-origin hypothesis gains more weight.

To determine typicality, it is necessary to rely on external data sources that describe the distribution, in the relevant population, of the parameters observed in the examination—both linguistic (e.g., regional or idiosyncratic lexical use, morphosyntactic patterns) and phonetic-acoustic (e.g., fundamental frequency, vowel formant values). A common source of data for establishing such distributions are language and speech data repositories[1], such as speech corpora or linguistic atlases, whose linguistic products (e.g., audio or text samples, linguistic maps) can serve to extract prevalent linguistic and phonetic patterns in a population of interest. To illustrate more concretely how data repositories can contribute to forensic linguistics and phonetics, we present two examples below.

Maps from a linguistic atlas can indicate prevalent lexical variants in specific geographic areas. Thus, if speakers in a particular area commonly use the lexical

[1] We use the term "language and speech data repositories" in a broad sense, to include speech corpora and linguistic atlases, as well as publications derived from them, even though the products generated by these data sources may differ. The rationale is that both the raw data collected by a project and products derived from the data—such as the linguistic maps of an atlas, for instance—are potentially relevant for establishing the typicality of linguistic and phonetic patterns.

variant *"mormaço"* (sultry weather, in Brazilian Portuguese) and both the QS and RS (whose producers, based on independent evidence, can be assumed to come from the same area) contain one of the variants *"normaço"* or *"bormaço"*[2], for example, this similarity strengthens the same-origin hypothesis for QS and RS due to low typicality and high similarity. If both samples contain the prevalent variant, this high-typicality pattern would lend more support to the different-origin hypothesis.

High-quality audio samples from speech repositories can be used to generate statistical summaries of phonetic parameters, indicating their distribution in the speaker population. Population distributions of phonetic parameters allow for determining the degree of typicality of the value of an acoustic parameter measured in the RS. Suppose the focus of a particular SC examination is the mean fundamental frequency (*f0*) of the compared samples. Knowing, for example, that the mean *f0* in the Brazilian male population is 135 Hz[3] and the mean measured in the RS is 132 Hz, this result—characterizing a high-typicality scenario—strengthens the different-origin hypothesis, especially if there is also low similarity between the values found in the RS and QS. If the mean observed in the RS diverged more substantially from what is typical in the population, this result would weigh in favor of the same-origin hypothesis for RS and QS, assuming high similarity between their values, given the low probability of coincidence between the pattern found in the RS and speech samples randomly drawn from the reference population.

[2] The hypothetical example was inspired by Question 64 of the Semantic-Lexical Questionnaire (QSL) in the Atlas Linguístico-Etnográfico da Região Sul do Brasil (Linguistic-Ethnographic Atlas of Southern Brazil) – ALERS (Altenhofen; Klassmann, 2002, p. 144).

[3] This value was taken from Cunha (2023), who used speech samples from the Corpus Forense do Português Brasileiro (Forensic Corpus of Brazilian Portuguese) (Portaria No. 934-DITEC/PF, July 30th, 2020) to generate population statistics on f0 distribution for Brazilian male speakers. This is a concrete example of repurposing a speech repository to support forensic phonetics practice.

2 Justification and Objectives

There are many speech data repositories of Brazilian Portuguese compiled to support research in sociolinguistics, dialectology, corpus linguistics, and other fields of linguistics. Some of the most notable include *Programa de Estudos sobre o Uso da Língua* (PEUL), a pioneer in creating data repositories for sociolinguistic purposes (Freitag, 2016), *Projeto da Norma Urbana Oral Culta* (NURC), and *Projeto Variação Linguística na Região Sul do Brasil* (Varsul). As previously discussed, establishing the typicality of linguistic and phonetic patterns in a population is a necessary component for applying LR calculations in SC examinations, and these repositories can serve this purpose.

Two issues arise regarding the potential use of language and speech repositories as resources to support SC examinations, justifying the present work. The first is the discoverability or visibility of existing repositories, as there is currently no comprehensive catalog or inventory that lists and characterizes the existing Brazilian language and speech repositories. The second issue is that most repositories were not created to serve as resources for forensic linguistics in general or SC examinations in particular, so it is important to assess whether they can adequately serve this purpose.

This paper has two main objectives, related to the justifications presented above. First, to conduct a census and characterization of existing language and speech data repositories in the Brazilian context. Second, to systematically evaluate the potential of each repository for generating population distributions of linguistic and phonetic patterns for use in SC examinations.

Section 3 presents the methodology used for the survey and the criteria for evaluating each repository's potential. Section 4 presents the evaluation results. Based on these, Section 5 identifies the repositories with the greatest potential for use as data sources in the context of LR-based SC examinations.

3 Methodology

In the initial stage of the project, we surveyed the data repositories to be evaluated in this research. For this purpose, we contacted specialists in sociolinguistics and dialectology from all regions of Brazil, asking them to indicate language and speech data repositories that could contribute to the objectives of the present work.[4] In this phase, 26 repositories were suggested. In addition to consulting specialists, we also used search engines such as Google and Google Scholar, as well as institutional academic repositories. The search terms used were: *"banco de dados linguístico"* (linguistic database), *"dados linguísticos"* (linguistic data), *"amostras de fala"* (speech samples), *"amostras linguísticas"* (linguistic samples), *"corpus linguístico"* (linguistic corpus), *"sociolinguística"* (sociolinguistics), *"atlas linguístico"* (linguistic atlas), *"atlas sociolinguístico"* (sociolinguistic atlas), *"atlas fonético"* (phonetic atlas), *"atlas geolinguístico"* (geolinguistic atlas), *"atlas linguístico regional"* (regional linguistic atlas), and *"repositório linguístico"* (linguistic repository). In this stage, 19 repositories were found.

The complete list of repositories found at this stage, including bibliographic references describing them and URLs for digital access (where available), can be found in the document *"info-acervos.ods"*, available at: https://osf.io/avu9e/files/osfstorage/.

Repositories that did not align with the research purposes were excluded from the initial survey. Additionally, we disregarded repositories that do not provide access to the collected data, do not disclose relevant information about the data, or where access to the material is only possible through physical media such as CD-ROMs or magnetic tapes. The excluded repositories and the justification for their exclusion are listed in Table 1.

[4] Specialists were sourced from participants of Grupo de Trabalho de Sociolinguística da ANPOLL (ANPOLL Sociolinguistics Working Group) list (https://anpoll.org.br/gt/sociolinguistica) and conducted searches using keywords such as "sociolinguistics," "dialectology," and "databases." A total of 47 researchers, active in all regions of Brazil, were contacted.

Table 1 – Excluded Data Repositories and Exclusion Criteria.

Excluded Data Repositories	Exclusion Criteria
NURC-SP audio samples under the custody of the "Alexandre Eulalio" Cultural Documentation Center, CEDAE/Unicamp	Inability to access the samples.
Pirá: A Bilingual Portuguese-English Dataset for Question-Answering about the Ocean	Lack of essential information for speaker comparison tasks, such as geographic coverage and gender/age classification of participants.
Corpus de Textos Orais do Português Santareno (CTOPS) (Corpus of Oral Texts from Santarém Portuguese)	Lack of general repository information (e.g., types of collected materials, participant characteristics) and physical-only publication (CD-ROM).
Aspectos linguísticos da fala londrinense: esboço de um atlas linguístico de Londrina (Linguistic Aspects of Londrina Speech: Draft of a Linguistic Atlas of Londrina)	Lack of general repository information (e.g., types of collected materials, participant characteristics)

Source: prepared by the authors.

After exclusion, 45 repositories remained. Table 2 presents the selected repositories grouped by Brazilian geographic region.

Table 2 – List of Surveyed Repositories by Brazilian Geographic Region.

Region	Language and Speech Data Repositories
South Region	1. *Atlas Linguístico-Etnográfico da Região Sul do Brasil (ALERS)* (Linguistic-Ethnographic Atlas of Southern Brazil)
	2. *LínguaPOA* (LanguagePOA)
	3. *Projeto Variação Linguística na Região Sul do Brasil (VARSUL)* (Linguistic Variation Project in Southern Brazil)
	4. *Atlas Linguístico do Paraná (ALPR)* (Linguistic Atlas of Paraná)
	5. *Atlas Geossociolinguístico de Londrina (AGeLO)* (Geosociolinguistic Atlas of Londrina)
Southeast Region	1. *ALIP - Iboruna (Amostra Censo)* (ALIP - Iboruna [Census Sample])
	2. *NURC RJ* (NURC Rio de Janeiro)
	3. *NURC SP* (NURC São Paulo)
	4. *Programa de Estudos sobre o Uso da Língua (PEUL)* (Language Use Studies Program)

	5. *Projeto SP2010* (SP2010 Project)
	6. C-ORAL-BRASIL (I)
	7. CORAA NURC-SP Minimal Corpus
	8. *Esboço de um Atlas Linguístico de Minas Gerais (EALMG)* (Draft of a Linguistic Atlas of Minas Gerais)
	9. *Atlas Semântico-Lexical da Região do Grande ABC* (Semantic-Lexical Atlas of the Greater ABC Region)
	10. *Atlas Semântico-Lexical de Caraguatatuba, Ilhabela, São Sebastião e Ubatuba* (Semantic-Lexical Atlas of Caraguatatuba, Ilhabela, São Sebastião, and Ubatuba)
	11. *Atlas Linguístico Pluridimensional do Português Paulista* (Multidimensional Linguistic Atlas of Paulista Portuguese)
Northeast Region	1. NURC Digital/Recife
	2. *Norma Oral do Português Popular de Fortaleza (NORPOFOR)* (Oral Norm of Popular Portuguese in Fortaleza)
	3. *PORTAL - Variação linguística no português alagoano* (PORTAL - Linguistic Variation in Alagoas Portuguese)
	4. *Projeto Variação Linguística no Estado da Paraíba (VALPB)* (Linguistic Variation Project in Paraíba)
	5. *Banco de dados falares sergipanos* (Database of Sergipe Speech Varieties)
	6. *A língua portuguesa do semiárido baiano* (The Portuguese Language of the Bahian Semi-Arid Region)
	7. *Programa de Estudos sobre o Português Popular de Salvador (PEPP)* (Study Program on Popular Portuguese in Salvador)
	8. *Estudos da Língua Oral do Cariri* (Studies of Cariri Oral Language)
	9. *Dialetos Sociais Cearenses* (Social Dialects of Ceará)
	10. *Português Oral Culto de Fortaleza (PORCUFORT)* (Cultured Oral Portuguese of Fortaleza)
	11. *Atlas Linguístico da Paraíba (ALPB)* (Linguistic Atlas of Paraíba)
	12. *Atlas Linguístico de Sergipe (ALS)* (Linguistic Atlas of Sergipe)
	13. *Atlas Linguístico de Sergipe II (ALS II)* (Linguistic Atlas of Sergipe II)
	14. *Atlas Linguístico da Mata Sul de Pernambuco (ALMASPE)* (Linguistic Atlas of Southern Pernambuco Forest Zone)

	15. *Atlas Linguístico do Estado do Ceará (ALECE)* (Linguistic Atlas of Ceará)
	16. *Atlas Linguístico de Pernambuco (ALiPE)* (Linguistic Atlas of Pernambuco)
Central-West Region	1. *Atlas Linguístico de Mato Grosso do Sul (ALMS)* (Linguistic Atlas of Mato Grosso do Sul)
	2. *Atlas Linguístico da Mesorregião Sudeste de Mato Grosso (ALMESEMT)* (Linguistic Atlas of Southeastern Mato Grosso)
North Region	1. *Atlas Linguístico Sonoro do Pará (ALISPA)* (Sound Linguistic Atlas of Pará)
	2. *Atlas Geolingüístico do Litoral Potiguar (ALiPTG)* (Geolinguistic Atlas of the Potiguar Coast)
	3. *Atlas Linguístico do Amazonas (ALAM)* (Linguistic Atlas of Amazonas)
	4. *Atlas linguístico do Amapá* (Linguistic Atlas of Amapá)
Multi-Regional[5]	1. *ALIP - Iboruna (Amostra de Interação)* (ALIP - Iboruna [Interaction Sample])
	2. *Projeto Atlas Linguístico do Brasil (Projeto ALiB)* (Linguistic Atlas of Brazil Project)
	3. *Discurso & Gramática* (Discourse & Grammar)
	4. *BrasilData* (BrazilData)
	5. *Corpus Forense do Português Brasileiro (CFPB)* (Forensic Corpus of Brazilian Portuguese)
	6. *Atlas Prévio dos Falares Baianos (APFB)* (Preliminary Atlas of Bahian Speech Varieties)
	7. Mozilla Common Voice

Source: prepared by the authors.

Once the list of identified repositories was compiled, we began evaluating their applicability in forensic phonetics tasks. The evaluation was based on a questionnaire designed to identify the linguistic information that can be extracted from each repository, particularly those may be useful for supporting tasks in forensic linguistics

[5] The "multi-regional" category refers to repositories covering more than one Brazilian region, such as the Atlas Prévio dos Falares Baianos (Preliminary Atlas of Bahian Speech Varieties, APFB), for example, which includes areas adjacent to Bahia, encompassing cities in the states of Sergipe, northern Minas Gerais, eastern Goiás, and present-day Tocantins, thus spanning the North, Northeast, Southeast, and Central-West regions of Brazil.

and phonetics, especially SC. The questionnaire can be seen in the document *"questionario.pdf"* file, available at https://osf.io/avu9e/files/osfstorage/. The questions are organized into six main sections, as listed below.

General characteristics of the repository, such as the project name, types of materials collected and provided, collection period, number of samples, and conditions for accessing and using the collected material.

Sociodemographic characteristics of the participants.

Geographic regions covered by the repository.

Repository characteristics, such as the conditions under which the samples were collected and whether there is a field diary record, for example.

Information about audio samples (considered only for repositories providing access to audio materials), such as the perceptual quality of the audio, file formats, the type of communicative/interactional situation (e.g., dialogue, monologue, directed content), how sample duration was recorded, and vocal characteristics of participants (e.g., use of orthodontic appliances, health conditions, surgical history).

Details about transcriptions of the collected samples, if available: information about the transcribed material, the type of transcription provided (phonetic, orthographic, orthographic with conversational markers, etc.), and the transcription system used.

4 Results and Discussion

The results of the analysis are presented here, following the sections of the questionnaire used for their evaluation. It is important to note that the information presented on the vertical axis (*y*-axis) of the bar graphs in this section is the raw number of repositories in each response category, and the number inside each bar is the percentage this value represents relative to the total number of repositories evaluated in that analysis. Due to space limitations, we present here a selection of the results. Responses to all questionnaire items for all analyzed repositories can be found in the document *"info-acervos.ods"*, available at: https://osf.io/avu9e/files/osfstorage/.

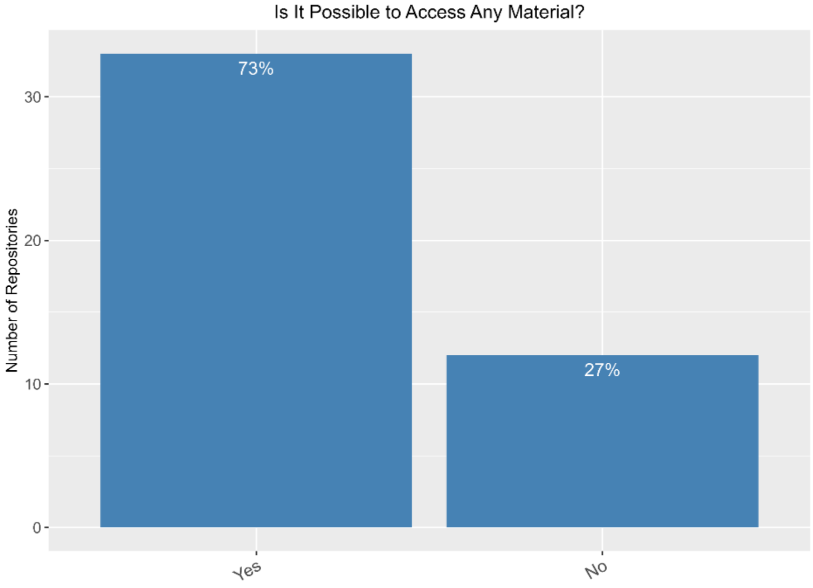
4.1 General Characteristics

This section presents general information about the repositories, focusing on essential data relevant to their repurposing as resources for SC tasks. The results of the repository analysis regarding general aspects, collection and availability of materials, and the period during which they were collected and accessed are presented in the form of graphs.

4.1.1 Collection and Availability of Materials

All analyzed repositories collected language or speech materials, but not all make them publicly available. Availability of materials is shown in Graph 1. We classify linguistic material within repositories in two categories: collected material and material that is made available to the public. This distinction will be explained below using examples.

Graph 1 – Access to Repository Materials.

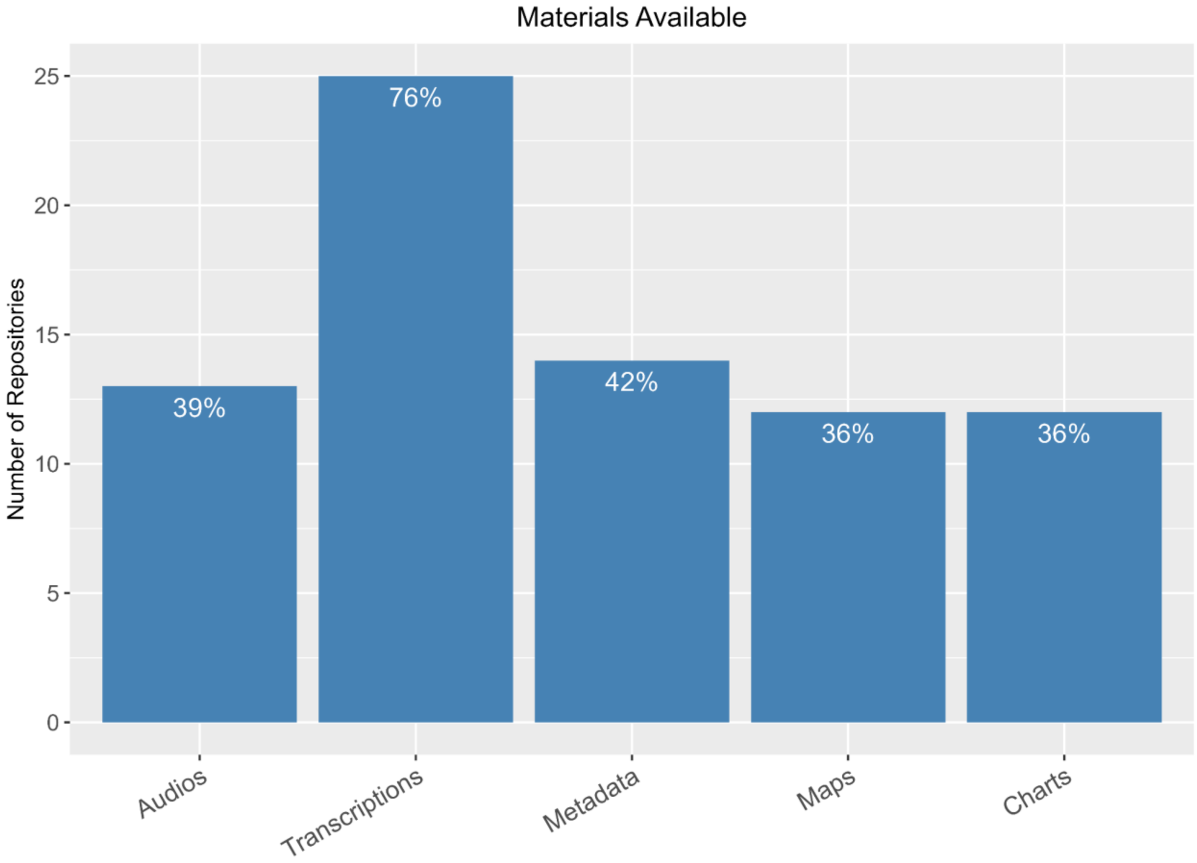


Source: prepared by the authors.

Of the 45 analyzed repositories, 73% (33) provide full or partial access to collected material. In the 12 cases where the material is unavailable, some offer only descriptions of the data collected in the form of journal papers or other academic publications, but provide no information about how someone may have access to the materials present in the repository. In other cases, the repository websites are under maintenance, and the samples are temporarily unavailable.

After identifying how many repositories provide access to materials, we look at what type of material each repository makes available. This information is shown in Graph 2. The raw values and percentages in Graph 2 are based on the 33 repositories that provide materials.

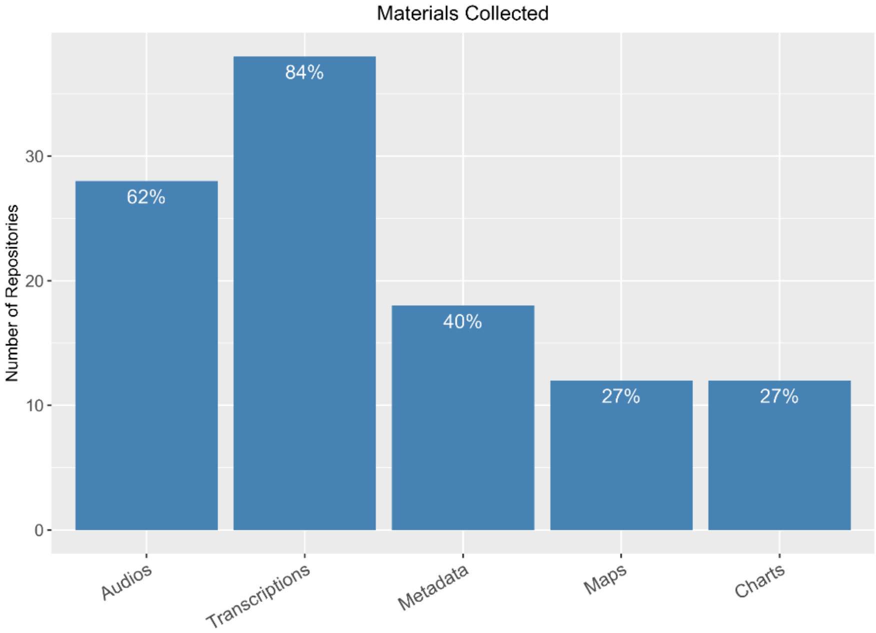
Graph 2 – Types of Materials Available in Repositories.



Source: prepared by the authors.

Graph 2 shows the types of materials provided by the repositories, while Graph 3 highlights the difference between the total number of repositories that collected each type of material and those that provide access to them. The raw values and percentages in Graph 3 are based on the analysis of all 45 repositories evaluated in this work.

Graph 3 – Types of Materials Collected.



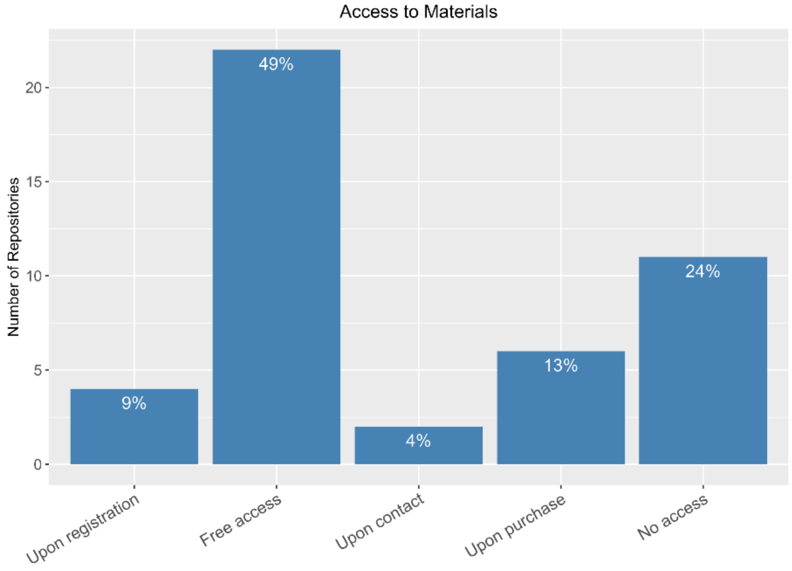
Source: prepared by the authors.

It is possible to observe that, except for maps and graphs, the number of provided materials decreased compared to the collected materials. This indicates that many of the repositories collected language and speech materials but did not aim to make them publicly available. In many cases, the organizers of repositories that do not provide the collected materials are not explicit about why the data collected and/or products derived from the data is not shared with the public. We can assume this may be due to legal restrictions related to privacy that may prevent public release, or that some repositories were collected to support specific studies without plans for public dissemination. Finally, some repositories may aim to release transcriptions of collected speech materials without releasing the audio files themselves. For forensic phonetics purposes, it is preferable for repositories to provide access to both audio materials and transcriptions. As explained in section 3, audio samples and transcriptions can be useful as sources of information about typical linguistic and phonetic patterns of the language.

4.1.2 Access to Materials and Terms of Use

The conditions for accessing repository materials vary. Graph 4 illustrates the different ways materials in repositories may be accessed.

Graph 4 – Procedures for Accessing Materials in Repositories.



Source: prepared by the authors.

Graph 4 shows that most surveyed data repositories make their samples freely available, either on their websites or through downloadable files. For repositories that do not provide access, the reasons vary, as discussed below.

Some repositories collected and stored their data exclusively in physical media (e.g., CDs, floppy disks, magnetic tapes) and have not digitized them, preventing their sharing through the internet. In other cases, repositories collected materials solely to support specific research projects, with no intention of public release. Examples include the *Projeto Descrição do Português Oral Culto de Fortaleza (PORCUFORT)* and the *Projeto A língua portuguesa falada no semiárido baiano*.

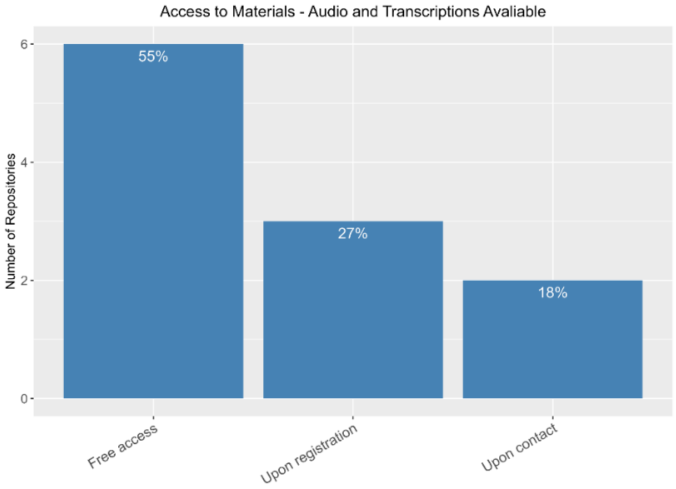
Repositories that provide materials upon contacting the responsible parties include *Projeto SP2010* and *LínguaPOA*. Payment is required to access materials from the following repositories: *Projeto Atlas Linguístico do Brasil (Projeto ALiB)*, *Atlas Prévio dos Falares Baianos (APFB)*, *Esboço de um Atlas Linguístico de Minas Gerais (EALMG)*, *Atlas Linguístico de Mato Grosso do Sul (ALMS)*, *Atlas Linguístico do Estado do Ceará (ALECE)*, and *Atlas linguístico do Amapá*.

There are also cases where access is granted only through registration, either via email sign-up (e.g., the two repositories under the umbrella of *Projeto Amostra Linguística do Interior Paulista (ALIP)*) or immediate access after registration (e.g., *C-ORAL-BRASIL*).

Another notable case is the *Corpus Forense do Português Brasileiro (CFPB)*, specifically designed for forensic purposes by *Instituto Nacional de Criminalística (INC)*, a forensics body connected to the Brazilian Federal Police. Researchers from universities can ask for permission to use the materials in this repository to conduct projects aligned with *INC*'s interests. The repository's description, purpose, and sharing conditions are outlined in *Portaria Nº 934-DITEC/PF*, dated July 30, 2020.

Given the importance of accessing speech samples for forensic phonetics, Graph 5 presents information about the access procedure exclusively for the 11 repositories that provide both speech samples and transcriptions.

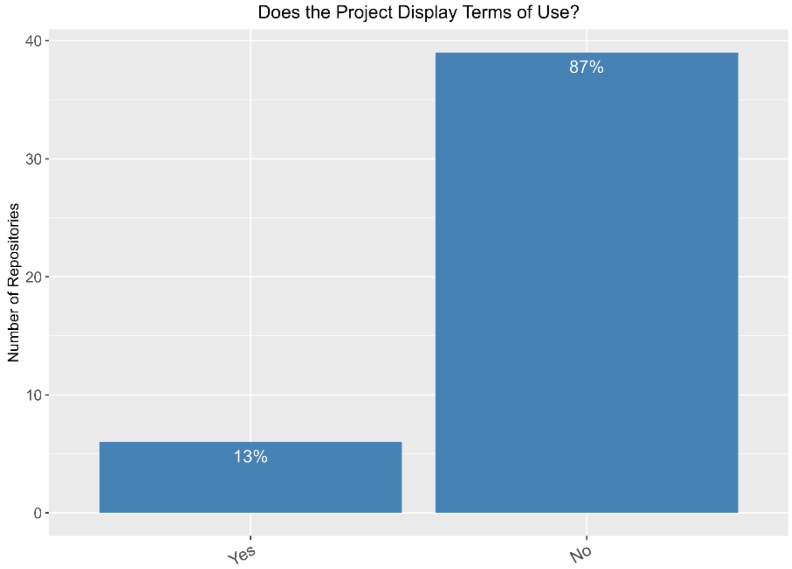
Graph 5 – Access Procedure for Repositories Providing Audio and Transcription.



Source: prepared by the authors.

Finally, Graph 6 shows the presence or absence of terms of use for repositories that provide access to materials in some form.

Graph 6 – Presence of Terms of Use.



Source: prepared by the authors.

The low number of Brazilian language and speech data repositories with terms of use is noteworthy. The repositories with terms of use identified in the present research all date from the 2000s onward, a period marked by growing concerns about ethical issues related to the collection and dissemination of research data involving human subjects.

This shift can be partly explained by the enactment of legislation in 2016 requiring researchers to provide terms of use for data collected in human and social sciences research, per *Resolução nº 510* (Brasil, 2016) of the *Conselho Nacional de Saúde*, as well as *Lei Geral de Proteção de Dados Pessoais (LGPD)* in 2018, which mandates that data collection in human sciences adhere to essential legal principles, such as specific purpose, transparency, necessity, security, anonymization, and consent.

Lastly, it is worth noting that all repositories listed in the "Yes" column of Graph 6 provide access to audio materials.

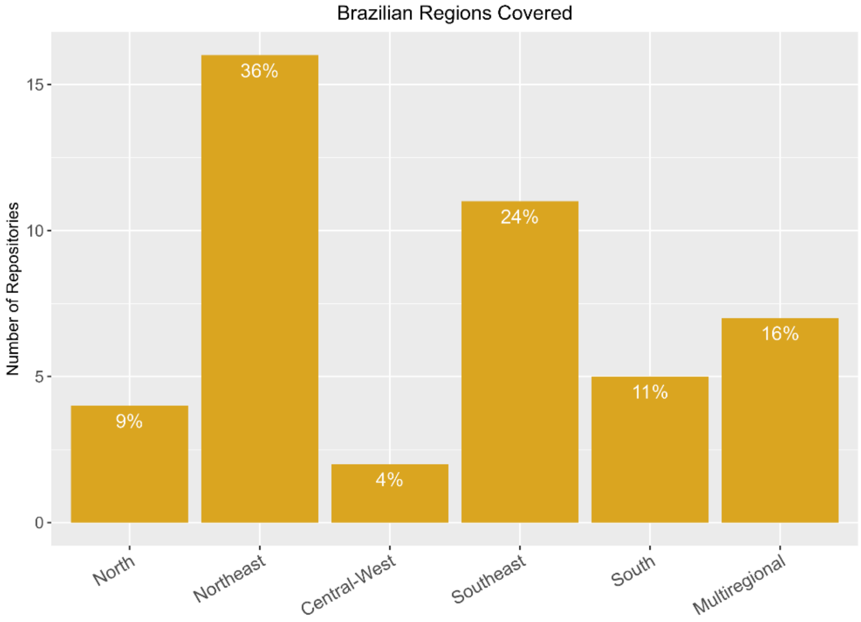
4.2. Coverage and Sociolinguistic Characteristics

This section addresses aspects related to the participants whose data were collected by the surveyed language and speech repositories. These include information about collection regions, states covered, and other details of the language and speech samples.

4.2.1. Brazilian Regions Covered

All Brazilian regions (according to the IBGE[6]'s regional classification, which divides the country into five major regions – North, Northeast, Central-West, Southeast, and South) are covered by the repositories surveyed in this research, though the distribution of repositories per region is not uniform. Some repositories span multiple regions, meaning they gathered data from locations in more than one region of the country. Graph 7 represents the regional coverage of the surveyed data repositories

Graph 7 – Regional coverage.



Source: prepared by the authors.

[6] Instituto Brasileiro de Geografia e Estatística (Brazilian Institute of Geography and Statistics).

Graph 7 shows that the Northeast and Southeast regions have the highest coverage in terms of repositories. In the Northeast, most repositories are linguistic atlases. For forensic purposes, the use of such repositories—especially for generating population statistics of phonetic-acoustic features—is more limited, as the atlases in our survey do not directly provide audio materials, as their primary purpose is to generate maps and graphs showing the spatial variation of linguistic phenomena, without the intention of providing complete transcriptions or audio recordings. However, even without audio access, linguistic atlases can be useful in SC tasks for mapping socio-phonetic evidence relevant to determining typicality, as proposed by Gonçalves and Brescancini (2020). Such proposals justify the inclusion of atlases in this work.

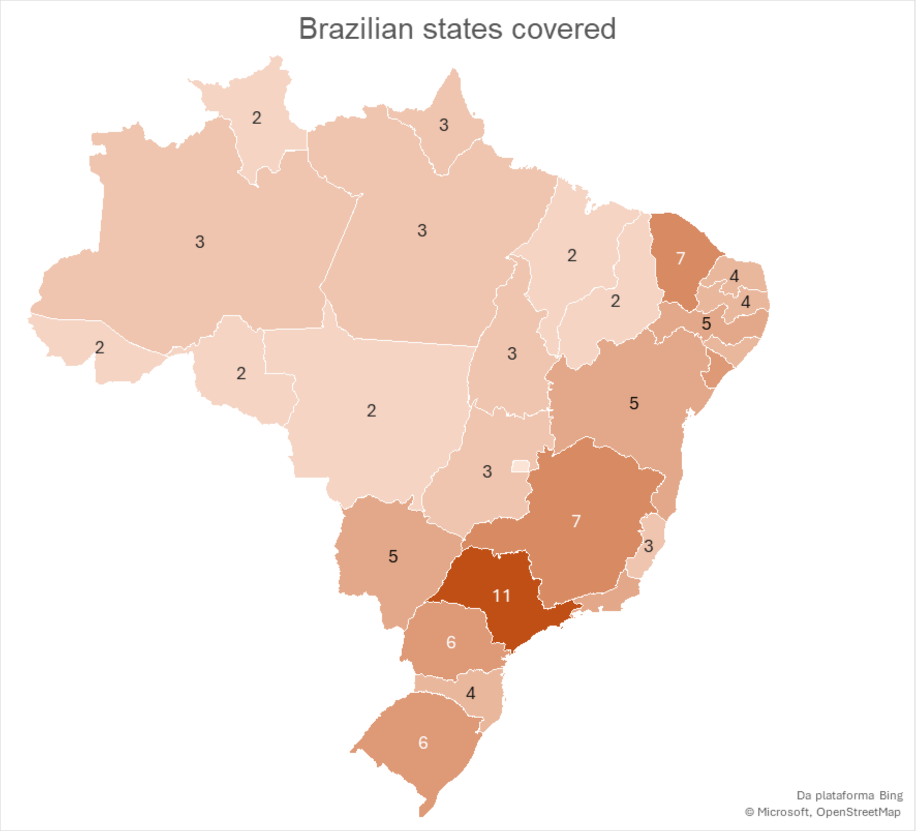
Among the multi-regional repositories mentioned, few aimed to collect samples nationwide: these are *Projeto Atlas Linguístico do Brasil (ALiB)*, *BrasilData*, and the *Corpus Forense do Português Brasileiro (CFPB)*. Other repositories collected materials in border areas between regions without aiming for nationwide coverage, such as *Atlas Prévio dos Falares Baianos (APFB)* and the *ALIP-Iboruna* project.

4.2.2. Brazilian States Covered

The map of Brazil in Figure 1 illustrates the distribution of language and speech data repositories for each state.

Figure 1 shows the distribution of repositories that collected material in each Brazilian state. Six states have the highest concentrations of repositories: São Paulo, Minas Gerais, Ceará, Sergipe, Paraná, and Rio Grande do Sul. The research revealed the notable absence of relevant repositories representing the Federal District, a metropolitan region currently home to around 3 million inhabitants surrounding the federal capital.

Figure 1 – Brazilian states covered.



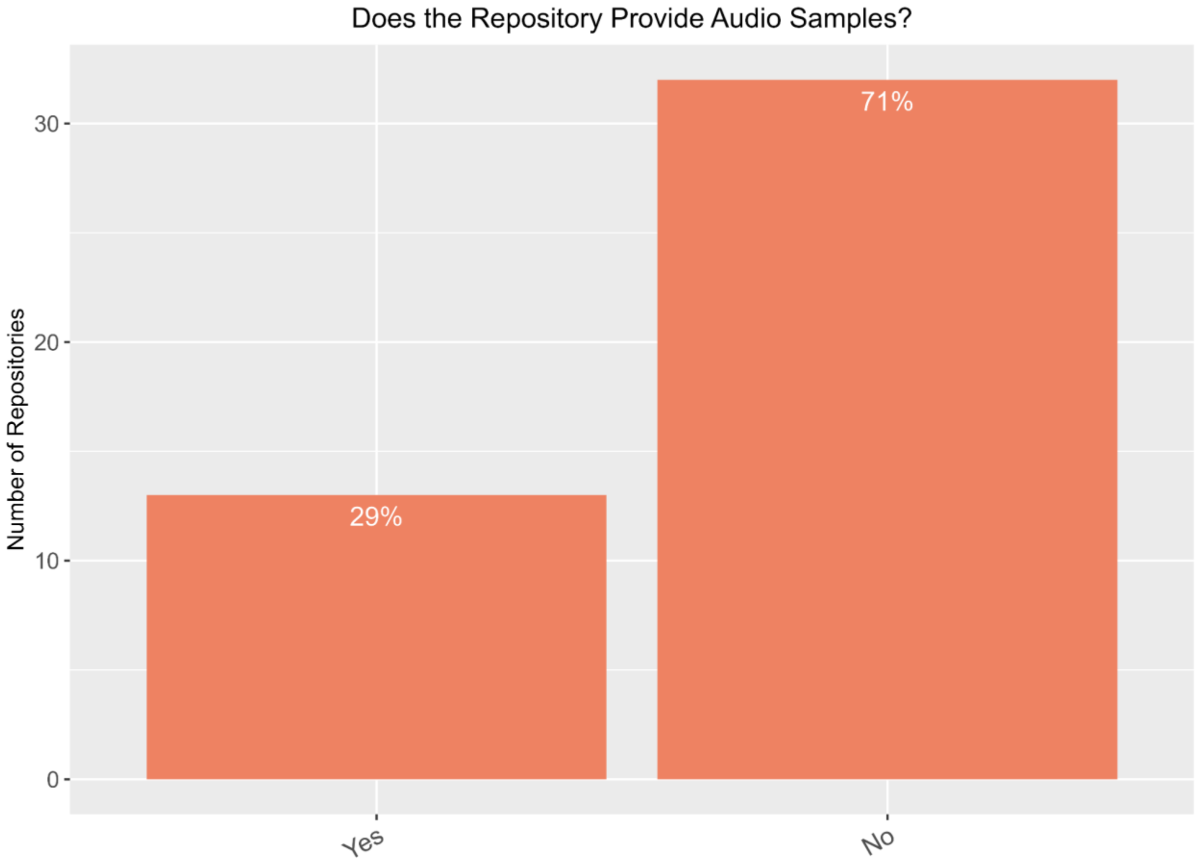
Source: prepared by the authors.

4.3. Audio

The *Audio* section of the questionnaire evaluates general characteristics of the audio samples provided by the repositories. First, it is important to visualize how many of the surveyed repositories provide audio materials, as shown in Graph 8.

We can see that 13 repositories, or 29% of the total, provide audio samples. This classification applies to audio samples that can be accessed—i.e., that are available at the time of this study by a researcher not connected with the repository. Thus, the "Yes" response does not include repositories under maintenance or temporarily unable to provide the data to interested parties.

Graph 8 – Availability of Audio Samples.

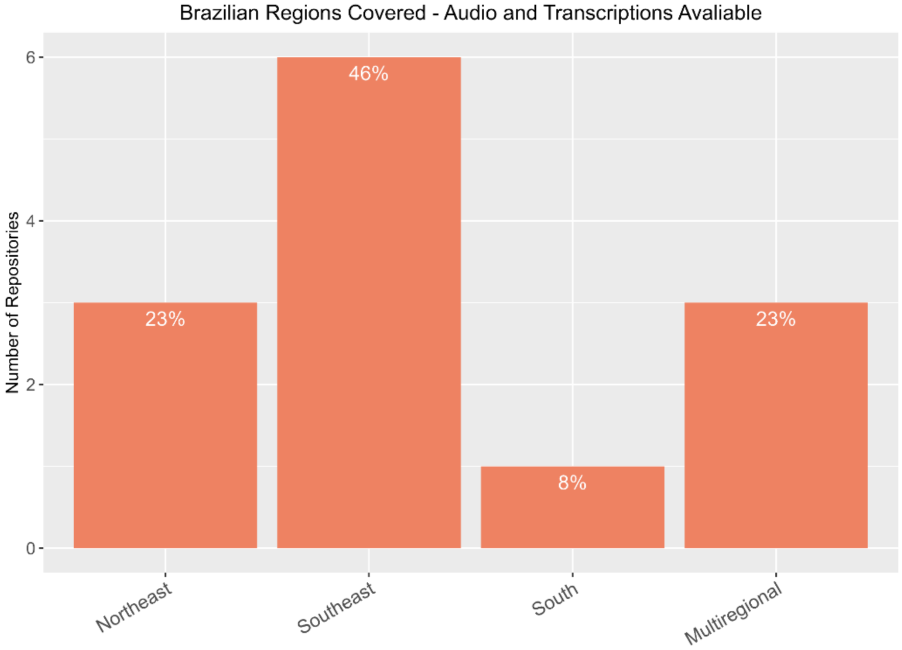


Source: Prepared by the authors.

The fact that only 29% of repositories provide audio samples highlights a limitation in their use for generating population statistics of phonetic-acoustic parameters, given that acoustic analyses are crucial for producing statistics about linguistic characteristics of interest for determining the typicality of patterns observed in the analyzed samples in SC examinations. Knowing how many repositories provide access to audio samples, it is relevant to understand their distribution across the country's regions, as shown in Graph 9.

The comparison between Graph 9 and Graph 7 is significant, as both deal with information about the regional coverage of repositories that provide their samples. Notably, most repositories providing audio samples are concentrated in the Southeast, whereas, considering the total number of language and speech repositories surveyed in this work, there is an abundance of repositories in the Northeast, explained by the large number of regional linguistic atlases produced there.

Graph 9 – Regions Covered by Repositories Providing Audio.

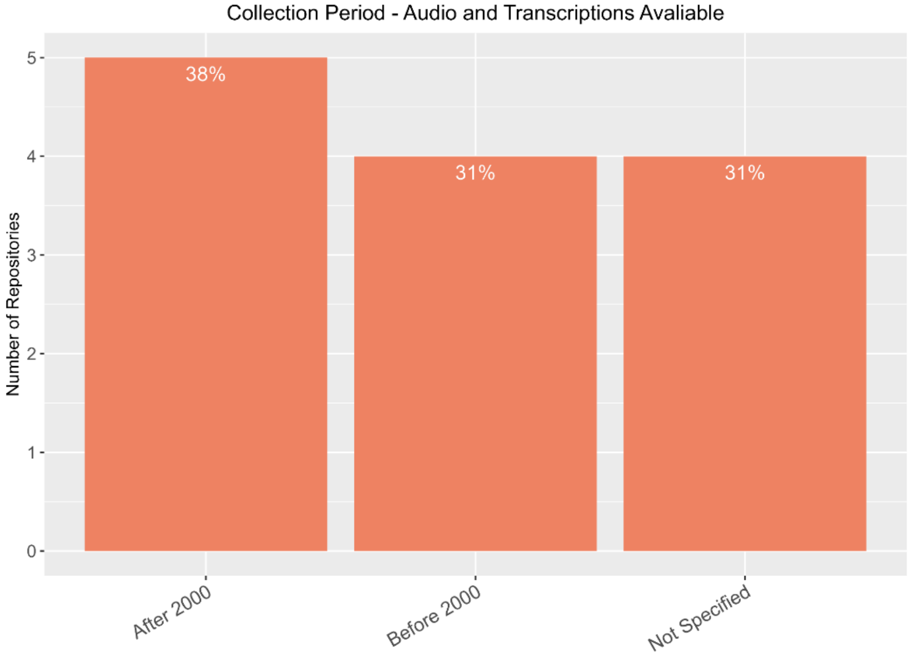


Source: prepared by the authors.

The repositories in the "Multi-regional" category are *ALIP - Iboruna (Amostra de Interação)*, covering cities in the northwest of São Paulo and a small part of Mato Grosso; the *Corpus Forense do Português Brasileiro (CFPB)*, with nationwide coverage; and *Mozilla Common Voice*, which collects data from across Brazil. Thus, no repositories providing audio samples from speakers in the Central-West and North regions were identified, limiting the regional coverage of available data that can be used to support the compilation of population distributions of linguistic parameters.

Graph 10 presents the time frame in which the repositories were collected, distinguishing between those gathered before and after the year 2000.

Graph 10 – Collection Period of Repositories Providing Audio and Transcriptions.



Source: Prepared by the authors.

Graph 10 shows that 5 repositories collected samples from 2000 onward, meaning they are more recent, which suggests that the acoustic quality of the speech samples was not compromised by the degradation of audio recorded on non-digital media, such as cassette tapes.

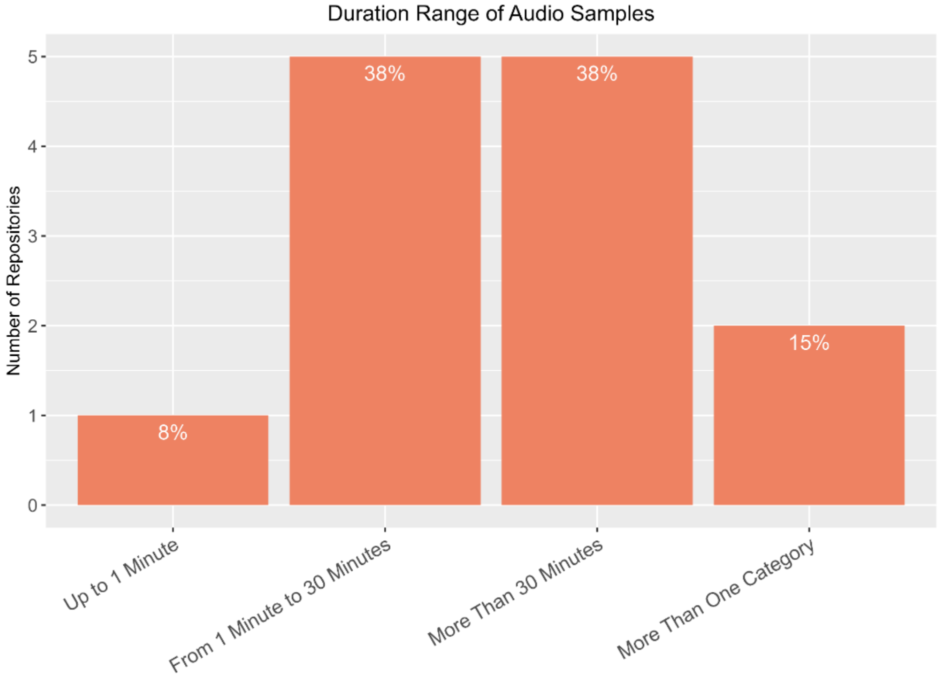
The preference for repositories with acoustic samples collected from the 2000s onward in SC tasks reflects generational variation (Eckert, 1997), as language use is shaped by the historical and sociocultural context in which the data were recorded.

The 5 repositories that collected samples after 2000 are *ALIP - Iboruna (Amostra Censo)*, *ALIP - Iboruna (Amostra de Interação)*, *Projeto SP2010*, *C-ORAL-BRASIL (I)*, and *LínguaPOA*.

4.3.1. Duration Range of Audio Samples

Graph 11 shows the duration range of samples provided by the surveyed repositories.

Graph 11 – Duration Range of Audio Samples.



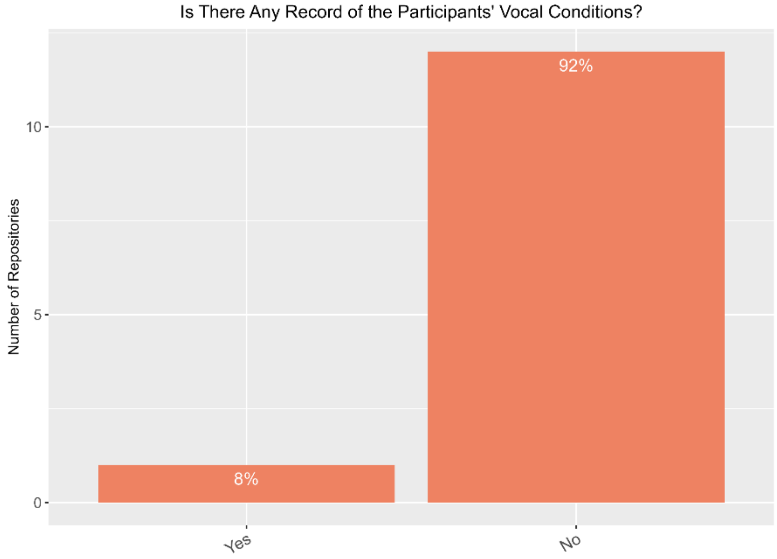
Source: prepared by the authors.

Graph 11 shows that most repositories offer audio samples longer than 1 minute or even 30 minutes. This is crucial for SC tasks, as very short samples limit the extraction of acoustic parameters that capture long-term speech patterns. Longer recordings are therefore preferable.

4.3.2. Recording of Vocal Conditions

In forensic acoustic analysis, it may be relevant to have information about factors that can systematically affect different components of the speech production system and alter participants' vocal production, such as a history of respiratory diseases, vocal tract surgery, smoking, or the use of orthodontic appliances, among others. Graph 12 shows that, of the 13 repositories providing access to audio materials, only one documents the vocal conditions of participants, indicating that this is not typically a concern in the design of the repositories in our sample. The repository in question is the *Corpus Forense do Português Brasileiro (CFPB)*, which was specifically designed to support forensic phonetics.

Graph 12 – Number of Repositories Documenting Conditions Affecting Participants' Vocal Production.

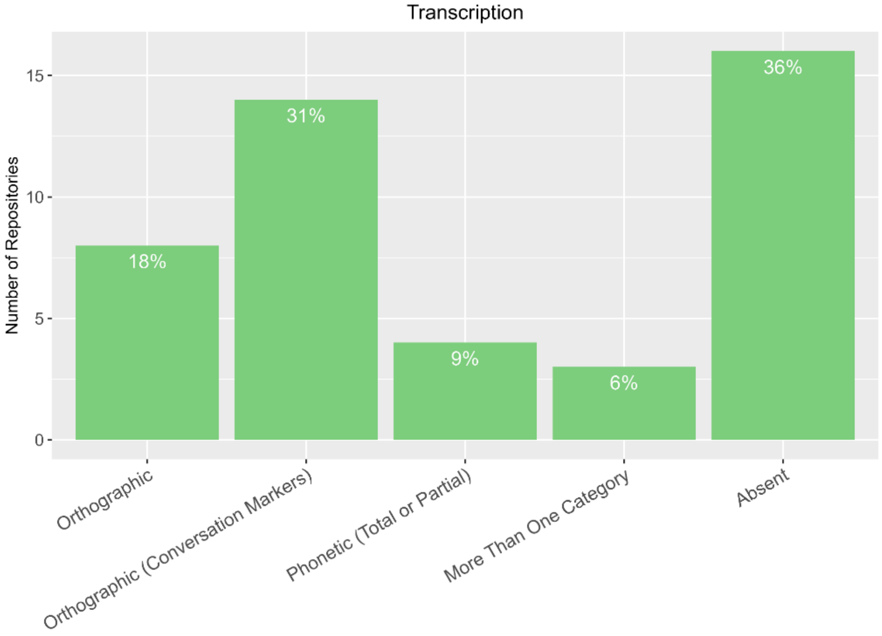


Source: prepared by the authors.

4.4. Transcription

Here we analyze the types of transcriptions collected and/or provided by the 45 surveyed repositories. Graph 13 presents the distribution of this information. Below, we explain the classification of transcription types we adopted.

Graph 13 – Type of Transcription Available in Repositories.



Source: prepared by the authors.

In orthographic transcriptions, the linguistic content of a speech sample is represented following the orthographic norms of Brazilian Portuguese. That is, the words in the collected samples are transcribed as they appear in dictionaries, without recording possible systematic or idiosyncratic variations in their pronunciation by the speaker.

In orthographic transcriptions with conversational markers, in addition to the linguistic content itself, special graphic conventions are used to indicate elements of speech organization, such as pauses, emphasis, interrupted or incompletely pronounced words, and other conversational features.

In phonetic transcriptions, the specific sounds used by the sample producer are recorded using conventional symbols, such as those of the International Phonetic Alphabet. Phonetic transcriptions may be complete (all words are transcribed) or partial (only some words or even isolated sounds within them are transcribed phonetically). Partial transcriptions mainly appear in linguistic atlases, where the goal is to present phonetic variations in the pronunciation of specific words using phonetic symbols that record the sounds produced by the interviewed speakers. For example, possible phonetic variations of the word *″sete″* could be transcribed as [ˈsɛtɪ] or [ˈsɛt͡ʃĩ]. If the goal is to highlight a specific sound within the word, the phonetic transcription may focus only on the phoneme in question, as in *″se[t]e″* and *″se[t͡ʃ]e″*.

It is worth noting that, although phonetic transcription is the least common among the surveyed repositories, it is the most relevant for forensic phonetics. This is because phonetic transcription enables the precise identification of systematic allophonic variations in the pronunciation of certain phonemes in the language. This level of detail allows for meticulous analysis of acoustic and articulatory characteristics that may be essential in forensic contexts.

5. Key Repositories of Interest for Speaker Comparison Tasks

Based on the results presented in the previous sections, Table 3 compiles the repositories that, given the criteria present in the questionnaire, can be considered most suited as good resources of data in the context of forensic phonetics, particularly for generating population statistics of phonetic and acoustic-phonetic parameters in LR-based SC examinations. With this in mind, we prioritized repositories that collected their samples from the 2000s onward and/or provide audio samples with good acoustic quality. The criterion of relative contemporaneity of the samples is relevant because repositories with data over three decades old may no longer reflect the synchronic state of the language due to various factors that can cause linguistic change, both internal and external to the system (Passetti *et al*., 2024). Good acoustic quality of the samples is important for extracting acoustic parameters, many of which are sensitive to the recording conditions of the acoustic signal, such as determining vowel formant values and the observation of specific phone classes, like fricatives (Barbosa *et al*., 2020).

Table 3 – Key Repositories of Interest for Speaker Comparison Tasks in Forensic Phonetics.

Repository	ALIP - Iboruna (Amostra Censo) (ALIP - Iboruna [Census Sample])	ALIP - Iboruna (Amostra de Interação) (ALIP - Iboruna [Interaction Sample])	Projeto SP2010 (SP2010 Project)	C-ORAL-BRASIL (I)	PORTAL - Variação linguística no português alagoano (PORTAL - Linguistic Variation in Alagoas Portuguese)	LínguaPOA (Language POA)	Corpus Forense do Português Brasileiro (CFPB) (Forensic Corpus of Brazilian Portuguese)
Collection Period	2004-2007	2005-2006	2011-2013	2006-2011	Not specified	2015-2019	Not specified
Access Procedure	Registration required	Registration required	Contact responsible parties	Registration required	Open access	Contact responsible parties	Project agreement with Federal Police
Has Terms of Use?	Yes	Yes	No	Yes	No	No	Yes
Terms Allow Forensic Use?	Yes	Yes	Not specified	Yes	Not specified	Not specified	Yes
Covered Brazilian Regions	Southeast	Multi-regional	Southeast	Southeast	Northeast	South	Multi-regional
Collection Metadata Available?	Yes	Yes	Yes	Yes	Yes	No	No
Metadata Scope	By sample	By sample	By sample	By sample	General info	Not specified	Not specified

Audio File Format	.mp3	.mp3	.wav/.mp3	.wav	.wav	.mp3	.wav
Speech Style	Directed content	Free content	Directed content	Directed content	Directed/free content	Directed content	Word/phrase reading + directed content
Sample Duration Range	1–30 min	1–30 min	>30 min	1–30 min	1–30 min	>30 min	1–30 min
Vocal Condition Records?	No	No	No	No	No	No	Yes
Transcriptions Available?	Yes	Yes	Yes	Yes	Yes	Yes	No
Transcription Type	Orthographic with conversational markers	Orthographic with conversational markers	Orthographic with conversational markers	Orthographic with conversational markers	Orthographic	Orthographic	N/A

Source: prepared by the authors.

6 Conclusion

This paper surveys Brazilian Portuguese language and speech data repositories, assessing their suitability for forensic phonetic applications—particularly as sources of population statistics for linguistic and phonetic parameters in likelihood-ratio-based speaker comparison (SC).

Our findings reveal a limited number of repositories in Brazil that can effectively support SC tasks. Of the 45 repositories examined, only seven provide adequate data for generating population statistics on linguistically relevant features, which are crucial for evaluating the typicality of speech patterns.

Regarding regional representation, our results highlight gaps in coverage, particularly in the Central-West and North regions, where no suitable repositories were found. Only one repository per region was identified for the South and Northeast, underscoring the need for broader data accessibility. We recommend that future projects include terms of use permitting the broader dissemination and repurposing of audio samples beyond their original scope.

While the surveyed repositories successfully document lexical, syntactic, and conversational variation, their utility for forensic phonetic remains relatively limited, given the specific needs of SC examinations. Future research will examine in greater detail the phonetic-acoustic parameters that can be extracted from available audio samples.

We hope this work aids forensic linguists, phoneticians, and other language researchers, encouraging the development of new repositories that support population-level linguistic statistics in Brazilian Portuguese.

Translation by the authors.

Acknowledgments

The first author thanks the *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)* (PIBIC scholarship, process no. 142760/2023-4). The second author thanks

CAPES (Postdoctoral scholarship, process CAPES 88887.804443/2023-00). The authors jointly thank Professor Lívia Oushiro (UNICAMP) for her valuable comments and suggestions on a preliminary version of the paper and the journal's anonymous reviewers for their feedback.

References

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