UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS

JOSE MANOEL SIQUEIRA DA SILVA

COVARIAÇÃO MORFOSSINTÁTICA NO PORTUGUÊS BRASILEIRO: IDENTIFICAÇÃO DIALETAL DE ESTUDANTES DA UNIVERSIDADE FEDERAL DE SERGIPE

SÃO CRISTÓVÃO – SE
2025

JOSE MANOEL SIQUEIRA DA SILVA

COVARIAÇÃO MORFOSSINTÁTICA NO PORTUGUÊS BRASILEIRO: IDENTIFICAÇÃO DIALETAL DE ESTUDANTES DA UNIVERSIDADE FEDERAL DE SERGIPE

Tese de doutorado apresentada ao Programa de Pós-Graduação em Letras como parte dos requisitos necessários à obtenção do título de Doutor em Letras pela Universidade Federal de Sergipe.

Orientadora: Profa. Dra. Raquel Meister Ko. Freitag.

SÃO CRISTÓVÃO – SE
2025

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL

UNIVERSIDADE FEDERAL DE SERGIPE

Silva, José Manoel Siqueira da

S586c Covariação morfossintática no português brasileiro : identificação dialetal de estudantes da Universidade Federal de Sergipe / José Manoel Siqueira da Silva ; orientadora Raquel Meister Ko. Freitag. – São Cristóvão, SE, 2025.

186 f. ; il.

Tese (Doutorado em Letras) – Universidade Federal de Sergipe, 2025.

1. Linguística. 2. Linguagem e línguas - Variação. 3. Língua portuguesa - Morfologia. 4. Língua portuguesa - Sintaxe. 5. Dialetos urbanos - Sergipe. I. Universidade Federal de Sergipe. II. Freitag, Raquel Meister Ko., orient. III. Título.

CDU 81’282(813.7)

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO EM PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS
MESTRADO E DOUTORADO EM LETRAS

Ata de Exame de Defesa da Tese de Doutorado apresentada por **JOSE MANOEL SIQUEIRA DA SILVA** em 21 de fevereiro de 2025.

No vigésimo primeiro dia do mês de fevereiro do ano de dois mil e vinte e cinco, às treze horas, reuniu-se, na sala 302 da didática VII da Universidade Federal de Sergipe, a comissão para o Exame de Defesa da tese de doutorado intitulada:**Covariação morfossintática no português brasileiro: identificação dialetal de estudantes da Universidade Federal de Sergipe,** composta por Raquel Meister Ko Freitag, Presidente e Orientadora, Livia Oushiro, da Universidade Estadual de Campinas, Elyne Giselle de Santana Lima Aguiar Vitório, da Universidade Federal de Alagoas, Fabricio Paiva Mota, da Universidade Federal de Sergipe, Marta Deysiane Alves Faria Sousa, do Instituto Federal de Sergipe. A presidente da comissão examinadora deu início ao exame de defesa, facultando ao candidato a exposição oral em até vinte minutos. Em seguida, passou a palavra a cada examinador, por igual tempo, para arguição do trabalho. Terminada a arguição, a comissão examinadora se reuniu em particular para proceder à avaliação final. Retornando à sala, a presidente da comissão examinadora anunciou a aprovação do trabalho de **JOSE MANOEL SIQUEIRA DA SILVA** na atividade EXAME DE DEFESA do Programa de Pós-Graduação em Letras. Nada mais havendo a tratar, a presidente encerrou a sessão e lavrou a presente ata, aprovada e assinada pela comissão.

Documento assinado digitalmente
gov.br RAQUEL MEISTER KO FREITAG
Data: 21/02/2025 16:26:05-0300
Verifique em https://validar.iti.gov.br

RAQUEL MEISTER KO FREITAG
Universidade Federal de Sergipe

Documento assinado digitalmente
gov.br FABRICIO PAIVA MOTA
Data: 22/02/2025 19:21:59-0300
Verifique em https://validar.iti.gov.br

FABRICIO PAIVA MOTA
Universidade Federal de Sergipe

Documento assinado digitalmente
gov.br MARTA DEYSIANE ALVES FARIA SOUSA
Data: 23/02/2025 09:08:38-0300
Verifique em https://validar.iti.gov.br

MARTA DEYSIANE ALVES FARIA SOUSA
Instituto Federal de Sergipe

Documento assinado digitalmente
gov.br LIVIA OUSHIRO
Data: 21/02/2025 16:32:57-0300
Verifique em https://validar.iti.gov.br

LIVIA OUSHIRO
Universidade Estadual de Campinas

Documento assinado digitalmente
gov.br ELYNE GISELLE DE SANTANA LIMA AGUIAR VITO
Data: 21/02/2025 19:34:35-0300
Verifique em https://validar.iti.gov.br

ELYNE GISELLE DE SANTANA LIMA AGUIAR VITÓRIO
Universidade Federal de Alagoas

Cidade Universitária Prof. José Aloísio de Campos
Didática VII , sala 308. Jardim Rosa Elze s/n - São Cristóvão (SE) CEP 49.100-000
Tel. (79) 3194-6860 – e-mail:ppgl@academico.ufs.br

Uma língua é um dialeto com um exército e marinha
Max Weinreich

AGRADECIMENTOS

Como disse Isaac Newton, “se eu pude enxergar longe, foi por estar sobre ombros de gigantes”. Essa frase reflete a minha gratidão a todas as pessoas que me apoiaram ao longo desta jornada acadêmica.

Em primeiro lugar, gostaria de expressar minha gratidão à minha orientadora, Profa. Dra. Raquel Meister Ko. Freitag, cuja orientação, desde 2019, foi essencial para o desenvolvimento não só desta pesquisa, mas também para o meu desenvolvimento como pesquisador e professor. Todos os seus conselhos estarão para sempre guardados em minha memória.

Este estudo é o resultado do apoio de muitas pessoas, mas, acima de tudo, é uma conquista que dedico ao meu pai Gilvan e à minha mãe Ana, que sempre me apoiaram, me financiaram e foram compreensíveis em todos os momentos em que estive ausente. Obrigado por acreditarem em mim. Sou grato também à minha irmã, Joseane, que, mesmo entre tapas, está ao meu lado.

Agradeço aos amigos que fiz ao longo desta jornada, em especial aos amigos do GELINS, em nome de Paloma e Lucas, cujas amizades quero levar para a vida. Vocês tornaram esses anos de estudo uma experiência que vale a pena, mesmo com a minha distância.

Aos meus amigos de fora do mundo acadêmico, que têm grande contribuição nesse processo, e espero que sigam ao meu lado em nossas conquistas, em especial Luiz Felipe, Cleissiane, Cleison, Francisco, Viviane, Adriano, Jakelyne, Rui e Franciele.

À Profa. Dra. Marta Deysiane Alves Faria Sousa, que me ajudou em grande parte deste trabalho, com leituras, comentários, sugestões, construção de códigos de extração e ouvindo minhas reclamações.

Às professoras Doutoras Elyne Vitório, Josilene Mendonça, Lívia Oushiro e Roana Rodrigues e ao professor Doutor Fabrício Mota, cujos apontamentos nas bancas de qualificação e de defesa contribuíram significativamente para a construção deste trabalho.

Aos(às) professores(as) e colegas do Programa de Pós-graduação em Letras (PPGL) da Universidade Federal de Sergipe (UFS), agradeço pelas valiosas contribuições, discussões e pelo ambiente de colaboração que tanto enriqueceu este trabalho.

À Fundação de Apoio à Pesquisa e à Inovação Tecnológica do Estado de Sergipe (FAPITEC), pelo financiamento desta pesquisa.

Por fim, a todos os participantes e colaboradores desta pesquisa, meu sincero agradecimento por suas contribuições essenciais.

RESUMO

Variedades dialetais são formadas por um conjunto de traços linguísticos nos diferentes níveis da língua, que as tornam distintas de outras variedades (Chambers; Trudgill, 2004; Mansfield; Leslie-O'neill; Li, 2023). Estudos sobre variedades dialetais do português brasileiro têm se concentrado em descrever padrões de fenômenos variáveis isoladamente. Se por um lado esta abordagem tem permitido uma compreensão mais profunda do comportamento das regras variáveis e seus efeitos na organização da língua, por outro lado há pouco conhecimento da interação entre diferentes variáveis linguísticas – a covariação (Guy, 2013; Oushiro, 2015a; 2016a; Beaman, 2022; Freitag, 2022). Mesmo dentre as abordagens de covariação, o recorte de fenômenos tem priorizado o nível fonético-fonológicos (Labov, 2006[1966]; Tamminga, 2019). Assumindo que a análise de covariação é relevante para a identificação da origem dialetal dos falantes, e a ampliação para variáveis morfossintáticas pode contribuir para revelar padrões recorrentes de uso linguístico em regiões geográficas específicas, o que pode ser verificado empiricamente, neste trabalho investigamos a seguinte questão: a descrição da covariação morfossintática na fala de grupos de falantes de diferentes regiões geográficas identifica sua origem dialetal? Exploramos a hipótese de que a identificação da origem dialetal do falante a partir de variáveis morfossintáticas pode ser realizada utilizando técnicas de análise de covariação. Assim, o objetivo é descrever covariação entre quatro variáveis morfossintáticas geograficamente distintas do PB: i) uso variável de artigo definido antes de possessivos pré-nominais (*sua casa* x *a sua casa*) (ART); ii) pronomes pessoais de segunda pessoa do singular (2PS) em função de sujeito (*tu anda* x *você anda* x *cê anda*) (pro2PS); iii) pronomes clíticos de 2PS (*te vi* x *lhe vi*) (cli2PS); e iv) pronomes possessivos de 2PS (*tua casa* x *sua casa*) (pos2PS). Para alcançar esse objetivo, empregamos técnicas de análise de covariação para identificar padrões dialetais, tais como teste de correlação, padrões de agrupamento social e análise de *cluster*. Os dados utilizados provêm de amostras sociolinguísticas com a fala de 181 estudantes universitários da Universidade Federal de Sergipe (UFS) – Deslocamentos (2019), Deslocamentos (2020) e Linguagem Corporificada (2023) – considerando as variáveis sociais deslocamento, tempo no curso e gênero do falante. Os resultados mostram, nas análises individuais, a predominância da ausência de artigo antes de possessivos, do pronome *você* como sujeito, do clítico *te* e do possessivo *seu*. Variáveis sociais, como deslocamento, tempo no curso e gênero, correlacionam-se com as variáveis morfossintáticas. A análise de covariação revelou correlações significativas entre os pares ART~pro2PS, pro2PS~pos2PS, cli2PS~pos2PS e pro2PS~cli2PS, bem como um padrão de agrupamento social em que mais de 30% dos falantes de determinados grupos compartilham frequências similares. A análise de *cluster* identificou três agrupamentos naturais com base nos usos linguísticos, considerando similaridades e diferenças. Embora as técnicas de covariação tenham evidenciado padrões relevantes, a identificação da origem dialetal dos falantes só foi possível na análise de *cluster*, na qual falantes baianos apresentaram comportamento distinto de outros falantes. Apesar das limitações, como a baixa frequência de variáveis morfossintáticas nos dados coletados, este estudo contribui metodologicamente para a descrição de padrões covariáveis e reforça a importância do estudo de covariação para a caracterização de padrões dialetais.

Palavras-chave: Covariação. Identificação Dialetal. Morfossintaxe.

ABSTRACT

Dialectal varieties are formed by a set of linguistic features at different language levels, making them distinct from other varieties (Chambers; Trudgill, 2004; Mansfield; Leslie-O'neill; Li, 2023). Studies on dialectal varieties of Brazilian Portuguese (BP) have focused on describing patterns of variable phenomena in isolation. While this approach has allowed for a deeper understanding of the behavior of variable rules and their effects on the organization of the language, on the other hand, there is little knowledge of the interaction between different linguistic variables – covariation (Guy, 2013; Oushiro, 2015a; 2016a; Beaman, 2022; Freitag, 2022). Even among covariation approaches, the focus on phenomena has prioritized the phonetic-phonological level (Labov, 2006[1966]; Tamminga, 2019). Assuming that covariation analysis is relevant for identifying the dialectal origin of speakers, and that expanding it to morphosyntactic variables can help reveal recurring patterns of language use in specific geographic regions, which can be verified empirically, in this work we investigate the following question: does the description of morphosyntactic covariation in the speech of groups of speakers from different geographic regions identify their dialectal origin? We explore the hypothesis that the identification of the dialectal origin of the speaker from morphosyntactic variables can be performed using covariation analysis techniques. Thus, our goal is to describe covariation between four geographically distinct morphosyntactic variables of BP: i) variable use of the definite article before prenominal possessives (*sua* casa x *a sua* casa) (ART); ii) second-person singular personal pronouns (2PS) as a function of subject (*tu* anda x *você* anda x *cê* anda) (pro2PS); iii) 2PS clitic pronouns (*te* vi x *lhe* vi) (cli2PS); and iv) 2PS possessive pronouns (*tua* casa x *sua* casa) (pos2PS). We used covariation analysis techniques to identify dialectal patterns, such as correlation tests, social grouping patterns, and cluster analysis. The data used come from sociolinguistic samples with the speech of 181 university students from the Federal University of Sergipe (UFS) – Deslocamentos (2019), Deslocamentos (2020), and Linguagem Corporificada (2023) – considering the social variables displacement, time in the course, and gender of the speaker. In the individual analyses, the results show the predominance of the absence of an article before possessives, of the pronoun *você* as the subject, of the clitic *te*, and of the possessive *seu*. Social variables, such as displacement, time in the course, and gender, correlate with the morphosyntactic variables. Covariation analysis revealed significant correlations between the pairs ART~pro2PS, pro2PS~pos2PS, cli2PS~pos2PS, and pro2PS~cli2PS, as well as a pattern of social clustering in which more than 30% of speakers from certain groups share similar frequencies. The cluster analysis identified three natural groupings based on linguistic uses, considering similarities and differences. Although covariation techniques revealed relevant patterns, the identification of the dialectal origin of the speakers was only possible in cluster analysis, in which speakers from Bahia presented distinct behavior from other speakers. Despite limitations, such as the low frequency of morphosyntactic variables in the collected data, this study contributes methodologically to the description of covariation patterns and reinforces the importance of covariation studies for the characterization of dialectal patterns.

Keywords: Covariation. Dialectal Identification. Morphosyntax.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O advento da Sociolinguística Variacionista – aporte teórico-metodológico proposto por Weinreich, Labov e Herzog (2006[1968]) e Labov (2006[1966]; 2008[1972]) que busca descrever os usos reais da língua em seu contexto social – permitiu a descrição da diversidade linguística do português brasileiro (PB) e de suas relações com os aspectos históricos, culturais e sociais do Brasil. Assim, variedades linguísticas, entendidas, segundo Chambers e Trudgill (2004), como entidades únicas que podem ser caracterizadas por pronúncia, léxico e morfossintaxe distintos de outras variedades, passaram a ser descritas por meio de estudos de produção sociolinguística. Estudos de produção, conforme Freitag (2018a), observam como e por que determinadas formas linguísticas e significados sociais se vinculam, como essas formas são produzidas e quais fatores interferem em seus usos, evidenciando que o PB é variável, isto é, apresenta diferentes meios de se dizer a mesma coisa em um mesmo contexto (Labov, 1972, p. 164).

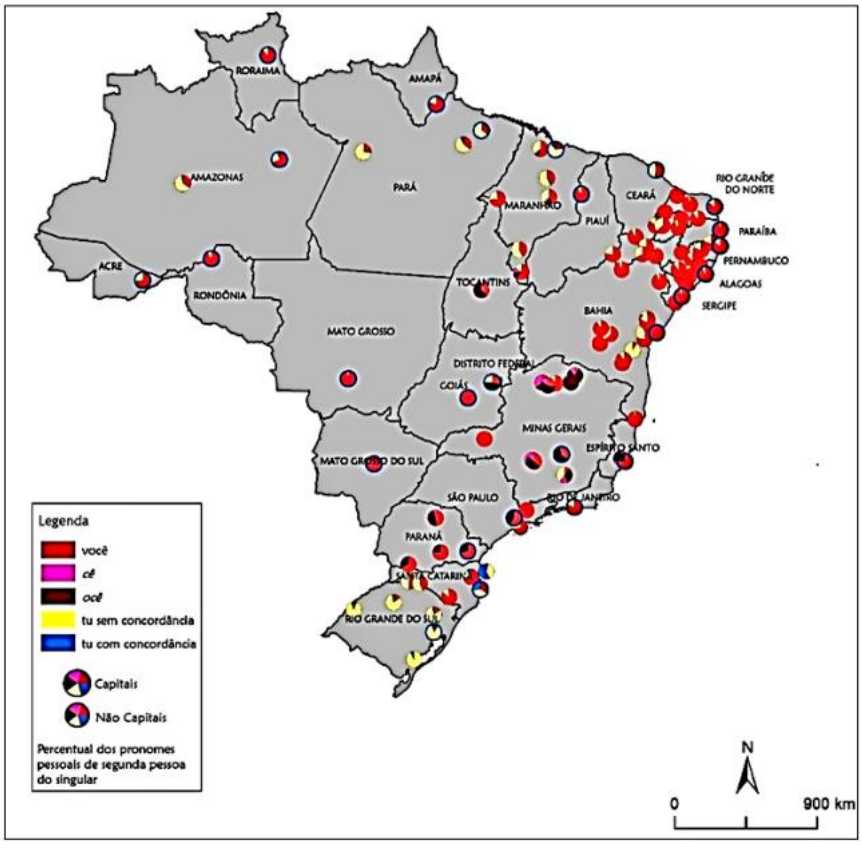
Dentre formas variáveis observadas no PB, estudos de produção, como Callou, Moraes e Leite (1996) sobre a pronúncia do /R/ em final de sílaba (p.ex., ma**r**, po**r**ta), e Callou e Moraes (1996) sobre a pronúncia do /S/ pós-vocálico (p.ex., pa**s**ta, te**s**ta), evidenciam a existência de fenômenos linguísticos que descrevem como falantes de uma comunidade geograficamente localizada fazem uso da língua, as normas linguísticas compartilhadas constitutivas de sua variedade, distintas de outras comunidades. Nesse sentido, em consonância com Mansfield, Leslie-O'Neill e Li (2023), consideramos que esse tipo de variedade da língua, composta por variáveis geograficamente estratificadas, é dialetal, quando compartilha sua gramática, fonologia e léxico, mas que são diferentes o suficiente para serem reconhecidas como variedades distintas, baseadas na geografia – esta entendida no sentido de espaço geográfico, em seu uso corrente, sinônimo de local, relacionando-se à noção cartográfica, isto é, no sentido de apontar onde está alguém ou algo.

Falantes tendem a associar a variação dialetal a formas superficiais da língua (p.ex., variáveis fonético-fonológicas). Freitag (2023) pontua que a associação a variedades dialetais é comumente feita a partir de diferenças linguísticas ao nível fonológico ou suprassegmental. Por exemplo, o uso de certas variantes fonético-fonológicas do PB pode contribuir para o reconhecimento da região de origem do falante, o que, popularmente, conhecemos como "sotaque": a palatalização de oclusivas alveolares (acréscimo dos fones [ʃ] e [ʒ] aos fones [t] e [d] antes de /i/, em ['pẽtʃɪ] – *pente* – e ['kɾɛdʒɪtʊ] – *crédito*), associada a variedades do Sudeste e Sul, e a aproximante retroflexa alveolar em coda silábica (['pɔɻtə] para *porta*),

associada a variedades mineiras e paulistas, são formas linguísticas que falantes, ao escutarem, podem atribuir determinadas características dialetais para quem as produziu. Essa associação é comumente corroborada por descrições sociolinguísticas. Oushiro (2015a), em consonância com Cristófaro-Silva (2007), ao discutir o uso de /R/ em final de sílaba, aponta que estudos de produção evidenciam a variante tepe [ɾ] como variante "paulistana", a aproximante retroflexa [ɻ] como variante dos "paulistas do interior", a fricativa velar [x ɣ] como dos "cariocas", e a fricativa glotal [h ɦ] de "belo-horizontinos".

Variáveis morfossintáticas, por outro lado, tendem a ser menos associadas a variedade dialetais (Labov, 1993; 2001; Meyerhoff; Walker, 2013; Moore, 2021). No entanto, mesmo a variação na morfossintaxe pode estar sujeita à distinção dialetal. O estudo de Scherre, Andrade e Catão (2021) sobre os pronomes pessoais de segunda pessoa do singular (2PS) (*tu* x *você* x *cê*) em posição de sujeito é exemplo para essa distinção (Figura 1). Para a descrição do fenômeno e construção do mapa, os autores utilizaram não só dados provindos da metodologia variacionista, mas também da metodologia dialetológica, integradas.

Figura 1 – **Percentual de cinco construções com pronomes pessoais de segunda pessoa do singular (você, cê, ocê, tu sem concordância e tu com concordância) no português brasileiro: capitais e não capitais**



Fonte: Scherre, Andrade e Catão (2021, p. 181).

O mapeamento da variação nos pronomes pessoais de 2PS em posição de sujeito permite a visualização de diferentes padrões de uso a depender da região geográfica do falante. Ainda que o pronome *você* ocorra com maior frequência, ainda há usos da variante *tu* sem concordância em diferentes regiões do Brasil, como também o uso da variante em implementação *cê*, da variante *ocê* e de *tu* com concordância. Especificamente, na região Nordeste, há predominância de *você* em todos os estados, variando com *tu* sem concordância. Nesse sentido, dialetos da região podem ser definidos por meio do uso de pronomes pessoais de 2P em função de sujeito.

No entanto, descrever somente um traço morfossintático pode não ser o bastante para definir e identificar a origem dialetal de grupos de falantes: a descrição conjunta de traços morfossintáticos, através de suas frequências de uso, contribui para a identificação da origem dialetal do falante. É possível que, ao descrever um padrão morfossintático, entendam-se outros padrões. Por exemplo, a mudança no paradigma pronominal de 2PS em posição de sujeito, evidenciada no mapa anterior, resultou em outras modificações no sistema pronominal do PB (Duarte; Varejão, 2013; Scherre; Duarte, 2016), devido a pressões internas e à força de paralelismo, isto é, a tendência para a ocorrência de formas gramaticais similares conjuntamente, conforme (i) e (ii):

i) o uso do clítico *lhe*, oriundo de terceira pessoa (3P), variante de *te*, pronome clítico canônico de 2PS; e
ii) o uso do possessivo *seu*, oriundo do paradigma de 3P, variante de *teu*, pronome canônico de 2PS.

É o que Weinreich, Labov e Herzog (2006[1968]) colocam como o problema do encaixamento: como as mudanças linguísticas se encaixam no sistema linguístico de uma comunidade. Assim, a introdução do pronome *você* no paradigma de 2PS provocou uma reorganização no quadro pronominal do PB que, por sua vez, levou à adoção do pronome clítico *lhe*, forma de objeto indireto de terceira pessoa do singular (3PS), como uma forma alternante ao pronome *te* na 2PS em contextos de objeto direto e indireto. Já no final da década de 1990, Ramos (1999) propõe a existência de padrões gramaticais distintos para a variação, a depender da região geográfica da comunidade. Os seus resultados evidenciam que no PB falado em capitais do Nordeste, como Maceió, Recife, Salvador e João Pessoa, emprega-se *você* em posição de sujeito e o clítico *lhe* como objeto de 2PS, comportamento distinto do Maranhão e de estados do Norte, em que há uma distinção entre *tu* e *você*, sendo o primeiro

utilizado em situações íntimas e familiares e o segundo em contextos formais e de respeito, cuja distinção se reflete nos clíticos *te* e *lhe* respectivamente.

Esse paralelismo e efeito do encaixamento se reflete também nos possessivos de 2PS, dado que a inserção de *você* como pronome de 2PS resulta na adoção de *seu* como forma possessiva de 2PS, variando com o possessivo *teu*, cuja variação está correlacionada à região geográfica do falante. É o que mostra o estudo de Lopes (2008): em regiões onde se utiliza *você*, a forma de 3PS *seu* tende a ser preservada, enquanto em regiões nas quais se utiliza *tu* como forma predominante, falantes tendem a empregar mais frequentemente *teu*: *teu* se configura como uma forma relacionada ao uso de *tu*, ao passo em que *seu* se relaciona à forma *você*.

Fenômenos morfossintáticos aparentemente não relacionados também podem apresentar distinção dialetal para sua distribuição. É o caso do uso variável de artigo antes de possessivos pré-nominais, em "perdi *meu* celular" ~ "perdi *o meu* celular", cujo uso é dialetalmente distinto. Callou e Silva (1997), na década de 1990, evidenciam o comportamento dialetal da variável a partir de um estudo contrastivo entre capitais brasileiras a partir de dados do projeto Norma Urbana Linguística Culta (NURC): falantes das capitais de estados das regiões Sul e Sudeste (Porto Alegre, São Paulo e Rio de Janeiro) fazem maior uso do artigo definido antes de possessivos do que falantes das capitais de estados da região Nordeste (Recife e Salvador).

Se há diferença nos usos quanto à região geográfica dos falantes, podemos supor que há evidências de que as quatro variáveis linguísticas exibem distinção dialetal. Seus padrões de uso variam dependendo da região geográfica do falante, com múltiplas variantes coexistindo dentro dessas comunidades. Isso implica que associar apenas uma das variantes morfossintáticas a um dialeto pode não ser suficiente para sua identificação, dada a natureza heterogênea da língua e à possibilidade de diferentes comunidades adotarem formas variáveis de maneira semelhante. Há, por exemplo, grupos de falantes que podem fazer uso de *você+lhe+seu+ausência*, como também há grupos de falantes que podem fazer uso de *você+te+teu+presença*, *tu+te+teu+presença*, *tu+lhe+teu+ausência* e assim sucessivamente. Nesse sentido, é essencial considerar mais de uma variável ao traçarmos um entendimento de fenômenos morfossintáticos de dialetos da língua, dado que variantes podem funcionar conjuntamente na comunidade, por força dialetal, como é o caso das quatro variáveis, e também por força de paralelismo, como é o caso específico das variáveis de 2PS.

Para a observação de uso conjunto de variáveis linguísticas, a pesquisa em Sociolinguística Variacionista tem adotado técnicas de descrição pautadas na covariação, entendida, para Oushiro (2015a), como a correlação entre múltiplas variáveis nos usos de

falantes individuais, em que se observam, segundo Freitag (2022), padrões conjuntos de uso de mais de uma variável. O estudo de Labov (2006[1966]) para a estratificação social do inglês no Lower East Side, em Nova Iorque, Estados Unidos, é pioneiro na aplicação dessa técnica de descrição linguística. Nesse estudo, o autor observou, a partir de contagens de usos das variáveis, uma coerência estrutural para cinco variáveis fonológicas: falantes que tendiam a uma baixa frequência de alçamento de (aeh), como em *bad* (mau), também tendiam a uma baixa frequência de alçamento de (oh), como em *caught* (pego); similarmente, aqueles com uma alta frequência de alçamento de uma também tendiam a ter uma alta frequência para a outra. O estudo de Labov (2006[1966]) evidencia que a existência de correlações e a necessidade de observar usos linguísticos conjuntos torna imperativa a descrição de padrões sociolinguísticos que covariam para a descrição de variedades dialetais da língua. Ainda assim, como expõe Oushiro (2015a), esse tipo de observação pouco foi posto em prova.

Além disso, para a descrição de covariação em variedades dialetais da língua, é necessário um conjunto de dados que possibilite esse tipo de visualização. No estado de Sergipe, a descrição linguística tem sido desenvolvida por pesquisadores da Universidade Federal de Sergipe (UFS), dentro da linha de pesquisa *Produção e percepção sociolinguística* e do projeto *Como fala, lê e escreve o universitário?*, coordenados pela Profa. Dra. Raquel Meister Ko. Freitag. Fruto dessa linha de pesquisa e do projeto é a constituição das amostras *Deslocamentos* – Deslocamentos (2019) e Deslocamentos (2020)[1] – e *Linguagem Corporificada* (2023), que comportam a fala de estudantes universitários da UFS, *campus* Prof. José Aloísio de Campos, em São Cristóvão, considerando sua região de origem e seu acesso – em termos de mobilidade – à universidade (Corrêa, 2019; Cardoso, 2021; Novais, 2021; Pinheiro, 2021; Ribeiro, 2019; Rodrigues, 2021; Silva, 2020; Silva, 2021; Souza, 2022; entre outros), já que a UFS recebe estudantes de variados lugares (do próprio estado; de estados vizinhos; de estados mais distantes) que interagem entre si e desenvolvem práticas em conjunto no *campus*. As amostras contam com a fala de estudantes universitários que i) são naturais e residentes da região metropolitana de Sergipe; ii) são do interior do estado que fazem o percurso diário para o *campus* Prof. José Aloísio de Campos, em São Cristóvão; iii) são do interior do estado que residem na região circunvizinha ao *campus*; e iv) são externos ao estado que residem na região circunvizinha ao *campus*. A existência de falantes de diferentes regiões geográficas pode fornecer evidências para defender a hipótese de que há

[1] Há ainda a amostra Deslocamentos (UFS-Itabaiana2018), constituída por Araújo (2022), Mendonça (2022) e colaboradores.

diferenças dialetais em seu comportamento linguístico, tanto em formas superficiais da língua, quanto na morfossintaxe.

Essa previsão já foi posta em prova, já que estudos de descrição feitos no escopo do projeto com base nas amostras reforçam que fenômenos variáveis são dialetalmente distintos: i) em Corrêa (2019), falantes externos ao estado de Sergipe palatalizam mais as oclusivas alveolares /t d/ do que falantes internos aos estado, em [‘tʃia] para “tia”; e ii) Ribeiro (2019) descreve que estudantes do interior do estado utilizam mais a variante *ni* como preposição locativa do que estudantes da região metropolitana, em “vou *ni* São Cristóvão”. Em específico, na morfossintaxe, i) Rodrigues (2021) evidencia que estudantes externos a Sergipe fazem maior uso da preposição *em* com verbos de movimento, como “vou *em* Aracaju”; ii) Silva (2020) demonstra que universitários sergipanos tendem a usar menos o artigo definido antes de possessivos pré-nominais do que os externos ao estado, em “*a sua* mãe”; e iii) em Siqueira, Sousa e Rodrigues (2023), falantes do interior de Sergipe fazem maior uso da variante *cê* para a segunda pessoa do singular em posição de sujeito e da variante *lhe* como clítico de segunda pessoa do que falantes da capital do estado, em “*cê* chegou tarde” e “eu *lhe* comprei um livro”. Dado que fenômenos variáveis podem apresentar diferentes comportamentos a depender da região de origem do falante, a sua descrição em termos de covariação pode descrever aspectos de uma variedade dialetal do PB.

O estudo da covariação auxilia na verificação da possibilidade de múltiplas variáveis se correlacionarem nos usos de falantes individuais (Oushiro, 2015a) e combinar as variáveis “pode nos ajudar a entender padrões dialetais e os diferenciar de padrões sociais” (Freitag, 2022, p. 209).[2] A descrição de traços morfossintáticos dialetais, por sua vez, auxilia na identificação de traços linguísticos do dialeto do grupo de falantes, cujos resultados podem ser aplicado para a identificação de sua região geográfica. Nesse sentido, buscamos, neste trabalho, responder à seguinte questão: a descrição da covariação morfossintática na fala de grupos de falantes de diferentes regiões geográficas identifica sua origem dialetal? Investigamos a hipótese de que a identificação da origem dialetal do falante a partir de variáveis morfossintáticas pode ser realizada utilizando técnicas de descrição de covariação, isso porque, analisando padrões de variação na morfossintaxe por meio da fala de indivíduos de diferentes origens, é possível identificar características distintivas de determinadas regiões geográficas, associando o falante a áreas específicas. Além disso, o uso dessas técnicas

[2] No original: “can help us to understand dialectal patterns and differentiate them from social patterns”. Todas as traduções de citações ao longo deste trabalho são nossas.

permite observar relações entre variantes morfossintáticas em padrões conjuntos de uso, o que contribui para identificar a origem dialetal dos falantes.

Frente à necessidade de se observar a covariação entre variáveis morfossintáticas, tomamos como objetivo geral desta tese descrever a covariação entre variáveis morfossintáticas geograficamente distintas (i)-(iv), buscando identificar padrões que permitam a caracterização das variedades dialetais do PB e a associação entre falantes e suas respectivas regiões de origem.

i) Uso variável de artigo definido antes de possessivo pré-nominal:
    a. *meu irmão* é formado em Contabilidade então eu vi ah isso é massa (36ent.UFS-SaoCristovao2020_desl3_inicio_ama_administracao.fs.20).[3]
    b. *o meu curso* que é Jornalismo ele só tinha aqui na universidade (35ent.UFS-SaoCristovao2020_desl3_inicio_fer_jornalismo.fs.18).

ii) Pronomes pessoais de segunda pessoa do singular (2PS) em posição de sujeito:
    a. rouba teu telefone e *tu* num tá com teu telefone ele vai e mata *tu* ou um dos teus (65ent.UFS-SaoCristovao2020_desl4_inicio_bia_biologia.fs.19).
    b. outros clichezinhos *você* além de *você* se sentir bem ou *você* pode seguir até os conselhos que têm no próprio livro (72ent.UFS-SaoCristovao2020_desl4_final_aur_engquimica.fs.22).
    c. por quê? eu amo dormir vei *cê* é doido é faço isso porque (02ent.UFS-SaoCristovao2020_desl1_inicio_ant_geo.ms.18).

iii) Pronomes clíticos de 2PS:
    a. mas tem professor que tá ali e *lhe* ajuda os meus orientadores mesmo (15ent.UFS-SaoCristovao2020_desl1_final_dan_qui.fs.22).
    b. porque acho que (hes) a vida acadêmica *te* cobra isso (11ent.UFS-SaoCristovao2020_desl1_final_ali_fono.fs.23).

[3] Os exemplos que apresentam essa codificação ao final são retirados do banco de dados Falares Sergipanos (Freitag, 2013). O número inicial corresponde ao número da entrevista (36) e *ent* indica que é uma entrevista. Seguindo, informa-se o *campus* onde foi feita a coleta (UFS-SaoCristovao) e o ano (2020). Em seguida, apresenta-se o perfil de deslocamento do falante, identificando seu acesso à UFS em termos de mobilidade (desl1, desl2, desl3 e desl4). Após, identifica-se o tempo de curso do falante (início e final), as três primeiras letras de seu nome (ama), seu curso (no caso da amostra de 2020), sexo (f ou m), a etapa de ensino (s – superior) e a idade (20).

iv) Pronomes possessivos de 2PS:

a. a escola dizer "oh se você trabalhar você vai ganhar *seu* dinheiro num vai precisar roubar num vai fazer mal a ninguém" (74ent.UFS-SaoCristovao2020_desl4_final_cat_engflorestal.fs.21).

b. e tu num tá com *teu* telefone ele vai e mata tu ou um dos *teus* (65ent.UFS-SaoCristovao2020_desl4_inicio_bia_biologia.fs.19).

Os fenômenos foram escolhidos seguindo três critérios: i) são variáveis morfossintáticas cujos resultados de estudos observacionais apontam uma distinção dialetal, em que grupos de falantes de diferentes regiões apresentam diferentes padrões de uso, conforme evidenciam Callou e Silva (1997), Lopes (2008), Ramos (1999) e Scherre, Andrade e Catão (2021); ii) descrições prévias feitas com base no PB falado por universitários da UFS já evidenciaram sua produtividade em amostras sociolinguísticas do banco de dados Falares Sergipanos (Freitag, 2013) e sua distinção dialetal, conforme Araújo (2022), Araújo e Borges (2021) e Siqueira, Sousa e Rodrigues (2023); e iii) há similaridade estrutural entre as variáveis, dada a presença de pronomes do PB: três dos fenômenos envolvem pronomes de 2PS, enquanto o quarto, ainda que se atenha à variação no uso de artigo, envolve pronomes possessivos.

Tamminga (2019, p. 12) argumenta que "um contexto no qual a covariação pode surgir é quando há alguma relação estrutural entre dois fenômenos. Outro contexto é quando fenômenos pertencem a dialetos distintos".[4] Diante disso, apresentamos duas hipóteses para a observação de covariação entre os fenômenos selecionados. A primeira, frente à força do paralelismo, é que haverá correlação forte e significativa entre os fenômenos que envolvem a 2PS, uma vez que há uma similaridade estrutural entre as variáveis, enquanto haverá correlação fraca, mas significativa, com o uso variável de artigo definido antes de possessivo pré-nominal. A segunda, com base no que apontam estudos observacionais (Callou; Silva, 1997; Lopes, 2008; Ramos, 1999; Scherre; Andrade; Catão, 2021), é que, uma vez que são variáveis dialetais e que contamos com falantes de regiões diversas, haverá diferentes padrões de uso com base nas frequências de realização de cada fenômeno.

A descrição de agrupamento de traços sociolinguísticos se inicia a partir da descrição de fenômenos individuais da língua e, em seguida, procura por padrões de covariação. Nesse sentido, como objetivos específicos, buscamos (i) descrever fenômenos morfossintáticos variáveis do PB com base em fatores sociais; (ii) investigar a covariação entre quatro

[4] No original: "One context where covariation can arise is when there is some structural relationship between two features. Another is when features belong to distinct dialects".

fenômenos morfossintáticos variáveis, de modo a sistematizar padrões de uso conjuntos; e (iii) identificar as diferenças e semelhanças nos usos de traços sociolinguísticos morfossintáticos por estudantes universitários da UFS.

Como *corpora* para a descrição dos dados, trabalhamos com três amostras sociolinguísticas desenvolvidas com base no falar de estudantes da UFS oriundos de diferentes regiões, que estudam no *campus* Prof. José Aloísio de Campos, em São Cristóvão, SE, as amostras *Deslocamentos* – Deslocamentos (2019) (N= 64 falantes) e Deslocamentos (2020) (N= 60 falantes) –, e amostra *Linguagem Corporificada* (2023) (N= 57 falantes), por já terem evidenciado distinção dialetal em variáveis linguísticas do PB. O padrão etário dos estudantes é entre 18 e 30 anos (*Média* = 21).

Nosso trabalho se insere em um projeto conjunto alinhado ao Grupo de Estudos em Linguagem, Interação e Sociedade (GELINS), na Linha de Pesquisa de Produção e Percepção Sociolinguística, coordenado pela Profa. Dra. Raquel Meister Ko. Freitag na UFS. O projeto, desde 2007, tem contribuído profusamente para a descrição linguística no estado de Sergipe (Araujo, 2014; 2022; Barreto, 2014; Cardoso, 2021; Correa, 2019; Mendonça, 2016; 2022; Novais, 2021; Pinheiro, 2021; Ribeiro, 2019; Rodrigues, 2021; Santana, 2019; Silva, 2020; Silva, 2021; Souza, 2016; Souza, 2022; entre outros).

Esta pesquisa é financiada pela Fundação de Apoio à Pesquisa e à Inovação Tecnológica do Estado de Sergipe (FAPITEC/SE/FUNTEC Nº 04/2021) com base no projeto *Variação morfossintática e a sensibilidade sociolinguística dos falantes*. Os resultados da descrição dos traços morfossintáticos foco desta tese serão aplicados para a produção de um banco de dados linguístico que possa ser aplicado ao desenvolvimento de um algoritmo de identificação geográfica (e etária) de usuários de *chatbots* de atendimento. Sua aplicação decorre do fato de que a descrição linguística propiciada pela Sociolinguística Variacionista, através da sistematização de traços linguísticos, geográficos e sociais, oferece material de treinamento e teste de algoritmos para a identificação de características sociais ao *chatbot*, de modo a potencializar suas interações com os humanos, como proposto por Freitag (2021).

Este trabalho também se propõe a contribuir não apenas teoricamente, mas também metodologicamente à agenda de descrição linguística, focando especialmente na observação conjunta de traços linguísticos e na covariação desses traços. A ênfase metodológica recai sobre o uso de diferentes abordagens estatísticas para a análise da covariação, incluindo a aplicação de testes de correlação, padrões de agrupamento social e técnicas de análise de *cluster*. Essas ferramentas contribuem para uma descrição mais robusta da covariação, especialmente no que diz respeito aos traços morfossintáticos dialetais do PB. Além disso, a contribuição metodológica deste trabalho é

ampliada pelo desenvolvimento e disponibilização pública dos *scripts* utilizados nas análises, por meio do OSF e GitHub[5], permitindo que outros pesquisadores possam reproduzir os procedimentos adotados, garantindo a transparência e a reprodutibilidade dos resultados. Com isso, nesta pesquisa não só ampliamos o escopo de abordagens para a descrição da covariação, mas também provemos uma metodologia que pode ser aplicada para a identificação da região dialetal de um falante – um campo ainda não amplamente explorado no contexto do PB. Assim, esta tese busca contribuir com a agenda do Eixo 4 – Questões teóricas e metodológicas do GT de Sociolinguística da Associação Nacional de Pós graduação e Pesquisa em Letras e Linguística (ANPOLL).

Organizamos esta tese em sete capítulos, de forma a abarcar o que propomos acima. O primeiro dos capítulos é a Introdução, na qual apresentamos o nosso objeto de estudo e encaminhamentos, como também apresentamos a nossa questão de pesquisa, tese e objetivos. No segundo capítulo, discutimos sobre variação dialetal na morfossintaxe. Para tanto, em um primeiro momento, falamos sobre a informação indexical e distinção dialetal na morfossintaxe, por meio de reflexões propostas em Labov (1972; 1978; 1993; 2001), Sankoff (1973), Lavandera (1978), Freitag (2020; 2022) e Mainsfield, Leslie-O'neill e Li (2023). Em seguida, apresentamos uma revisão integrativa dos fenômenos morfossintáticos dialetalmente distintos mobilizados nesta pesquisa.

No terceiro capítulo, abordamos a covariação, compreendendo reflexões que levaram a sua proposição como também sua conceituação, a partir de discussões em Labov (2006[1966]; 2008[1972]), Weinreich, Labov e Herzog (2006[1968]), Guy (2013) e Guy e Hinskens (2016). A partir disso, abordamos três tipos de técnicas utilizadas para a descrição da covariação, iniciando pela técnica que utiliza testes inferenciais de correlação, por meio de estudos de Guy (2013), Oushiro (2015a; 2016a) e Tamminga (2019) e, em seguida, pelas técnicas de agrupamento de dados, a partir de Guy (2013), Oushiro (2015a; 2016a), Horvath e Sankoff (1987), Beaman (2021) e Freitag (2022).

O quarto capítulo é dedicado à metodologia empregada para a extração e análise de nossos dados. Para tanto, descrevemos o conjunto de dados utilizados na pesquisa, por meio das amostras Deslocamentos (2019), Deslocamentos (2020) e Linguagem Corporificada (2023), explicitamos os processos necessários para a sistematização dos dados por meio de

[5] OSF (*Open Science Framework*) e GitHub são plataformas gratuitas e abertas para compartilhar informações. O OSF é usado principalmente por pesquisadores para disponibilizar dados, códigos e materiais de estudos, facilitando a colaboração e a transparência na ciência. Já o GitHub é mais voltado para programadores, permitindo que eles armazenem, compartilhem e trabalhem juntos em códigos de software. Ambos ajudam a promover a colaboração e o acesso aberto a informações.

estudo de produção, discorrendo sobre o processo de extração de dados, a partir de Sousa (2023), e também apresentamos as etapas de análise estatística adotadas.

O quinto capítulo tem como objetivo descrever a distribuição das quatro variáveis morfossintáticas selecionadas com base em dados extraídos de nossas amostras. Para tanto, realizamos a descrição e análise dos dados a partir da observação de padrões globais de uso, prosseguindo para o cotejamento dos dados a partir da i) amostra, em ordem de observar se os resultados são sensíveis ao tipo de coleta; ii) deslocamento, com vistas a observar se há diferença notável entre grupos geograficamente distintos; iii) tempo no curso, buscando inferir, por meio da distribuição, efeito de exposição à comunidade; iv) indivíduo, para descrever o alcance da variação; e v) gênero e idade, uma vez que fenômenos variáveis podem se correlacionar com características sociais de seus falantes.

No sexto capítulo, avançamos em relação à descrição da variação e realizamos análises de covariação entre os quatro fenômenos morfossintáticos, com vistas a responder nossa pergunta de pesquisa. A descrição de covariação se inicia a partir da observação de possíveis relações entre os pares de variáveis, por meio de análise de correlação. Após isso, realizamos análises de padrões de agrupamento social, o que culmina na análise de *cluster*, de modo a observar possíveis agrupamentos dos falantes em grupos naturais. Por fim, o sétimo capítulo apresenta nossas considerações finais, na qual respondemos às questões lançadas ao longo da pesquisa e sintetizamos os principais achados.

2. VARIAÇÃO DIALETAL NA MORFOSSINTAXE

Os conceitos de variedade e dialeto nem sempre são claros na literatura, uma vez que há diferentes focos. Por exemplo, a sociolinguística brasileira tende a lidar com o termo "variedade" no lugar de "dialeto", uma vez que a descrição de variedades no Brasil não se baseia em hierarquia, com uma língua oficial e as demais variedades como "dialetos inferiores", como é corrente com a adoção do termo "dialeto". Com vistas na distinção, consideramos, a partir de Chambers e Trudgill (2004), variedade como uma entidade única caracterizada por pronúncia, léxico e morfossintaxe distinta de outras variedades, e dialeto, segundo Matthews (1997) e Mansfield, Leslie-O'Neill e Li (2013), como qualquer variedade de uma língua que compartilha sua gramática, fonologia e léxico, mas que é diferente o suficiente para ser reconhecida como distinta, baseada na geografia (p.ex. diferente regiões, estados, zonas etc.), não reconhecendo a existência de hierarquias.

Dialetos da língua são comumente agrupados por falantes através de fenômenos variáveis aos níveis fonético-fonológico e lexical, uma vez que, conforme Labov (1993; 2001), falantes são mais conscientes da variação em formas mais superficiais da língua. Variáveis morfossintáticas, entretanto, também podem apresentar atribuição de significado dialetal, uma vez que há fenômenos morfossintáticos do PB que são dialetalmente distintos. Neste capítulo, objetivamos analisar a variação dialetal no nível morfossintático, com vistas a visualizar padrões morfossintáticos dialetalmente distintos.

No que segue, iniciamos nossa análise abordando sobre a informação indexical e a distinção dialetal na morfossintaxe, por meio de reflexões propostas em Labov (1972; 1978; 1993; 2001), Sankoff (1973), Lavandera (1978), Freitag (2020; 2022) e Mainsfield, Leslie-O'neill e Li (2023). Em seguida, apresentamos uma revisão integrativa de fenômenos morfossintáticos dialetalmente distintos, os quais são descritos nesta tese, a saber: i) uso variável de artigo definido antes de possessivos; ii) pronomes pessoais de 2PS em posição de sujeito; iii) pronomes clíticos de 2PS; e iv) pronomes possessivos de 2PS. Por fim, discutimos sobre a necessidade de agregar diferentes variáveis morfossintáticas na descrição linguística, cujo objetivo é encaminhar nossa tese para a descrição de covariação, em ordem de obter pistas mais robustas para a descrição de dialetos da língua.

2.1 SIGNIFICADO DIALETAL MORFOSSINTÁTICO

Ao fazer uso da língua, conforme defende Freitag (2020), o que é produzido apresenta dois tipos de informação: a informação linguística (aquilo que está sendo dito) e a informação indexical (informações sobre quem fala, seus atributos sociais). Ao ouvir diferentes formas linguísticas, falantes podem atribuir diferentes informações indexicais a elas – e a seus usuários –, isto é, informações sociais nelas codificadas, como sexo/gênero, faixa etária, escolaridade e classe socioeconômica.

Além da atribuição de informações sociais, a língua também pode indexar características dialetais, isto é, informações da região dialetal do falante, resultando em um significado dialetal para a variação: a associação de certas formas linguísticas a determinadas regiões geográficas ou comunidades linguísticas específicas. Exemplos clássicos são o que se convencionou chamar de falares/dialetos nordestino, sulista, baiano, carioca, paulista etc., que se originam a partir da indexação de informações dialetais a formas relativamente específicas de variáveis linguísticas produzidas por falantes oriundos de determinadas comunidades.

A atribuição de informações dialetais, contudo, não é realizada de forma similar em todos os níveis da língua. As discussões aventadas por Labov (1972; 1978), Sankoff (1973), Lavandera (1978) e Romaine (2017[1981]) suscitaram as primeiras reflexões em relação à existência da variação em níveis gramaticais mais altos, como na morfossintaxe. Ainda que atualmente exista consenso em relação à existência de variação morfossintática, como também à atribuição de informação indexical a ela, conforme Moore (2021), a indexação de informações à língua é difícil de dissociar da fonologia uma vez que, como defendem Labov (1993; 2001) e Meyerhoff e Walker (2013), variáveis morfossintáticas são menos sujeitas a indexação do que variáveis fonológicas.

Sobre isso, Freitag (2023, p. 211) pontua que a atribuição de informações dialetais à variação linguística é comumente observada no que "se entende na sociedade como sotaque, a percepção das diferenças da língua no nível fonológico ou suprassegmental, e associada a características de um grupo". Simpson (1994, p. 8), por exemplo, define sotaque como "uma variedade falada de língua realizada nos sons da fala: qualquer sistema desses sons e suas possibilidades combinatórias constitui um sotaque dessa variedade de linguagem".[6] Tal definição não se distancia muito do senso comum, no qual um falante que caracteriza alguém como falante de "sotaque paulista", por exemplo, faz isso através do uso de certas variantes

[6] No original: "a spoken variety of language is realised in speech sounds: any one system of such sounds and their combinatorial possibilities constitutes an accent of that variety of language".

fonético-fonológicas, como a palatalização de oclusivas alveolares (acréscimo dos fones [ʃ] e [ʒ] aos fones [t] e [d] antes de /i/, em [‘pẽtʃɪ] – *pente* – e [‘kɾɛdʒɪtʊ] – *crédito*) e a aproximante retroflexa alveolar em coda silábica ([‘pɔɻtə] para *porta*), mas não por meio de uma variável morfossintática, como o uso de artigo antes de possessivos pré-nominais, em *eu vi a sua moto*, variável apontada por Guedes (2019) como dialetal e presente na fala de paulistas.

Considerar apenas o sotaque e sua concepção corrente, evidentemente, desconsidera variáveis linguísticas em níveis gramaticais mais altos, uma vez que enfoca apenas em sons e suas possibilidades combinatórias. Para a consideração de significados dialetais é necessária a observação de outros níveis linguísticos, dado que a noção de dialeto, como explica Britain (2004), envolve todos os níveis da linguagem, incluindo, mas não se restringindo ao fonológico. Fenômenos morfossintáticos, por exemplo, podem carregar informação indexical de quem e de onde se fala, isto é, significado dialetal, o que Freitag *et al*. (2016) chamam de “sotaques sintáticos”.

Essa noção é especialmente verdadeira em contextos de contato linguístico. Para Trudgill (1986), o contato linguístico é o processo pelo qual falantes de diferentes dialetos e/ou diferentes línguas interagem entre si, podendo trocar formas linguísticas por meio dessa interação. Em uma nova comunidade, falantes são expostos a variáveis linguísticas que até então não haviam sido expostos, como um traço distintivo entre sua comunidade de origem e a nova. Conforme Mainsfield, Leslie-O’neill e Li (2023), informações dialetais podem se desenvolver sem qualquer outra indexação social: quando falantes de variedades dialetais têm contato reduzido com outras variedades a cujas variantes os indivíduos não têm exposição suficiente, a exposição pode resultar na percepção de diferenças linguísticas como resultado da distinção geográfica. Assim, é possível que o falante atribua um significado dialetal a essa variante. É o que Freitag (2022) caracteriza como um fenômeno variável dialetalmente saliente, já que o reconhecimento da variação é “suscetível à circunstância geográfica, assumindo caráter de distinção dialetal” (Freitag, 2018, p. 8). Ao considerarmos que a inserção de um falante em uma comunidade linguisticamente diversificada permite que o falante seja exposto a variáveis que até então não reconheciam, entendemos a variação morfossintática dialetalmente saliente se apresenta como um traço distintivo, perceptível através do contato com falantes de diferentes origens, frente à diferença entre o padrão de uso de sua comunidade e de outras.

A observação do caráter dialetal de uma variável e de seu significado dialetal é feita por meio de técnicas de descrição linguística. Por exemplo, é comum que pesquisadores recorram a uma etapa de documentação chamada de avaliação subjetiva (Cardoso, 2015), na

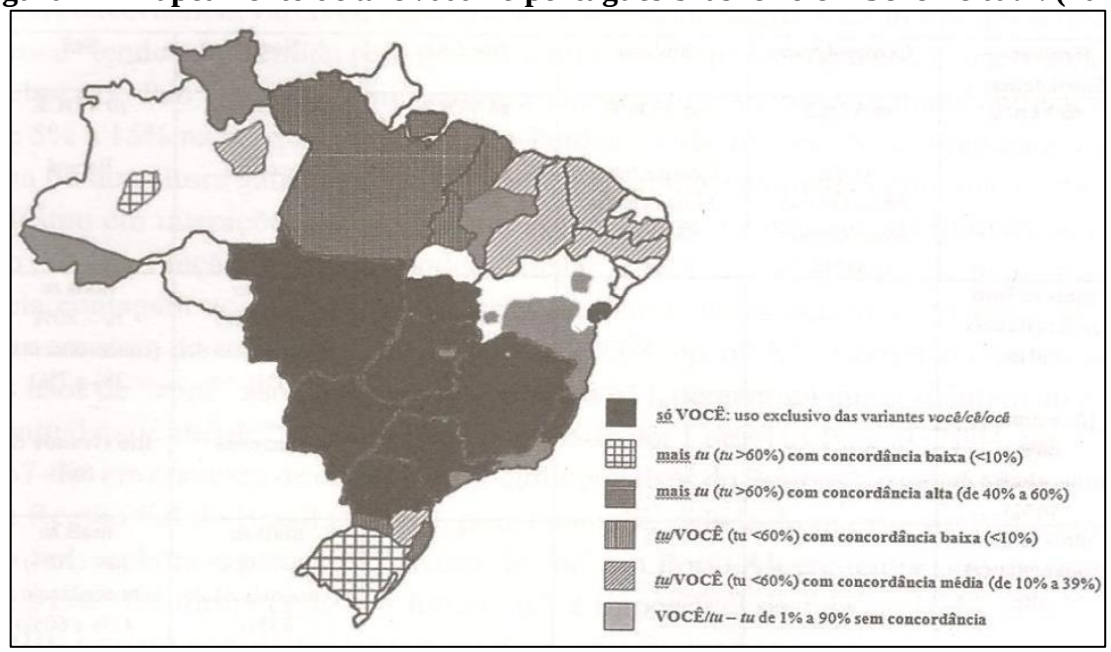
qual se observam inferências dos falantes de uma língua feitas ao serem expostos a um estímulo linguístico (escrito ou oral), que podem ou não ser conscientes. Na ausência de estudos de percepção, uma alternativa para observar diferenças deletais e, consequentemente, significados dialetais, é a observação da distribuição das variáveis em diferentes áreas geográficas, uma vez que, conforme Mainsfield, Leslie-O'neill e Li (2023), podemos dizer que uma variável é dialetal observando que ela não requer uma divisão categórica entre grupos, mas apenas uma diferença notável entre grupos geograficamente localizados em relação à variável. Para ilustrar isso, utilizamos os fenômenos morfossintáticos mobilizados nesta pesquisa.

No PB, há diferentes formas para se referir ao interlocutor durante uma interação. Algumas regiões, por exemplo, utilizam a forma *você*, enquanto outras utilizam a forma *tu*. O uso de tais variantes pode evocar, como pontua Freitag (2022), presunção de origem do falante – são variáveis dialetalmente distintas, isto é, que apresentam comportamentos diferentes a depender da região geográfica do grupo de falantes. Freitag *et al*. (2016, p. 79), em estudo de percepção com 215 falantes universitários do Sul (Universidade Federal de Santa Catarina e Universidade Federal da Fronteira Sul) e do Nordeste (Universidade Federal do Rio Grande do Norte e Universidade Federal de Sergipe), observam que,

> o reporte do *tu* como marca que diferencia a fala se deu tanto entre respondentes do Sul como do Nordeste, com a peculiaridade de que no Nordeste este é um traço pouco produtivo; essa é uma pista sintática que pode auxiliar no desvelamento de fronteiras dialetais, do ponto de vista da percepção.

Há, nesse caso, uma atribuição de informação dialetal com base no uso de formas de uma variável morfossintática. Essa associação, contudo, pode ser vaga, uma vez que há regiões onde falantes podem fazer uso de ambas as formas, como também regiões onde se utilizam as formas reduzidas *cê* e *ocê*, conforme Figura 2.

Figura 2 – Mapeamento de *tu* e *você* no português brasileiro em Scherre *et al.* (2015)



Fonte: Scherre *et al.* (2015, p. 142).

O mapeamento da variação nos pronomes pessoais de 2PS em posição de sujeito feito por Scherre *et al.* (2015) permite a visualização de diferentes padrões de uso a depender da região geográfica do falante. Diferente do mapa na Figura 1 que apresenta os percentuais por área, o mapa da Figura 2 apresenta os subsistemas de uso para as variantes no Brasil. Ainda que o pronome *você* ocorra com maior frequência, observamos que ainda há usos da variante *tu* nas mais variadas regiões do Brasil, como também o uso de *cê* e *ocê*. Especificamente, vemos que, na região Nordeste, há predominância de *você* em todos os estados, variando com *tu* sem concordância.

Não obstante, apenas a descrição de um traço morfossintático pode não ser suficiente para a caracterização de um dialeto, mas é possível que a descrição de um padrão morfossintático auxilie no entendimento de outros. É sabido, por exemplo, que a mudança nos pronomes pessoais de 2PS resultou em outras mudanças no sistema pronominal do PB. Por exemplo, a reorganização no quadro pronominal do PB em decorrência da implementação do pronome *você* no paradigma de 2PS resultou na adoção do clítico *lhe*, originalmente de 3P em função de objeto indireto (dativa), como forma variante ao pronome *te*, na 2PS, alternando entre ambas as formas tanto em função de objeto direto (acusativa) – p.ex., *vou sempre lhe amar* x *vou sempre te amar* –, quanto em função dativa – p.ex., *eu já lhe dei o dinheiro* x *eu já te dei o dinheiro*. O caráter dialetal dessa variação, por sua vez, é discutido desde a década de 90 por Ramos (1999), que argumenta para a existência de três gramáticas quanto aos usos dos clíticos em 2PS:

a) usa-se *você* como pronome pessoal de 2PS, o *lhe* como clítico para relações de respeito e o *te* em contextos familiares e informais – gramática do eixo Rio-São Paulo;
b) utiliza-se *você* como pronome de 2PS e o clítico *lhe* como substituto à forma *te* – gramática do português falado em capitais do Nordeste: Maceió, Recife, Salvador e João Pessoa;
c) há distinção *tu* e *você*, aquela no tratamento íntimo/familiar e esta no tratamento respeitoso; os clíticos *te* e *lhe* seguem a mesma distinção, respectivamente – gramática dos estados do Norte e do Maranhão.

Scherre e Duarte (2016) argumentam que o pronome *te* ainda é consistentemente utilizado no Brasil independentemente da região geográfica, enquanto *lhe* tem aparecido em variação com *te*, processo iniciado na região Nordeste, sendo também encontrado em outros locais. Há, então, regiões nas quais falantes fazem uso de *te*, e há regiões nas quais os falantes fazem uso de *lhe*, mas também há regiões nas quais ambas as formas coexistem, com maior predomínio (frequência de uso) de uma.

A entrada do pronome *você* no paradigma de 2PS ainda resulta no uso do pronome possessivo *seu*, originalmente de 3P, em referência à 2PS – *você viu sua irmã?* –, contrastando com o possessivo de 2PS *teu* – *você viu tua irmã?*. Segundo Lopes (2008), em variedades que optam pelo *você*, há a preservação da forma de 3PS *seu*, ao passo que o grupo social em que *tu* é mais produtivo tende a fazer maior uso de *teu*. Nesse sentido, *teu* se torna uma forma reservada a comunidades nas quais ainda há predomínio para a variante *tu*, mas ainda há regiões nos quais falantes fazem uso de ambas as formas, com diferença nas frequências.

A variação dialetal na morfossintaxe, contudo, não é restrita ao paradigma pronominal de 2PS. É o caso do uso variável de artigo antes de possessivos pré-nominais. No PB, os possessivos que precedem nomes – também chamados de possessivos pré-nominais –, em *trouxe as suas coisas*, podem vir antecedidos por artigos definidos à sua esquerda, em *a nossa escola*, ou vir sem ele, em *nossa escola*. Lucchesi (1993, p. 91), na década de 90, pontuou que diferentes comunidades geograficamente localizadas apresentam diferentes padrões de uso do artigo antes de possessivo: “em Portugal e no Sul do Brasil, o artigo definido e o possessivo coocorrem normalmente, enquanto no Norte e Nordeste do Brasil o artigo é

normalmente ausente".[7] Em outro anterior (Silva, 2020), já destacamos que, embora a variável seja, de fato, uma variável dialetalmente distinta, uma associação para um dialeto específico é vaga, pois falantes tendem a usar ambas as variantes (ausência/presença) em uma mesma interação.

Todas essas variáveis apontam, em algum nível, significado dialetal, dado que seus usos podem ser associados a determinadas regiões geográficas ou comunidades linguísticas específicas. É necessário, contudo, a consideração de mais estudos. Na seção que segue, buscamos observar o comportamento dessas variáveis mais detalhadamente por meio de uma revisão integrativa, que pode nos ajudar a identificar tendências nos padrões dialetais a partir de uma gama de estudos prévios sobre o PB. Nesse sentido, descrevemos sua distribuição no PB de modo a obter pistas indiretas para a existência de um significado dialetal, como também com vistas a observar se os fenômenos são dialetalmente distintos. Consideramos, para tanto, as frequências percentuais para observar a diferenciação dialetal.

2.2 MAPEANDO FENÔMENOS MORFOSSINTÁTICOS DIALETAIS

Para a busca e seleção dos trabalhos da revisão integrativa[8] para a observação da distribuição dos fenômenos morfossintáticos, utilizamos o indexador *Google Scholar* – por possuir uma ampla disponibilidade de material –, com auxílio do *Publish or Perish* (Harzing, 2007), um *software* que recupera e analisa produções acadêmicas. Ele usa uma variedade de fontes de dados, como *Google Scholar* e *CrossReference*, para obter as citações brutas, analisá-las e apresentar uma série de métricas de citação, incluindo o número de artigos, o total de citações e o índice h (Harzing, 2007).

Na busca, analisamos pesquisas que se encontravam até o resultado de número 100, sem delimitação de período temporal e de tipo de trabalho, podendo incluir teses, dissertações, artigos, anais etc., desde que digitalmente disponíveis. Os critérios de inclusão para todos os fenômenos foram: (i) ser um estudo de produção, que use como base dados espontâneos de fala; e (ii) ser com base na variedade brasileira do português. Excluímos, para tanto, pesquisas feitas com base na metodologia da Dialetologia, como pesquisas que usam dados do ALiB, uma vez que consideramos que os dados produzidos são resultados do uso de questionários

[7] No original: "in Portugal and the south of Brazil, the definite article and the possessive normally co-occur, whereas in northern and northeast Brazil the article is normally absent".
[8] A revisão foi conduzida no segundo semestre de 2022. É possível que outros trabalhos sobre os fenômenos tenham sido feitos.

com perguntas-gatilho (Comitê Nacional do Projeto Alib, 2001), com vistas direta na variável-alvo, não sendo dados espontâneos.

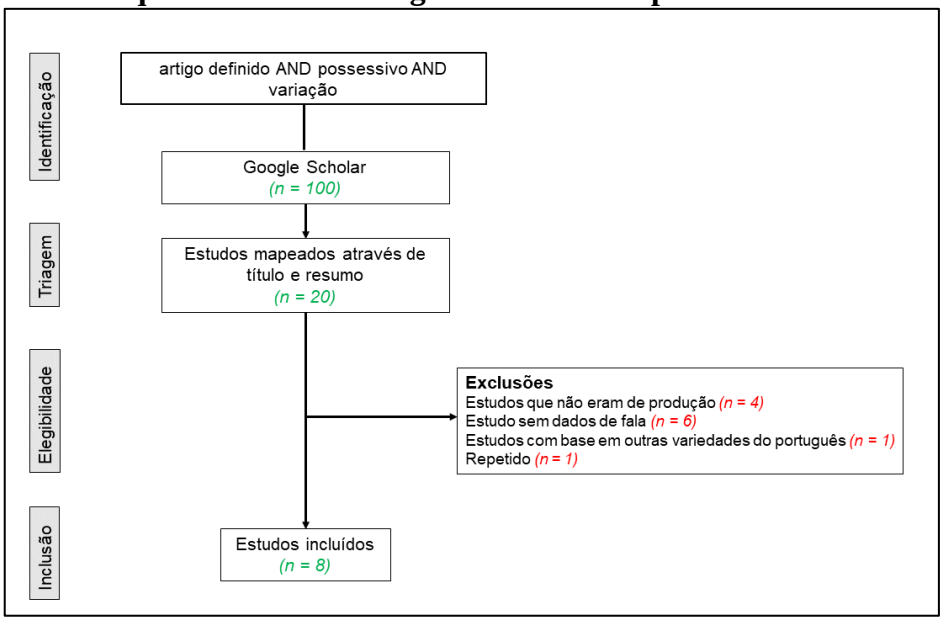
Os resultados de busca foram sistematizados em uma planilha para facilitar a seleção dos trabalhos considerando os critérios de inclusão e exclusão. Uma vez que os fenômenos selecionados, como informado no capítulo introdutório desta tese, já haviam sido descritos em outras pesquisas com base no português universitário da UFS, trabalhos anteriormente utilizados por essas pesquisas foram acrescidos, conforme informado no corpo da revisão.

Nas subseções seguintes, apresentamos as revisões considerando (i) uso variável de artigo antes de possessivos pré-nominais; (ii) pronomes pessoais de 2PS; (iii) pronomes clíticos de 2PS; e (iv) pronomes possessivos de 2PS.

2.2.1 Artigo antes de possessivos no português

Para a busca desse fenômeno, utilizamos as palavras-chave *artigo definido, possessivo* e *variação*, todas com o Operador Booleano AND (Figura 3). Operadores Booleanos são palavras que informam ao sistema de busca, como o *Google Scholar*, como combinar os termos de sua pesquisa, são eles: AND (e), OR (ou) e NOT (não). A utilização de AND informa que é necessário que todas as palavras ocorram no resultado.

Figura 3 – Esquema de busca artigo definido AND possessivo AND variação



Fonte: elaboração própria.

A busca retornou pesquisas derivadas de outras pesquisas, que utilizam os mesmos dados: o trabalho de Sedrins, Pereira e Silva (2017) é derivado do trabalho de Pereira (2017); os trabalhos de Siqueira (2021) e Siqueira e Freitag (2022) são recortes da pesquisa de Silva (2020). Apenas os dados das pesquisas originais são apresentados. Ainda, tínhamos em mãos os trabalhos de Callou e Silva (1997), Silva (1982; 1998a; 1998b) e Guedes (2019), incluídos em nossa revisão para incrementar a visualização da distribuição. Um sumário sobre as informações das pesquisas é visualizado no Quadro 1.

Quadro 1 – Pesquisas sobre a variação no uso de artigo definido antes de possessivos no português brasileiro

Autores	**Amostras**	**Condicionantes (valor: ausência)**
Silva (1982)	Entrevistas com três alunos de curso superior e um do então segundo grau e cinco jovens semialfabetizados que faziam parte do Movimento Brasileiro de Alfabetização (MOBRAL)	**Variáveis linguísticas:** possessivo *seu* como impulsionador da ausência; *especificidade* e *sintagmas preposicionados* inibidores da ausência.
Callou e Silva (1997)	Inquéritos gravados na década de 70 pelo projeto Norma Urbana Culta (NURC), com falantes de ambos sexos, nascidos nas capitais de São Paulo, Rio de Janeiro, Rio Grande do Sul, Bahia e Pernambuco, com ensino superior completo, distribuídos em três faixas etárias (25 a 35 anos; 36 a 55 e acima de 56), gravados em aulas e conferências, diálogos informais e entrevistas.	**Variáveis linguísticas**: em *sintagma nominal* e *preposicionado* predominam a presença; na natureza do possuído, a ausência ocorre com *parentes* (56%).
Silva (1998a; 1998b)	Entrevistas coletas entre os anos 1980 e 1983, em diversos bairros do Rio de Janeiro, estratificado em sexo (masculino e feminino), faixa etária (15-25 anos, 26-49, e acima de 50 anos) e escolaridade (1º e 2º ciclos do ensino fundamental e ensino médio), sendo representativo da variedade popular, visto que não conta com informantes do nível superior.	**Variáveis linguísticas**: ausência é frequente com *elemento não novo* (1637/2615 = 63%); *parentes* (1467/2223 = 66%); *relações humanas* (121/202 = 60%); e *possuído não-inerente* (261/478 = 55%). **Variáveis sociais**: ausência é frequente nas escolaridades *primário* (892/1404 = 64%) e *ginásio* (764/1205 = 63%); faixas etárias *7-14 anos* (541/771 = 70%) e *15-25 anos* (610/965 = 64%); sexo *masculino* (862/1324 = 52%) e feminino (1160/1939 = 60%); atuação da mídia *média* (742/1216 = 61%) e *fraca* (217/341 = 64%); atuação do mercado ocupacional *média* (759/1315 = 58%) e *fraca* (420/732 = 57%)
Campos Jr. (2012)	20 entrevistas extraídas da amostra PortVix, com falantes nascidos e residentes em Vitória-ES, estratificados em gênero (masculino e feminino), faixa etária (7-14 anos, 15-25, 26-49, e acima de 50 anos), escolaridade (fundamental, médio e universitário).	**Variáveis linguísticas**: ausência frequente em preposição *com* (43/62 = 69%); *3PS (seu/sua)* (11/13 = 85%), *reduplicado (mamãe/papai)* (54/59 = 92%), *2PS (seu/sua)* (53/87 = 61%); *parentesco* (489/644 = 76%), *não-*

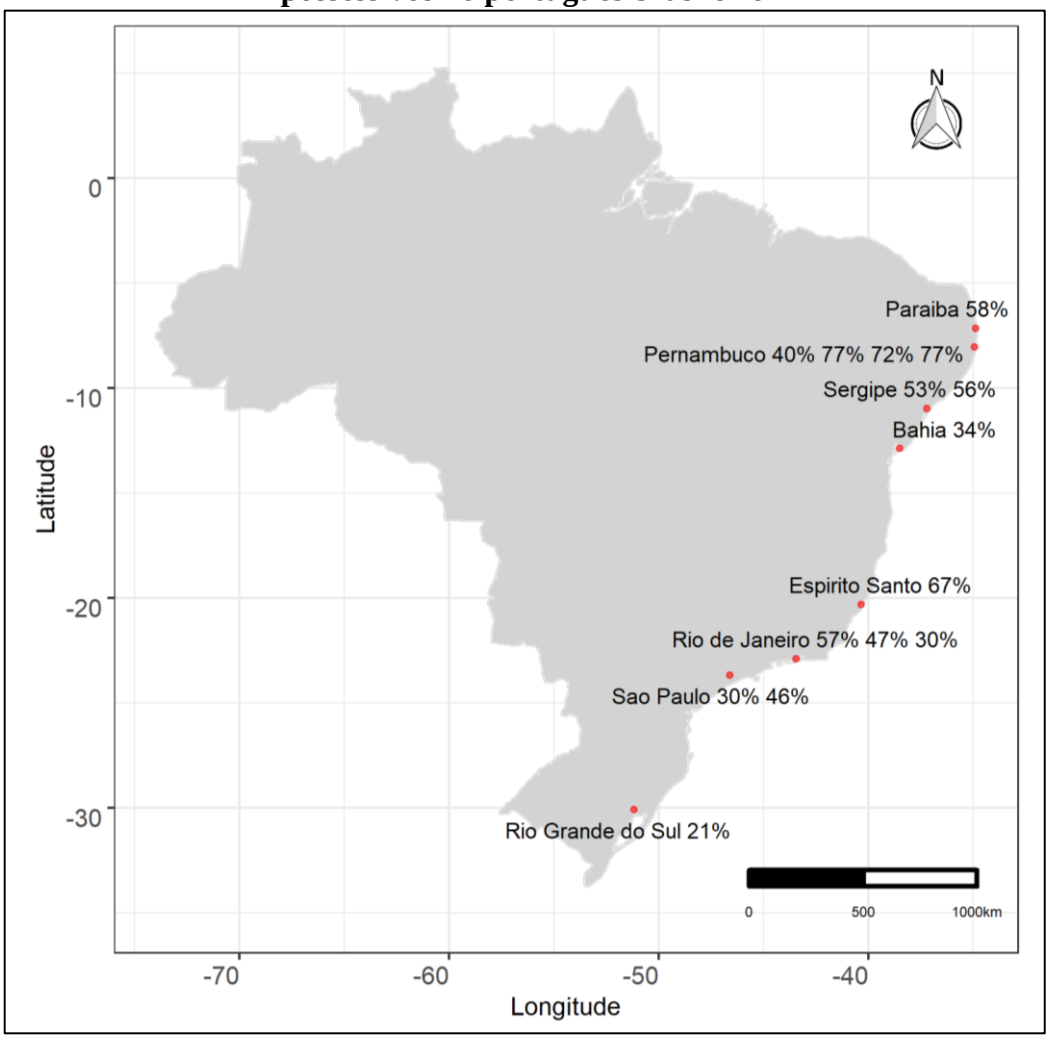
		parente (50/83 = 60%), *partes do corpo* (14/23 = 61%), *abstrações não-únicas* (48/88 = 55%), *objetos não-inerentes* (29/57 = 51%) e *abstrações únicas* (30/59 = 51%). **Variáveis sociais:** ausência frequente em *ensino fundamental* (367/507 = 72%), *ensino médio* (158/256 = 62%), *ensino universitário* (160/253 = 63%).
Siqueira (2014)	12 entrevistas sociolinguísticas com falantes de Serra Talhada-PE, estratificadas em sexo (feminino e masculino) e faixa etária (10 anos, de 20 a 39 anos e acima de 55 anos).	**Variáveis linguísticas**: ausência frequente em *sintagmas nominais* (91,8% 201/219).
Pereira (2017) e Sedrins, Pereira e Silva (2017)	24 entrevistas informais em Serra Talhada-PE e 24 entrevistas informais em Carnaíba (PE), com duração mínima de 10min cada, estratificadas em sexo (masculino e feminino), faixa etária (6-17 anos, 18-35, e acima de 35 anos) e escolaridade (fundamental, médio e superior).	**Variáveis linguísticas:** ausência frequente em funções sintáticas de *sujeito* (63/76 = 83%), *antitópico* (33/37 = 89%), *tópico* (26/41 = 63%), *objeto direto* (24/35 = 67%), *objeto indireto* (18/33 = 54,5%), *adjunto de nome* (67/89 = 76%) e *predicativo* (22/26 = 85%); preposição *com* (47/63 = 75%) e *por* (1/1 = 100%).
Guedes (2019)	8 informantes paraibanos migrantes (Amostra PBSP) estratificados em sexo (masculino e feminino), faixa etária (15-34 anos, 35-49, e acima de 50 anos), e escolaridade (até ensino médio e superior); 12 informantes paraibanos (Amostra PB) e 12 informantes paulistanos (Amostra SP), seguindo a mesma estratificação.	**Variáveis linguísticas:** ausência frequente em *sintagmas nominais* (63%); gênero do possessivo *masculino* (58%); funções de *objeto direto* (69%), *predicativo* (67%), *sujeito* (57%) e *tópico* (52%).
Siqueira (2020)	32 entrevistas extraídas da amostra Deslocamentos (2019), estratificadas em gênero (masculino e feminino), tempo no curso (do 3º período para baixo e do 7º para cima) e deslocamento (naturais e residentes de Aracaju que vão e voltam para a UFS todo dia; naturais e residentes do interior de Sergipe que vão e voltam para a UFS todo dia; naturais do interior de Sergipe que se mudaram para a Grande Aracaju; e naturais de outros estados que se mudaram para a Grande Aracaju).	**Variáveis linguísticas:** *sintagma nominal* (71,3%); traços semânticos de *humano* (65,8%); contexto de *dado* (55,1%). **Variáveis sociais**: falantes do Deslocamento 3 (55,4%), Deslocamento 1 (52,6%), Deslocamento 2 (51,7%) Deslocamento 4 (51,4%); alunos do *início* do curso (57,6%).
Silva (2020), Siqueira (2021) e Siqueira e Freitag (2022)	60 entrevistas extraídas da amostra Deslocamentos (2020), estratificadas em gênero (masculino e feminino), tempo no curso (do 4º período para baixo e do 5º para cima) e deslocamento (naturais e residentes de Aracaju que vão e voltam para a UFS todo dia; naturais e residentes do interior de Sergipe que vão e voltam para a UFS todo dia; naturais do interior de Sergipe que se mudaram para a Grande Aracaju; e naturais de Alagoas e Bahia que se mudaram para a Grande Aracaju).	**Variáveis linguísticas:** *possessivo no masculino* (60,2% 676/1123) *e feminino* (52,6% 633/1203), possessivo no *plural* (62,9% 299/364) e *singular* (55% 1080/1962), *sintagma nominal* (76,9% 1117/1453), preposição que *não contrai* (74,9% 143/191), funções sintáticas de *SN isolado* (86,4% 51/59), *sujeito* (80,4% 633/787), *acusativo* (72,3% 240/332), *tópico* (67,6% 119/176) e *predicativo* (57% 49/86).

		Variáveis sociais: falantes do Deslocamento 3 (63% 324/514) Deslocamento 1 (62,7% 315/502), Deslocamento 2 (56,8% 210/370); e Alagoas (53,9% 253/469); *início* do curso (58,4% 652/1116) e *final* (54,3% 657/1210).

Fonte: elaboração própria.

Os trabalhos selecionados não abarcam todas as regiões do país, mas o Nordeste (Callou; Silva, 1997; Siqueira, 2014; Pereira, 2017; Sedrins; Pereira; Silva, 2017; Guedes, 2019; Siqueira, 2020; 2021; Silva, 2020; Siqueira; Freitag, 2022), Sudeste (Silva, 1982; 1998a; 1998b; Campos Jr., 2012; Guedes, 2019) e Sul (Callou; Silva, 1997). Os dados na Figura 4 são apresentados considerando a ausência de artigo definido.

Figura 4 – Distribuição da ausência de artigo definido (vs. presença) antes de possessivos no português brasileiro



Fonte: elaboração própria.

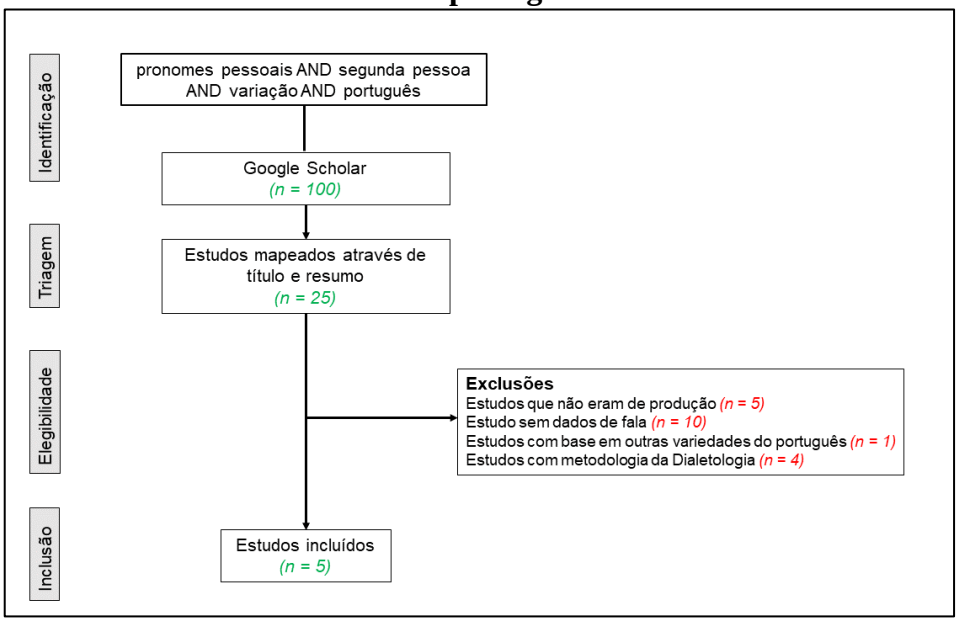
Na região Nordeste, os dados referentes à Paraíba são os de Guedes (2019); em Pernambuco, os dados são, respectivamente, de Callou e Silva (1997), Siqueira (2014) e Pereira (2017); Sergipe apresenta dados de Siqueira (2020) e Silva (2020); Bahia, Callou e Silva (1997). No Sudeste, os dados do Espírito Santo são de Campos Jr. (2012); em São Paulo, Callou e Silva (1997) e Guedes (2019), respectivamente; no Rio de Janeiro, de Silva (1982), Silva (1998a; 1998b) e Callou e Silva (1997). O único dado da região Sul, no Rio Grande do Sul, é de Callou e Silva (1997).

Na região Nordeste do Brasil, há maior frequência para a ausência de artigo antes do contexto observado. Com exceção dos dados de Callou e Silva (1997), que utilizam dados do NURC para a análise – com entrevistas gravadas nas décadas de 70 e 90 –, todas as pesquisas no Nordeste evidenciam predomínio para sintagmas nos quais não há um artigo definido antes de possessivos pré-nominais. Por outro lado, no Sudeste, há predomínio da ausência nos dados do Espírito Santo (Campos Jr., 2012) e no Rio de Janeiro (Silva, 1982), o que evidencia um gradiente para a frequência da variação nas diferentes regiões do país. Esses resultados reforçam a hipótese de que a variação na presença/ausência de artigo definido antes de possessivos pré-nominais em PB é um fenômeno morfossintático dialetal, dadas as diferentes frequências para diferentes regiões.

2.2.2 Pronomes pessoais de 2PS no português

Para a busca por pesquisas sobre a variação nos pronomes pessoais de 2PS em posição de sujeito, utilizamos as palavras-chave *pronomes pessoais, segunda pessoa, variação* e *português*, todas com o Operador Booleano AND (Figura 5).

Figura 5 – Esquema de busca pronomes pessoais AND segunda pessoa AND variação AND português



Fonte: elaboração própria.

O fenômeno é um dos mais investigados no PB, mas o baixo número de pesquisas retornadas não corresponde à realidade sociolinguística da descrição do fenômeno, resultado, possivelmente, de as plataformas nas quais os trabalhos foram publicados não estarem indexadas ao Google Scholar ou os trabalhos serem apenas em versão impressa. As pesquisas são as de Oliveira (2007), Franceschini (2015), Alves (2015), Fleck e Simioni (2016) e Nascimento e Paim (2016), acrescidas às análises as pesquisas que já tínhamos de Nogueira (2013), Silva e Vitório (2017), Guimarães (2019) e Siqueira, Sousa e Rodrigues (2023), conforme Quadro 2.

Quadro 2 – Pesquisas sobre a variação nos pronomes pessoais de segunda pessoa no português brasileiro

Autores	**Amostras**	**Condicionantes (valor = você)**
Oliveira (2007)	Amostras naturais de fala espontânea das comunidades de Santo Antônio de Jesus e Poções, ambas na Bahia: 48 entrevistas, 12 entrevistas para cada comunidade e subdivididas em duas localidades do município: sede e Rural, estratificadas em sexo, escolaridade (analfabeto ou semianalfabeto), estada fora da comunidade (ausência ou não da comunidade por pelo menos seis meses) e a faixa etária (faixa 1 – 20 a 40, faixa 2 – 40 a 60 e faixa 3 – acima de 60 anos).	**Variáveis linguísticas**: pronome *você* com referencial *determinado* (75% 378/507) e *indeterminado* (99% 612/621); quando antecedido por *você na oração anterior* (99% 186/188). **Variáveis discursivas**: pronome *você* em interações com *membro da comunidade* (72% 226/315), *entrevistador* (76% 81/107) e *indivíduo de fora* (89% 71/80); efeito gatilho de *você – última forma empregada* (89% 516/577).

		Variáveis sociais: pronome *você* com falantes de *20 a 40 anos* (88% 299/340), de *41 a 60 anos* (91% 486/534) e *mais de 60 anos* (81% 205/254); falantes do gênero *masculino* (89% 610/687) e *feminino* (86% 380/441).
Nogueira (2013)	Dados de falas culta e popular da cidade de Feira de Santana e Salvador, ambas na Bahia. Para Feira de Santana, 24 entrevistas pertencentes ao banco de dados do Projeto A Língua Portuguesa no Semiárido Baiano, sendo 12 do português culto e 12 do popular. Para Salvador, 12 inquéritos do Projeto Norma Urbana Culta de Salvador – NURC/SSA e 12 inquéritos do Programa de Estudos sobre o Português Popular Falado em Salvador – PEPP. Todos os inquéritos são do tipo diálogo entre informante e documentador (DID).	**Variáveis linguísticas**: pronome *você* em função de *sujeito* (97,8% 1454/1502), *não sujeito* (80,4% 119/148) e *sem verbo* (90,5% 57/63); tipo de frase *declarativa* (95,4% 1541/1615) e *não declarativa* (90,8% 89/98); tempo verbal *passado* (98,6% 137/13) e *não passado* (94,85% 1493/1574); tipo de discurso *direto* (96,1% 1297/1349) e *relatado* (91,5% 333/364); tipo de referência *específica* (86% 512/595) e *genérica* (100% 1118/1118). **Variáveis sociais**: pronome *você* com falantes do sexo *masculino* (93,7% 726/775) e *feminino* (96,4% 904/938); falantes do *português culto* (95,7% 1033/1079) e do *português popular* (94,2% 597/634); faixas etárias Faixa I (92,7% 575/620), Faixa II (95,9% 662/690) e Faixa III (97,5% 393/403).
Franceschini (2015)	24 entrevistas de falantes de Concórdia-SC, coletadas entre os anos de 2007 e 2010, e distribuídas por duas faixas etárias (26 a 45 anos; 50 anos ou mais); sexo (masculino; feminino) e três níveis de escolaridade (fundamental I; fundamental II; ensino médio).	**Variáveis sociais**: pronome *você* com falantes do *ensino médio* (53% 300/569); do sexo *masculino* (50% 231/458); e de *26 a 45 anos* (50% 290/582).
Alves (2015)	15h 43min de gravação de interações livres de dois colaboradores com pessoas de suas redes sociais, naturais de São Luís-MA e que da cidade não tenham se afastado por mais de 1/3 de sua vida, com residência e trabalhos fixos na localidade objeto do estudo.	**Variáveis linguísticas**: pronome *você não 1ª da série, precedido de você* (75,4% 43/57) e *não 1ª da série, precedido de cê* (100% 18/18); referência *genérica* (48,1% 102/212). **Variáveis sociais:** não houve predomínio do pronome *você*.
Fleck e Simioni (2016)	Dados de fala de 24 falantes de Itaqui-RS, estratificados em sexo, idade e escolaridade.	**Variáveis sociais**: pronome *você* predominou entre os maiores de 25 anos e entre falantes do Ensino Fundamental.
Nascimento e Paim (2016)	24 inquéritos do português falado na Bahia: 12 inquéritos do Programa de Estudos do Português Popular Falado de Salvador (PEPP) e 12 inquéritos gravados em Amargosa no ano de 2016. Os informantes são homens e mulheres em igual número, distribuídos em três faixas etárias (I: 15 a 24	**Variáveis linguísticas**: pronome *você* predomina no *tempo passado* (92,7%); em *relato atual* (91,7%), *relato próprio* (71,4%) e em *relato de outrem* (94,7%); em enunciação *declarativa* (88,9%), *narrativa* (93,6%), *questionamento* (69,2%),

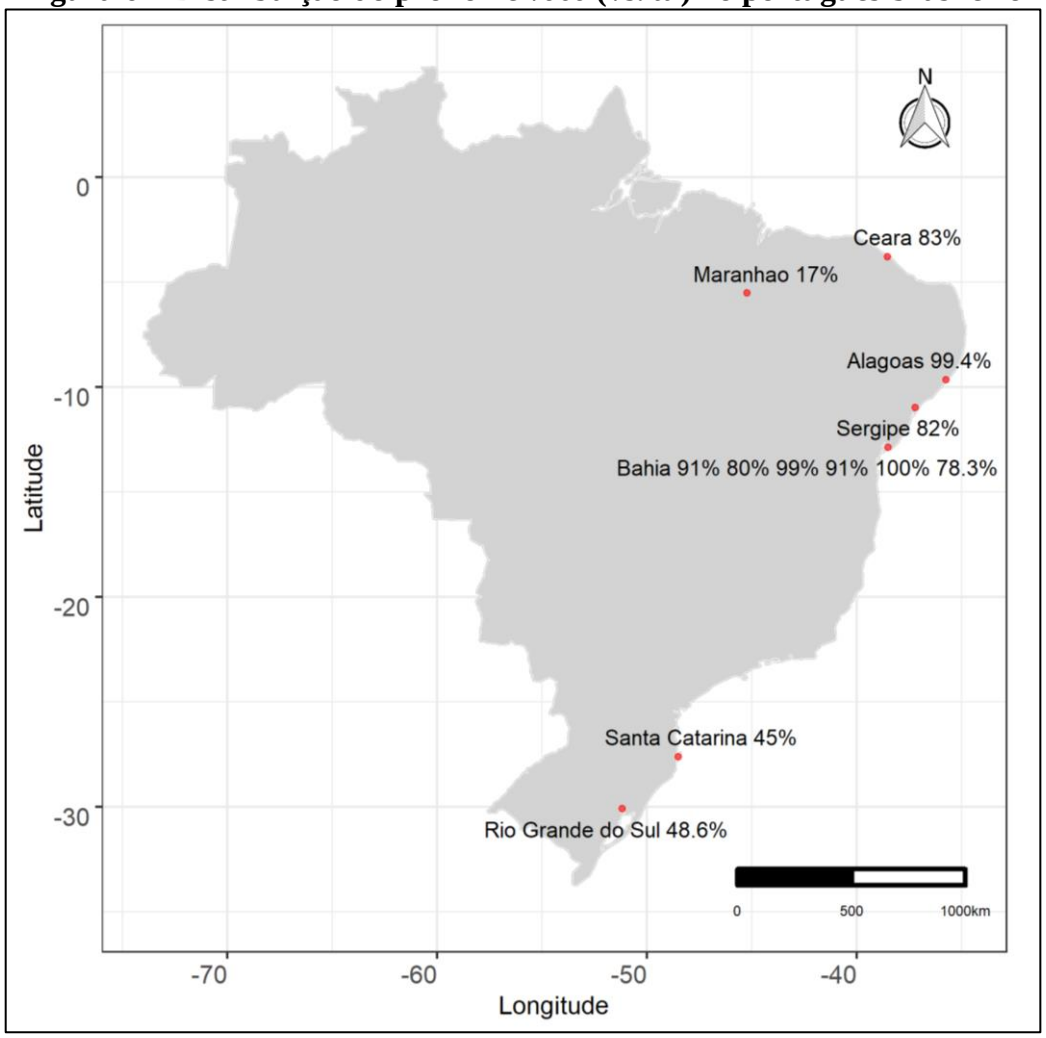
	anos; III: 45 a 55 anos; IV: 65 anos em diante).	de *advertência* (100%), *ordem* (100%) e *afirmação* (100%);
Silva e Vitório (2017)	96 entrevistas extraídas do banco de dados do projeto A Língua Usada no Sertão Alagoano (LUSA) com informantes naturais de cidades do sertão alagoano estratificadas de acordo com as variáveis sociais sexo/gênero (masculino e feminino), faixa etária (15-29 anos, 30-44 anos e acima de 44 anos) e escolaridade (analfabeto, ensino fundamental, ensino médio e ensino superior).	**Variáveis linguísticas**: pronome *você* predomina em *realização isolada* (91% 64/70), *primeiro da série* (94% 99/105), *antecedido por você* (95% 299/314) e *antecedido por cê* (68% 13/19). **Variáveis sociais**: pronome *você* predomina em falantes *analfabetos* (93% 38/41), do *ensino fundamental* (96% 81/84), *ensino médio* (86% 126/146) e *ensino superior* (97% 230/237).
Guimarães (2019)	18 inquéritos extraídos do PORCUFORT (Português Culto Falado em Fortaleza – CE), estratificados em sexo/gênero (homem e mulher) e faixa etária (22 a 35 anos; 36 a 55 anos; 56 anos em diante). Todos os inquéritos são do tipo D2 (Diálogo entre Dois Informantes).	**Variáveis linguísticas**: pronome *você* predomina em sequência discursivas *dialogal* (58,3%), *narrativa* (75%), *argumentativa* (89,6%) e *explicativa* (95,6%); em atuação da entonação *declarativa/exclamativa* (89,9%) e *interrogativa* (61,3%); tipo de referente *genérico* (97,3%) e *específico* (76,1%); em interações *simétricas* (63,9%), *parcialmente assimétricas* (89%) e *totalmente assimétricas* (85,2%); no paralelismo formal, em contexto *isolado* (77,7%), *primeiro da série* (89,4%), *você precedido de cê* (92,3%) e *você precedido de você* (95,9%). **Variáveis sociais**: pronome *você* predomina em todas as faixas etárias e em ambos os gêneros
Siqueira, Sousa e Rodrigues (2023)	60 entrevistas sociolinguísticas extraídas da amostra Deslocamentos (2020), com universitários da Universidade Federal de Sergipe do campus Prof. José Aloísio de Campos, estratificados em sexo (masculino e feminino), tempo no curso (4º período para baixo e 5º período para cima) e deslocamento (naturais e residentes de Aracaju que vão e voltam para a UFS todo dia; naturais e residentes do interior de Sergipe que vão e voltam para a UFS todo dia; naturais do interior de Sergipe que se mudaram para a Grande Aracaju; e naturais de Alagoas e Bahia que se mudaram para a Grande Aracaju).	**Variáveis sociais**: pronome *você* predomina em todos os perfis de Deslocamento: Deslocamento 1 (87,2% 572/656), Deslocamento 2 (73,9% 391/529), Deslocamento 3 (86,2% 274/318), Alagoas (78,9% 377/478) e Bahia (87,9% 400/455).

Fonte: elaboração própria.[9]

[9] Uma vez que a maioria das pesquisas lida apenas com *você* e *tu*, para a apresentação da distribuição, amalgamamos as reduções fonéticas de *você* (*ocê/cê*) à variante *você* e calculamos o percentual apenas com base nas variantes *você* e *tu*. No caso da pesquisa de Silva e Vitório (2018), que só apresenta percentual de *você* e *cê*, a frequência foi calculada considerando as três ocorrências de *tu* mencionadas pelas autoras.

Os estudos se concentram majoritariamente no Nordeste e no Sul. No Nordeste, os estudos de Oliveira (2007), Nogueira (2013) e Nascimento e Paim (2016) apresentam dados da Bahia, Alves (2015) do Maranhão, Silva e Vitório (2017) de Alagoas, Siqueira, Sousa e Rodrigues (2023) de Sergipe, e Guimarães (2019) do Ceará; no Sul, os dados de Franceschini (2015) são de Santa Catarina, e os de Fleck e Simioni (2016) do Rio Grande do Sul (Figura 6).

Figura 6 – Distribuição do pronome *você* (vs. *tu*) no português brasileiro



Fonte: elaboração própria.

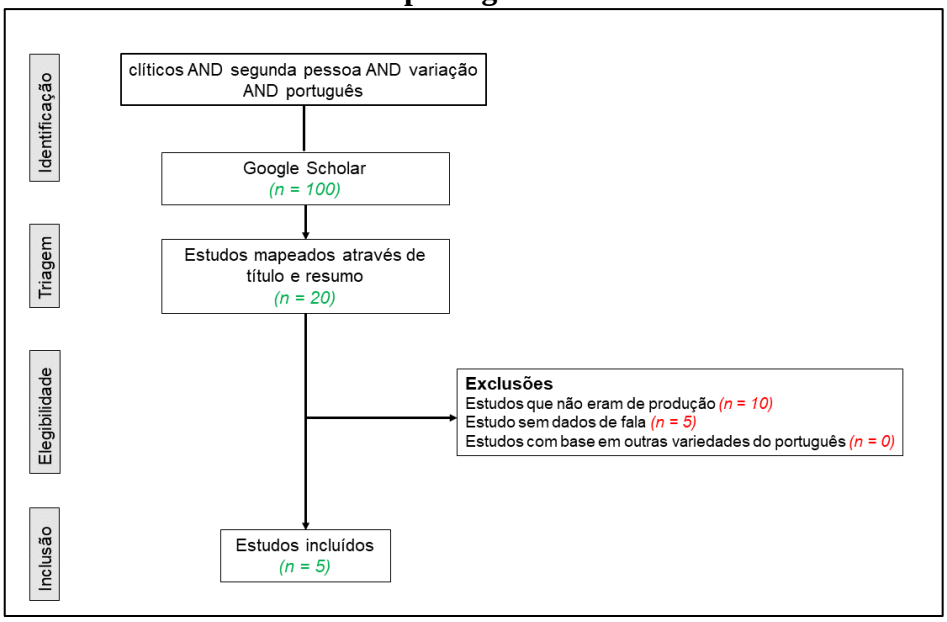
O único estado que apresenta mais de uma pesquisa é a Bahia, com dados, respectivamente, de Poções, Santo Antônio (Oliveira, 2007), Salvador, Feira de Santana (Nogueira, 2013), Salvador e Amargosa (Nascimento; Paim, 2016). Os dados do Nordeste, com exceção do Maranhão, apresentam alta frequência para o uso de *você* (<70%). No Sul, por outro lado, a distribuição é similar para ambas as variantes, com maior frequência para a variante *tu*. Os dados do Maranhão apresentaram a menor frequência para o *você*.

Há uma certa polaridade entre os dados do Nordeste e do Sul, em que no primeiro há tendência para o emprego do *você*, enquanto no segundo há tendência para o emprego de *tu*, com ambas as formas coexistindo em quase todos os dados. As frequências observadas na Figura 6 sugerem a existência de distinção dialetal do fenômeno, que apresenta diferentes taxas de uso a depender da região geográfica do falante. Os outros fenômenos variáveis que envolvem pronomes de 2PS podem seguir caminho similar, conforme vemos no que segue.

2.2.3 Clíticos de 2PS no português

Para a busca de pesquisas sobre a variação nos clíticos de 2PS, utilizamos as palavras-chave *clíticos, segunda pessoa, variação* e *português*, todas com o Operador Booleano AND (Figura 7).

Figura 7 – Esquema de busca clíticos AND segunda pessoa AND variação AND português



Fonte: elaboração própria.

Cinco estudos contemplam descrição do fenômeno considerando dados de produção de fala espontânea com base na variedade brasileira do português: Almeida (2009; 2014), Gama (2019), Bortoletto e Antonelli (2020) e Araújo e Borges (2021), acrescida a pesquisa de Dalto (2002), conforme Quadro 3.

Quadro 3 – Pesquisas sobre a variação nos clíticos de segunda pessoa no português brasileiro

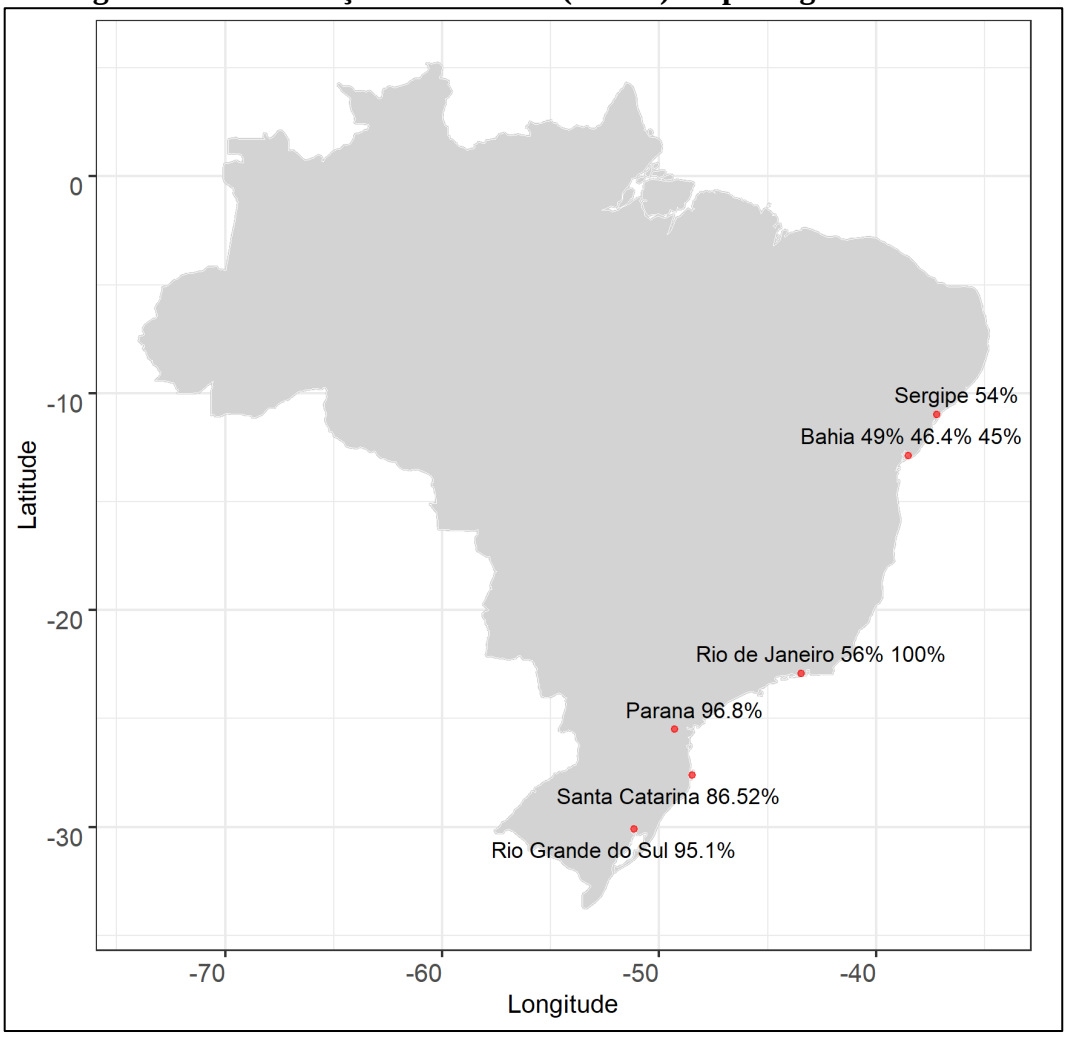
Autores	**Amostras**	**Condicionantes (valor = te)**
Dalto (2002)	Dados de fala de três capitais do Sul, Curitiba, Florianópolis e Porto Alegre, extraídas do Projeto VARSUL. A autora não explicita a quantidade utilizada por cidade.	A autora não faz distinção quanto ao tipo de pronome, mas apenas quanto à presença de objeto e objeto nulo, uma vez que esse era seu objetivo.
Almeida (2009)	36 informantes de Salvador-BA, estratificados em faixa etária (de 25 a 35 anos, de 45 a 55 anos e de 65 a 75 anos), escolaridade (Ensino Fundamental e Superior) e sexo (masculino e feminino).	Os dados são relativos apenas à função de objeto direto. **Variáveis linguísticas**: pronome *te* ocorre com *sujeito preenchido* (57% 193/341) e *não-preenchido* (70% 238/341); antecedido por *te* (94% 44/47); contexto *menos monitorado* (69% 227/330); e tipo de discurso *hipotético* (63% 167/264) e *real* (65% 251/384). **Variáveis sociais**: pronome *te* em faixa etária *1* (80% 50/523) e *2* (62% 163/264); sexo *feminino* (71% 249/350)
Almeida (2014)	36 informantes de Santo Antônio de Jesus-BA, estratificados em faixa etária (de 25 a 35 anos, de 45 a 55 anos e de 65 a 84 anos), escolaridade (Ensino Fundamental e Superior) e sexo (masculino e feminino).	**Variáveis linguísticas**: pronome *te* antecedido por *te* (90% 61/68); antecedido pela forma de tratamento subjetiva *tu* (80% 16/20); e *formas não-paralelas* (94,1% 30/32). **Variáveis sociais**: pronome *te* ocorre na faixa etária *1* (56% 192/343); falantes do gênero *feminino* (48% 241/501); falantes do *ensino superior* (46,5% 243/522).
Gama (2019)	60 entrevistas, sendo 36 de fala espontânea e 24 de inquéritos com falantes de Feira de Santana-BA. As entrevistas de fala espontânea compõem o banco de dados do projeto de pesquisa A língua portuguesa falada no semiárido baiano; os inquéritos foram organizados a partir do modelo desenvolvido Almeida (2009).	Os dados são relativos apenas à função de objeto direto. **Variáveis sociais**: pronome *te* em faixa etária *2* (52% 90/173); escolaridade *culto* (49,8% 115/231).
Bortoletto e Antonelli (2020)	Dados de fala extraídos do Projeto Norma Linguística Urbana Culta - RJ (NURC-RJ), coletados na década de 70, e do *corpus* do grupo PEUL (Programa de Estudos sobre o Uso da Língua), mais especificamente a amostra Banco de Dados Internacionais, coletados na década de 90. Ambos no Rio de Janeiro-RJ.	Amostra de 70 **Variável social:** pronome *te* no grupo etário *19-25* (94,44% 17/18).
Araújo e Borges (2021)	54 entrevistas sociolinguísticas extraídas da amostra Deslocamentos (2018/UFS-Itabaiana), com universitários da Universidade Federal de Sergipe do *campus* Itabaiana-SE.	As autoras não fazem distinção quanto ao tipo de pronome, mas apenas quanto ao uso de pronome lexical, clítico pronominal e categoria vazia, uma vez que esse era seu objetivo.

Fonte: elaboração própria.[10]

[10] Algumas das pesquisas abordam mais de uma variante (ex: pronome *você*, posição *vazia* e mais). Para a apresentação da distribuição, calculamos o percentual apenas com base nas variantes *te* e *lhe*.

As pesquisas dispõem sobre as regiões Nordeste (Almeida, 2009; 2014; Gama, 2019; Araujo; Borges, 2021), Sudeste (Bortoletto; Antonelli, 2020) e Sul (Dalto, 2002), concentradas em poucos estados dessas regiões (com exceção do Sul, onde todos os estados são contemplados). A distribuição dos dados pode nos dar indícios do comportamento da variável no PB (Figura 8).

Figura 8 – Distribuição do clítico *te* (vs. *lhe*) no português brasileiro



Fonte: elaboração própria.

No mapa, os dados referentes à Bahia são, respectivamente, Almeida (2009), Almeida (2014) e Gama (2001); em Sergipe, os dados são de Araújo e Borges (2021); Rio de Janeiro apresenta dados de Bortoletto e Antonelli (2020). Todos os dados do Sul são de Gama (2002).

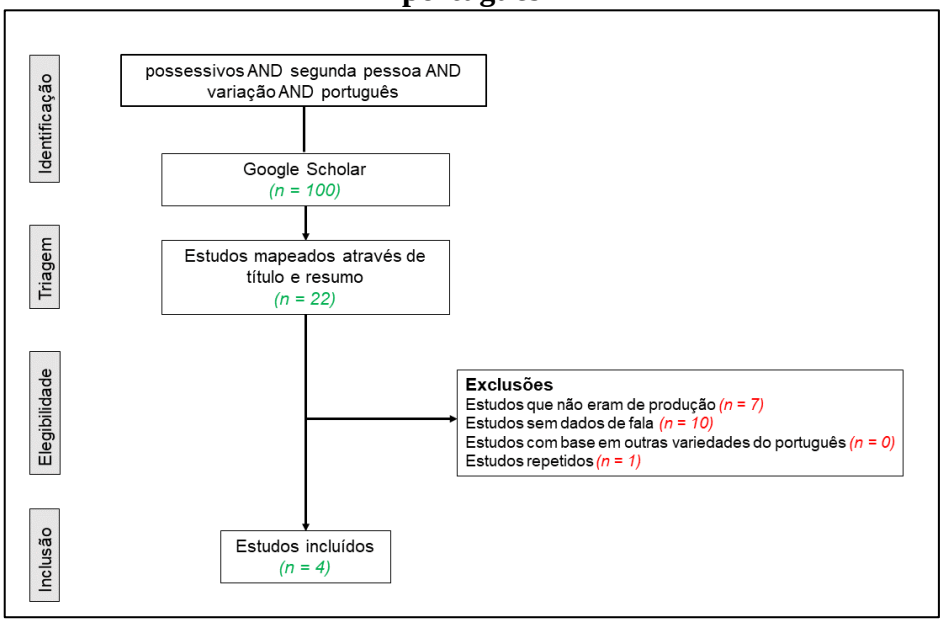
Os dados dispostos na Figura 8 apontam que ao Nordeste há menor uso do clítico *te*, oposto à alta frequência da forma ao Sul. Além disso, há uma drástica mudança na frequência entre os dados do Rio de Janeiro, que apresentam 56% do uso de *te* na década de 70 e 100%

na década de 90. A diferença observada entre os dados das diferentes regiões é uma evidência para o caráter dialetal da variação, uma vez que falantes de diferentes regiões parecem estar se comportando linguisticamente diferente para o uso dos clíticos, mas que, ainda assim, o pronome *te* segue consistentemente utilizado no Brasil independentemente da região geográfica, como apontam Scherre e Duarte (2016). Resta-nos considerar ainda a variação nos possessivos de 2PS.

2.2.4 Possessivos de 2PS no português

Para a busca de trabalhos sobre a variação nos possessivos de 2PS, utilizamos as palavras-chave *possessivos, segunda pessoa, variação* e *português*, todas com o Operador Booleano AND (Figura 9).

Figura 9 – Esquema de busca possessivos AND segunda pessoa AND variação AND português



Fonte: elaboração própria.

A busca resultou em apenas quatro pesquisas que se enquadram em nossos critérios: Arduin (2004; 2005), Mendes (2008) e Siqueira (2021). A pesquisa de Arduin (2004) é exploratória; por outro lado, os dados da pesquisa de Arduin (2005) são mais completos e usam a mesma amostra, por isso, são os apresentados. Há três (3) conjuntos de dados a serem apresentados, cujas informações são sintetizadas no Quadro 4.

Quadro 4 – Pesquisas sobre a variação nos possessivos de segunda pessoa no português brasileiro

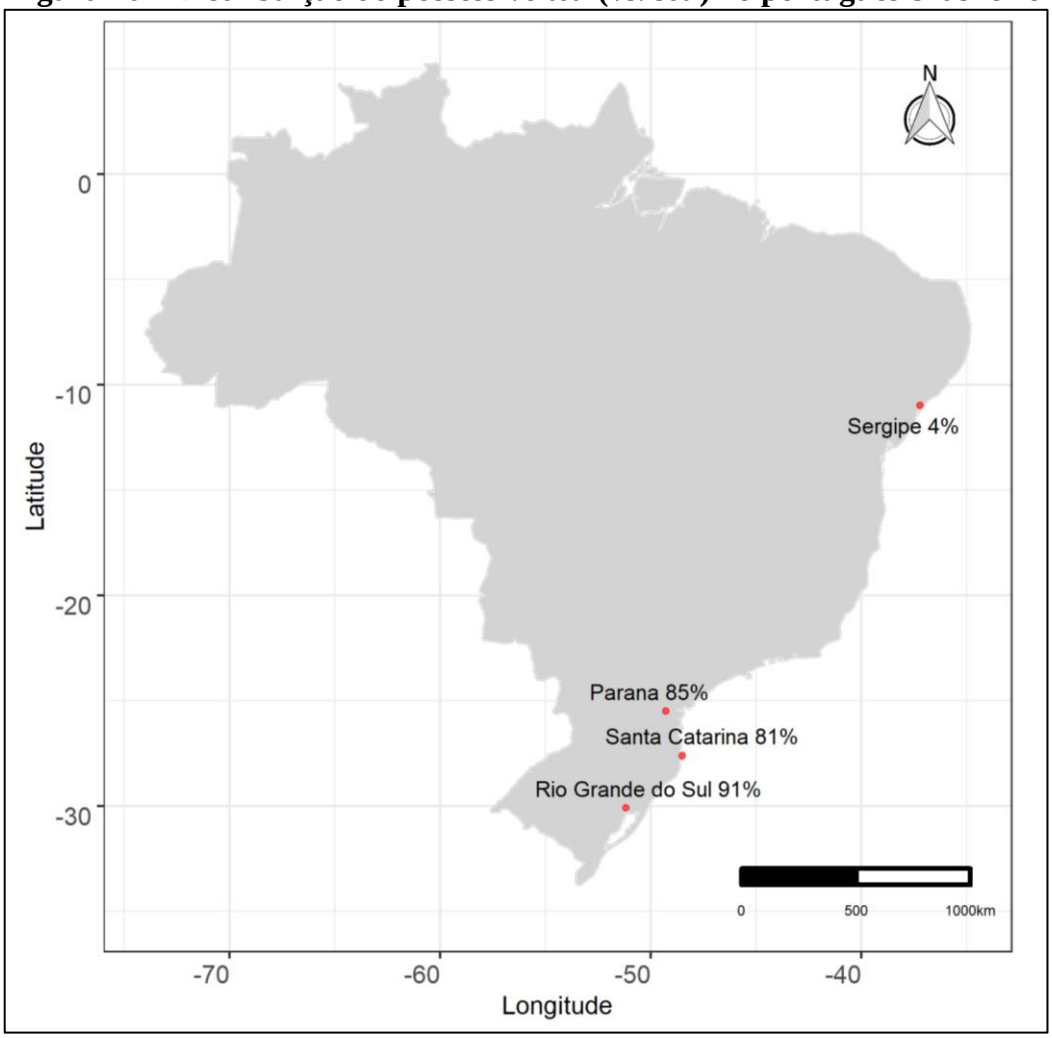
Autores	Amostras	Condicionantes (valor = teu)
Arduin (2004; 2005)	192 entrevistas extraídas do Banco de dados do projeto VARSUL, pertencentes às cidades catarinenses Florianópolis, Blumenau, Chapecó e Lages, e às cidades gaúchas Porto Alegre, Flores da Cunha, Panambi e São Borja, estratificadas considerando sexo (masculino e feminino), idade (25-49 anos e + de 50 anos), tempo de escolarização (até quatro anos, até oito anos e até doze anos) e região (as cidades citadas). Cada cidade é representada por 24 falantes.	**Variáveis linguísticas**: possessivo *teu* antecedido com sujeito tu (99% 143/144), *teu* com vocativo (86% 18/21), *teu* com sujeito nulo (77% 136/176) e *teu* com sujeito você (80% 59/74). **Variáveis estilísticas-discursivas:** possessivo *teu* em relação entre *superior* > *inferior* (91% 87/96) e *entre iguais* (91% 62/68); em pessoa a que se reporta, possessivo *teu* em *discurso de pessoa próxima* (96% 51/53), *discurso do próprio informante* (88% 67/76) e *discurso de pessoa não-próxima* (70% 42/60). **Variáveis sociais**: possessivo *teu* com falantes do gênero *feminino* (93% 229/245) e *masculino* (75% 127/170); falantes de *25 a 49 anos* (88% 177/200) e *mais de 50 anos* (83% 179/215); escolaridade *ginásio* (91% 106/116), *primário* (84% 137/163) e *colegial* (83% 113/136).
Mendes (2008)	32 entrevistas extraídas do Banco de dados do projeto VARSUL, pertencentes às cidades paranaenses de Curitiba, Irati, Londrina e Pato Branco, estratificadas considerando sexo (masculino e feminino), idade (25-49 anos e + de 50 anos), tempo de escolarização (até quatro anos, até oito anos e até doze anos) e região (as cidades citadas).	**Variáveis linguísticas**: possessivo *teu* antecedido por *tu* (100% 1/1) e *você* (85% 56/66); tipo de discurso *reportado* (88% 15/17) e *não reportado* (84% 42/50); pessoa do discurso reportado *não próxima* (78% 7/9), *falante* (100% 6/6) e *próxima* (100% 2/2); relação de *superior para inferior* (82% 40/49) e *igual para igual* (100% 17/17). **Variáveis sociais:** possessivo *teu* com falantes de *25 a 49 anos* (85% 29/34) e *mais de 50 anos* (85% 28/33); falantes do gênero *masculino* (88% 15/17) e *feminino* (84% 42/50).
Siqueira (2021)	60 entrevistas sociolinguísticas extraídas da amostra Deslocamentos (2020), com universitários da Universidade Federal de Sergipe do *campus* Prof. José Aloísio de Campos, estratificados em sexo (masculino e feminino), tempo no curso (4º período para baixo e 5º período para cima) e deslocamento (naturais e residentes de Aracaju que vão e voltam para a UFS todo dia; naturais e residentes do interior de Sergipe que vão e voltam para a UFS todo dia; naturais do interior de Sergipe que se mudaram para a Grande Aracaju; e naturais de Alagoas e	O autor não apresenta resultado de variáveis linguísticas. A única variável social apresentada não foi estatisticamente significativa.

	Bahia que se mudaram para a Grande Aracaju).	

Fonte: elaboração própria.

Neste fenômeno, as pesquisas são restritas à região Sul, os três estados, e ao Nordeste, apenas Sergipe. Os dados são apresentados considerando o nível *teu* (Figura 10).

Figura 10 – Distribuição do possessivo *teu* (vs. *seu*) no português brasileiro



Fonte: elaboração própria.

Ao Sul há maior a frequência de uso da variante *teu*, vide os dados do Paraná, Santa Catarina e Rio Grande do Sul, que apresentam frequências maiores do que 81%, enquanto em Sergipe a frequência é de apenas 4%. O uso de *teu* é quase inexistente na fala de universitários do estado de Sergipe que compõem a amostra, prevalecendo o uso da forma *seu*. Esse resultado, provavelmente, é reflexo do que pontua Lopes (2008): em variedades que utilizam *você* há a preservação da forma de 3PS *seu*; e em variedades que utilizam *tu* há preservação de *teu*. É necessário reconhecer, contudo, os poucos trabalhos encontrados sobre o fenômeno

no PB. Uma vez que o uso de *teu* se correlaciona ao uso da variante *tu*, e que o uso desta variante é dialetalmente distinto, endentemos que a distribuição observada na Figura 10 seja efeito disso.

2.3 AGREGANDO VARIÁVEIS MORFOSSINTÁTICAS

Neste capítulo, vimos que dialetos da língua são formados por um conjunto de fenômenos nos diferentes níveis linguísticos, mas que a atribuição de significado dialetal comumente é feita a partir de traços suprassegmentais da língua. Todavia, fenômenos morfossintáticos variáveis também podem ser dialetalmente distintos/salientes, uma vez que pode haver diferentes padrões de uso a depender da região de origem dos grupos de falantes e que a exposição a diferenças nos usos leva ao reconhecimento da variação. Frente a isso, a revisão integrativa conduzida na seção anterior nos permite observar algumas informações em relação aos fenômenos selecionados (Quadro 5).

Quadro 5 – Sistematização da revisão

Fenômeno	**Dialetal**	**Tipo de distribuição**	**Gradação**
Uso variável de artigo antes de possessivo	Sim	Binária (ausência/presença).	Sim, nenhuma pesquisa apresentou categoricidade.
Pronomes pessoais de 2PS	Sim	Quaternária (tu/você/cê/ocê). *Ocê* tem uso restrito.	Sim, apenas uma pesquisa apresentou categoricidade.
Pronomes clíticos de 2PS	Sim	Binária (te/lhe).	Sim, apenas uma pesquisa apresentou categoricidade.
Pronomes possessivos de 2PS	Sim	Binária (teu/seu).	Sim, nenhuma pesquisa apresentou categoricidade.

Fonte: elaboração própria

Temos quatro variáveis linguísticas que são dialetalmente distintas, o que pode indicar, em algum nível, existência de um significado dialetal para a variação: seus usos apresentam diferentes padrões a depender da região geográfica do falante, com diferentes variantes coexistindo nas comunidades descritas. Isso implica que, individualmente, a associação de apenas uma das variáveis morfossintáticas a um dialeto é insuficiente para sua descrição, uma vez que a informação dialetal das variáveis é relativa e que diferentes comunidades podem fazer uso das formas variáveis de forma similar, além de que dialetos da língua são formados por não só um traço linguístico, mas por vários. É necessário, então, que trilhemos um caminho no qual mais de uma variável seja considerada.

Existem, por exemplo, falantes que podem fazer uso de *você+lhe*, como também existem falantes que podem fazer uso de *você+te*, *tu+te*, *cê+lhe* e assim sucessivamente, existindo, pelo menos, 8 diferentes possibilidades para o uso dos pronomes pessoais (tu/você/cê/ocê) com os pronomes clíticos (te/lhe). A quantidade de possibilidades aumenta à medida que novas variáveis são inseridas. Se inserirmos os possessivos de 2PS (teu/seu), temos 16 possibilidades de uso; acrescendo o uso variável de artigo antes de possessivo (ausência/presença), temos, então, 32 possibilidades, conforme Quadro 6.

Quadro 6 – Possibilidades de uso das variantes

Pronomes pessoais de 2PS	Clíticos de 2PS	Possessivos de 2PS	Uso de artigo antes de possessivo
tu	te	seu	ausência
tu	te	seu	presença
tu	te	teu	ausência
tu	te	teu	presença
tu	lhe	seu	ausência
tu	lhe	seu	presença
tu	lhe	teu	ausência
tu	lhe	teu	presença
você	te	seu	ausência
você	te	seu	presença
você	te	teu	ausência
você	te	teu	presença
você	lhe	seu	ausência
você	lhe	seu	presença
você	lhe	teu	ausência
você	lhe	teu	presença
cê	te	seu	ausência
cê	te	seu	presença
cê	te	teu	ausência
cê	te	teu	presença
cê	lhe	seu	ausência
cê	lhe	seu	presença
cê	lhe	teu	ausência
cê	lhe	teu	presença
ocê	te	seu	ausência
ocê	te	seu	presença
ocê	te	teu	ausência
ocê	te	teu	presença
ocê	lhe	seu	ausência
ocê	lhe	seu	presença
ocê	lhe	teu	ausência
ocê	lhe	teu	presença

Fonte: elaboração própria.

Cada conjunto das 32 possibilidades ainda pode ter arranjos específicos a depender do sexo/gênero, escolarização, idade, zona de residência e outros atributos sociais dos falantes, aumentando as possibilidades de uso, dado que características extralinguísticas tendem a interferir

no uso efetivo da língua. Por exemplo, se considerarmos dois gêneros (masculino e feminino), quatro escolaridades (analfabeto, fundamental completo, médio completo e superior completo), três faixas etárias (15 a 30 anos, 31 a 45 anos e acima de 45 anos) e duas zonas de residências (zona rural e zona urbana), haverá 48 arranjos, que, multiplicados pelas 32 possibilidades combinatórias, resultariam em 1536 arranjos entre perfil do falante e uso linguístico das quatro variantes. Uma decisão metodológica para a redução desses arranjos é a de restringir a um grupo menor e relativamente homogêneo quanto a idade e escolarização, o que é possibilitado por nossas amostras – apresentadas no capítulo seguinte –, dado que permite uma melhor compreensão da utilização de padrões conjuntos de uso por perfis sociais específicos.

Além disso, a observação de padrões mais específicos de uso, a partir de uso conjunto de variantes linguísticas, pode agregar mais evidências em relação à presunção de origem do falante. Consideremos, contudo, que contamos com a fala de estudantes universitários de diferentes regiões geográficas, que estão constantemente em processo de interação com pessoas que podem ser falantes de outros dialetos. Expostos a variantes morfossintáticas distintas das suas a partir do contato linguístico, é possível que haja mudança linguística, já que, como discorre Ribeiro (2019, p. 27), o contato "promove variação e mudança de modo a satisfazer as necessidades de convívio linguístico, ou seja, para que os indivíduos possam se inserir em um local comum (região de contato), eles precisam se adequar uns aos outros". Nesse sentido, o padrão de fala do indivíduo pode não mais ser tão similar ao de sua região de origem, mas ao de falantes em sua nova comunidade.

Para uma compreensão mais ampla de como os dialetos são organizados, em termos morfossintáticos, é importante considerar não só uma variável, mas várias variáveis, descrevendo como elas interagem entre si. O agrupamento de traços morfossintáticos dialetalmente distintos pode revelar padrões dialetais de covariação, proposta de descrição linguística apresentada e detalhada no capítulo que segue.

3. PADRÕES CONJUNTOS DE VARIAÇÃO

Na Sociolinguística Variacionista, a descrição de variedades da língua tem se pautado no isolamento de um único fenômeno variável, de modo a observar seu uso na comunidade e os fatores que interferem nesse uso, o que resulta na desconsideração de padrões conjuntos de variação. Este capítulo tem como objetivo discutir uma proposta de descrição linguística pautada na covariação entre diferentes fenômenos variáveis da língua, com vistas a compreender diferentes técnicas mobilizadas para a descrição de usos conjuntos de fenômenos linguísticos variáveis.

Na primeira parte, discorremos sobre a covariação de maneira geral, compreendendo reflexões que levaram a sua proposição como também sua conceituação. A partir disso, abordamos dois tipos de técnicas utilizadas para a descrição da covariação, iniciando pela técnica que utiliza testes inferenciais de correlação e, em seguida, por técnicas de agrupamento de dados.

3.1 COVARIAÇÃO NA LÍNGUA

Ao discutir os "Fundamentos empíricos para uma teoria da mudança linguística", Weinreich, Labov e Herzog (2006[1968], p. 188) propuseram a tese de que "idioletos não fornecem a base para gramáticas autocontidas ou internamente consistentes". A descrição do comportamento linguístico de indivíduos isoladamente deveria dar lugar à descrição da gramática de toda uma comunidade de fala – grupos de indivíduos que compartilham as mesmas atitudes sobre traços linguísticos (Labov, 2008[1972) –, regida por diversos fatores, que reflete regularidade e coerência. Nos termos de Guy e Hinskens (2016, p. 2), comunidades de fala são sociolinguisticamente coerentes: "espera-se que as variáveis ordenadas que definem a comunidade se comportem coletivamente em paralelo: variantes (ou taxas de uso de variantes) que indexam um determinado estilo, *status* ou uma característica social devem ocorrer simultaneamente".[11]

Frente a isso, a pesquisa sociolinguística traçou como objetivo descrever o comportamento linguístico da comunidade, uma vez que, como propõe Labov (2006[1966], p. 5), o estudo da comunidade é mais importante do que o indivíduo quando se descreve a língua, dado que "a língua dos indivíduos não pode ser compreendida sem o conhecimento

[11] No original: "the orderly variables that define the community should collectively behave in parallel: variants (or rates of use of variants) that index a given style, status, or a social characteristic should co-occur".

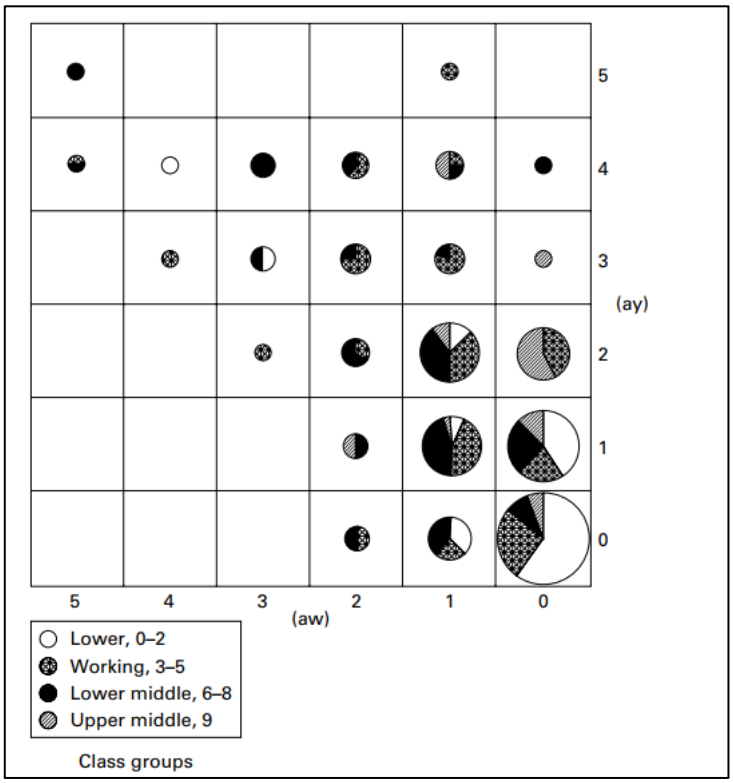
da comunidade da qual eles são membros".[12] Para o autor, usos linguísticos de uma comunidade de fala refletem a sua estratificação social, já que formas linguísticas sofrem o efeito de pressões sociais – por exemplo, formas de prestígio são mais usadas por falantes de classe socioeconômica e nível de escolarização mais elevados, de uma forma aproximadamente consensual em todas as comunidades de fala (Labov, 2006[1966]. Guy (2013) argumenta, a partir disso, que toda comunidade de fala já descrita pela pesquisa sociolinguística apresenta, pelo menos, um conjunto de variáveis linguísticas que são socialmente estratificadas e estilisticamente variáveis, o que pode resultar na formação de diferentes variedades linguísticas. Exemplo clássico é o estudo seminal de Labov (2006[1966]), a estratificação social do -r em Nova Iorque.

Ao descrever o comportamento linguístico de 122 falantes do Lower East Side, área da cidade de Nova Iorque, estratificados em grupo étnico (afro-americano, judeu – ortodoxo –, judeu – conservador e reformista –, católico e protestante) e classe socioeconômica (classe baixa, classe trabalhadora, classe média), Labov (2006[1966]) apresenta duas abordagens para a estratificação da língua. Na primeira abordagem, com base em um agrupamento social, o autor observa que os falantes convergiam quanto à realização de cinco variáveis fonológicas: variantes mais favorecidas por falantes de maior estrato social também eram favorecidas por todos os falantes em estilos de fala mais cuidadosos. A posição social do falante na comunidade era refletida em seus usos linguísticos, dado que seu comportamento era coerente com o comportamento linguístico esperado para o estrato social.

Na segunda abordagem, com base em um agrupamento linguístico, Labov (2006[1966]) observou uma coerência estrutural: falantes que tendiam a uma baixa frequência de alçamento de (aeh), como em *bad* (mau)*, bag* (bolsa)*, cash* (dinheiro), também tendiam a uma baixa frequência de alçamento de (oh), como em *caught* (pego)*, awed* (maravilhado)*, dog* (cachorro); similarmente, aqueles com uma alta frequência de alçamento de uma também tendem a ter uma alta frequência para a outra (Figura 11). Os dados apresentados por Labov (2006[1966]) mostraram que essas variáveis se correlacionavam entre si, com classe socioeconômica e com o estilo de fala do falante, emergindo o argumento de que há uma coerência para o comportamento linguístico da comunidade.

[12] No original: "the language of individuals cannot be understood without knowledge of the community of which they are members" (Labov, 2006, p. 05).

Figura 11 – Correlação de (ay) e (aw) por SEC para todos os informantes de Nova Iorque em Labov (2006[1966])



Fonte: Labov (2006[1966], p. 351).

As divisões dentro de cada círculo indicam a percentagem de cada grupo de classe dentro de cada célula: informantes de classe baixa predominam entre aqueles sem diferenciação de uso de (ay) e (aw), enquanto falantes da classe trabalhadora e classe média baixa mostram a maior tendência à diferenciação extrema dessas variáveis; a classe média alta não usa as duas variáveis de forma simétrica: a grande maioria dos falantes da classe média está localizada nas três células à direita que apresentam alguma diferenciação no uso de (ay), mas nenhuma para (aw). Para Labov (2006[1966]), uma vez que cada uma dessas variáveis apresenta correlação com classe e estilo de fala, esperava-se também que se correlacionem entre si, de modo que cada socioleto – variedade linguística própria de um grupo social – seria caracterizado por um agrupamento de variantes linguísticas. Surge, a partir das reflexões de Labov (2006[1966]), a importância de se descrever a covariação na língua – o que evidencia que covariação não é uma novidade, pois está presente desde o primeiro estudo, mas uma questão de escolha de pesquisa: verticalizar a compreensão do funcionamento de uma regra ou ver como as regras interagem entre si.

Covariação é um termo utilizado na linguística para verificar, como evidencia Oushiro (2015a, p. 69), "se múltiplas variáveis se correlacionam nos usos de falantes individuais e, se

sim, quais fatores sociais e linguísticos promovem a covariabilidade". Freitag (2022) acrescenta a ideia de que covariação é o tipo de abordagem que descreve padrões conjuntos de uso de mais de uma variável. Nesse sentido, segundo Oushiro e Guy (2015, p. 157), "a questão básica é se os falantes que tendem a empregar a variante x da variável A também tenderiam a empregar a variante y da variável B, ou se as variáveis estão independentemente incorporadas na língua e na sociedade".[13]

A partir da ideia de que covariação é a correlação entre duas ou mais variáveis, estudos de covariação, como Guy (2013) e Oushiro (2015a; 2016a), para o PB, e Tamminga (2019), para o inglês estadunidense, têm se amparado em análises de correlação para a descrição da covariação, seguindo uma abordagem aventada por Labov (2006[1966]), já que, como propõe Guy (2013, p. 64, acréscimo nosso), a pressuposição de que comunidades de fala são sociolinguisticamente coerentes resulta na expectativa de que há "algum grau de correlação entre as diferentes variáveis [ou variantes] presentes em uma comunidade".[14] Dentro desse escopo, a covariação muitas vezes é expressa como uma relação, que pode ser positiva (quando as variáveis mudam na mesma direção) ou negativa (quando as variáveis mudam em direções opostas). Na seção que segue, discutimos sobre esse tipo de abordagem.

3.2 COVARIAÇÃO COMO CORRELAÇÃO

Consideremos a seguinte situação: um linguista, interessado em descrever o comportamento do pronome *tu* em posição de sujeito no PB, observa que falantes que fazem uso dessa forma também tendem a empregar o pronome possessivo *teu* e o pronome clítico *te*. Diante disso, esse linguista formula a hipótese de que o uso dessas variáveis linguísticas está associado, dado que o aumento no uso de uma forma implica no também aumento no uso da(s) outra(s) forma(s), e que sua redução resulta na redução de uso da(s) outra(s). Para confirmar/refutar essa hipótese, o linguista necessitará de um teste estatístico de correlação.

Correlação é uma medida de associação do grau de relacionamento entre duas variáveis (Garson, 2009), que, como diz Moore (2003, p. 88), "mensura a direção e o grau da relação linear entre duas variáveis quantitativas".[15] Assim, quando nos interessamos em

[13] No original: "the basic question is whether speakers who tend to employ variant x of variable A would also tend to employ variant y of variable B, or if variables are independently embedded in language and society".
[14] No original: "some degree of correlation among the different variables present in a community".
[15] No original: "measures the direction and strength of the linear relationship between two quantitative variables".

observar o grau de associação entre duas variáveis quantitativas, realizamos uma análise de correlação, que calcula os coeficientes de associação (correlação) entre essas duas variáveis.

Um teste de correlação comum é o de Pearson (r), medida de associação linear entre variáveis. Na correlação de Pearson (r), "duas variáveis se associam quando elas guardam semelhanças na distribuição dos seus escores. Mais precisamente, elas podem se associar a partir da distribuição das frequências ou pelo compartilhamento de variância" (Figueiredo Filho; Silva Jr., 2009, p. 118). Esse tipo de correlação pressupõe, como dito por Moore e McCabe (2004), quatro condições:

1) A correlação exige que as variáveis sejam numéricas;
2) Os valores observados precisam estar normalmente distribuídos. A distribuição normal, ou distribuição gaussiana, é uma curva simétrica em torno do seu ponto médio, apresentando formato de sino. Testes estatísticos que lidam com distribuição normal são chamados de testes paramétricos;
3) Faz-se necessária uma análise de *outliers* – dados que se diferenciam drasticamente de todos os outros –, já que o coeficiente de correlação é fortemente afetado pela presença deles;
4) Faz-se necessária a independência das observações, ou seja, a ocorrência de uma observação X_1 não influencia a ocorrência de outra observação X_2, de modo a não ocorrer correlações espúrias.

Os coeficientes de correlação de Pearson (r) – assim como outros coeficientes (Spearman e Kendall) – variam de -1 a 1, cujo sinal indica correlação negativa ou positiva; quanto mais próximo de 1, maior a força da relação entre as variáveis. Valores próximos a zero indicam que não há relação entre as variáveis. A existência de gradiente nos valores (por exemplo -0,8, -0,35, 0,15, 0,5) indica diferentes forças de relação entre as variáveis. Dancey e Reidy (2006, p. 186) apontam para uma classificação segundo a qual um coeficiente de 0,10 até 0,30 é fraco, de 0,40 até 0,60 é moderado e de 0,70 até 1 é forte.

Em seu estudo, Labov (2006[1966]) utiliza coeficientes de r de Pearson[16] para descrever a covariação entre (æh) e (oh), entre (ay) e (aw), e (ah) *com* (oh). Estudos posteriores que visavam a observação de covariação, como Guy (2013) e Oushiro (2015a; 2016a), também utilizam esse tipo de correlação.

[16] Ainda que o autor não utilize o termo r de Pearson, seu reporte das análises evidencia a utilização desse tipo de correlação.

Guy (2013) demonstra que a covariação pode ocorrer entre variáveis de diferentes níveis linguísticos, como entre o nível fonético-fonológico – apagamento de (-s) em coda (*meno* para *menos*) e desnasalização (*garage* para *garagem*) – e o morfossintático – a concordância nominal e a concordância verbal de 3PP. Os fenômenos variáveis foram selecionados dado o fato de que, ainda que sejam produto de processos independentes, exibem um conjunto de relações linguísticas, cujas variantes apresentam oposição padrão ~ não padrão. Por exemplo, casos de concordância verbal de 3PP não padrão geralmente são feitos com a perda do morfema -m, similar a desnasalização; casos de concordância nominal não padrão geralmente são feitos com a perda do morfema -s, similar ao apagamento de (-s) em coda.

Em uma amostra com dados de fala de 20 falantes do Rio de Janeiro que pertenciam à classe trabalhadora inferior, sem educação formal e analfabetos, Guy (2013) utiliza pesos relativos extraídos de um modelo de regressão logística binária, por meio do Varbrul:[17] para cada falante, há pesos relativos para a variante escolhida como valor de aplicação de cada uma das quatro variáveis.

O resultado do teste para cada par de variáveis mostrou que o apagamento de (-s) em coda se correlaciona com a concordância nominal: quanto mais frequente o apagamento, menor a tendência à concordância (r = -0,740). Incrementalmente, a desnasalização está correlacionada à concordância verbal de 3PP: quanto mais frequente a desnasalização, menor a tendência à concordância verbal (r = -0,450). O resultado do teste de correlação de Pearson também indicou relação entre as variáveis morfossintáticas: falantes que fazem maior uso do morfema de plural na concordância nominal também o fazem na concordância verbal (r = 0,592). Na Figura 12, os dados são apresentados considerando os coeficientes de r de Pearson.

Figura 12 – Correlação entre quatro variáveis sociolinguísticas no português brasileiro em Guy (2013)

Denas			
-.450*	**Verb Agr**		
.256	-.371	**S del**	
-.436*	.592**	-.740***	**Nom Agr**

Significance: *p<.05, **p<.01, ***p<.005

Fonte: Guy (2013, p. 66).

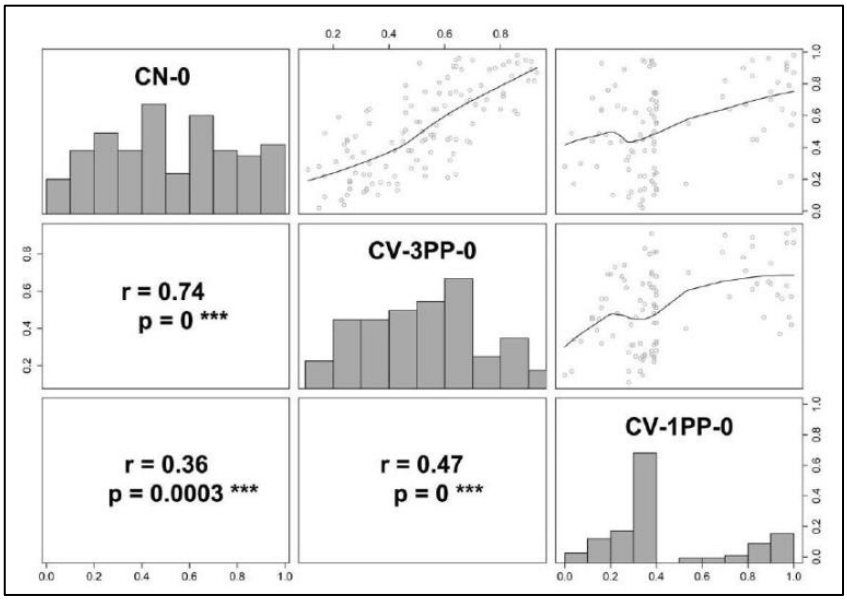
[17] O Varbul é pacote de *software* estatístico usado para analisar padrões de variação linguística, desenvolvido por David Sankoff e Randolph Henry na década de 1970 (Sankoff; Henry, 1975). Ele produz um modelo de regressão logística binária, que visa observar a relação entre a variável linguística binária (ex: concordância padrão x concordância não padrão) e múltiplos fatores (sociais, linguísticos, cognitivos etc.) simultaneamente.

Os resultados do estudo realizado por Guy (2013) sugerem que falantes com comportamento linguístico mais padrão marcam mais plural em substantivos e verbos, e desnasalizam e apagam (-s) menos. Os resultados do autor sugerem que a covariação é motivada linguisticamente e socialmente.

Em Oushiro (2015a), a autora analisa a relação entre a concordância nominal, a concordância verbal de primeira pessoa do plural (1PP) e a concordância verbal de 3PP, já que o mapeamento dessas variáveis é extenso no PB e que certos padrões de estratificação social e linguística são recorrentes nas mais variadas comunidades em que os fenômenos foram descritos. Por meio de 118 gravações de falantes nascidos em São Paulo – estratificados em sexo/gênero (feminino e masculino), faixas etárias (20-34 anos, 35-59 anos, e 60 anos ou mais), níveis de escolaridade (até Ensino Médio e Ensino Superior) e região de residência na cidade (bairros mais centrais e bairros mais periféricos) –, o estudo de Oushiro (2015a) visa investigar se há uma coesão dialetal em relação a essas variáveis no comportamento dos falantes que compõem a comunidade de fala paulistana. As três variáveis selecionadas pela autora são fenômenos linguísticos que sofrem efeito da prescrição, frente à alternância entre uma forma padrão e uma forma não padrão, e todas lidam com a concordância como elemento em comum.

Para descrever a covariação, Oushiro (2015a) utiliza o teste de correlação de Pearson a partir dos pesos relativos extraídos de modelos de efeitos mistos binários na plataforma R (R Core Team, 2018), com auxílio do pacote Rbrul (Johnson, 2009). Os testes evidenciam que todas as variáveis se correlacionam significativamente (Figura 13).

Figura 13 – Matriz de correlações entre variáveis em Oushiro (2015a)



Fonte: Oushiro (2015a, p. 81)

Na matriz, os quadros do canto inferior esquerdo mostram os coeficientes de correlação (r) junto a seus respectivos valores de significância (p). Todas as correlações são positivas: CN-∅ se correlaciona com o emprego de CV-3PP-∅ (r = 0,74) e com CV-1PP-∅ (r = 0,36), e as duas CV se correlacionam entre si (r = 0,47). Como aponta Oushiro (2015a, p. 81), a "correlação mais forte entre esses três pares de variáveis é aquela entre CN e CV-3PP, e não entre CV-1PP e CV-3PP, do modo como uma perspectiva puramente estrutural poderia prever". O maior emprego da forma não padrão da concordância verbal de 1PP covaria com um maior emprego na forma não padrão de concordância nominal ou de concordância verbal de 3PP.

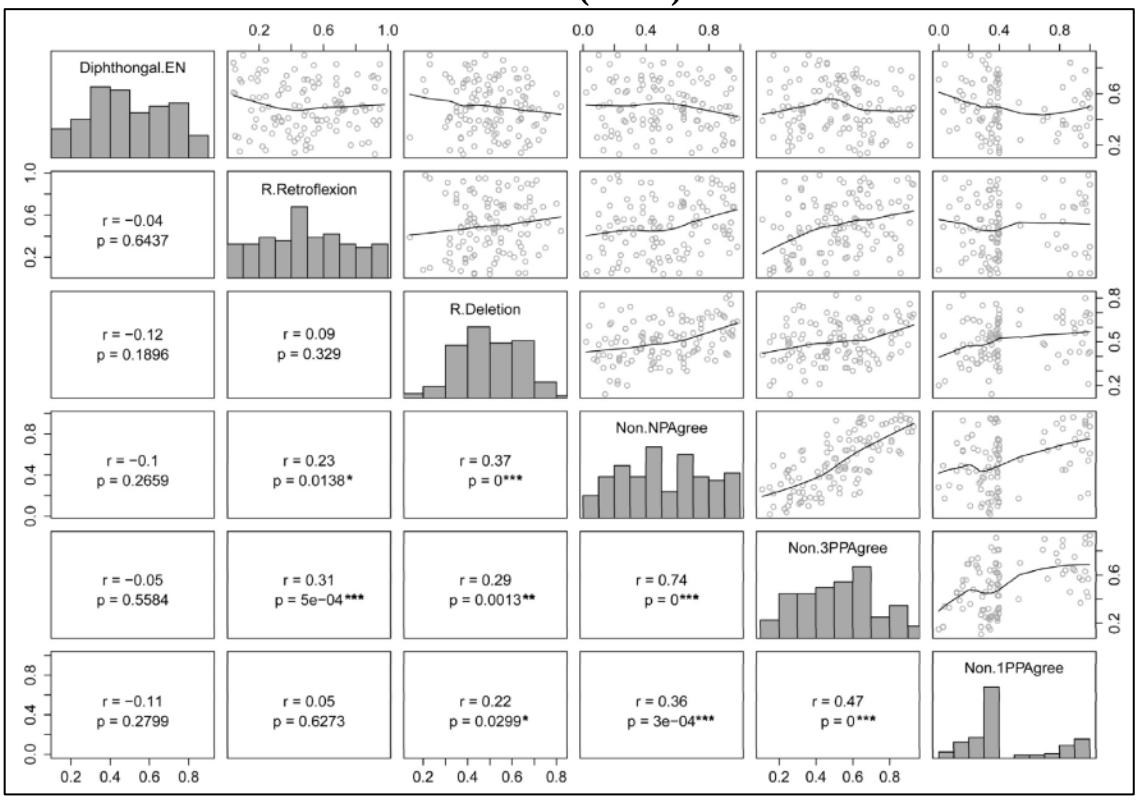
Ao observar o efeito de fatores sociais sobre os usos linguísticos, Oushiro (2015a) conclui que as correlações entre o uso das formas CN e CV-3PP são fortes em vários subgrupos da comunidade de São Paulo, indicando que as variantes padrão e não padrão geralmente são usadas em conjunto. Por outro lado, as correlações com a forma CV-1PP são mais fracas, já que nem todos os grupos apresentam o mesmo nível de coesão no uso dessa variante. Por fim, conclui a autora que a covariação está relacionada a grupos sociais que têm menor exposição a diferentes normas linguísticas, como moradores de bairros mais centrais, filhos de paulistanos e falantes de classes mais baixas. Além disso, Oushiro (2015a) demonstra que a covariação entre as formas depende não só das suas similaridades estruturais, como no caso das concordâncias de 3PP e 1PP, mas também de restrições linguísticas mais gerais que afetam múltiplas variáveis, como a saliência fônica.

Em estudo posterior, Oushiro (2016a) expande a quantidade de variáveis linguísticas, interagindo diferentes níveis linguísticos, como feito em Guy (2013): três variáveis morfossintáticas, descritas no estudo anterior (Oushiro, 2015a), e três fonológicas – a ditongação da vogal nasal [ẽ], o uso de -r retroflexo em coda e o apagamento do -r em coda. A escolha das variáveis foi feita considerando que, ao analisar variáveis fonológicas e morfossintáticas, como também variáveis estruturalmente (não) relacionados, seria possível contrastar padrões de covariação e separar o efeito do condicionamento social e linguístico na covariabilidade.

Ao utilizar 118 entrevistas com falantes paulistanos, e com base no método de análise estatística do estudo anterior (teste de Pearson a partir dos pesos relativos extraídos de modelos de efeitos mistos binários), Oushiro (2016a) observa que (i) a ditongação da vogal nasal [ẽ̞] não se correlaciona com nenhuma outra variável; (ii) o -r retroflexo se correlaciona com a concordância nominal e com a concordância verbal de 3PP, mas não se correlaciona com as outras

variáveis fonológicas; (iii) o apagamento de -r em coda se correlaciona com as três variáveis sintáticas; e (iv) todas as variáveis sintáticas se correlacionam entre si (Figura 14).

Figura 14 – Matriz de correlação entre seis variáveis no Português Brasileiro em Oushiro (2016a)



Fonte: Oushiro (2016a, p. 122).

Para Oushiro (2016a), variáveis estruturalmente não relacionadas, mas socialmente estratificadas, podem se correlacionar significativamente no uso dos falantes, como também as variáveis que possuem relação estrutural (concordância) se correlacionam por questões sociais e estruturais. A autora argumenta que a covariação não é motivada por categorias sociais macro, mas como resultado da densidade da comunicação: falantes tendem a exibir uma coesão social significativamente maior se interagirem mais com falantes do grupo interno do que com aqueles do grupo externo. Além disso, os dados da autora demonstram que a covariação é promovida tanto pela similaridade estrutural entre variáveis dependentes, quanto por restrições linguísticas mais gerais que se correlacionam com múltiplas variáveis, como a saliência fônica. É o caso, por exemplo, do que Duarte e Varejão (2013), segundo Oushiro (2015a), chamam de regularização paradigmática, no qual formas pertencentes ao mesmo paradigma, como ao paradigma pronominal de 2PS, tendem a ocorrer conjuntamente: falantes que usam *tu*, podem também usar as formas do paradigma de *tu*, como *te* e *teu*.

As pesquisas, contudo, lidam com teste de correlação que exigem a suposição de normalidade na distribuição dos dados. Como uma análise de correlação se comportaria com dados que violam a suposição de normalidade? Seria necessário, nesse caso, a adoção de testes não-paramétricos (cf. Guy; Oushiro; Mendes, 2022). O teste de correlação de rho de Spearman (ρ) é uma alternativa para estimar correlações lineares em situações nas quais há violação da suposição de normalidade (Zar, 2005). Similar ao teste de Pearson, o de Spearman também lida com variáveis numéricas (contínuas ou discretas, mas também ordinais) e segue uma interpretação de coeficientes igual ao r de Pearson, de -1 a 1.

Tamminga (2019), em face da problemática de utilizar um teste paramétrico em dados com distribuição não normal, utiliza dois tipos de coeficiente de correlação, o r de Pearson paramétrico e o ρ de Spearman não paramétrico, de modo a ter resultados mais apurados. Com o objetivo de descrever a covariação entre seis vogais (pronúncias de *price*, *down, goat, face, tooth* e *thought*) do inglês falado na cidade de Filadélfia, Estados Unidos, a partir da fala de 66 jovens mulheres entre 18 e 29 anos, que cresceram na região da cidade ou no subúrbio adjacente e se identificam como mulheres brancas, a autora conduz um modelo de regressão linear de efeitos mistos que inclui várias variáveis fixas e item lexical como aleatório. Por meio do modelo de regressão, Tamminga (2019) extraiu os resíduos, que refletem a variabilidade não capturada pelas variáveis de controle, e então calculou a média dos resíduos associados ao conjunto de observações de cada falante para chegar a uma medida de tendência central do falante. A Tabela 1 evidencia os quinze pares de correlação a partir dessas medidas de tendência central dos falantes.

Tabela 1 – Correlações de pares de médias residuais de falante em Tamminga (2019)

Vowel pair	Prsn r	Hlm p(r)	BH p(r)	Sprmn ρ	Hlm p(ρ)	BH p(ρ)
FACE ~ PRICE	0.125	1.000	0.592	0.078	1.000	0.835
FACE ~ TOOTH	0.168	1.000	0.381	0.094	1.000	0.835
FACE ~ DOWN	0.028	1.000	0.950	0.018	1.000	0.889
FACE ~ GOAT	0.082	1.000	0.762	0.063	1.000	0.835
FACE ~ THGHT	0.083	1.000	0.762	0.079	1.000	0.835
PRICE ~ TOOTH	0.073	1.000	0.762	0.059	1.000	0.835
PRICE ~ DOWN	0.170	1.000	0.381	0.205	1.000	0.265
PRICE ~ GOAT	0.008	1.000	0.950	−0.029	1.000	0.877
PRICE ~ THGHT	−0.017	1.000	0.950	−0.054	1.000	0.835
TOOTH ~ DOWN	0.222	0.873	0.219	0.245	0.567	0.177
TOOTH ~ GOAT	−0.014	1.000	0.950	−0.040	1.000	0.862
TOOTH ~ THGHT	0.222	0.873	0.219	0.201	1.000	0.265
DOWN ~ GOAT	0.519	<0.001	<0.001	0.446	0.003	0.003
DOWN ~ THGHT	0.506	<0.001	<0.001	0.378	0.025	0.013
GOAT ~ THGHT	0.352	0.048	0.018	0.359	0.040	0.015

Fonte: Tamminga (2019, p. 10).

Os dados mostram que a maioria das correlações entre pares não é estatisticamente significativa. Exceções são vistas nos pares DOWN~GOAT, DOWN~THOUGHT e GOAT~THOUGH, os quais são considerados processos de mudanças reversas: "as mudanças reversas covariam dentro de si; elas não covariam com as mudanças contínuas" (Tamminga, 2019, p. 10).[18] A utilização de diferentes testes paramétricos aponta que os resultados são similares nos coeficientes de correlação com r de Pearson e com ρ de Spearman, o que leva Tamminga (2019, p. 10) a concluir que "os resultados não dependem da tomada de uma decisão metodológica específica".[19]

Os estudos de correlação vistos confirmam as hipóteses de Guy (2013), Oushiro (2015a; 2016a) e Tamminga (2019) sobre a existência de relação entre duas ou mais variáveis, à medida que falantes que tendem a empregar dada variante da variável A também tenderiam a empregar dada variante da variável B, como é o caso das relações entre a concordância não padrão de 1PP, de 3PP e a concordância nominal no português paulista, em Oushiro (2016a), a relação entre o apagamento de (-s) em coda e a concordância nominal, e a desnasalização relacionada à concordância verbal de 3PP nos dados de falantes do Rio de Janeiro, em Guy (2013). Os mesmos estudos, todavia, evidenciam que nem sempre pares de variáveis apresentam correlação entre si, como é mais explicitamente visto em Tamminga (2019).

Testes de correlação são uma técnica estatística importante para confirmar/refutar hipóteses sobre a relação entre duas variáveis a partir de um conjunto de dados resumido numericamente. Mas não é possível, a partir deles, que possamos observar o comportamento de cada falante. Uma proposta que permita a visualização do comportamento por falante pode acrescentar informações à análise de covariação aventada pela correlação, como é o caso de uma análise pautada no agrupamento, uma vez que, como proposto por Freitag (2022), a covariação também observa padrões conjuntos de uso de mais de uma variável. Na seção que segue, apresentamos estudos que lidam com o agrupamento de falantes através de seus usos individuais.

3.3 COVARIAÇÃO COMO AGRUPAMENTO

Guy (2013), ao reconhecer a necessidade de ir além das comparações de pares das variáveis, propôs a observação de padrões amplos de agrupamento. A ideia do autor era, a partir de padrões recorrentes de classificação de uso das variantes por falante, verificar se

[18] No original: "the reversing changes covary within themselves; they do not covary with the continuing changes".
[19] No original: "the results are not contingent on making a particular methodological decision".

haveria um padrão mais frequente que sugerisse uma coerência entre o grupo. Para tanto, o autor transformou valores contínuos (taxas de uso da variante de prestígio de cada variável) em valores categóricos, de forma ternária: taxa alta (h), média (m) e baixa (l), classificando cada falante de acordo com suas taxas de utilização das quatro variáveis. Desse modo, cada falante tinha uma classificação como hhmh (alta-alta-média-alta), hmml etc. Nos resultados, 25% dos falantes fazem uso das variantes com o mesmo padrão (hhhh, mmmm etc.); 25% fazem uso de três variantes na mesma faixa de uso; e 20% dos falantes não mostram nenhum agrupamento significativo destas variáveis.

Similar a Guy (2013), Oushiro (2015a) também analisa padrões amplos de agrupamento de cada indivíduo por meio dos pesos relativos de uso das variantes: alta (A), média (M) ou baixa (B). Para a autora, "haveria um alto grau de coesão dialetal se todas as variantes se encaixarem na mesma categoria (p.ex., AAA ou BBB) para a maioria dos indivíduos da amostra" (Oushiro, 2015a, p. 79). Na análise, Oushiro (2015a) observou a ocorrência de 21 dos 27 padrões possíveis, em que o mais frequente é aquele no qual as tendências de emprego de marca zero são baixas para as três variáveis (BBB, 22 falantes), seguido pelo padrão AAA (15 falantes), de altas tendências de emprego da variante. Para a autora, a grande distribuição de uso por outros padrões tende a sugerir a existência de uma baixa coesão dialetal na comunidade em relação aos usos linguísticos das variáveis descritas.

A partir de seu estudo anterior, Oushiro (2016a) classifica os pesos relativos de uso individual das variantes em altos (H) ou baixos (L) para cada uma das seis variáveis. Na análise, HLLLLL e LHHHHH são os padrões mais frequentes, cada um com 8 falantes, que tendem a empregar as variantes de maneira semelhante, favorecendo-as ou desfavorecendo-as. Os dois padrões em que todas as seis variáveis covariam na mesma frequência (LLLLLL e HHHHHH) também estão entre os mais frequentes, ocorrendo para 6 e 4 falantes respectivamente. Para a autora, falantes tendem a não se agrupar em grupos sociais significativos.

A consideração de usos individuais acrescenta informações em relação ao comportamento das variáveis no indivíduo e na comunidade. Observemos a seguinte situação hipotética: falantes de Aracaju tendem a fazer maior uso da variante *você* em posição de sujeito de 2PS. Contudo, é possível que alguns falantes destoem desse padrão e façam maior uso da variante *tu*; adicionalmente, falantes que fazem maior uso de *você* podem fazer também maior uso de *seu* como possessivo de 2PS, mas alguns fazem maior uso de *teu*. Isso implica na ideia de que, embora se espere, como propõem Guy e Hinskens (2016), que as variáveis que definem a comunidade se comportem coletivamente em paralelo, é possível que no uso individual isso não ocorra. Para tanto, técnicas estatísticas, como análise de componentes

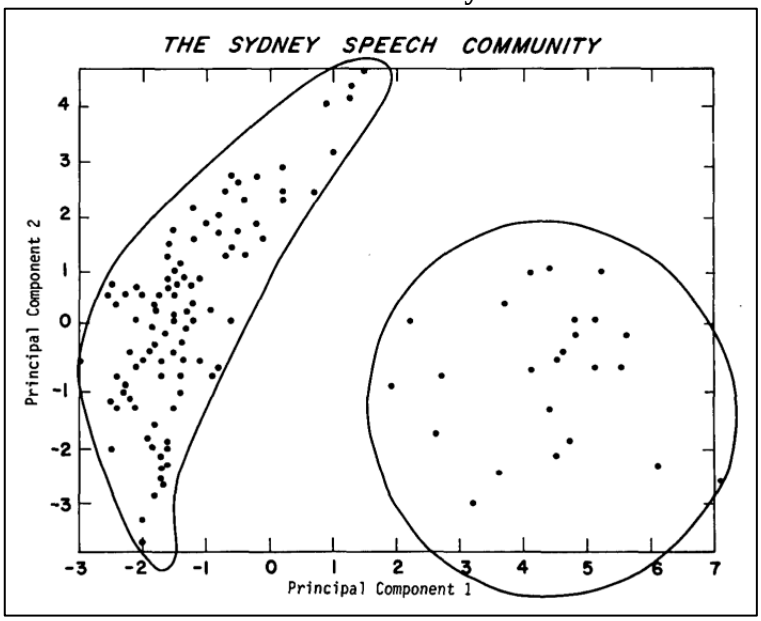
principais e análise de *clustering*, podem ser utilizadas para a visualização do agrupamento de falantes por usos individuais.

Horvath e Sankoff (1987) são pioneiros na descrição da covariação a partir de agrupamento. Para tanto, os autores utilizam a técnica de análise de componentes principais (PCA, *principal component analysis*), um tipo de modelagem multivariada utilizada na sociolinguística como forma de agrupar falantes com base em seu comportamento linguístico. Segundo Varella (2008, p. 3), os componentes principais apresentam algumas propriedades importantes:

i) Cada componente principal é uma combinação linear de todas as variáveis originais;
ii) Os componentes principais são independentes entre si e estimados com o propósito de reter o máximo de informação, em termos da variação total contida nos dados;
iii) A análise de componentes principais é associada à ideia de redução de massa de dados, com perda mínima possível da informação;
iv) A análise agrupa os indivíduos de acordo com sua variação, isto é, os indivíduos são agrupados segundo suas variâncias, ou seja, segundo seu comportamento dentro da população, representado pela variação do conjunto de características que define o indivíduo, ou seja, a técnica agrupa os indivíduos de uma população segundo a variação de suas características.

Para a análise, Horvath e Sankoff (1987) consideraram uma amostra que inclui 117 falantes de origem anglo-céltica e italiana residentes em Sidney, Austrália, considerando em sua organização outras características sociais, como classe socioeconômica, sexo e duas faixas etárias – adolescentes e adultos. A análise mobilizou quatro variáveis fonológicas, todas incluindo variação vocálica: pronúncia de iy, ey, ow e ay. Os dados para a PCA consistiram, para cada falante, no número de ocorrências de cada variante das quatro variáveis. Nos resultados dos autores, quatro componentes principais foram identificados, dos quais apresentamos dois (Figura 15).

Figura 15 – A comunidade de fala de Sidney em Horvath e Sankoff (1987)

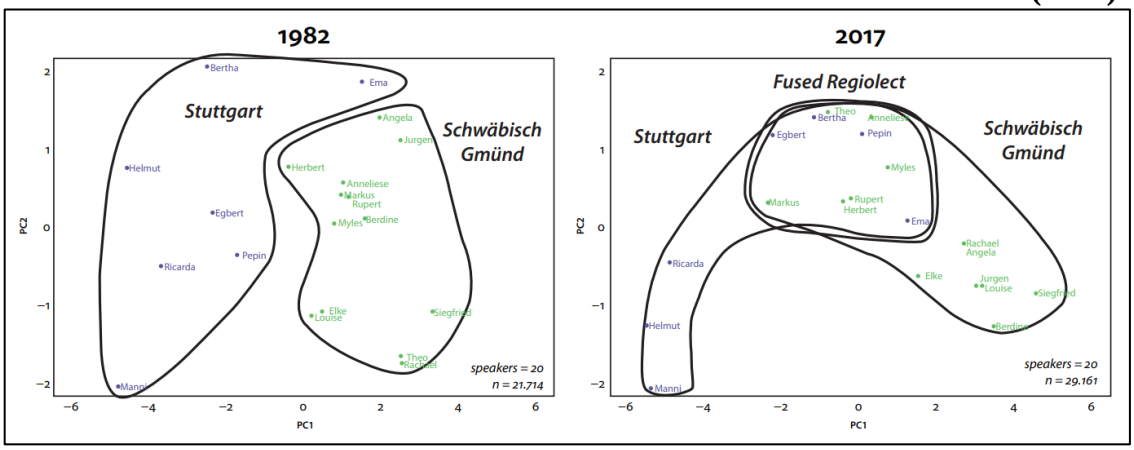


Fonte: Horvath e Sankoff (1987, p. 190)

A Figura 15 apresenta o Componente Principal 1 (31,9% da variância), na horizontal, e o Componente Principal 2 (15,0%), na vertical. Nela, os dados sugerem a existência de pelo menos duas variedades de inglês falado Sydney, que são relativamente distintas, nos quais o primeiro grupo – em forma de lua crescente, como chamam os autores – possui o maior quantitativo de falantes, enquanto o segundo grupo – formato de lua cheia – possui menos falantes. Para Horvath e Sankoff (1987), o primeiro grupo ainda pode ser dividido entre três subgrupos, que formam um contínuo, indicando a probabilidade de mais sobreposições no comportamento linguístico dos falantes.

Beaman (2021), ao apresentar uma abordagem exploratória para modelar e medir o conceito de coerência letal (de "leto" – sistema linguístico), analisa doze traços fonológicos e morfossintáticos do Suábio Central, uma variedade do alemão falado no sudoeste da Alemanha. A autora examina dados de duas comunidades (Stuttgart e Schwäbisch Gmünd) em dois pontos no tempo (1982 e 2017) – considerando idade (18-29, 30-60 e 61-88), sexo (masculino e feminino) e educação (alta ou baixa) –, e utiliza a técnica de PCA para revelar agrupamentos significativos de falantes do Sábio Central, dispostos na Figura 16. A autora não especifica quais tipos de dados são utilizados e o modelo de extração.

Figura 16 – Análise de componentes principais (PCA) representando Stuttgart e Schwäbisch Gmünd letos em 1982 e a fusão de letos em 2017 em Beaman (2021)



Fonte: Beaman (2021, p. 151)

Na Figura 16, o Componente Principal 1 (PC1) está no eixo horizontal e PC2 no eixo vertical. Em 1982, PC1 e PC2 representam juntos 62% da variação, e em 2017 PC1 e PC2 representam juntos 82% da variação. Em 1982, vemos dois letos distintos, falantes de Stuttgart à esquerda e falantes de Schwäbisch Gmünd à direita. O PCA para 2017 retrata uma imagem em mudança da situação dialetal do Suábio, no qual há três divisões e a do meio apresenta uma "fusão" no leto.

Em ambos os estudos, não visualizamos correlações entre variáveis, mas agrupamentos através dos usos linguísticos pelos indivíduos, de acordo com sua variação, o que apresenta evidências para o comportamento linguístico de grupos de falantes e como eles se organizam em suas respectivas comunidades, relevando possíveis agrupamentos a partir de uso de variantes linguísticas específicas. Para Horvath e Sankoff (1987), a importância da técnica PCA reside na sua utilidade na identificação de falantes que se enquadram na comunidade de fala e em permitir que os grupos sejam determinados pelo comportamento linguístico e não somente pelas características sociais. A técnica PCA, então, agrupa os falantes quanto ao seu comportamento linguístico individual, o que torna a análise por agrupamento significativa para a descrição da covariação.

Análise de componentes principais, todavia, não é a única técnica estatística existente para a observação de agrupamento. É possível, também, a adoção de técnica de *clustering*, como feito por Freitag (2022) para o PB. *Clustering* ou análise de *cluster* é uma técnica estatística para agrupar pontos de dados em *clusters* (grupos) com base em sua similaridade. O objetivo do agrupamento é encontrar grupos de pontos de dados semelhantes entre si e

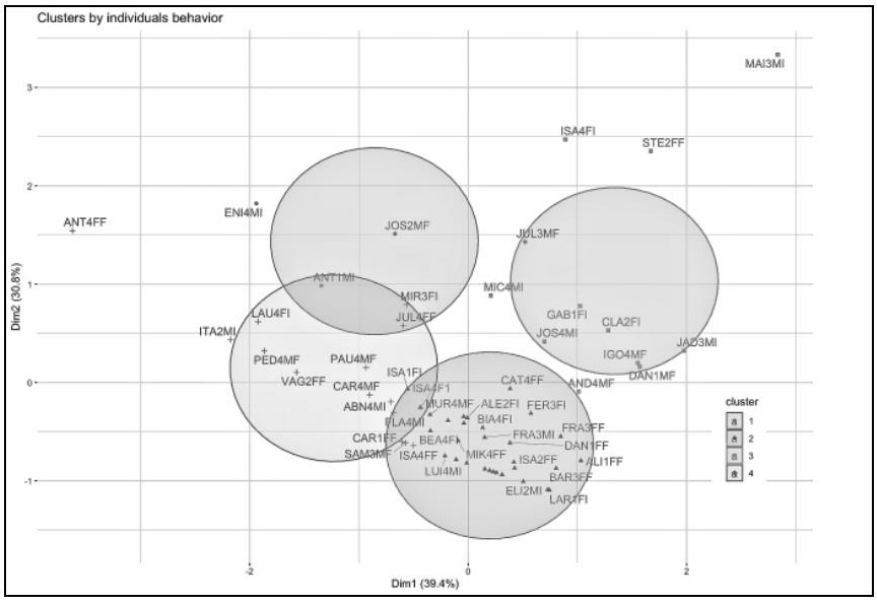
diferentes dos pontos de dados em outros grupos. Para Valli (2012, p. 78), são características de uma análise de *cluster*:

i) Encontrar os agrupamentos naturais de indivíduos;
ii) Um agrupamento de indivíduos devidamente caracterizados pode formar uma população completa ou pode ser uma amostra de alguma população maior;
iii) A análise de *cluster* objetiva alocar indivíduos em grupos de elementos mutuamente exclusivos, semelhantes, isto é, agrupa-se tal que os elementos pertencentes a um grupo são mais parecidos quanto possível uns com outros, enquanto indivíduos em grupos diferentes são dissimilares;
iv) Isto permite a medição da semelhança (ou diferença) de todo par de indivíduos. As semelhanças, às vezes, são observadas diretamente, enquanto em outros casos elas são derivadas de uma matriz de dados de um modo apropriado.

Ao explorar a situação de contato entre diferentes normas linguísticas na fala de estudantes universitários da Universidade Federal de Sergipe (UFS), estratificados em gênero (masculino e feminino), tempo no curso (início e final) e deslocamento (natural e residente da Grande Aracaju; natural e residente do interior do estado, que faz o percurso diário; natural e residente do interior do estado que residem na Grande Aracaju; natural e residente de Alagoas ou Bahia que reside na Grande Aracaju), Freitag (2022) seleciona três fenômenos morfossintáticos previamente descritos com base na amostra Deslocamentos (2020): concordância verbal de 3PP (Novais, 2021) – *os carros quebrou x os carros quebraram*; preposições locativas em verbos de movimento (Rodrigues, 2021) – *vou a festa x vou pra festa x vou na festa*; e uso de determinante antes de possessivos (Silva, 2020) – *achei o seu pai x achei seu pai*. Os fenômenos foram selecionados considerando seu padrão de comportamento entre os falantes universitários, tanto em questão de sensibilidade dialetal, quanto em questão de adequação à norma.

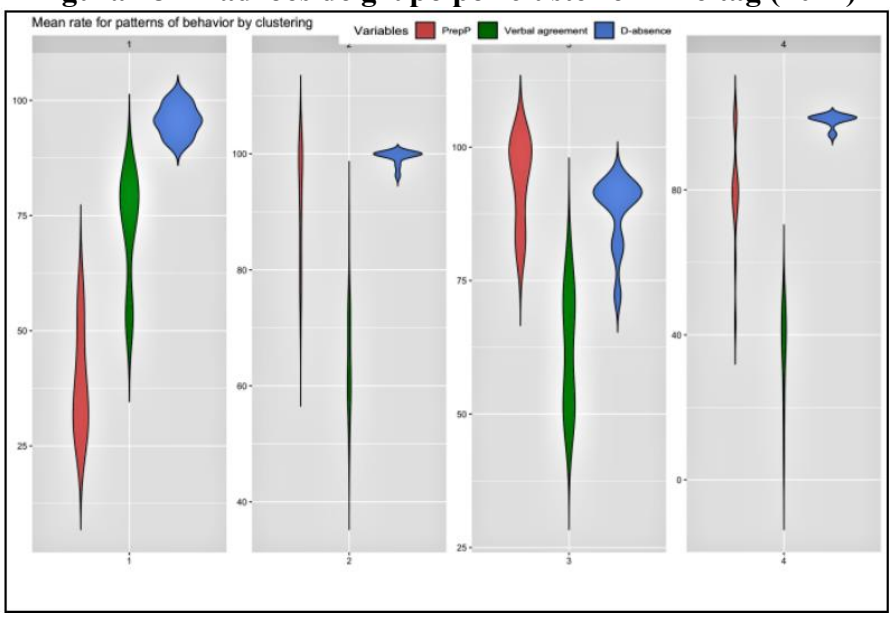
Para a técnica de *cluster*, a autora utiliza o agrupamento por meio de k-*medoids* (Qiao *et al.*, 2011), algoritmo de agrupamento de dados que identifica grupos não sobrepostos que ocorrem naturalmente dentro de uma população de dados, em que cada agrupamento possui um objeto representante mais próximo do centro – um *medoid*. A utilização da técnica possibilitou a observação de quatro grupos distintos para o uso das variantes (Figura 17).

Figura 17 – Análise de *clustering* em Freitag (2022)



Fonte: Freitag (2022, p. 213).

Figura 18 – Padrões de grupo por cluster em Freitag (2022)



Fonte: Freitag (2022, p. 214).

Dos grupos, o Grupo 1 é o menor (n = 4), com falantes unidos pela menor média de uso de preposição padrão (*a*); ausência de determinante e concordância padrão acima da média geral. A característica social comum entre os falantes é serem calouros e nenhum alagoano. O Grupo 2 é o maior (n = 30): os universitários são unidos por uma média alta de preposição padrão, com a ausência de artigo acima da média geral e média de concordância padrão quase categórica. O grupo é equilibrado em gênero e tempo no curso e a maioria dos alunos é de Sergipe (Freitag, 2022).

O Grupo 3 (n = 11) apresenta a maior média de preposição padrão, com a ausência de artigo acima da média global e a menor média de concordância verbal padrão. Predominam os alagoanos iniciantes do sexo masculino. O Grupo 4 (n = 14) apresenta a menor média de ausência de artigo, a menor média de preposição padrão e média quase categórica de concordância verbal padrão. A maioria dos alunos são baianos e veteranos (Freitag, 2022).

Os resultados de Freitag (2022) evidenciam que a utilização da técnica de *clustering* permite observar a organização de falantes em grupos naturais, a partir de seus usos linguísticos quanto às variantes de aplicação dos fenômenos variáveis, mais precisamente, permite a observação de agrupamento de falantes com características morfossintáticas semelhantes, o que pode corresponder a variedades regionais específicas ou maior adequação à norma em seu comportamento linguístico.

3.4 DESCRIÇÃO DA COVARIAÇÃO

Este capítulo objetivou discutir uma proposta de descrição linguística pautada na covariação entre diferentes fenômenos variáveis da língua, com vistas a compreender diferentes técnicas mobilizadas para a descrição de usos conjuntos de fenômenos linguísticos variáveis.

Vimos que a covariação é frequentemente tomada como uma relação entre variáveis, que pode ser positiva ou negativa. Nesse sentido, é comum a adoção de testes de correlação, como a correlação de Pearson (r), conforme feito por Labov (2006[1966]) para o inglês, e Guy (2013) e Oushiro (2015a; 2016a) para o português brasileiro. Os resultados dessas pesquisas apontam para uma relação nos usos das variáveis linguísticas, uma vez que o uso de variante x da variável A pode se associar ao uso de variante y da variável B. A existência de distribuições não normais pode ser contornada através da utilização de testes não paramétricos, como a correlação de rho de Spearman (ρ), conforme feito por Tamminga (2019).

Um problema para a análise de correlação, contudo, é a existência de variáveis não tão frequentes na língua. Embora esse não seja o problema observado em Guy (2013), Oushiro (2015a; 2016a) e Tamminga (2019), fenômenos variáveis, especialmente morfossintáticos, como pontua Cheshire (1999), podem apresentar baixa contagem em análises linguísticas, o que pode interferir na análise de correlação, uma vez que a existência de zeros interfere na distribuição de dados. Além disso, testes de correlação analisam associação entre duas ou mais variáveis, o que não aborda as características individuais dos falantes. Para contornar isso, autores tendem a usar técnicas distintas, como o agrupamento social apresentado em Guy (2013) e Oushiro (2015a; 2016a), para o PB, e a utilização de técnicas estatísticas que

lidem com o agrupamento natural de falantes, conforme apresentado em Horvath e Sankoff (1987), para o inglês falado em Sidney, Beaman (2021), para o Suábio Central, e Freitag (2022), para o PB.

Ao analisar os agrupamentos de características linguísticas, é possível identificar traços comuns compartilhados por falantes, bem como traços distintivos que os diferenciam uns dos outros. Essa análise pode fornecer maior detalhamento em relação ao comportamento dos indivíduos e o que os leva a serem agrupados em dado grupo. Além disso, representações visuais dos agrupamentos, como os gráficos vistos neste capítulo, facilitam a compreensão dos padrões linguísticos.

Não temos, contudo, como pontuar qual o melhor caminho para a descrição da covariação, dado que a modelagem dos estudos é diferente entre si (Quadro 7), que cada técnica tem objetivos distintos e que, enquanto algumas pesquisas usam testes inferenciais, outras usam técnicas de agrupamento de dados.

Quadro 7 – Organização das pesquisas que lidam com covariação

Estudo	Variáveis	Tipo de variáveis	Amostra	Técnica
Testes inferenciais de correlação				
Labov (2006[1966])	5 variáveis	Fonológicas	**122 falantes** nativos do Lower East Side que viveram na área por pelo menos dois anos antes da pesquisa ALS, estratificados em grupo étnico (afro-americano, judeu – ortodoxo –, judeu – conservador e reformista –, católico e protestante); e classe socioeconômica (classe baixa, classe trabalhadora, classe média).	Teste de correlação paramétrico
Guy (2013)	4 variáveis	Fonológicas e morfossintáticas	**20 falantes** (11 homens e 9 mulheres) residentes no Rio de Janeiro entrevistados para o estudo 'Competências Básicas' do português brasileiro popular, falantes do vernáculo popular, de classe trabalhadora inferior, sem educação formal e analfabetos.	Teste de correlação paramétrico
Oushiro (2015a)	3 variáveis	Morfossintáticas	**118 falantes** nascidos na cidade de São Paulo, estratificados de acordo com sexo/gênero (feminino e masculino), faixas etárias (20-34 anos, 35-59 anos, e 60 anos ou mais), níveis de escolaridade (até Ensino Médio e Ensino Superior) e região de residência na cidade (bairros mais centrais e bairros mais periféricos).	Teste de correlação paramétrico
Oushiro (2016a)	6 variáveis	Fonológicas e morfossintáticas	**118 falantes** nascidos na cidade de São Paulo, estratificados de acordo com sexo/gênero (feminino e masculino), faixas etárias (20-34 anos, 35-59 anos, e 60 anos ou mais), níveis de escolaridade (até Ensino Médio e Ensino Superior) e região de residência na cidade (bairros mais centrais e bairros mais periféricos).	Teste de correlação paramétrico
Tamminga (2019)	6 variáveis	Fonológicas	**66 falantes** jovens mulheres entre 18 e 29 anos, que cresceram na cidade de Filadélfia ou em um subúrbio adjacente da Pensilvânia, que se identificaram como mulheres brancas.	Teste de correlação paramétrico e não paramétrico
Modelos de agrupamento de dados				
Horvath e Sankoff (1987)	4 variáveis	Fonológicas	**117 falantes** de origem anglo-céltica e italiana, selecionados considerando classe socioeconômica, sexo e duas faixas etárias – adolescentes e adultos.	Técnica de PCA
Beaman (2021)	12 variáveis	Fonológicas e morfossintáticas	**20 falantes** e 40 entrevistas, sendo 20 coletadas em 1982 e 20 em 2017. As amostras consideram idade (18-29, 30-60 e 61-88), sexo (masculino e feminino), educação (alta ou baixa) e cidade (Stuttgart e Schwäbisch Gmünd).	Técnica de PCA
Freitag (2022)	3 variáveis	Morfossintáticas	**60 falantes** estudantes da Universidade Federal de Sergipe, estratificados em gênero (masculino e feminino), tempo no curso (início e final) e deslocamento (natural e residente da Grande Aracaju; natural e residente do interior do estado, que faz o percurso diário; natural e residente do interior do estado que residem na Grande Aracaju; natural e residente de Alagoas ou Bahia que reside na Grande Aracaju).	Técnica de *clustering* por *k-medoids*

Fonte: elaboração própria.

No que pesem as considerações apresentadas até aqui, este trabalho visa testar três dessas abordagens para um mesmo conjunto de variáveis, uma vez que a mobilização de diferentes técnicas ajuda na resolução de nossa questão de pesquisa. Na primeira delas, adotamos a análise de correlação, a partir de Labov (2006[1966]), Guy (2013), Oushiro (2015a; 2016a) e Tamminga (2019), por ser frequente e utilizada em descrições de covariação no PB; na segunda, utilizamos os padrões amplos de agrupamento, conforme desenvolvido por Guy (2013), Oushiro (2015a; 2016a), por permitir observar o agrupamento de falantes a partir de suas taxas de uso; na terceira, o agrupamento por *cluster*, conforme Freitag (2022), por já ter sido aplicada ao PB e, mais precisamente, por já ter sido utilizada na descrição do português falado por universitários da UFS, por meio da amostra Deslocamentos (2020), que permite a visualização do português falados por estudantes de diferentes áreas geográficas, em diferentes períodos no curso e em níveis de integração ao *campus* distintos.

As análises são feitas com quatro variáveis morfossintáticas que apresentam diferentes gradientes de frequência na amostra, por meio da fala de 181 indivíduos. No capítulo que segue, apresentamos, detalhadamente, nosso *corpus* e os procedimentos metodológicos adotados nesta tese.

4. PROCEDIMENTOS METODOLÓGICOS

Este capítulo tem como objetivo apresentar os procedimentos metodológicos adotados nesta tese, visando a descrição de (co)variação de traços sociolinguísticos morfossintáticos, a saber: i) uso variável de artigo definido antes de possessivos; ii) pronomes pessoais de 2PS em posição de sujeito; iii) pronomes clíticos de 2PS; e iv) pronomes possessivos de 2PS. Para tanto, este capítulo considera (i) a caracterização das amostras sociolinguísticas que compõem o banco de dados Falares Sergipanos, as amostras Deslocamentos (2019), Deslocamentos (2020) e Linguagem Corporificada (2023), que permitem a visualização de diferentes padrões dialetais; ii) a extração dos dados de produção a partir das amostras; e (iii) a análise e descrição de dados por meio da utilização de estatística descritiva e inferencial, contribuindo para a observação do uso conjunto das variáveis morfossintáticas selecionadas por falantes universitários da UFS.

4.1 AMOSTRAS SOCIOLINGUÍSTICAS

Nas seções anteriores, discutimos sobre a existência de fenômenos morfossintáticos que apresentam significado dialetal, isto é, que indexam informações dialetais em relação aos seus falantes e sua região de origem. Discutimos também a necessidade de se descrever usos conjuntos de variáveis linguísticas com vistas a ter uma maior compreensão do comportamento linguístico dos grupos de falantes em termos dialetais, dado que a descrição de apenas um traço linguístico não é suficiente para a compreensão do dialeto do falante. Para integrar esses dois pontos discutidos, é necessário que utilizemos amostras que permitam tanto a observação de diferentes padrões dialetais quanto que possuam um quantitativo significativo de falantes, de modo com que haja a ocorrência também significativa das variáveis morfossintáticas, já que, como pontuado por Cheshire (1999), fenômenos morfossintáticos podem apresentar baixa contagem em análises linguísticas, o que pode interferir em análises estatísticas, frente à possibilidade de existir zeros na distribuição dos dados.

Em um ambiente como a UFS, que recebe estudantes das mais variadas regiões, é possível que haja uma profusão de dialetos: falantes com um comportamento linguístico comum à região Nordeste, Sudeste, Sul etc., e comportamentos mais específicos a determinada sub-região, cidade, bairro, povoado etc. Grupos de falantes do interior do próprio estado podem falar de forma relativamente distinta do comportamento de grupos de falantes da região metropolitana, como também grupos de falantes de estados vizinhos podem

apresentar um padrão de uso distinto, conforme evidenciam pesquisas desenvolvidas no escopo do projeto *Como fala, lê e escreve o universitário?* (Freitag, 2018), coordenado pela Profa. Dra. Raquel Meister Ko. Freitag.

Para observar padrões distintos de comportamento de grupos de falantes de diferentes regiões, é necessária a utilização de amostras sociolinguísticas que possibilitam esse tipo de visualização. Assim, a seleção e/ou estratificação da amostra deve considerar indivíduos de múltiplas áreas geográficas, com base no interesse do pesquisador. Nesta tese, trabalhamos com dados de fala semiespontâneos de três amostras do banco de dados Falares Sergipanos (Freitag, 2013), que permitem esse tipo de visualização: a amostra Deslocamentos (2019), a amostra Deslocamentos (2020) e a amostra Linguagem Corporificada (2023), o que nos fornece cento e oitenta e uma (181) entrevistas, totalizando 1.002.535 de palavras proferidas apenas pelos informantes.

As amostras consideram a fala de estudantes universitários da Universidade Federal de Sergipe (UFS), do *campus* Prof. José Aloísio de Campos, localizado em São Cristóvão, SE, com faixa etária entre 18 e 30 anos (*Média* = 21). As amostras foram escolhidas por possuírem falantes de diferentes regiões, como também por considerarem o acesso dos estudantes ao *campus* em termos de mobilidade, o que permite a observação indireta do significado dialetal da variação a partir de diferenças nos usos dos falantes que pertencem a grupos distintos, como também a descrição de diferentes dialetos do português brasileiro.

A amostra Deslocamentos (2019), constituída entre 2018-2019 e organizada por Corrêa (2019), Ribeiro (2019) e colaboradores, é estratificada considerando (1) o *deslocamento* do falante (Quadro 8), (2) o *tempo no curso* do estudante, segmentado em *início* (3º período ou antes) e *final* (7º período ou depois), e (2) *sexo/gênero* do estudante, dividido entre *masculino* e *feminino* (Tabela 2), e é composta por 64 entrevistas sociolinguísticas.

Quadro 8 – Deslocamentos na amostra de 2019

Deslocamento 1	Estudantes da UFS nascidos na Grande Aracaju (Aracaju, Nossa Senhora do Socorro, São Cristóvão e Barra do Coqueiros) e que residem nela.
Deslocamento 2	Estudantes da UFS nascidos no interior de Sergipe que fazem o trajeto diário para a UFS.
Deslocamento 3	Estudantes da UFS nascidos no interior de Sergipe que residem na Grande Aracaju.
Deslocamento 4	Estudantes da UFS nascidos em outros estados que atualmente residem na Grande Aracaju.

Fonte: elaboração própria a partir de Corrêa (2019) e Ribeiro (2019).

Tabela 2 – Estratificação da amostra Deslocamentos (2019)

	Início		Final	
	Masculino	Feminino	Masculino	Feminino
Deslocamento 1	4	4	4	4
Deslocamento 2	4	4	4	4
Deslocamento 3	4	4	4	4
Deslocamento 4	4	4	4	4

Fonte: elaboração própria a partir de Corrêa (2019) e Ribeiro (2019).

No Deslocamento 4 da amostra de 2019, dos 16 falantes que o compõem, treze (13) são da Bahia, dois (2) de São Paulo e um (1) do Mato Grosso do Sul. A diferença na origem do falante deste deslocamento pode interferir na distribuição dos resultados para o perfil.

A amostra Deslocamentos (2020), organizada por Cardoso (2021), Novais (2021), Pinheiro (2021), Rodrigues (2021), Silva (2020), Silva (2021) Souza (2022) e colaboradores, apresenta algumas diferenças em relação à amostra Deslocamentos (2019): (i) o Deslocamento 4 foi restringido a estudantes oriundos de Alagoas e da Bahia (Quadro 9), estados que fazem divisa com Sergipe; (ii) na variável *tempo no curso*, os períodos de abrangência foram reformulados, uma vez que ficava um "vazio" entre o 3º e o 7º período, para que o *início* abrangesse do 4º período para baixo e o *final* do 5º período para cima; e (iii) reduziu-se o número de participantes por célula (3), resultando em 60 informantes (Tabela 3).

Quadro 9 – Deslocamento 4 na amostra Deslocamentos (2020)

Deslocamento 4	Estudantes da UFS nascidos e criados em Alagoas e Bahia, que atualmente residem na Grande Aracaju.

Fonte: elaboração própria.

Tabela 3 – Estratificação da amostra Deslocamento (2020)

	Início		Final	
	Masculino	Feminino	Masculino	Feminino
Deslocamento 1	3	3	3	3
Deslocamento 2	3	3	3	3
Deslocamento 3	3	3	3	3
Deslocamento 4				
Alagoas	3	3	3	3
Bahia	3	3	3	3

Fonte: elaboração própria.

A amostra Linguagem Corporificada (2023), desenvolvida por Cardoso *et al*. (2024), é composta por sessenta e duas (62) entrevistas, das quais utilizamos cinquenta e sete (57), e

segue um caminho distinto das amostras Deslocamentos (2019 e 2020): ela não é estratificada considerando nenhum critério sociodemográfico, mas ainda considera estudantes universitários da UFS *campus* Prof. José Aloísio de Campos. O objetivo da amostra é captar, a partir da gravação simultânea de uma câmera frontal e outra lateral, produção linguística e gestos corporais que têm função comunicativa, partindo do pressuposto de que a linguagem é multimodal. Os trabalhos que a amostra subsidia (Cardoso, 2025; Assis, 2025) são de descrição pragmática e, consequentemente, têm como escopo de análise o efeito comunicativo, em contextos específicos, de determinados itens ou construções linguísticas.

Apesar de a amostra não ter um desenho que considere fatores sociodemográficos, é possível aplicá-los a partir das informações coletadas, considerando o local e o público de execução (Tabela 4), com células preenchidas de forma desbalanceada, mas que, dado o modelo de análise estatística utilizado (Seção 4.3), podemos acomodar esse desbalanceamento.

Tabela 4 – Estratificação da amostra Linguagem Corporificada (2023)

	Início			Final	
	Masculino	Feminino	Não-binário	Masculino	Feminino
Deslocamento 1	3	9	-	6	7
Deslocamento 2	-	6	-	4	1
Deslocamento 3	2	1	1	2	1
Deslocamento 4	1	7	1	2	3

Fonte: elaboração própria.[20]

A coleta das amostras segue o protocolo definido para o banco de dados Falares Sergipanos (Freitag, 2013), com entrevistas de cerca de 40-60min a partir de um roteiro de questões variadas: as primeiras perguntas são de checagem, fatos em relação ao falante; as demais são perguntas voltadas a questões sociais, como educação, segurança, saúde, igualdade de gênero etc. Além disso, na amostra Linguagem Corporificada (2023), há dois blocos específicos: um bloco sobre negação (Cardoso, 2025) e outro sobre *schardenfreude* (Assis, 2025) – o sentimento de prazer ou satisfação derivado da infelicidade ou do fracasso

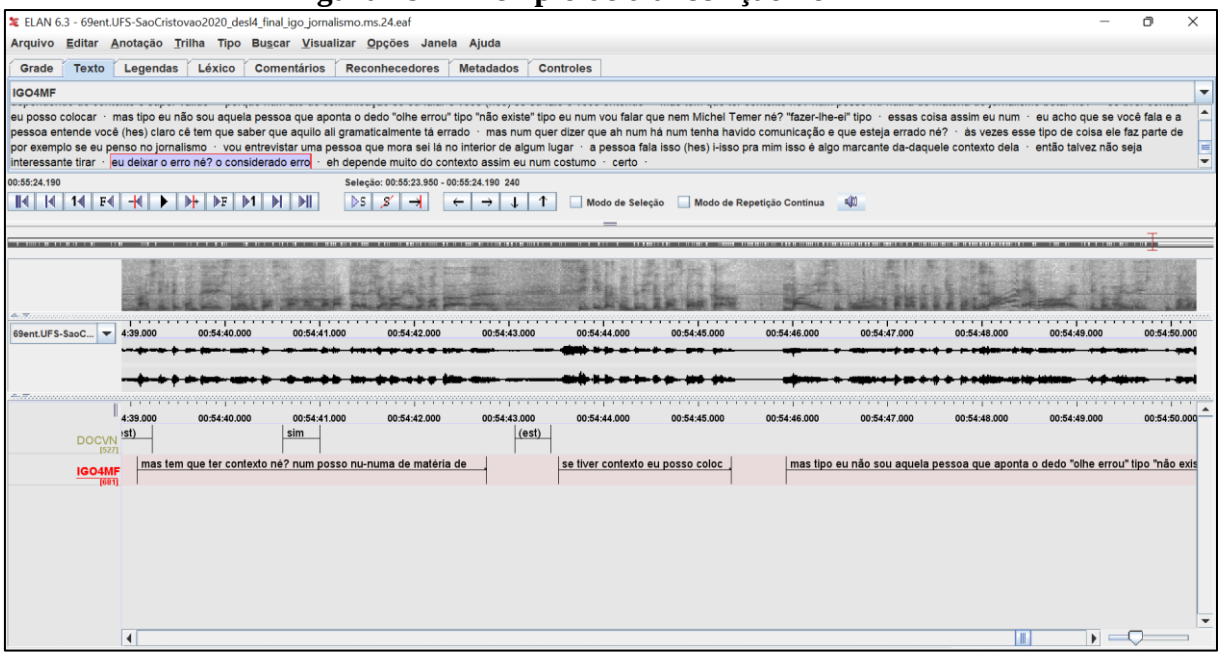
[20] A introdução de falantes do gênero não binário na sociolinguística desafia a tradicional estratificação de gênero entre masculino e feminino – presente nas amostras Deslocamentos (2019 e 2020) –, o que tende a impactar a forma como a língua é descrita e utilizada. Freitag (2025), em "Não existe linguagem neutra!", discute como a língua reflete e reforça estruturas sociais, incluindo as normas de gênero. A introdução de gênero não binário evidencia isso, o que leva a inovações linguísticas que refletem uma demanda por representação e inclusão, mas também geram resistência, pois desafiam normas estabelecidas.

de outra pessoa. Ambos são estruturados em situações hipotéticas, nas quais os participantes devem dizer como reagiriam a x situação.

As amostras Deslocamentos (2019 e 2020) foram constituídas seguindo uma amostragem aleatória estratificada, na qual recorremos de forma aleatória a estudantes do *campus*, e também de conveniência, visto a colaboração de estudantes voluntários e sua acessibilidade (Buchstaller; Khattab, 2013). Com isso, as amostras são rotuladas como de cota, ou de julgamento, visto o "número fixo de falantes em cada uma das categorias, que são preenchidas pelo pesquisador de campo por conta da conveniência e/ou julgamento" (Freitag, 2018, p. 671). A amostra Linguagem Corporificada (2023), dado o seu objetivo, segue um modelo de conveniência, em que os participantes ou elementos da amostra foram escolhidos com base em sua acessibilidade ou disponibilidade, em vez de critérios de representatividade ou aleatoriedade.

Todas as entrevistas foram transcritas com o suporte do software ELAN (Hellwig; Geerts, 2013) – que facilita a sincronização do áudio com a transcrição/anotação, de modo simultâneo (Figura 19) –, seguindo as normas de transcrição do Banco de Dados Falares Sergipanos, conforme apresentas em Sousa e Souza (2022), e revisadas, a fim de verificar e corrigir erros e/ou desalinhamento. Após este processo, todas as transcrições foram exportadas no formato do programa e em .txt (documento de texto padrão que contém texto não formado) para futuro uso na extração dos dados.

Figura 19 – Exemplo de transcrição no ELAN



Fonte: elaboração própria.

Esperamos que a utilização das amostra possibilite a extração de dados que demonstrem, em algum nível, diferenças entre os perfis de deslocamento quanto aos usos dos fenômenos selecionados (presença/ausência de artigo antes de possessivo pré-nominal; os pronomes *tu, você* e *cê* em função de sujeito; *te* e *lhe* em posição oblíqua; e possessivos *teu* e *seu*) – dado que outras pesquisas que lidam com esses fenômenos em alguma das duas amostras já evidenciam esse comportamento –, como também que apresentem padrões conjuntos de uso linguístico, contribuindo para o mapeamento dos fenômenos no PB.

Nas seções que seguem, visualizamos como extraímos e trabalhamos com os dados oriundos das amostras Deslocamentos (2019 e 2020) e Linguagem Corporificada (2023).

4.2 EXTRAÇÃO DOS DADOS

A Sociolinguística Variacionista lida com grande volume de dados, coletados em ordem de se observar o padrão linguístico de diferentes grupos de falantes. Ao longo dos anos, pesquisadores do ramo têm buscado maneiras de automatizar a análise desse grande volume de dados, ora por meio de desenvolvimentos próprios dentro da área (a exemplo do Varbrul e outras versões, p.ex. GoldVarb X, para análise estatística), ora adotando ferramentas computacionais de outras áreas, como da Linguística de Corpus,[21] com a utilização de concordanciadores, etiquetadores etc. A inclusão de ferramentas possibilita uma análise mais automatizada dos dados, evitando o amplo dispêndio de tempo por parte do pesquisador, como também auxilia na replicabilidade do método para outras pesquisas (Sousa, 2023).

Frente ao desenvolvimento dessas ferramentas, em nossa pesquisa utilizamos para a extração dos dados o spaCy 3.6 (Honiibal *et al*., 2020), biblioteca aberta de Processamento de Linguagem Natural (PLN), com base na linguagem de programação Python, com suporte para mais de 70 línguas, incluindo o português. O spaCy 3.6 permite lematização, tokenização, anotação POS (Part-of-Speech) e sintática, e recursos para buscas automáticas, como a função *Matcher* (busca de tokens) e o *Dependency Matcher* (que realiza buscas sintáticas) (Sousa, 2023).

Sousa (2023), em sua tese de doutorado com objetivo de estabelecer um protocolo de sistematização e etiquetagem de dados linguísticos, demonstra que o spaCy 3.6, em comparação com o *LancsBox* 6.0 (Brezina; Weill-Tessier; Mcenery, 2021), oferece o melhor

[21] A Linguística de Corpus trabalha com a coleta e exploração de *corpora*, ou conjuntos de dados linguísticos textuais coletados com o propósito de servirem para a descrição de uma língua ou de uma variedade. Além disso, dedica-se à exploração da linguagem por meio de computador (cf. Sardinha, 2000).

desempenho para a etiquetagem automática, frente ao seu conjunto de etiquetas POS mais simplificado. A autora também evidencia que "as buscas automáticas também foram melhores para o spacy 3.5, uma vez que essa biblioteca oferece mais recursos nesse quesito" (Sousa, 2023, p. 114), além de haver menores taxas de erros no resultado, maior quantidade de resultados e maior refinamento dos atributos de busca, o que torna essa biblioteca uma poderosa ferramenta para a busca automática por fenômenos linguísticos, auxiliando na descrição linguística. O protocolo desenvolvido por Sousa (2023) serviu como base para a extração de dados desta pesquisa.

Para o uso do spaCy 3.6, é necessária a utilização de um ambiente compatível com versão Python 3.6+, como o Google Colaboratory (Google Colab), ambiente de desenvolvimento integrado (IDE) *online* hospedado pelo próprio Google, ou o Jupyter Notebook (Kluyver *et al*., 2016), ambiente *offline,* com um visual simples e muito fácil de utilizar. Para o desenvolvimento de nossa pesquisa, utilizamos o Google Colab, por (i) permitir um acesso direto ao Google Drive, espaço em nuvem no qual depositamos nossos arquivos em .txt; e (ii) por ter uma interface mais amigável e mais veloz, uma vez que conta com suporte dos servidores Google, de forma gratuita. Nesse sentido, construímos, por meio do Google Colab, *notebooks* nos quais são informados os códigos que auxiliam no desenvolvimento de nossa tarefa. Os *notebooks* com seus códigos foram construídos com grande colaboração da Prof.ª Dr.ª Marta Deysiane Alves Faria Sousa.

Nosso foco, para o uso do spaCy 3.6, é a exploração de textos usando a função *Matcher*, que permite encontrar palavras e frases usando regras de busca com base no termo exato, em etiqueta POS ou ambos. A função gera uma lista de todas as instâncias de um ou mais *tokens* – cada ocorrência individual da palavra ou lema – pesquisados em *corpora*, como um concordanciador. Um concordanciador, na Linguística de Corpus, é utilizado para listar as ocorrências de uma palavra ou frase, com uma quantidade definida de contextos, tanto a sua esquerda quanto a sua direita. Em nossa pesquisa, realizamos a extração dos dados considerando a quantidade de 20 termos antes e 20 depois da variante.

Por fim, por ser uma biblioteca em linguagem Python, após a extração dos dados pelo spaCy 3.6, usamos a função *data.frame* da biblioteca pandas para armazenar os dados em planilha que, no futuro, pode ser utilizada em outras plataformas, como o R, para análises estatísticas descritivas e inferenciais. As colunas da planilha possuem as seguintes informações: ordem de ocorrência nos arquivos, nome do arquivo .txt do qual foi extraído o termo buscado, contexto anterior e seguinte (cuja quantidade já foi explicitada) e o contexto completo onde se encontra o termo buscado (Figura 20).

Figura 20 – Exemplo de extração dos dados em planilha

	A	B	C	D	E	F
1		Numero_ent	Contexto anterior	Ocorrencia	Contexto Seguinte	Contexto
2	0	02ent.UFS-SaoCrist	vai ter um na sala qu	querer te	zuarfazer uma piadin	vai ter um na sala q
3	1	02ent.UFS-SaoCrist	ter um na sala que v	te zuar	fazer uma piadinhae	ter um na sala que
4	2	03ent.UFS-SaoCrist	simeu acho que o fil	te dar	oportunidade de voc	simeu acho que o f
5	3	03ent.UFS-SaoCrist	comprava livrose m	te trazer	uma vai te trazer nov	comprava livrose n
6	4	03ent.UFS-SaoCrist	me incentivava ness	te trazer	novas perspectivase	me incentivava nes
7	5	03ent.UFS-SaoCrist	novas tecnologiaseu	posso te	falar é que parece qu	novas tecnologiase
8	6	03ent.UFS-SaoCrist	tecnologiaseu como	te falar	é que parece que a e	tecnologiaseu com
9	7	03ent.UFS-SaoCrist	ser quem você é eu	te passe	um mínimo de segura	ser quem você é eu
10	8	06ent.UFS-SaoCrist	eles dão uma datapi	te encaminham	para umum laboratór	eles dão uma data
11	9	09ent.UFS-SaoCrist	ela é a única aí tipo	lhe ajudou	que sempre puxou se	ela é a única aí tipo
12	10	11ent.UFS-SaoCrist	mas no formal tem (	te cobra	issoacho que (hes)o j	mas no formal tem
13	11	13ent.UFS-SaoCrist	do do (hes) do temp	hes) te	faz- (hes)ter mais ter	do do (hes) do tem
14	12	14ent.UFS-SaoCrist	local escurodo que v	te parar	e te assaltarmas voc	local escurodo que
15	13	14ent.UFS-SaoCrist	do que vocêque é ho	te assaltar	mas você não vai ter	do que vocêque é l
16	14	14ent.UFS-SaoCrist	te parar e te assalta	te assaltar	de alguém de moto t	te parar e te assalt
17	15	14ent.UFS-SaoCrist	mas você não vai te	te parar	e te estuprarde algué	mas você não vai t
18	16	14ent.UFS-SaoCrist	vai ter medo de algu	te estuprar	de alguém de motop	vai ter medo de alg
19	17	17ent.UFS-SaoCrist	(hes)ah sim vou fala	te falei	que tenho um tio qu	(hes)ah sim vou fal
20	18	21ent.UFS-SaoCrist	que realmente você	te botam	pra cápro João Alves	que realmente voc
21	19	23ent.UFS-SaoCrist	do armamento né u	pode lhe	matar ou matar uma	do armamento né
22	20	23ent.UFS-SaoCrist	armamento né uma	lhe matar	ou matar uma uma u	armamento né um
23	21	24ent.UFS-SaoCrist	crescendo e eu fico	puderem te	dar alguma coisa já c	crescendo e eu fic

Fonte: elaboração própria.

É a partir do uso da função *Matcher* do spaCy 3.6 que fazemos a extração de nossos dados. Contudo, nem todas as ocorrências se adequam àquilo que buscamos. O próprio spaCy 3.6 possibilita a "limpeza" dos dados indesejáveis, retornando apenas as ocorrências que são compatíveis ao contexto de uso da variável, com base nos comandos apresentados, conforme (a-d) e Quadro 10. Ainda que o spaCy apresente essa possibilidade, a "limpeza" dos dados também foi feita manualmente, após a extração dos dados em planilha.

a) Artigo antes de possessivo pré-nominal: apenas quando os possessivos precederem nomes;[22]
b) Pronomes pessoais de 2PS: apenas quando as variantes estiverem em posição de sujeito, foneticamente explícitas, precedendo ou sucedendo o verbo com o qual estabelece concordância.
c) Clíticos de 2PS: quando o clítico estiver em função dativa ou acusativa, em ênclise ou próclise;
d) Possessivos de 2PS: quando os possessivos precederem e sucederem nomes;

[22] Foram excluídos (i) contextos com demonstrativos, como em *essa minha irmã*; (ii) vocativos: *Meu irmão, como é que pode isso?*; (iii) expressões cristalizadas: *Meu Deus do Céu!*; (iv) expressões idiomáticas: *Cada macaco no seu galho*; (v) contextos em que o nome não aparece representado foneticamente na sentença: *trouxe meu casaco e o seu*; (vi) preposição para + a: *eu trouxe isso pra a minha mãe*; (vii) outros contextos em que é impreciso saber se há artigo: *quando o meu pai estava vivo*; e (viii) contextos em que o falante repete a pergunta do entrevistador: DOC: *qual a sua opinião sobre x?* FALANTE: *a minha opinião?*

Quadro 10 – Regras de busca para os fenômenos

Fenômeno	Regras	Contexto
Uso variável de artigo antes de possessivos	1. [{'LEMMA': {'NOT_IN': ['meu','teu', 'seu', 'nosso']}},[23] {'POS': 'DET', 'MORPH':{'IS_SUPERSET': ['PronType=Prs']}}, {'POS': 'NOUN'}]	1. Possessivos meu, teu, seu e nosso (e suas flexões) em contextos nos quais precedem nomes.
Pronomes pessoais de 2PS	1. [{'ORTH': {'IN': ['você', 'cê', 'tu']}}, {'POS': 'VERB', 'MORPH': {'IS_SUPERSET': ['VerbForm=Fin']}}] 2. [{'ORTH': {'IN': ['você', 'cê', 'tu']}}, {}, {'POS': 'VERB', 'MORPH': {'IS_SUPERSET': ['VerbForm=Fin']}}]	1. Pronomes *você*, *tu* e *cê* antecedendo verbos. 2. Pronomes *você*, *tu* e *cê* antecedendo verbos com material interveniente entre pronome e verbo.
Pronomes clíticos de 2PS	1. [{'ORTH': {'IN': ['te', 'lhe']}}, {'POS': 'VERB'}] 2. [{'POS': 'VERB'}, {'ORTH': {'IN': ['te', 'lhe']}}] 3. [{'POS': 'VERB'}, {'IS_PUNCT': True}, {'ORTH': {'IN': ['te', 'lhe']}}]	1. Pronomes *te* e *lhe* antecedendo verbos, em posição proclítica. 2. Pronomes *te* e *lhe* procedendo verbos, em posição enclítica. 3. Pronomes *te* e *lhe* procedendo verbos, em posição enclítica, ligados por algum sinal de pontuação (-).
Pronomes possessivos de 2PS	1.[{"LEMMA": {"IN": ["seu", "teu"]}}, {"POS": "NOUN"}] 2.[{"POS": "NOUN"}, {"LEMMA": {"IN": ["seu", "teu"]}}]	1. Pronomes *teu* e *seu* (e flexões) antecedendo nomes. 2. Pronomes *teu* e *seu* (e flexões) procedendo nomes.

Fonte: elaboração própria.

É sabido que há a possibilidade de existir material interveniente. Nos pronomes pessoais de 2PS, por exemplo, o verbo não necessariamente vem seguido do pronome, como "você não sabe o que aconteceu", em que há o advérbio *não* entre o pronome *você* e o verbo *saber*. O spaCy 3.6 possibilita que esses contextos não sejam excluídos da extração, por meio da função coringa (*wildcard* = {}), possibilitando que contextos nos quais há elementos entre o pronome e o verbo sejam também contabilizados.

Passada a extração dos dados e limpeza, é necessária a codificação dos dados inserindo as informações sociais do falante que produziu determinada ocorrência da variável (Figura

[23] Esta linha em questão desconsidera ocorrências de repetições de pronomes, como "o meu meu pai chegou".

21), como informações relativas à amostra, perfil de deslocamento, tempo no curso, idade e gênero, que são utilizadas na análise estatística.

Figura 21 – Exemplo de planilha com informações da amostra e do falante

	A	B	C	D	E	F
1	informante	amostra	deslocamento	tempo	idade	genero
2	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
3	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
4	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
5	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
6	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
7	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
8	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
9	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
10	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
11	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
12	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
13	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
14	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
15	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
16	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
17	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
18	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
19	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
20	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino
21	01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24	Deslocamentos 2019	Deslocamento 1	final	24	masculino

Fonte: elaboração própria.

Organizadas as planilhas com as ocorrências dos fenômenos morfossintáticos selecionados nesta pesquisa, passamos para a etapa de análise estatística dos dados.

4.3 ANÁLISE DOS DADOS

As análises estatísticas dos dados foram feitas na plataforma R (R Core Team, 2018), na interface RStudio (Rstudio Team, 2015), por meio de pacotes de visualização gráfica e estatísticos, como o `ggplot2` (Wickham, 2016) e `ggstatsplot` (Patil, 2021), que apresentam a distribuição da variável dependente em relação às variáveis independentes controladas por meio de gráficos. Os gráficos gerados já apresentam testes de associação. Para todas as análises conduzidas, o nível α (alfa) foi fixado em $p<0.05$, de acordo com o estipulado nas pesquisas da área das ciências sociais, humanas e cognitivas. O teste apresenta um p-valor, que é comparado com nosso p-valor pré-determinado, o α, que significa que, se repetirmos um teste 100 vezes, cinco dessas vezes o resultado pode ser diferente do obtido inicialmente. Com isso, o teste estatístico inferencial busca rejeitar a hipótese nula (H_0), segundo a qual os dados são efeito do acaso, não existe relação entre as variáveis, quando p-valor igual ou maior que 0,05; enquanto a H_1 é a hipótese alternativa, quando o p-valor for menor que 0,05: rejeitamos a H_0, existe uma relação entre duas variáveis. O resultado do teste

é significativo. Significância se refere à probabilidade de que uma associação, relação ou diferença entre variáveis seja real e não apenas uma coincidência, efeito do acaso.

A prévia observação do comportamento dos fenômenos morfossintáticos variáveis no Capítulo 2 possibilitou um maior conhecimento em relação à sua variação no PB. Assim, consideramos um valor de aplicação para as variáveis, com base no que foi apresentado como forma predominante em estudos prévios realizados em Sergipe (i-iv).

(i) Para a variação no uso de artigo antes de possessivo pré-nominal, o valor de aplicação é a ausência de artigo;
(ii) Para a variação nos pronomes pessoais de 2PS, o valor de aplicação é *você*;
(iii) Para a variação nos clíticos de 2PS, o valor de aplicação é *te*; e
(iv) Para a variação nos possessivos de 2PS, o valor de aplicação é *seu*.

Para a descrição e explicação dos dados, seguimos duas etapas: uma considerando cada fenômeno individualmente, outra considerando padrões conjuntos nos usos de mais de uma variável, conforme vemos no que segue.

4.3.1 Análise univariada

Cada um dos fenômenos variáveis foi descrito por meio de análise univariada: observa-se, individualmente, a possível relação entre a variável dependente (o fenômeno) e as variáveis independentes (condicionantes extralinguísticos). A motivação para este tipo de análise é observar se há relação entre as variáveis sociais – como deslocamento, tempo no curso, gênero e idade –, e a distribuição de cada uma das variáveis morfossintáticas.

Uma vez que partimos da hipótese de que os fenômenos morfossintáticos são dialetalmente distintos e salientes, a observação da possível relação entre a distribuição do fenômeno e as variáveis deslocamento e tempo no curso ajudam-nos a obter evidências para essa hipótese. Além disso, considerando que a covariação é socialmente motivada, como argumentam Guy (2013) e Oushiro (2015a), descrever o comportamento dos fenômenos por meio de condicionantes sociais permite-nos observar o comportamento das variáveis nos grupos de falantes.

Realizamos a análise univariada por meio do Teste de qui-quadrado. Quando as condições de qui-quadrado não foram cumpridas, utilizamos o Teste exato de Fisher para verificar se duas variáveis categóricas são independentes. Rejeitamos a H_0 – os dados são independentes – quando p-valor for menor que o nosso α – há associação entre os dados.

Usamos o V^2 de Cramer para medir a associação entre as variáveis, quando houve relação significativa. Essa medida de associação varia de 0 a 1, em que: i) 0 é a ausência da associação; ii) valores altos do V^2 de Cramer indicam uma relação mais forte entre as variáveis; iii) e os valores menores indicam uma relação fraca.

A visualização da variável numérica idade foi feita por meio de regressão logística (regressão linear generalizada), que permite a predição de valores categóricos a partir de uma série de variáveis independentes, tanto categóricas quanto numéricas, com base em *log odds* (*logs* de razões de chance de a variável de interesse ocorrer).

4.3.2 Análise de covariação

A análise univariada nos fornece informações essenciais sobre as variáveis envolvidas na descrição. A partir da compreensão do comportamento dos fenômenos morfossintáticos selecionados nesta pesquisa e sua (não) relação com os condicionantes sociais (deslocamento, tempo no curso, gênero e idade), é possível a obtenção de informações sobre cada variável individualmente. Com isso, a análise univariada dá suporte à interpretação dos resultados que podem ser obtidos através de uma análise de usos conjuntos de dados linguísticos, isto é, uma análise de covariação, pois fornece informações sobre o efeito individual de cada variável sobre cada variável dependente, já apontando evidências do que pode ser obtido por meio de uma análise de covariação.

A descrição da covariação foi feita considerando diferentes métodos de análise estatística. Em todos os métodos de análise utilizados, as variáveis são numéricas contínuas, extraídas de duas formas: i) taxas de uso da variante eleita como valor de aplicação: calculamos a taxa de realização por indivíduo. Por exemplo, se na fala de falante X há 100 ocorrências para a 2PS em posição de sujeito, e que dessas ocorrências 60 são para *você*, a taxa de realização para a variante é de 60% ((60*100)/100); ii) *log odds* de uso das variantes eleitas como valor de aplicação extraídos de modelos de efeitos mistos com falante como efeito aleatório (não inserimos as variáveis sociais no modelo), através do pacote `lme4` (Bates *et al*., 2015). A análise de covariação foi feita seguindo três etapas.

Na primeira etapa, adotamos métodos estatísticos de correlação de rho de *Spearman*, de modo a observar se a distribuição dos fenômenos linguísticos se correlaciona entre si, similar ao que foi feito em Guy (2013) e Oushiro (2015a; 2016a). Para observar a distribuição dos dados, utilizamos duas ferramentas em conjunto: teste estatístico de Shapiro-Wilk – cujo p-valor < 0.05 indica distribuição não normal –, e visualização gráfica por histograma.

Na segunda etapa, realizamos análises de agrupamento social, a partir de Guy (2013) e Oushiro (2015a; 2016a). Para tanto, utilizamos a taxa de frequência de uso individual das formas – os mesos dados da análise de correlação, conforme (i). Realizamos, *a priori*, uma divisão ternária, classificando frequências abaixo de 40% como B (baixas); entre 40% e 60% como M (médias), e acima de 60% como A (altas). Também realizamos uma classificação binária, adotando A para taxas iguais ou superiores a 50%, e B para taxas inferiores a 50%.

Na terceira etapa, conforme desenvolvido por Freitag (2022), utilizamos uma análise de *cluster,* por meio de k-*medoids*: *medoids* são objetos representativos de um conjunto de dados ou um *cluster* dentro de um conjunto de dados cuja soma das distâncias a outros objetos no cluster é mínima. K-*medoids* formam *clusters* com base na distância para *medoids*. Uma vez que o estudo da autora permitiu observar padrões conjuntos de uso, observando quais perfis sociais fazem determinados usos linguísticos quanto aos fenômenos variáveis analisados, consideramos que a metodologia empregada, em conjunto com análise de correlação, permite uma descrição satisfatória em relação à covariação de variáveis morfossintáticas dialetais no PB.

Os resultados das análises, tanto univariadas quanto de covariação, são apresentados no que segue.

5. DESCRIÇÃO E ANÁLISE DE VARIÁVEIS INDEPENDENTES

No capítulo 2, discutimos sobre como variáveis linguísticas podem indexar significado dialetal: a associação de certas formas linguísticas a determinadas regiões geográficas ou comunidades linguísticas específicas. Além disso, argumentamos que fenômenos morfossintáticos também carregam significado dialetal, isto é, informação indexical de quem e de onde se fala, o que Freitag *et al*. (2016) chamam de "sotaques sintáticos". Para embasar esse argumento, demonstramos, por meio de levantamento bibliográfico, o comportamento de quatro variáveis morfossintáticas:

i) uso variável de artigo antes de possessivo pré-nominal;
ii) pronomes pessoais de 2PS em posição de sujeito;
iii) pronomes clíticos de 2PS; e
iv) pronomes possesivos de 2PS.

Este capítulo visa atestar esse argumento, cujo objetivo é descrever a distribuição das variáveis morfossintáticas (i)-(iv) a partir de dados extraídos de nossas amostras, Deslocamentos (2019) (doravante D2019), Deslocamentos (2020) (doravante D2020) e Linguagem Corporificada (2023) (doravante LC2023). Para guiar as análises, lançamos as seguintes questões: i) qual o padrão de realização das variáveis em nossas amostras?; ii) há diferenças na frequência de uso das quatro variáveis morfossintáticas entre os perfis de deslocamento dos falantes?; iii) o tempo no curso interfere no modo como os falantes fazem uso das variáveis?; iv) como os falantes se comportam individualmente quanto ao uso das variáveis morfossintáticas?; v) fatores extralinguísticos, como gênero e idade, interferem nos usos linguísticos dos falantes quanto às variáveis morfossintáticas descritas?

Como hipóteses, a partir do discutido em seções anteriores, consideramos que i) há predomínio da ausência de artigo, do pronome *você* em posição de sujeito de 2PS, do pronome *te* como clítico de 2PS e do possessivo *seu* de 2PS; ii) a frequência de uso das quatro variáveis morfossintáticas apresenta diferenças ao longo dos perfis de deslocamento dos falantes, dado que as variáveis evidenciam comportamentos distintos a depender da região geográfica dos falantes; iii) a inserção de um falante em uma nova comunidade, diversa quanto à origem de seus falantes, desencadeia um processo de mudança linguística quanto às variáveis morfossintáticas analisadas; iv) falantes apresentam variação em seus usos linguísticos, mas

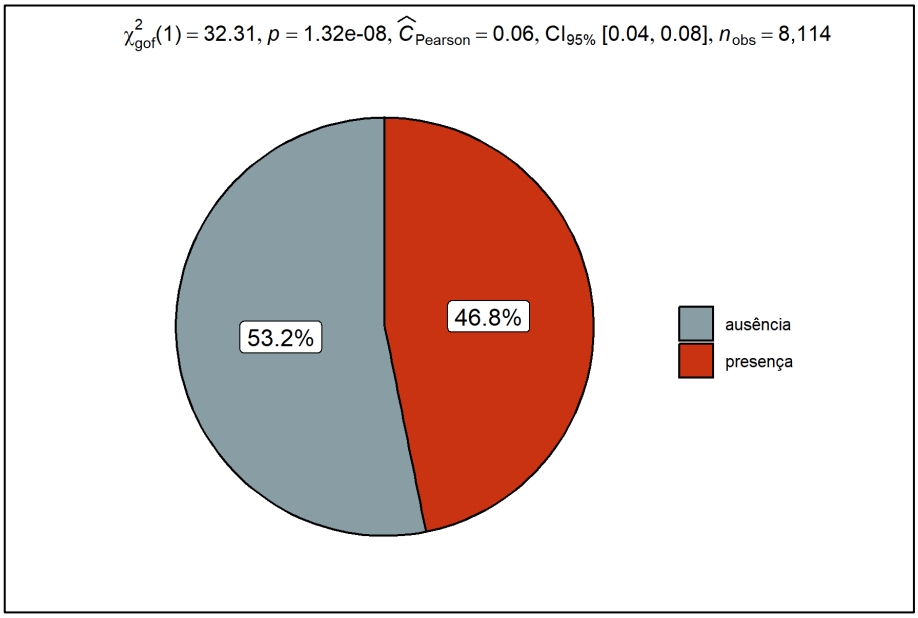
tendem a fazer uso de uma das variantes com maior frequência; e v) há relação entre gênero e idade na distribuição dos fenômenos variáveis.

No que segue, iniciamos a descrição e análise dos dados a partir da observação de padrões globais de uso. Em seguida, cotejamos nossos dados considerando a amostra, em ordem de observar se os resultados são sensíveis ao tipo de coleta. Após isso, observamos a possível associação dos dados com base em outros fatores, na ordem: i) deslocamento, com vistas a observar se há diferença notável entre grupos geograficamente distintos; ii) tempo no curso, buscando inferir, por meio da distribuição, efeito de exposição à comunidade; iii) indivíduo, para descrever o alcance da variação; e iv) gênero e idade, uma vez que fenômenos variáveis se correlacionam com características sociais de seus falantes.

5.1 DISTRIBUIÇÃO GERAL DOS DADOS

As Figura 22-Figura 25 apresentam a distribuição global dos quatro fenômenos morfossintáticos variáveis após a extração, organização e quantificação de todas as ocorrências nas 181 entrevistas que compõem as amostras utilizadas. A observação dos padrões gerais de uso auxilia na compreensão do comportamento das variáveis no português da comunidade da UFS, permitindo a visualização de como os falantes universitários de nossas amostras se comportam linguisticamente quanto ao uso dos fenômenos morfossintáticos. Iniciamos a descrição pela distribuição do uso variável de artigo antes de possesivo (Figura 22).

Figura 22 – Distribuição do uso variável de artigo antes de possessivos pré-nominais nas amostras



Fonte: elaboração própria.

Das 8.114 ocorrências de possessivos pré-nominais no *corpus*, 53,2% (4313/8114)[24] apresentam a ausência de artigo definido os antecedendo, como (1). Essa distribuição é estatisticamente significativa ($X^2(1, N= 8,114) = 32.31$ $p < 0.001$). Nos dados das amostras utilizadas, predomina a ausência de artigo definido antes de possessivos pré-nominais.

(1) eu sou de Minas por exemplo no meu caso especificamente no meu caso não mas **minha mãe meus irmão**s falando as pessoas identificam logo como de Sergipe falam eu acho a linguagem do interior mais (34ent.UFS-SaoCristovao2020_desl3_inicio_mir_jornalismo.fs.18).

(2) assaltado você não pode sair mais à noite se você sai à noite você fica preocupado se **o seu filho** precisa sair à noite pra estudar as pessoas ficam sempre presas e vivem amedrontadas eu já fui assaltado (22ent.UFS-SaoCristovao2018__desl. II_final_elv.ms.20).

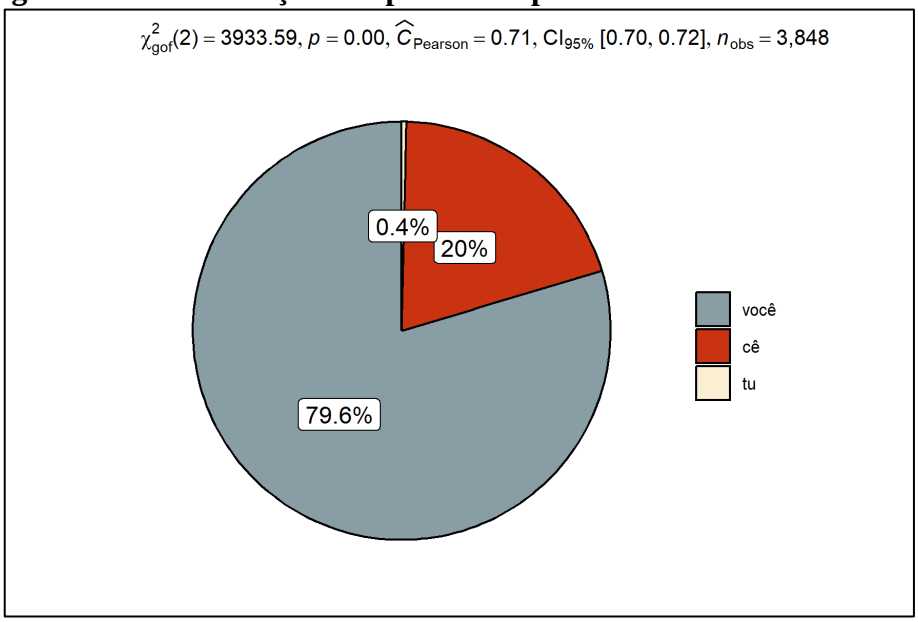
Esses resultados são similares aos dos estudos descritivos apresentados no Capítulo 2 feitos com base no PB, que apontam para uma tendência na qual na região Nordeste há predomínio para a ausência de artigo em sintagmas com possessivos (Pereira, 2017; Guedes,

[24] Os dados são apresentados considerando percentual (p.ex. 53,2%), ocorrências da variante (p.ex. 4313) e total de ocorrências da variável (p.ex. 8114).

2019). Em termos de comportamento dialetal, a distribuição geral dos dados converge com o comportamento da região nordestina para a ausência.

A Figura 23 apresenta a distribuição dos pronomes pessoais de 2PS.

Figura 23 – Distribuição dos pronomes pessoais de 2PS nas amostras



Fonte: elaboração própria

O pronome pessoal *você*, como (3), apresenta a maior frequência de uso (79,6% 3064/3848), seguido pela variante *cê* (21,5% 770/3848), em (4), e pela variante *tu* (0,4% 14/3848), em (5). Não houve ocorrência de *ocê* em todas as 3.848 realizações da variável. A distribuição dos dados é estatisticamente significativa (X^2(2, N= 3,848) = 3933.59 p = 0.00).

(3) a própria universidade tipo a própria universidade tem eventos cria eventos pra que **você** saia pra que **você** se distraia e tudo o mais pra que **você** tenha uma vida separada da vida acadêmica claro **você** tem (01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24).

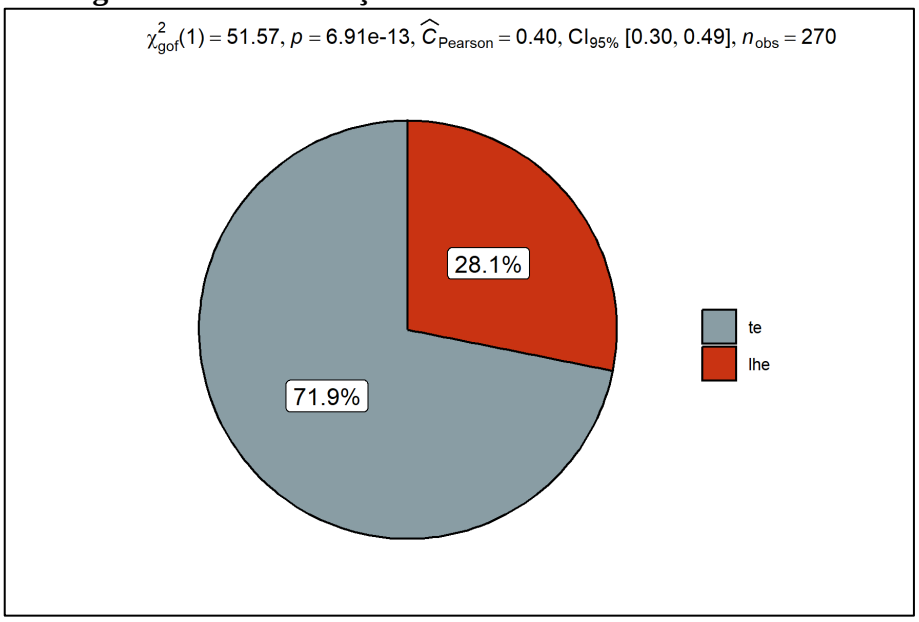
(4) parque tipo a Disneylândia não mas tipo um lugar que é/ você possa sair com seus filhos **cê** pode tipo **cê** pode sair com s/ **cê** possa respirar o ar puro **cê** possa ver (01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24).

(5) sabe? ninguém pode falar aí não assim e/ a/ quando marca fora né? aí **tu** não fica tão à vontade quando a a gente quando a gente fica na casa de alguém aí geralmente marca (11ent.UFS-SaoCristovao2018__desl. I_final_evi.fs.21).

Considerando o mapa proposto por Scherre *et al.* (2015, p. 142) (Figura 2), podemos dizer que o comportamento dos universitários da UFS dialoga com o sexto grupo (VOCÊ/*tu* – *tu* de 1% a 90% sem concordância), uma vez que há predomínio das variantes *você/cê*, e apenas 12 ocorrências de *tu*. Além disso, nossos dados são similares aos de outras pesquisas feitas com base no português falado no Nordeste, em que a forma *você* é predominante (Oliveira, 2007; Nogueira, 2013; Nascimento; Paim, 2016; Silva; Vitório, 2019; Guimarães, 2019).

Com falantes das amostras utilizadas, a variação tende a ser restrita às formas *você* e *cê*, dada a baixa frequência do pronome *tu* para o quantitativo de dados (N = 3.848). Se há predomínio de *você* e *cê* e uso extremamente reduzido de *tu*, é de se esperar que formas pertencentes ao paradigma pronominal de 2PS (*te* e *teu*) também ocorram de forma reduzida, uma vez que a implementação de *você* resulta na adoção de formas de 3PS (*lhe* e *seu*) no paradigma de 2PS. A Figura 24 apresenta a distribuição dos pronomes clíticos de 2PS.

Figura 24 – Distribuição dos clíticos de 2PS nas amostras



Fonte: elaboração própria.

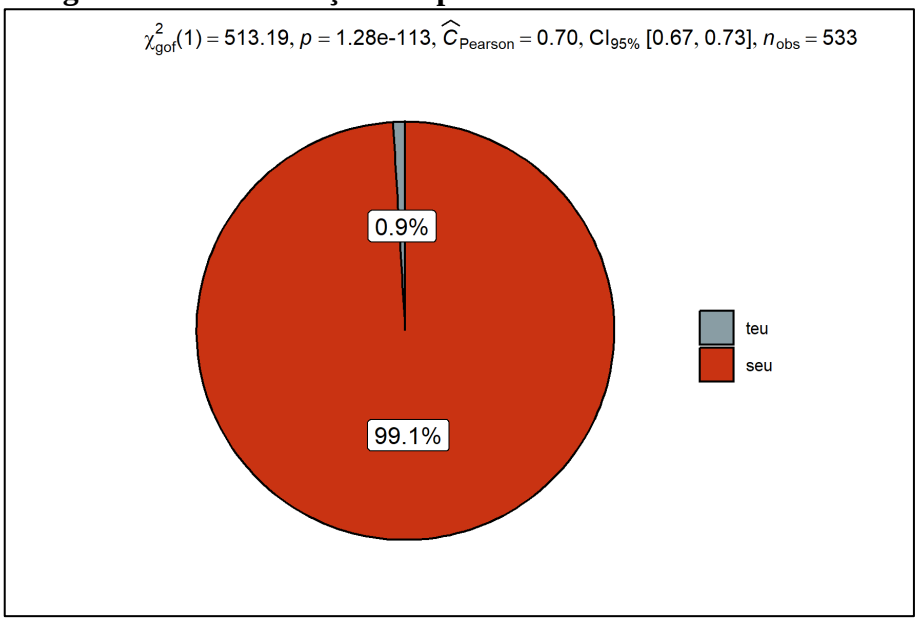
O quantitativo de dados para clíticos de 2PS é baixo (N= 270), o que pode interferir em análises estatísticas futuras. Desses dados, predomina o pronome *te* (71,9% 194/270), em (6), do paradigma de 2PS, para 28,1% (76/270) de *lhe*, em (7). A distribuição é estatisticamente significativa ($X^2(1, N= 270) = 51.57$ $p < 0.001$).

(6) eu fico sempre lendo tem minha mãe muito porque primeiro porque é conhecimento segundo porque **te** ajuda a se comunicar melhor o livro é como se você tivesse viajando mesmo que não seja pra um lugar (34ent.UFS-SaoCristovao2020_desl3_inicio_mir_jornalismo.fs.18).

(7) né? antigamente era muito incomum você tipo sei lá sair com alguma coisa na rua e alguém pegar e **lhe** roubar muito difícil eu amava minha cidade que era uma cidade pequena só que toda cidade já foram pequena (40ent.UFS-SaoCristovao2020_desl3_final_dey_engcivil.ms.23).

Os resultados vão contra o que pontuamos acima, dado o predomínio de *te*, e corroboram o que propuseram Scherre e Duarte (2016), segundo as quais o pronome *te* ainda é consistentemente utilizado no Brasil independentemente da região geográfica, enquanto *lhe* tem aparecido em variação com *te*, processo iniciado na região Nordeste. Além disso, diferem-se do que foi pontuado por Ramos (1999) sobre a utilização de *você* como pronome de 2PS e o clítico *lhe* como substituto à forma *te* em capitais do Nordeste. Ainda que haja predomínio de *você*, o pronome *lhe* tem frequência relativamente baixa em nossos dados. Os resultados também se distanciam de outros dados para o Nordeste, nos quais há predomínio de *lhe* ou frequência próxima entre as variantes (Almeida, 2009; 2014; Gama, 2019; Araujo; Borges, 2021) e aproximam-se de dados do Sul (Dalto, 2002), nos quais há predomínio de *te*. A Figura 25 apresenta os dados em relação à variação nos possessivos de 2PS.

Figura 25 – Distribuição dos possessivos de 2PS nas amostras



Fonte: elaboração própria.

O baixo quantitativo de dados também se observa no possessivo *teu*, em (8), que representa 0,9% (5/533) dos dados das amostras, resultado, possivelmente, de um paralelismo entre os usos de *tu* ~ *teu* – princípio de que formas levam a formas similares. É possível que a utilização de *você* como pronome pessoal de 2PS resultou na também utilização de *seu* (98,6% 528/533), em (9), como pronome de 2PS, frente ao seu predomínio em nossos dados, cuja distribuição é estatisticamente significativa ($X^2(1, N= 533) = 513.19$ $p < 0.001$).

(8) estagiando na **sua** área em outro país e aí quando você voltar tipo você vai ter lá no **seu** currículo "estagiei eh no Japão na minha área" então cê vai derrubar qualquer outro currículo (59ent.UFS-SaoCristovao2018__desl. IV_início_mun.fs.19).

(9) bem complicado imagina tá lá imagina você tá no ônibus com sua família aí vai lá um ladrão rouba **teu** telefone e tu num tá com **teu** telefone ele vai e mata tu ou um dos **teus** cê vai (65ent.UFS-SaoCristovao2020_desl4_inicio_bia_biologia.fs.21).

A distribuição é similar à de outras pesquisas, uma vez que segue a tendência apresentada em Siqueira (2021) para falantes da UFS e se distingue dos dados do Sul, apresentados em Arduin (2004; 2005) e Mendes (2008), nos quais há predomínio de *teu*. Em nossos dados, uma vez que falantes tendem a empregar o pronome *você* como sujeito de 2PS, e que a implementação desse pronome resultou em uma reestruturação no paradigma pronominal, é possível que o uso de *seu* seja resultado dessa reestruturação. Isso, evidentemente, pode ser validado através de análises de correlação, o que é feito no capítulo de covariação.

A distribuição global de nossos dados apresenta quatro tendências:

1. Predomínio da ausência de artigo antes de possessivos pré-nominais;
2. Predomínio do pronome *você* em posição de sujeito de 2PS, variando com *cê* e baixa frequência de *tu*;
3. Predomínio do clítico *te*;
4. Predomínio do possessivo *seu*, com baixa frequência de *teu*.

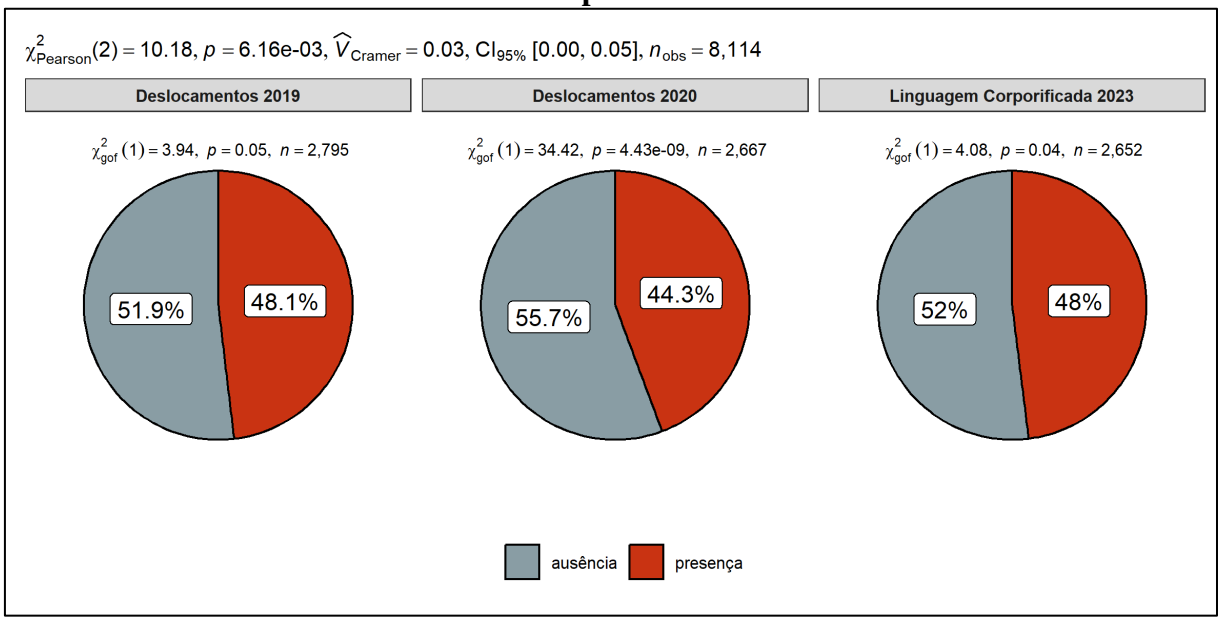
A observação dos padrões gerais fornece informações em relação ao comportamento das variáveis na comunidade da UFS, porém não nos dá indícios do comportamento a partir da fatores extralinguísticos. Nas seções que seguem, observamos como esse comportamento se relaciona com as variáveis controladas nesta pesquisa, iniciando pelo tipo de amostragem.

5.1.1 Distribuição por amostra

Os dados apresentados na seção anterior não discriminam as amostras. Ainda que as três amostras lidem com estudantes universitários, com a mobilidade do falante, tempo no curso, gênero e idade, há diferenças metodológicas em relação a sua constituição, como a quantidade de falantes por célula, a abrangência dos tempos no curso e a delimitação do Deslocamento 4. Assim, é possível que haja diferenças quanto aos usos linguísticos ao comparar as três amostras, uma vez que, como argumentam Freitag e Rost-Snichelotto (2015), diferenças/similaridades encontradas em análises e descrições podem ser resultantes da metodologia amostral.

Nesta seção, observamos o comportamento dos fenômenos variáveis considerando a amostra, com vistas a observar se os resultados são sensíveis ao tipo de coleta. Iniciamos a descrição pelo uso variável de artigo definido antes de possessivo (Figura 26).

Figura 26 – Distribuição do uso variável de artigo definido antes de possessivo pré-nominal por amostra



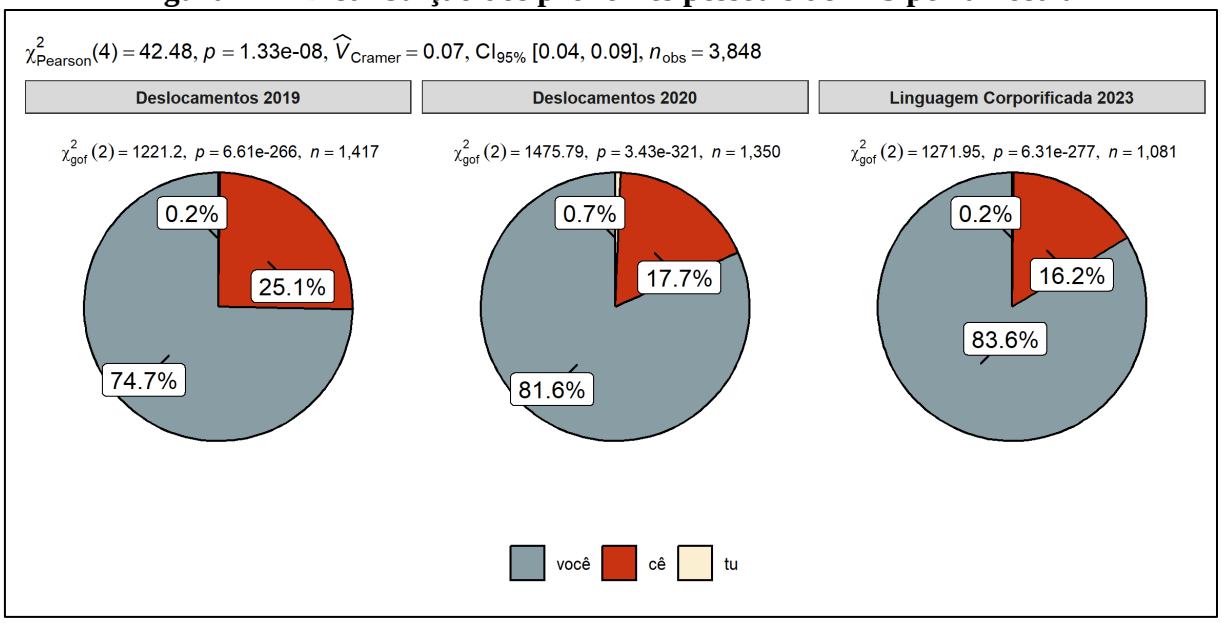
Fonte: elaboração própria.

Há predomínio da ausência de artigo em todas as amostras. Na amostra D2019, a frequência para a ausência é de 51,9% (1450/2795), na amostra D2020 a frequência é de 55,7% (1485/2667) e na amostra LC2023 a frequência é de 52% (1378/2652). Rejeitamos a hipótese nula: a diferença entre as amostras é estatisticamente significativa (X^2(2, N= 8114) = 10.18 p < 0.001). Falantes da amostra D2020 fazem maior uso da ausência de artigo antes de possessivos pré-nominais do que falantes das amostras D2019 e LC2023. Ainda que as

amostras apresentem frequências distintas para a variação, o predomínio da ausência em todos os *corpora* evidencia uma tendência para a não utilização do artigo definido antes de possessivos por estudantes universitários da UFS.

A Figura 27 apresenta os resultados por amostra para os pronomes pessoais de 2PS em posição de sujeito.

Figura 27 – Distribuição dos pronomes pessoais de 2PS por amostra

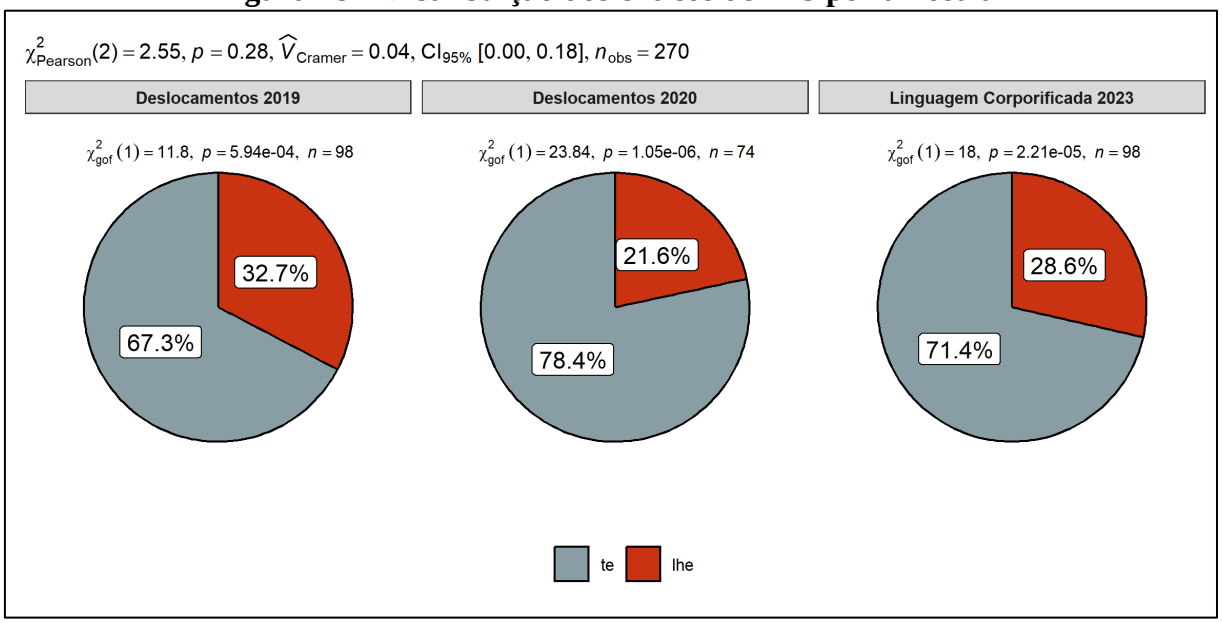


Fonte: elaboração própria.

Rejeitamos a H_0 ($X^2(4, N= 3,848) = 42.48\ p < 0.001$) e apresentamos a H_1: há diferença significativa entre as três amostras. A amostra LC2023 apresenta a maior frequência para o pronome *você* (83,6% 904/1081), seguida pela amostra D2020 (81,6% 1102/1350), enquanto a amostra D2019 apresenta a maior frequência da variante reduzida *cê* (25,1% 356/1417). O maior quantitativo do pronome *tu* é observado na amostra D2020, com 9 das 12 ocorrências de todo o *corpus*, correspondendo a 0,7% dos dados. Em todas as amostras, os falantes apresentam comportamento predominante para *você* e *cê*.

A Figura 28 dispõe a distribuição dos clíticos de 2PS por amostra.

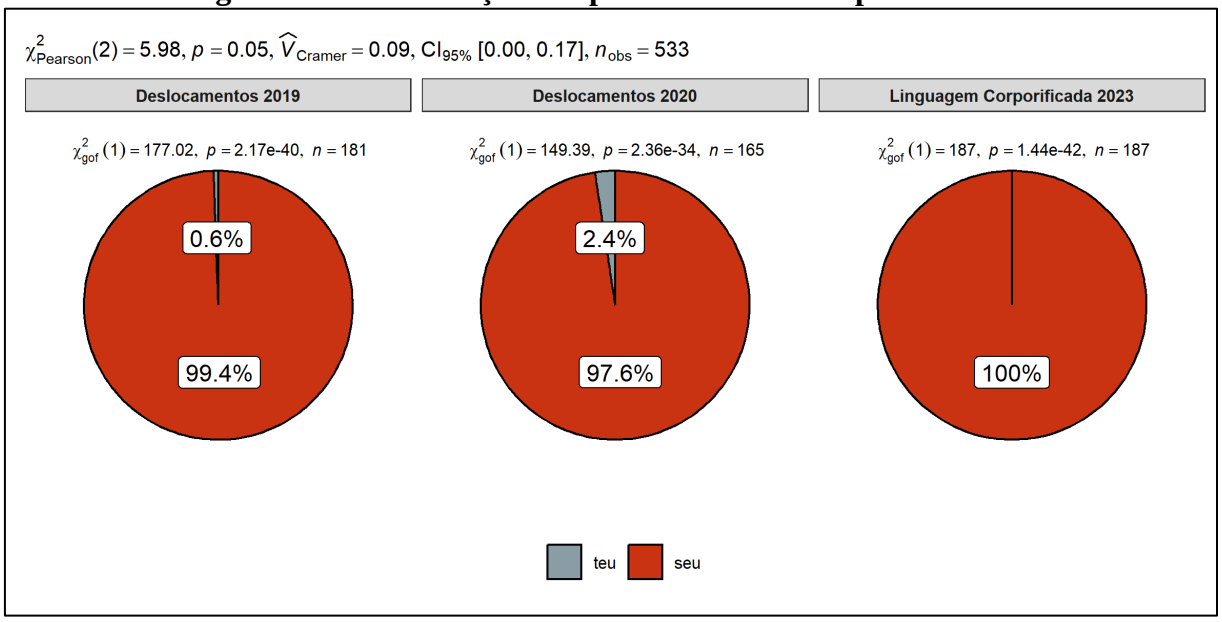
Figura 28 – Distribuição dos clíticos de 2PS por amostra



Fonte: elaboração própria.

A amostra D2019 apresenta a menor frequência para o uso do clítico *te* (67,3% 66/98) comparada às amostras D2020 (78,4% 58/74) e LC2023 (71,4% 70/98); D2020 possui menos dados de clíticos (N= 74) em comparação à D2019 (N= 98) e LC2023 (N= 98). Contudo, falhamos em rejeitar a H_0 (X^2(2, N= 270) = 2.55 p = 0.28), uma vez que não há diferença significativa na distribuição dos dados entre as amostras. Os dados por amostra referentes aos possessivos de 2PS podem ser visualizados na Figura 29.

Figura 29 – Distribuição dos possessivos de 2PS por amostra



Fonte: elaboração própria.

O possessivo *seu* apresenta um percentual alto em todas as amostras. A maior frequência é observada na fala de universitários da amostra LC2023, em que todos os 187 dados correspondem ao pronome *seu* (100% 187/187), seguida pela amostra D2019, com percentual de 99,4% (180/181), contra os 97,6% (161/165) da amostra D2020. A similaridade entre as amostras e a baixa ou nenhuma frequência de *teu* evidenciam que a diferença entre as amostras não é estatisticamente significativa ($X^2(2, N= 533) = 5.98$ $p = 0.05$). A existência de apenas cinco (5) casos para a variante *teu* é indício de que a maior parte dos falantes exibe um comportamento categórico para a variação.[25]

A distribuição dos dados é sensível ao tipo de coleta apenas quanto ao uso variável de artigo antes de possessivos e à variação nos pronomes pessoais de 2PS em posição de sujeito, frente aos resultados do teste de qui-quadrado. Na variação dos clíticos de 2PS e dos possessivos de 2PS, falhamos em rejeitar a H_0, o que sugere que não há interferência do tipo de amostra sobre a distribuição dos dados. Independente do teste estatístico, entre amostras os dados seguem a mesma tendência, isto é, predomínio da ausência de artigo, do pronome *você*, do clítico *te* e do possessivo *seu*.

Considerando o tipo de coleta, podemos passar para a visualização da possível relação com variáveis extralinguísticas, iniciando pela variável deslocamento.

5.1.2 Distribuição por deslocamento

Temos assumido uma posição de que há fenômenos morfossintáticos dialetalmente distintos, isto é, que apresentam comportamentos diferentes a depender da região geográfica de grupos de falantes. Grupos de falantes de comunidades geográficas diferentes podem possuir padrões linguísticos diferentes: usos linguísticos específicos de regiões distintas resultam em uma variedade ou dialeto da língua, representando o padrão de realização da língua de uma determinada comunidade. Com isso, além de condicionantes sociais macro (p.ex. escolaridade, faixa etária e gênero) que podem interferir no uso linguístico de um grupo de falantes, a região geográfica também atua sobre a constituição de sua variedade, conforme pesquisas de Corrêa (2019), Ribeiro (2019), Silva (2020) entre outras.

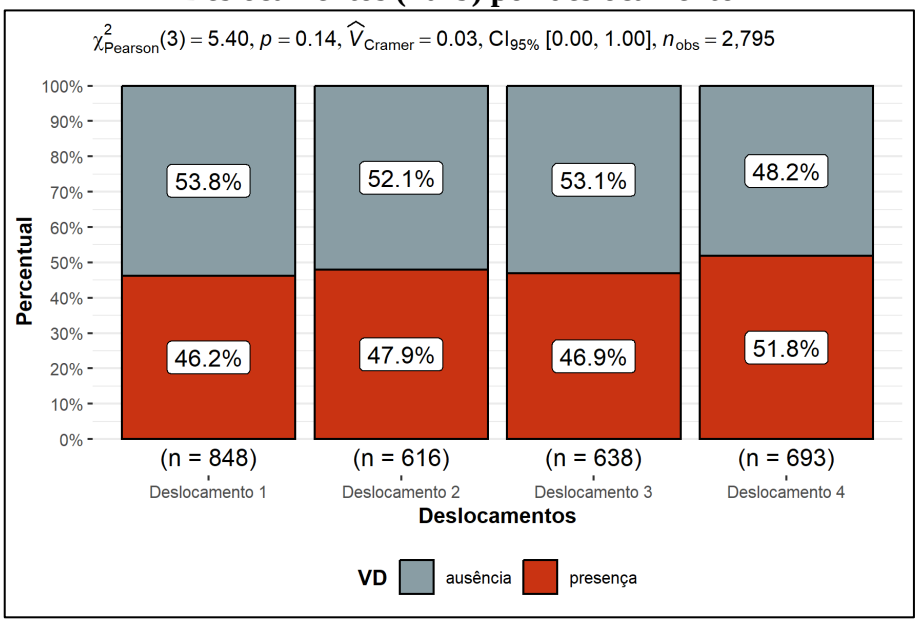
Retomando Mainsfield, Leslie-O'neill e Li (2023), uma variável é dialetal observando que ela não requer uma divisão categórica entre grupos, mas apenas uma diferença notável

[25] Não entraremos em discussões do que se configura como variação a partir de frequências. Contudo, é importante evidenciar a visão de Labov (2008[1972]) de que a variação é uma constante nos usos da língua, independentemente da frequência. O autor enfatiza que mesmo variantes de baixo uso podem ter significados sociais ou identificar grupos.

entre grupos geograficamente localizados em relação à variável. Ao considerar que uma variável dialetal não requer uma divisão categórica rígida, mas uma gradação entre regiões, questionamos se há diferenças na frequência de uso das quatro variáveis morfossintáticas entre os perfis de deslocamento dos falantes. A partir do observado em pesquisas anteriores sobre os fenômenos, tomamos como hipótese que a frequência de uso das quatro variáveis morfossintáticas apresenta diferenças ao longo dos perfis de deslocamento dos falantes, dado que as variáveis evidenciam comportamentos distintos a depender da região geográfica do falante, o que as sinalizam como variáveis dialetais. Assim, esperamos observar diferenças na frequência de uso de cada variável à medida que consideramos diferentes perfis de deslocamento, de modo a evidenciar que os fenômenos variáveis são dialetalmente distintos em nossos dados.

Para tanto, é necessária a utilização de amostra(s) que possibilita(m) esse tipo de visualização, o que nos leva a utilizar as amostras Deslocamentos (2019 e 2020) e Linguagem Corporificada (2023). A utilização desses dados possibilita visualizar se grupos de falantes que pertencem a regiões distintas apresentam, em seu repertório linguístico, diferentes frequências para os fenômenos variáveis morfossintáticos que descrevemos. Iniciamos a discussão pelo uso variável de artigo antes de possessivo pré-nominal (Figura 30).

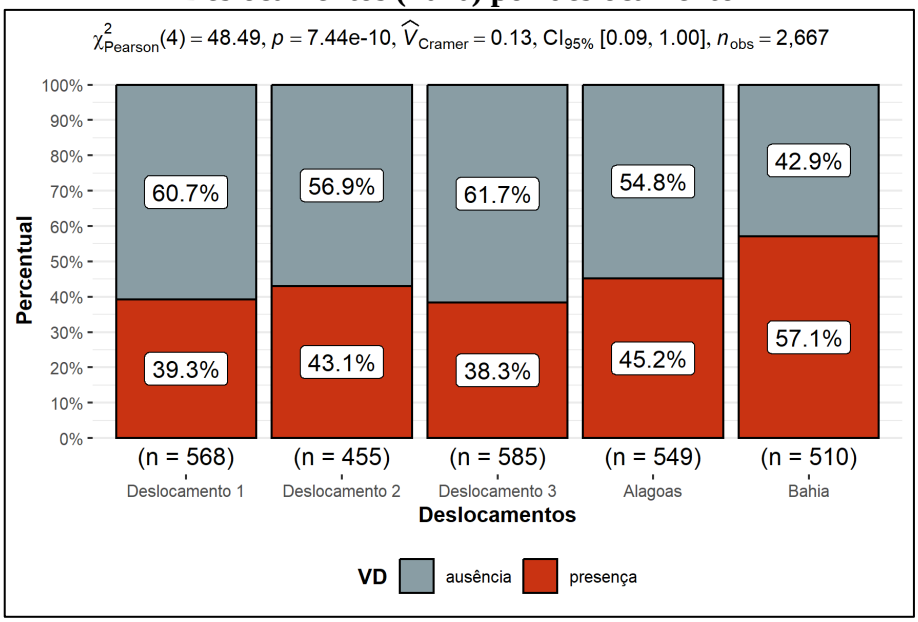
Figura 30 – Distribuição do uso variável de artigo antes de possessivo na amostra Deslocamentos (2019) por deslocamento



Fonte: elaboração própria.

Na amostra D2019, a maior frequência para a ausência de artigo é observada com falantes do Deslocamento 1 (53,8% 456/848), nascidos e residentes na região metropolitana do estado, seguidos por falantes do Deslocamento 3 (53,1% 339/638), do interior do estado que residem na região metropolitana, e por falantes do Deslocamento 2 (52,1% 321/616), do interior do estado que fazem o trajeto diário. O Deslocamento 4 (48,2% 334/693) é o único que apresenta predomínio para a presença de artigo antes de possessivo (51,8% 359/639), falantes externos ao estado de Sergipe. Contudo, falhamos em rejeitar a hipótese nula (X^2(3, N= 2,795) = 5.40 p = 0.14). Ainda que a frequência para a ausência se distribua de diferentes formas para os grupos, estatisticamente essa diferença não é significante. Observemos os dados na amostra D2020 (Figura 31).

Figura 31 – Distribuição do uso variável de artigo antes de possessivo na amostra Deslocamentos (2020) por deslocamento

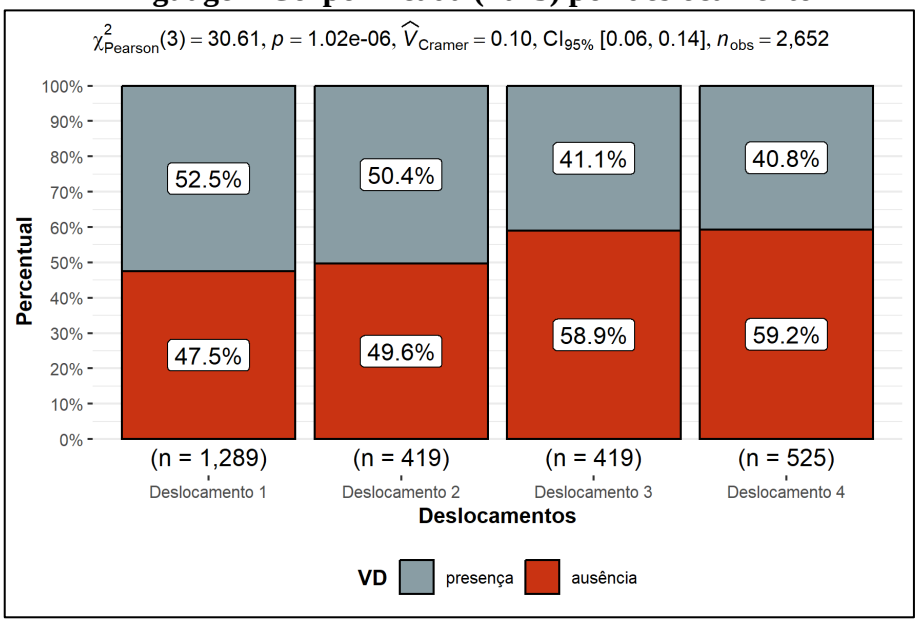


Fonte: elaboração própria.

Em D2020, as frequências da ausência nos Deslocamentos 1-3 são relativamente maiores em comparação à amostra de 2019: a maior frequência é observada nos dados do Deslocamento 3 (61,7% 361/585), seguido por falantes do Deslocamento 1 (60,7% 345/568) e Deslocamento 2 (56,9% 259/455). Similar aos dados de D2019, são os falantes externos a Sergipe que apresentam as menores frequências para a ausência: com dados de falantes de Alagoas, ainda que predomine a ausência (54,8% 301/549), sua frequência é relativamente menor comparada aos dados de Sergipe; os dados da Bahia, por sua vez, são os únicos que apresentam predomínio para a presença (57,1% 291/510), similar aos dados de D2019, em

que o Deslocamento 4 possui treze (13) falantes baianos. O p-valor nos permite rejeitar a H_0 ($X^2(4, N= 2,667) = 48.49$ $p < 0.001$), há relação entre o deslocamento e a variável dependente, com efeito pequeno ($V^2 = 0.13$). A Figura 32 apresenta a distribuição na amostra LC2023.

Figura 32 – Distribuição do uso variável de artigo antes de possessivo na amostra Linguagem Corporificada (2023) por deslocamento



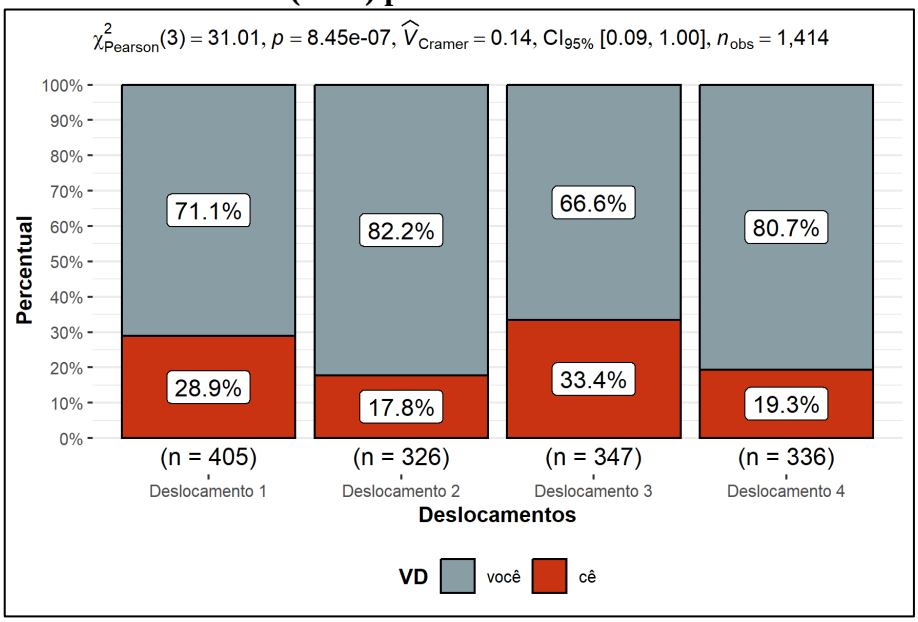
Fonte: elaboração própria.

A distribuição da amostra LC2023 para o uso variável de artigo antes de possessivo se distancia do observado nas outras amostras. Isso porque nos Deslocamentos 1 (52,5% 612/1289) e 2 (49,6% 208/419) há predomínio para a presença de artigo, enquanto o Deslocamento 4, que nas outas amostras apresenta maior frequência para a presença, evidencia maior uso da ausência (59,2% 311/525). A diferença observada entre os perfis de deslocamento é estatisticamente significativa ($X^2(3, N= 2,652) = 30.615$ $p < 0.001$).

Os dados de D2020 e LC2023 corroboram com a hipótese de que essa variável é dialetalmente distinta: grupos de falantes de diferentes regiões apresentam diferentes padrões linguísticos de uso da ausência de artigo.

Passemos para a observação da variação nos pronomes pessoais de 2PS. Conforme discutimos na seção anterior, a variação em nossos dados é restrita aos pronomes *você* e *cê*, com uso extremamente baixo de *tu*. Manter essa variante em nossas análises interfere em análises estatísticas. Frente a isso, removemos a variante *tu*, mantendo *você* e *cê* (Figura 33).

Figura 33 – Distribuição dos pronomes pessoais de 2PS na amostra Deslocamentos (2019) por deslocamento

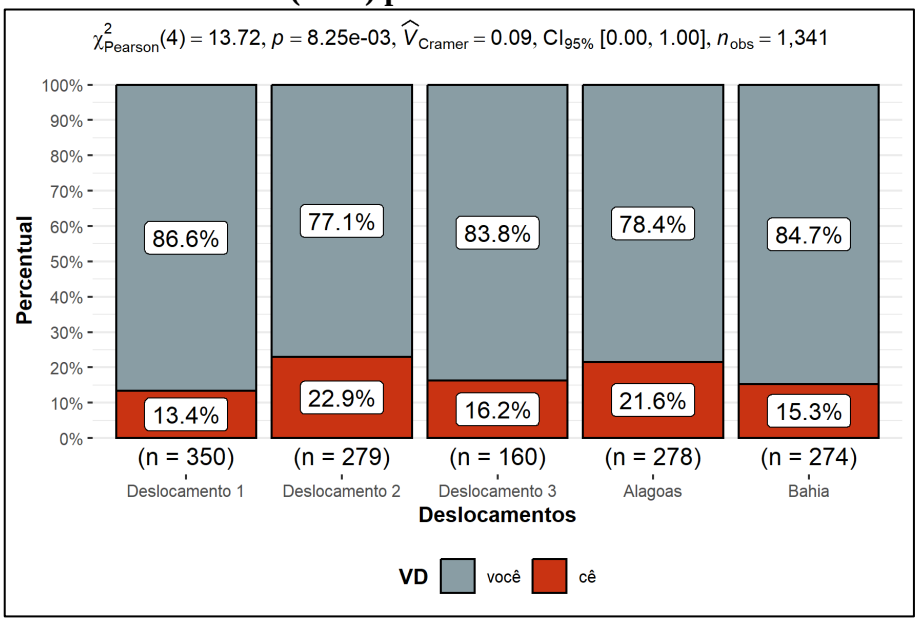


Fonte: elaboração própria.

O uso do pronome *você* apresenta a maior frequência nos Deslocamento 2 (82,2% 268/326) e Deslocamento 4 (80,7% 271/336). O comportamento de falantes externos a Sergipe está mais próximo ao de falantes do interior do estado que fazem o percurso diário para o *campus*. As menores frequências são observadas com falantes dos Deslocamento 3 (66,6% 231/347) e Deslocamento 1 (71,1% 288/405). Rejeitamos a H_0 ($X^2(3, N= 1,414) = 31.01$ $p < 0.001$) e apesentamos a H_1: há relação entre o perfil de deslocamento do falante e o uso dos pronomes pessoais de 2PS, com efeito pequeno ($V^2 = 0.14$).

Ainda que em todos os deslocamentos haja predomínio de *você*, as diferentes frequências a depender da região de origem do falante podem ser resultado do padrão de uso de sua comunidade origem, localizada em pontos distintos uma das outras. É possível que na amostra D2020 haja um comportamento similar (Figura 34).

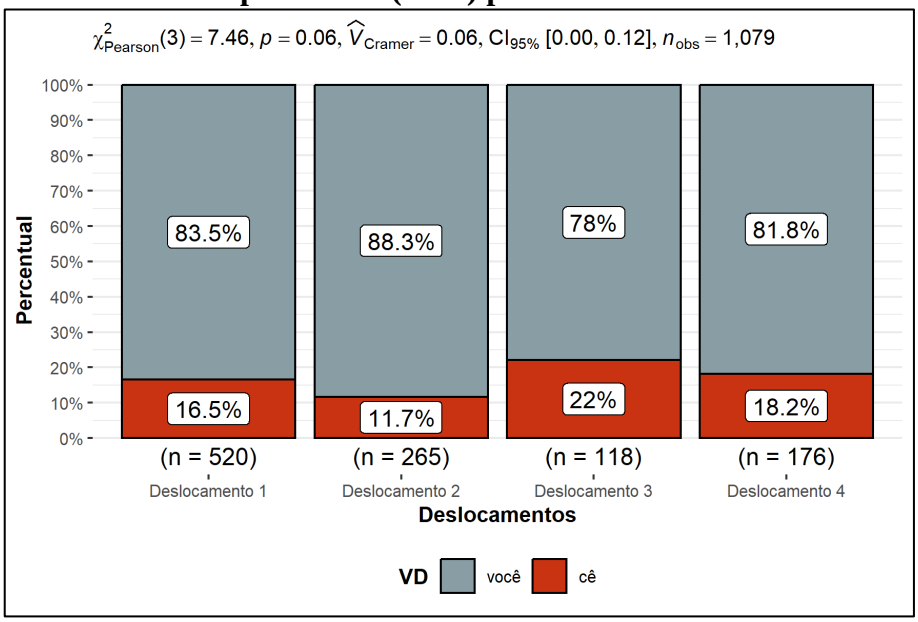
Figura 34 – Distribuição dos pronomes pessoais de 2PS na amostra Deslocamentos (2020) por deslocamento



Fonte: elaboração própria.

Nos dados de D2020, as maiores frequências de *você* são observadas nos Deslocamento 1 (86,6% 303/350), Bahia (84,7% 232/274) e Deslocamento 3 (83,8% 134/160), enquanto os Deslocamento 2 (77,1% 215/279) e Alagoas (78,4% 218/278) apresentam as menores frequências. Rejeitamos a H_0 ($X^2(4, N= 1,341) = 13.72$ $p < 0.001$), há uma relação entre as variáveis, com efeito pequeno ($V^2 = 0.09$). Aqui houve uma inversão, uma vez que na amostra D2019 são os Deslocamentos 1 e 3 que apresentam as menores frequências de *você*, o que refuta nossa previsão lançada no parágrafo anterior. Um ponto a se considerar são os dados da Bahia nessa amostra e os dados do Deslocamento 4 na amostra de 2019. Como sabemos, a amostra D2019 possui treze (13) falantes no Deslocamento 4 que são da Bahia. A similaridade entre os dados pode ser efeito de um comportamento na Bahia para alto uso de *você*. A Figura 35 apresenta os dados na amostra LC2023.

Figura 35 – Distribuição dos pronomes pessoais de 2PS na amostra Linguagem Corporificada (2023) por deslocamento



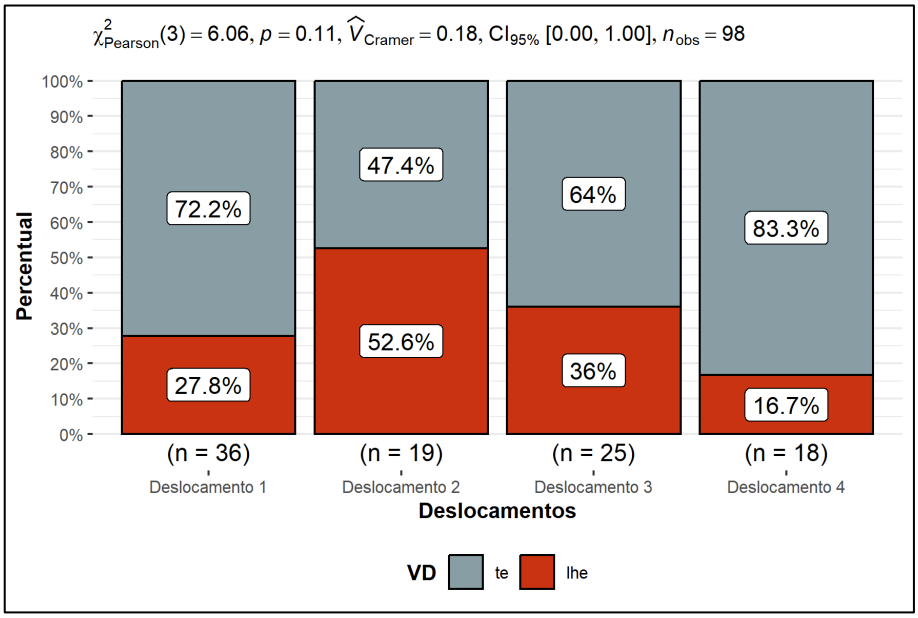
Fonte: elaboração própria.

Falantes do Deslocamento 2 fazem o maior uso da variante *você* (88,3% 234/265), seguido por falantes do Deslocamento 1 (83,5% 434/520) e falantes do Deslocamento 4 (81,8% 144/176); são os falantes do Deslocamento 3 que menos usam o pronome *você* (78% 92/118), similar aos dados de D2019. A diferença entre os perfis de deslocamento, contudo, não é estatisticamente significativa ((X^2(3, N= 1,079) = 7.46 p = 0.06).

As diferenças em relação ao comportamento dos perfis de deslocamento em D2019 e D2020 evidenciam que a variável apresenta, em algum nível, estratificação quanto à região geográfica do falante, uma vez que falantes de diferentes perfis de deslocamento apresentam frequências distintas para as variantes *você* e *cê*. Nesse sentido, em nossos dados, os pronomes pessoais de 2PS são variantes dialetalmente distintas.

Passamos para observar o comportamento variável de clíticos de 2PS (Figura 36).

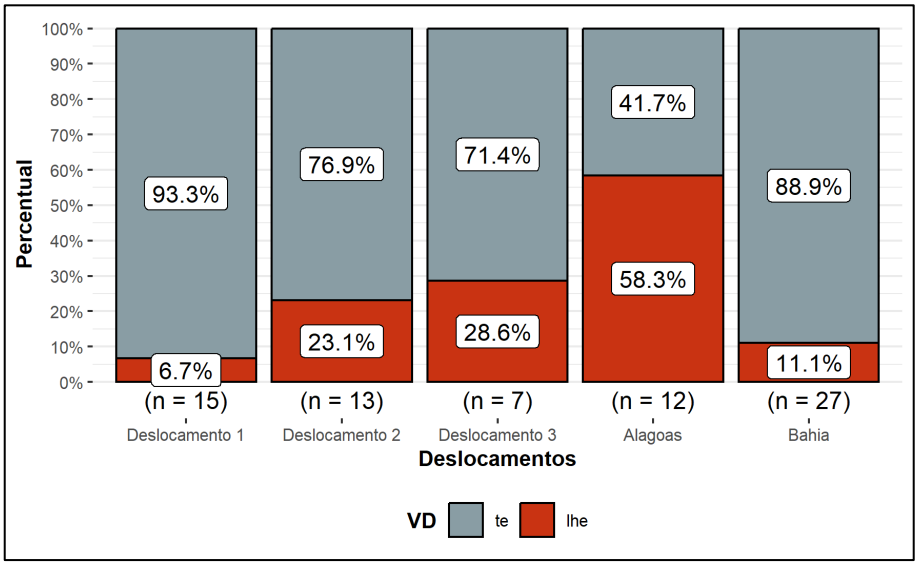
Figura 36 – Distribuição dos clíticos de 2PS na amostra Deslocamentos (2019) por deslocamento



Fonte: elaboração própria.

A frequência do pronome *te* é maior com dados do Deslocamento 4 (83,3% 15/18), no qual há o menor quantitativo de ocorrências da variável, seguido pelo Deslocamento 1 (72,2% 26/36), no qual há o maior quantitativo de dados, e pelo Deslocamento 3 (64% 16/25). No Deslocamento 2, o predomínio é para a variante *lhe* (52,6% 10/19). Falhamos em rejeitar a hipótese nula ($X^2(3, N= 98) = 6.06$, $p = 0.11$), os dados são independentes. Observamos o comportamento da variável na amostra D2020 (Figura 37).

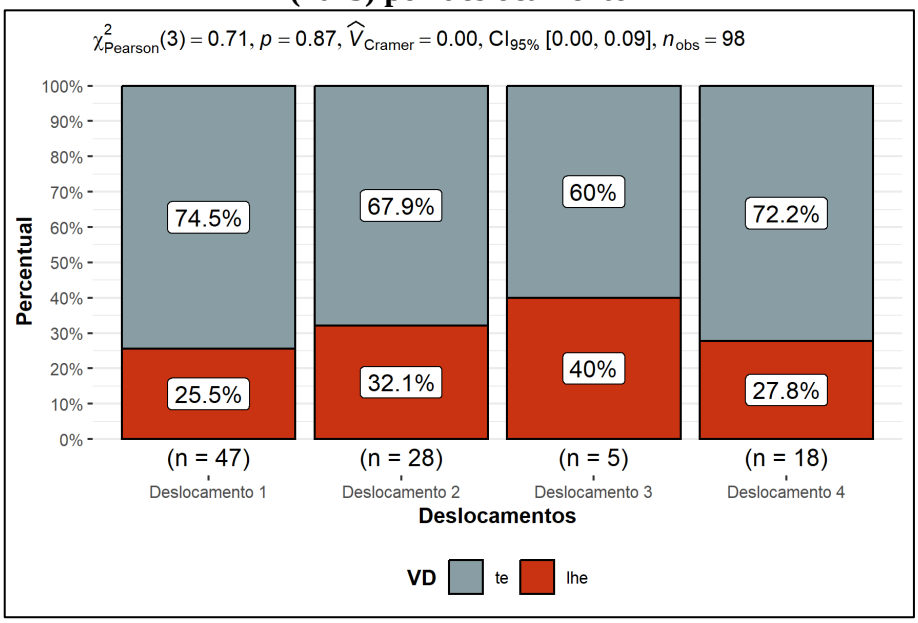
Figura 37 – Distribuição dos clíticos de 2PS na amostra Deslocamentos (2020) por deslocamento



Fonte: elaboração própria.

Nos dados da amostra D2020, as maiores frequências de *te* são observadas nos Deslocamento 1 (93,3% 14/15) e Bahia (88,9% 24/27). O Deslocamento 2, que na amostra D2019 apresentou predomínio de *lhe*, aqui apresenta predomínio de *te* (76,9% 10/13), seguida pelo Deslocamento 3 (71,4% 5/7). O Deslocamento Alagoas foi o único que apresentou predomínio de *lhe* (58,3% 7/5). Notemos, contudo, que as células nesta amostra apresentam poucos dados: apenas nos dados da Bahia há um quantitativo relativamente considerável. Por isso, recorremos ao Teste Exato de Fisher para observar a distribuição dos dados, que não apresentou muita diferença em relação ao Teste de qui-quadrado. Com base no p-valor (p= 0.009), rejeitamos a hipótese nula, os dados não são independentes: falantes do interior de Sergipe fazem um menor uso da variante *te*; falantes da região metropolitana e da Bahia apresentam a maior proporção de uso da variante; e falantes alagoanos tendem a usar mais o clítico *lhe*. A Figura 38 dispõe a distribuição na amostra LC2023.

Figura 38 – Distribuição dos clíticos de 2PS na amostra Linguagem Corporificada (2023) por deslocamento



Fonte: elaboração própria.

A variante *te* predomina em todos os perfis de deslocamento e tem as maiores frequências nos Deslocamento 1 (74,5% 35/47) e 4 (72,2% 13/18), seguidos pelo Deslocamento 2 (67,9% 19/28) e 3 (60% 3/5), este que apresenta a menor frequência e também a menor contagem de dados. A distribuição observada prevê a não significância estatística da relação entre a variável dependente e o deslocamento dos falantes, comprovada pelo teste Exato de Fisher (p = 0.83).

A distribuição dos clíticos de 2PS nas amostras apresenta distinções em relação ao perfil de deslocamento, mas os dados com as amostras D2019 e LC2023 não apresentaram significância. Ainda assim, podemos visualizar a existência de uma diferença dialetal na amostra D2020, na medida em que grupos de falantes de diferentes regiões tendem a apresentar frequências distintas em relação ao uso dos clíticos no PB. Temos evidências de que a variação nos clíticos de 2PS é dialetalmente distinta.

Apresentar a distinção por deslocamento da variável possessivos de 2PS não é necessário, uma vez que há predomínio da variante *seu*, com dados escassos de *teu*. A variável se comporta (semi)categoricamente por meio do possessivo *seu* em nossos dados: há uma (1) ocorrência de *teu* no Deslocamento 3 da amostra D2019; uma (1) no Deslocamento 2 e três (3) no Deslocamento Alagoas na amostra D2020. Podemos dizer que é uma variável dialetalmente distinta? Depende. Se considerarmos, numa perspectiva macro, que regiões dialetais do Nordeste tendem a fazer maior uso de *seu*, informação corroborada em nossos dados, sim, podemos dizer que é dialetal. Contudo, com base apenas em nossos resultados, não podemos dizer que a variável é dialetalmente distinta, uma vez que não visualizamos uma diferença notável entre grupos geograficamente localizados em relação à variável.

Hipotetizamos, no início desta seção, que a frequência de uso das quatro variáveis morfossintáticas apresentaria diferenças ao longo dos diferentes perfis de deslocamento dos falantes, uma vez que os dados com os quais contamos pertencem a falantes de diferentes regiões geográficas. O que nossos dados permitem observar é que essa hipótese é verdadeira em partes. O uso variável de artigo antes de possessivo, a variação nos pronomes pessoais de 2PS, excluído o *tu*, e a variação nos clíticos de 2PS mostraram, em ao menos uma das amostras, variação gradual ao longo dos diferentes perfis de deslocamento dos falantes, frente às diferenças nos padrões de uso de cada fenômeno considerando os grupos de falantes controlados. A variação nos possessivos de 2PS mostrou comportamento categórico para *seu* na maioria dos perfis de deslocamento, além de não ter nenhum caso na amostra LC2023.

A hipótese para a variável deslocamento, contudo, é apenas um ponto de partida para a investigação. É possível que outros fatores influenciem no comportamento das variáveis morfossintáticas, como a exposição a diferentes normas, observada indiretamente pela variável tempo no curso.

5.1.3 Distribuição por tempo no curso

Variáveis morfossintáticas podem ser reconhecidas por grupos de falantes, sendo percebidas e passíveis da atribuição de significados dialetais, conforme discutimos no Capítulo 2. Isso porque o reconhecimento da variação também é suscetível à circunstância geográfica na qual se localiza o falante, o que leva Freitag (2023) a pontuar que há variáveis salientes por força dialetal. Ao considerar que a inserção de um falante em uma nova comunidade permite a sua exposição a variáveis que até então não reconhecia, a variação morfossintática dialetal se apresenta como um traço distintivo, frente à diferença entre o padrão de uso de sua comunidade e de outras.

Temos ao menos três variáveis morfossintáticas que são dialetalmente distintas em nossos dados. Seus usos apresentam diferentes frequências a depender da região geográfica do falante. Adicionalmente, sabemos que falantes universitários estão em um contínuo processo de exposição a variantes morfossintáticas distintas das suas, o que possibilita o contato linguístico. Seus colegas de classe, de *campus*, seus professores e outros segmentos socias que compõem a universidade podem ser de regiões distintas e possuírem comportamento linguístico distinto. A exposição a novas variantes pode resultar no reconhecimento de novas formas e a força dialetal pode ser fator determinante para esse reconhecimento. Ao reconhecer formas novas, é possível que ocorra mudança em detrimento do novo padrão reconhecido, o que na literatura sociolinguística tem se observado sob o rótulo de acomodação dialetal, conforme evidenciam estudos de contato como Corrêa (2019), Guedes (2019), Oushiro (2016b), Ribeiro (2019). Nesse sentido, a observação da mudança linguística evidenciaria, em algum nível, existência de reconhecimento da variação. Mas como observar a mudança?

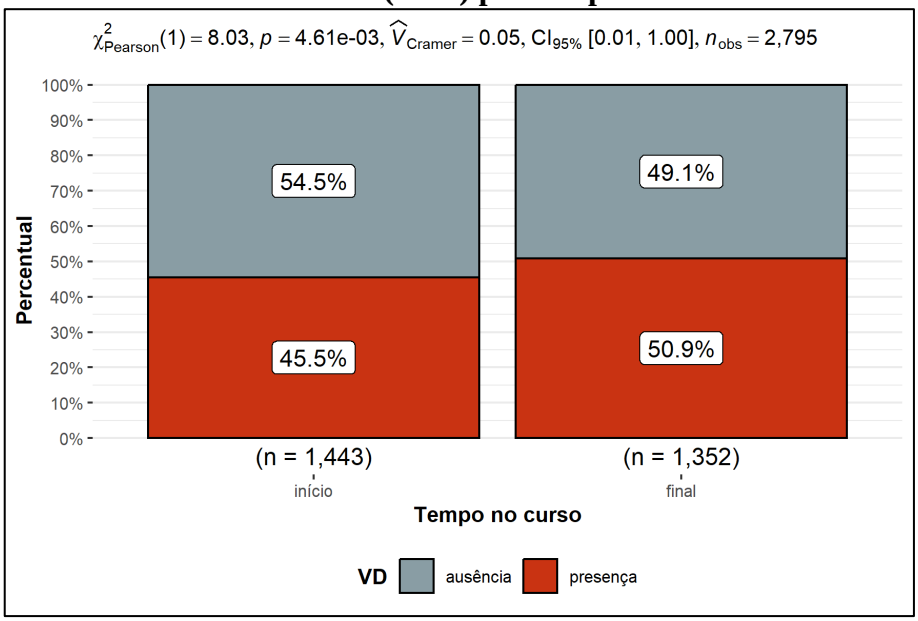
Comumente, na sociolinguística, a variável faixa etária/idade é utilizada para se observar a mudança, "pois tais resultados podem referendar generalizações sobre o andamento do processo de variação e mudança" (Freitag, 2000, p. 106). Para tanto, muitas vezes, recorre-se, considerando um período de tempo reduzido (Labov, 1994), a uma observação em tempo aparente, na qual se analisa o comportamento linguístico dos falantes considerando diferentes faixas etárias: "essa saída metodológica pressupõe que a idade cronológica dos indivíduos represente uma 'passagem no tempo'" (Freitag, 2005, p. 110). Mas também é possível que outras alternativas sejam postas em prática para observar a mudança em tempo aparente.

Em nosso estudo, amparados em discussões previamente aventadas por Côrrea (2019), Ribeiro (2019), Novais (2021), Rodrigues (2021) e Silva (2020), utilizamos a variável tempo no curso, já apresentada em nossos procedimentos metodológicos, para observar possíveis mudanças. Por meio dela, temos evidências indiretas para a mudança dialetal no comportamento linguístico dos falantes, e evidências indiretas para o reconhecimento da variação por grupos de falantes.

Ao utilizar a variável tempo no curso, questionamos se a exposição à variedade da comunidade leva a um processo de acomodação dialetal. Como hipótese, consideramos que a inserção de um falante em uma nova comunidade, diversa quanto à origem de seus falantes, desencadeia um processo de mudança linguística quanto às variáveis morfossintáticas analisadas, uma vez que, como propõe Freitag (2018), o reconhecimento da variação é suscetível à circunstância geográfica, o que pode culminar na adoção de novas formas linguísticas por parte dos falantes – a acomodação dialetal.

No que segue, fazemos a descrição dos fenômenos variáveis a partir do tempo no curso, que em nossas amostras se subdivide em início e final. Iniciamos pelo uso variável de artigo antes de possessivos pré-nominais (Figura 39).

Figura 39 – Distribuição do uso variável de artigo antes de possessivo na amostra Deslocamentos (2019) por tempo no curso



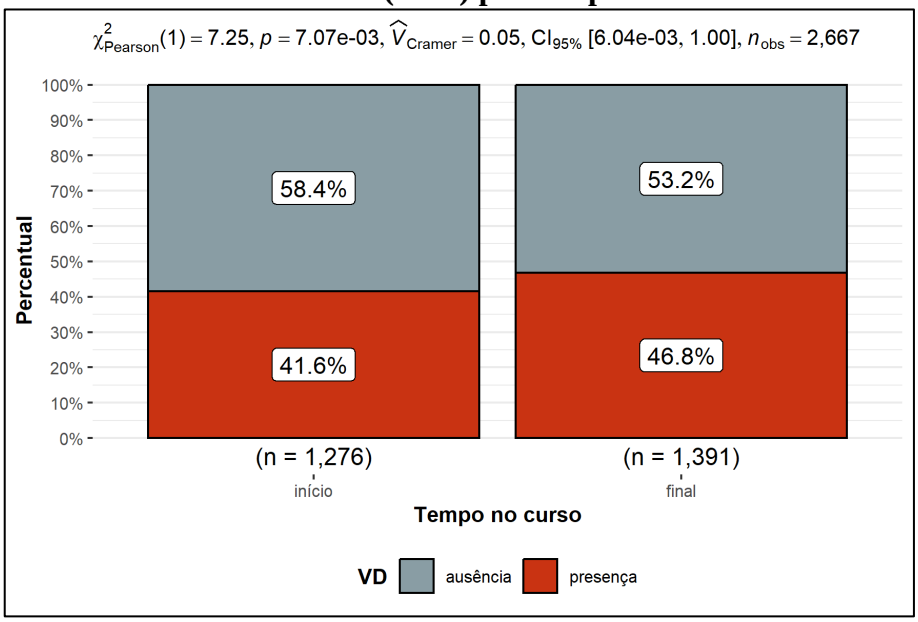
Fonte: elaboração própria.

Na amostra D2019, falantes ao início do curso (54,5% 786/1443) fazem maior uso da ausência de artigo antes de possessivos do que falantes ao final do curso (49,1% 664/1352),

em que a frequência da presença é maior. Rejeitamos a hipótese nula (X^2(1, N= 2,795) = 8.03 p < 0.001), há relação entre o tempo no curso e o uso variável de artigo antes de possessivos, com efeito pequeno (V^2 = 0.05).

Adicionalmente, conduzimos uma ANOVA (Análise de Variância) fatorial 4x2, por meio da taxa de frequência da ausência de artigo por indivíduo, para examinar o efeito do deslocamento e do tempo de curso sobre a distribuição da ausência de artigo antes de possessivos. Não houve interação significativa entre deslocamento e tempo (F(3, 56) = 0.34, p = 0.798). Os dados com base na amostra D2020 são dispostos na Figura 40.

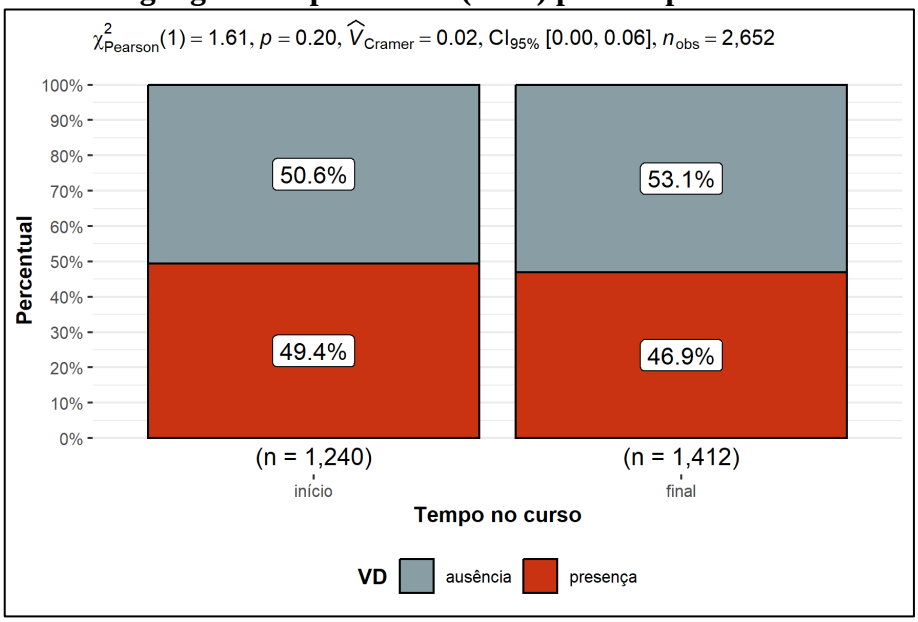
Figura 40 – Distribuição do uso variável de artigo antes de possessivo na amostra Deslocamentos (2020) por tempo no curso



Fonte: elaboração própria.

Na amostra D2020, a frequência da ausência também é maior em dados de falantes do início do curso (58,4%745/1276), à medida em que esse número decresce ao final (53,2% 740/1391), com ainda predomínio da ausência. Rejeitamos a H_0 (X^2(1, N= 2,667) = 7.25 p < 0.001), há relação entre as variáveis, com efeito pequeno (V^2 = 0.05). Realizamos uma ANOVA fatorial 5x2 para examinar o efeito do deslocamento e do tempo com base na amostra D2020, na qual não houve interação significativa (F(4, 50) = 1.17, p = 0.337). Na Figura 41, visualizam-se dados da amostra LC2023.

Figura 41 – Distribuição do uso variável de artigo antes de possessivo na amostra Linguagem Corporificada (2023) por tempo no curso



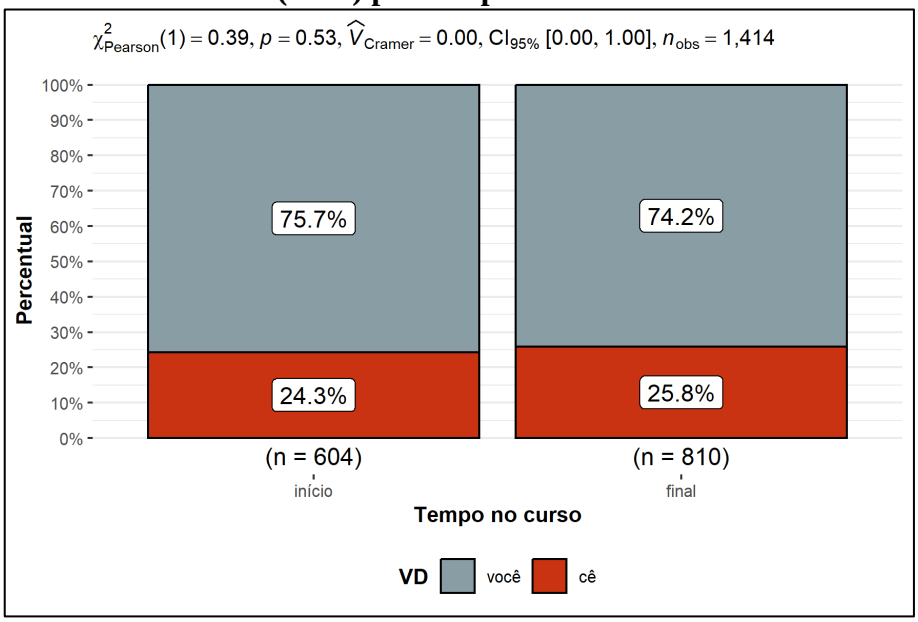
Fonte: elaboração própria.

A frequência da ausência é maior com falantes ao final do curso (53,1% 750/1412) do que ao início (50,6% 628/1240), diferente das outras amostras. A diferença entre os tempos no curso não é estatisticamente significativa ($X^2(1, N= 2,652) = 1.61$ p = 0.20). Adicionalmente, a análise de ANOVA fatorial 4x2 evidencia a interação não significativa entre tempo e deslocamento ($F(3, 49) = 1.56$, p = 0.212).

Nas amostras D2019 e D2020, a diferença entre os tempos de curso nos leva a hipotetizar que, no decorrer do curso, os falantes mudam seus usos quanto à variação no uso de artigo antes de possessivo, de predomínio de maior frequência de ausência para menor; no caso da amostra de 2019, maior frequência da presença de artigo. A mudança é evidência indireta de que existe, em algum nível, reconhecimento da variável, o que nos leva a inferir que o uso variável de artigo antes de possessivos é dialetalmente saliente.

A Figura 42 apresenta os resultados da variação nos pronomes pessoais de 2PS quanto ao tempo no curso na amostra D2019.

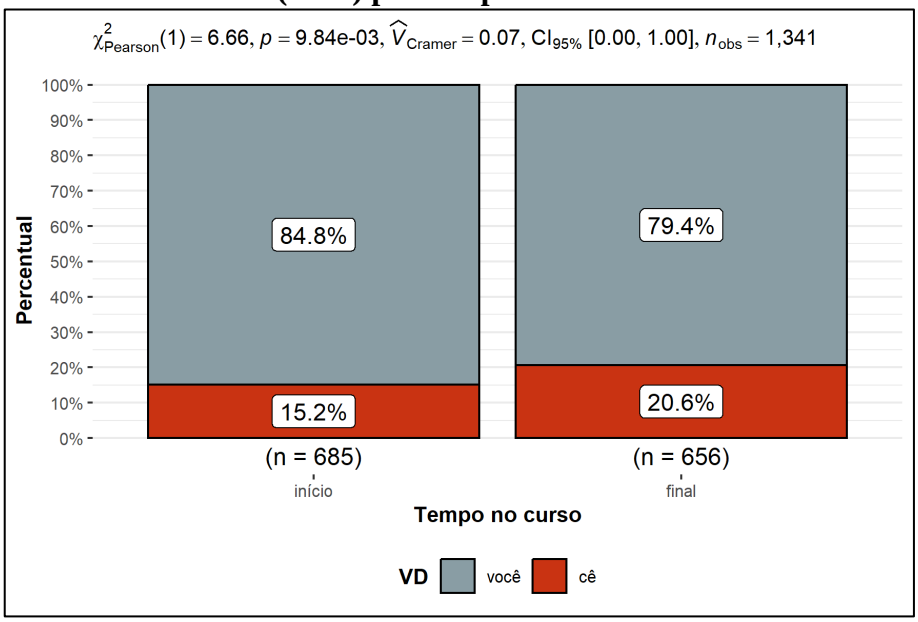
Figura 42 – Distribuição dos pronomes pessoais de 2PS na amostra Deslocamentos (2019) por tempo no curso



Fonte: elaboração própria.

Visualmente, percebemos que os percentuais para ambos os tempos são similares: 75,7% (457/604) para o início e 74,2% (601/810) para o final. O teste estatístico confirma que não há relação entre os usos dos pronomes pessoais de 2PS e o tempo no curso (X^2(1, N= 1,414) = 0.39 p = 0.53). Além disso, a interação entre deslocamento e tempo no curso não é estatisticamente significativa (*F*(3, 56) = 2,03, p = 0.120), calculada por meio de ANOVA com as taxas de uso de *você*. A Figura 43 apresenta a distribuição dos pronomes considerando a amostra D2020.

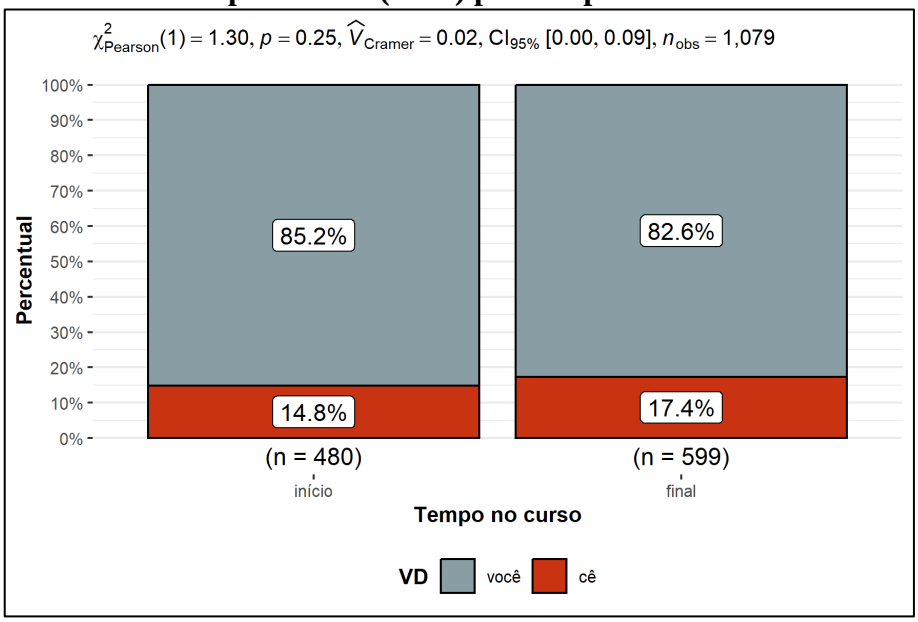
Figura 43 – Distribuição dos pronomes pessoais de 2PS na amostra Deslocamentos (2020) por tempo no curso



Fonte: elaboração própria.

Falantes ao início do curso fazem maior uso da variante *você* (84,8% 581/685) do que ao final do curso (79,4% 521/656). Rejeitamos a H_0 ($X^2(1, N= 1,341) = 6.66$ $p < 0.001$), há relação entre o tempo no curso e o uso dos pronomes pessoais de 2PS em posição de sujeito, com efeito pequeno ($V^2 = 0.07$). A interação entre deslocamento e tempo no curso, com base nas taxas de uso de *você* por indivíduo, não é estatisticamente significativa ($F(4, 50) = 1,78$, $p = 0,147$). Na Figura 44, visualizamos os dados em LC2023.

Figura 44 – Distribuição dos pronomes pessoais de 2PS na amostra Linguagem Corporificada (2023) por tempo no curso



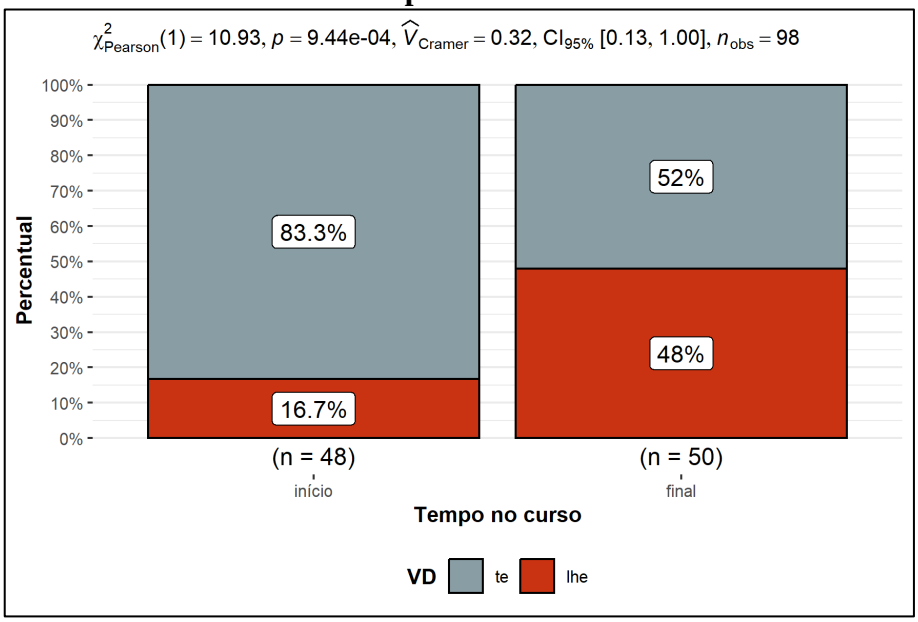
Fonte: elaboração própria.

Os dados desta amostra seguem o padrão observado nas demais: a frequência de *você* é maior ao início do curso (85,2% 409/480) do que ao final (82,6% 495/599), mas a diferença entre os tempos não é estatisticamente significativa (X^2(1, N= 1,079) = 1.30 p = 0.25), tampouco a interação entre tempo no curso e deslocamento (F(3, 49) = 0.65, p = 0.589).

No caso de D2020, a diferença significativa entre os dados e a diminuição na frequência de *você* com dados ao final do curso apontam que falantes podem apresentar um comportamento distinto para o uso da variável linguística ao passar do tempo. Considerando os dados dessa amostra, pontuamos que a mudança evidencia, indiretamente, a existência de um reconhecimento para a variável, em que grupos de falantes expostos a diferentes formas podem alterar seu comportamento.

Por último, temos a variação nos clíticos de 2PS. Vimos que a distribuição dos clíticos de 2PS apresenta associação com o deslocamento dos falantes em D2020, que evidencia seu caráter geográfico. Falta-nos, contudo, observar se há efeito do tempo no curso. A Figura 45 apresenta os dados de clíticos de 2PS por tempo na amostra D2019.

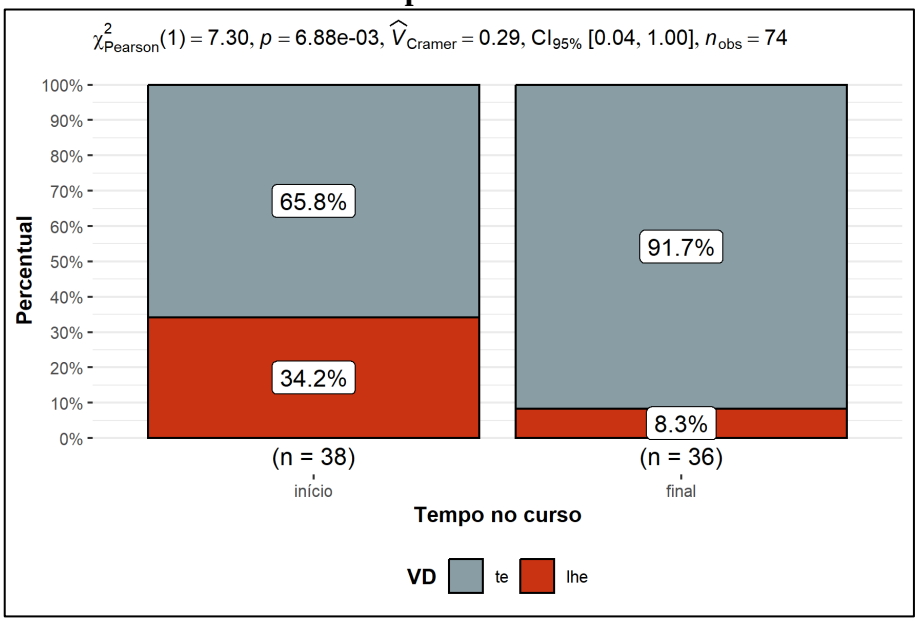
Figura 45 – Distribuição dos clíticos de 2PS na amostra Deslocamentos (2019) por tempo no curso



Fonte: elaboração própria.

A frequência do pronome clítico *te* é maior ao início do curso (83,3% 40/48) do que ao final (52% 26/50), uma diferença de 31,3%. Rejeitamos a hipótese nula (X^2(1, N= 98) = 10.93 p < 0.001), há relação entre o uso dos pronomes clíticos de 2PS e o tempo no curso, e a força de associação entre as variáveis é média (V^2 = 0.32). Falantes tendem, ao final do curso, a usar menos o clítico *te*, evidenciando, indiretamente, que há uma mudança em seu comportamento linguístico, possível resultado do contato no ambiente universitário. A interação entre deslocamento e tempo no curso, por meio da taxa de uso de *te* por indivíduo, não é estatisticamente significativa (F(3, 56) = 0,96, p = 0,417). Na Figura 46, apresentamos os dados com base na amostra D2020.

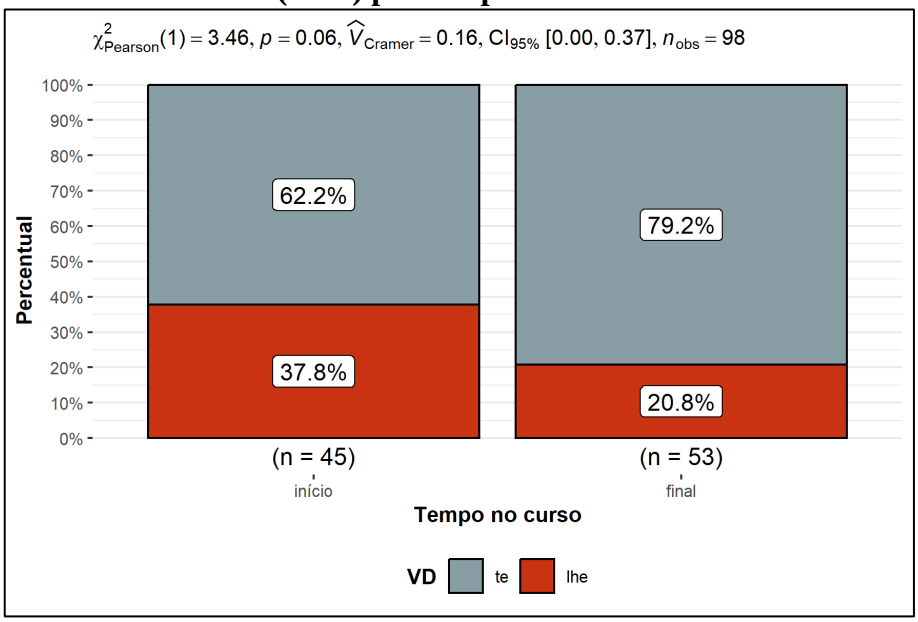
Figura 46 – Distribuição dos clíticos de 2PS na amostra Deslocamentos (2020) por tempo no curso



Fonte: elaboração própria.

Nos dados de D2020, há uma inversão em relação aos de D2019: a maior frequência de *te* é observada ao final do curso (91,7% 33/36), ainda que ao início haja predomínio do pronome (65,8% 25/38). Rejeitamos a H_0 ($X^2(1, N= 74) = 7.30$ $p < 0.001$), há relação entre as variáveis, com associação média ($V^2 = 0.29$). A interação entre deslocamento e tempo no curso não é estatisticamente significativa ($F(4, 50) = 1,13$, $p = 0,354$). A Figura 47 apresenta o comportamento da variável na amostra LC2023.

Figura 47 – Distribuição dos clíticos de 2PS na amostra Linguagem Corporificada (2023) por tempo no curso



Fonte: elaboração própria.

Falantes ao final do curso fazem maior uso de *te* (79,2% 42/53) do que falantes ao início (62,2% 28/45). Essa diferença, contudo, não é estatisticamente significativa, o que nos impossibilita de rejeitar a H_0 (($X^2(1, N= 94) = 3.46$ p = 0.06). A interação entre deslocamento e tempo no curso também não é estatisticamente significativa ($F(3, 49) = 0.88$, p = 0.459).

As mudanças em relação ao comportamento dos tempos em D2019 e D2020 são evidências de que grupos de falantes podem reconhecer os diferentes usos para as variantes e, a partir disso, mudar, inconscientemente, seu comportamento. As diferenças entre as amostras pode ser efeito dos diferentes grupos de falantes controlados. Na amostra D2020, comparada à D2019, há maior presença de falantes externos a Sergipe, enquanto em LC2023 há estratificação posterior de tempo, não sendo critério de constituição da amostra. Além disso, como vimos na variável deslocamento, a maior parte dos perfis apresenta padrão de uso para *te*, o que pode ter incidido sobre os dados. Temos então evidências indiretas da existência de um reconhecimento da variação, o que sugere que a variável pode ser dialetalmente saliente.

Para a variável possessivos de 2PS, os poucos dados não permitem a elaboração de hipóteses para possíveis mudanças. Por exemplo, na amostra D2019 há apenas uma (1) ocorrência de *teu* no início do curso, nenhuma ao final. Na amostra D2020, há uma (1) ao início e três (3) ao final; em LC2023, não há dados de *teu*. Evidentemente, não é possível rejeitar a hipótese nula com esses dados.

Hipotetizamos, no início desta seção, que a exposição à variedade da comunidade leva a um processo de acomodação dialetal. Os nossos dados nos permitem confirmar essa previsão, uma vez que de fato ocorre mudanças no comportamento linguístico de falantes ao final do curso, evidência indireta de que as variáveis são reconhecidas pelos falantes.

Esses resultados ajudam a compreender que o deslocamento de uma região dialetal para outra, juntamente com o contato com diferentes variantes e com a integração dos falantes à comunidade, podem interferir em seus usos linguísticos, uma vez que pode ocorrer mudanças quanto ao uso de alguma das variantes, já que variedades que estão em contato umas com as outras podem mudar, conforme teorizado por Trudgill (1986). Através do engajamento, em um espaço profuso em deslocamento e migração, aumenta-se a possibilidade de exposição a variantes diferentes.

Em suma, fenômenos variáveis morfossintáticos são condicionados por fatores extralinguísticos, como deslocamento e tempo no curso, e que há variáveis morfossintáticas dialetalmente distintas. Não vimos, contudo, como falantes se comportam individualmente quanto ao uso dos fenômenos morfossintáticos aqui descritos. Isso é feito na seção que segue.

5.1.4 Distribuição por indivíduo

As análises realizadas até aqui consideram a variação com base em grupos, como falantes de dado deslocamento ou de dado tempo no curso. Conforme Oushiro (2015a, p. 201),

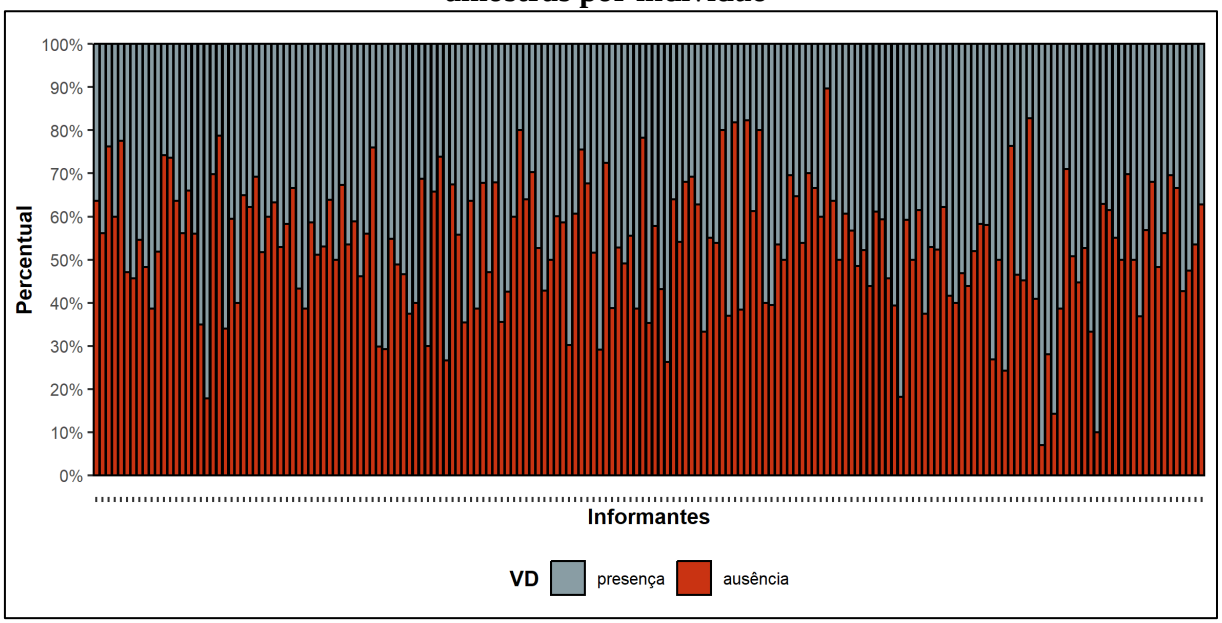
> pela necessidade metodológica de se analisar uma grande quantidade de dados, geralmente se volta à fala de agrupamentos de indivíduos, procedimento que normalmente não permite verificar se a variação observada de fato está presente na fala individual ou se simplesmente representa o amontoamento de comportamentos linguísticos divergentes.

Individualmente, falantes podem apresentar comportamentos específicos, que se distinguem entre si. Assim, ao considerar a variação com base no indivíduo, evitamos agrupar dados de maneira homogênea, mascarando padrões e diferenças importantes. Isso leva a resultados mais precisos quanto à variação, especialmente quando lidamos com populações diversas, já que, para Labov (2001, p. 87, tradução nossa), é "necessário rastrear e comparar a variação intrafalante e interfalante em vários contextos diferentes. Em qualquer caso, o estudo da variação intrafalante deve ser buscado sistematicamente"[26].

[26] No original: "necessary to track and compare both intra-speaker and interspeaker variation across several different contexts. In any case, the study of intra-speaker variation must be pursued systematically.

Além disso, considerar o indivíduo possibilita a observação do alcance de variação, o que revela a amplitude dos usos (maior e menor taxa). Para esta análise, lançamos a seguinte questão: falantes individuais variam quanto ao uso das variáveis morfossintáticas? Nossa hipótese é a de que falantes apresentam variação em seus usos linguísticos, mas tendem a fazer uso de uma das variantes com maior frequência. A presente seção buscar realizar a análise da distribuição por indivíduo. Na Figura 48, observamos essa distribuição com base no uso variável de artigo antes de possessivo.

Figura 48 – Distribuição do uso de artigo antes de possessivo pré-nominal nas amostras por indivíduo

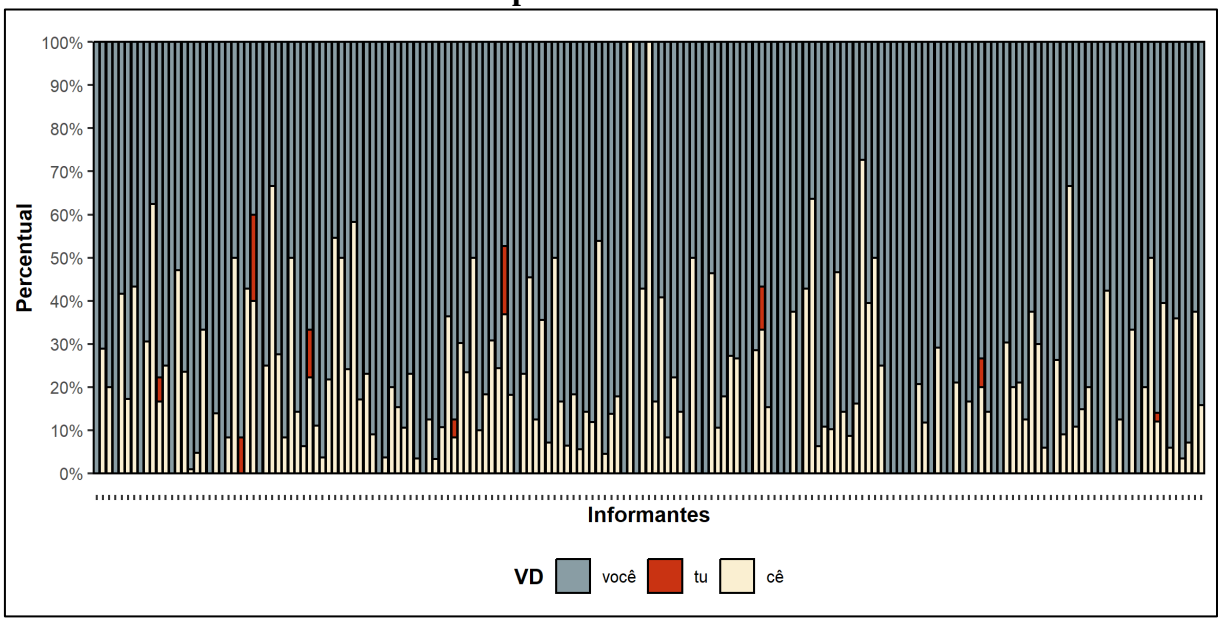


Fonte: elaboração própria.

Nesses dados, as taxas de ausência variam de 6,9% a 89,6% e todos os 181 falantes fazem uso da variável. Os dados para o uso variável de artigo antes de possessivo apontam que há variação intrafalante, ou seja, nenhum falante faz uso categórico para a variação e a variação é a regra de uso, que pode apresentar predomínio de uma ou outra forma. Em nossos dados, o conjunto apresenta predomínio da ausência.

Outras variáveis, contudo, podem não apresentar o mesmo comportamento. A Figura 49 apresenta resultados por falante dos pronomes pessoais de 2PS em posição de sujeito.

Figura 49 – Distribuição dos pronomes pessoais de 2PS nas amostras Deslocamentos por indivíduo



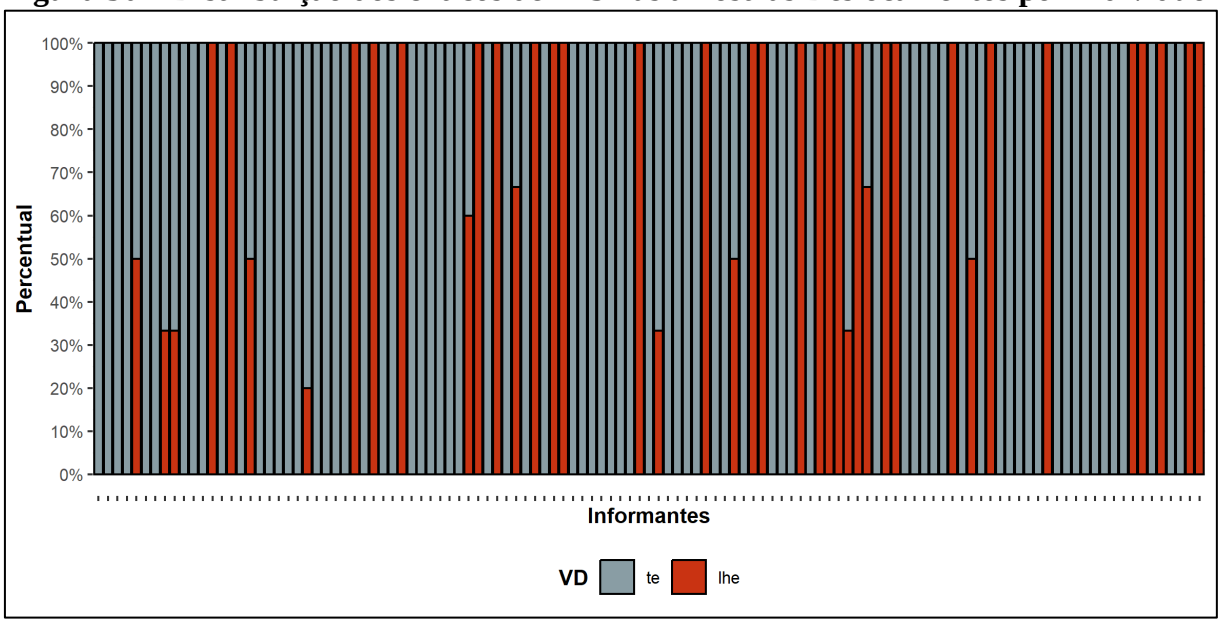
Fonte: elaboração própria.

Dos 181 falantes, 177 fazem uso de ao menos uma das variantes, 98% da amostra. As taxas de *você* variam de 0% para 100%. Dos que apresentam 0% (N= 2) para o pronome, nenhum faz uso de *tu*, mas fazem uso de *cê*. Nesse caso, há dois (2) falantes que fazem uso categórico da variante *cê*, o que representa menos de 1% da amostra, e trinta e nove (39) que fazem uso categórico de *você* – 22% da amostra –, além de quatro (4) que não usam nenhuma das variantes, 2% da amostra. Na fala dos demais informantes (136), ocorre variação entre as formas *você* e *cê*, 75% da amostra. A variante *tu* ocorre apenas na fala de nove (9) falantes, sempre variando com *cê* e/ou *você*.

Nos dados de universitários que compõem as amostras, a variação é restrita aos pronomes *você* e *cê* – com exceções –, enquanto o pronome *tu* não é uma variante corriqueira para o fenômeno variável.

A Figura 50 apresenta os dados por indivíduo dos clíticos de 2PS.

Figura 50 – Distribuição dos clíticos de 2PS nas amostras Deslocamentos por indivíduo

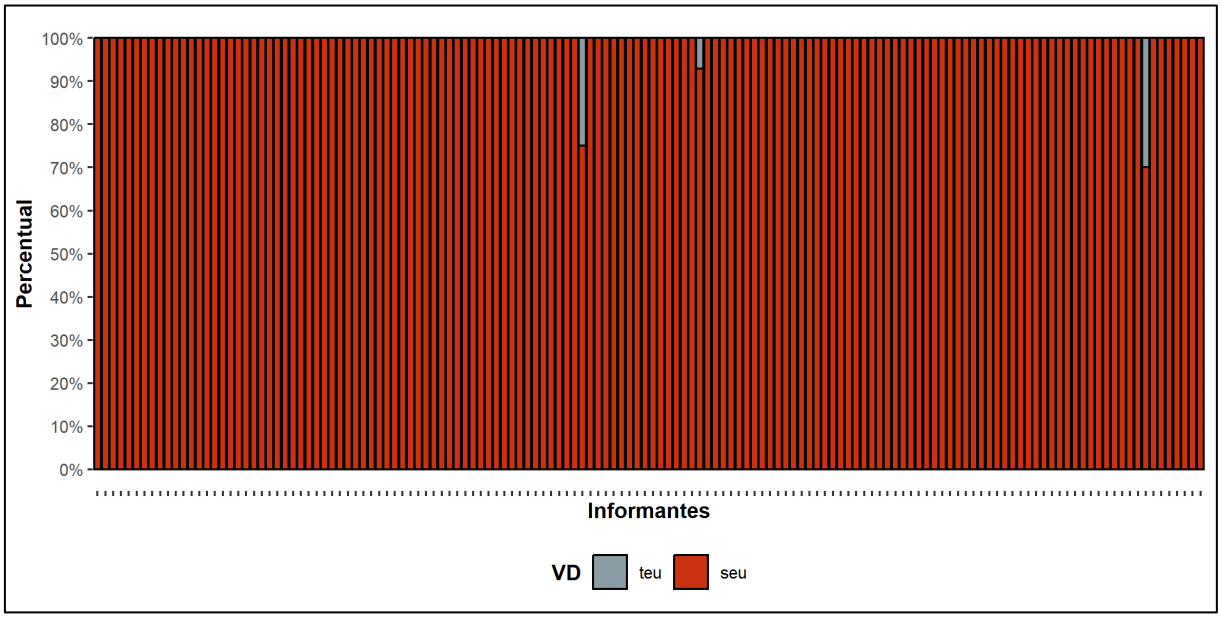


Fonte: elaboração própria.

O alcance do uso de *te* varia de 0% para 100%. Sessenta e quatro (64) falantes não fazem uso das variantes *te* ou *lhe*, o que corresponde a 35% da amostra, à medida em que a maior parte dos falantes são categóricos no uso de *te* (N= 76), 42% dos falantes. Há, contudo, exceções, dado que vinte e nove (29) falantes são categóricos no uso de *lhe*, 16% de todos os falantes. Por outro lado, temos doze (12) falantes que fazem uso de ambas as formas, 7% dos falantes da amostra. A variação nos usos linguísticos individuais quanto aos clíticos de 2PS, em nossos dados, constitui a exceção, não a regra.

Por fim, a Figura 51 apresenta a distribuição por indivíduo dos possessivos de 2PS.

Figura 51 – Variação nos possessivos de 2PS nas amostras Deslocamentos por indivíduo



Fonte: elaboração própria.

O alcance de variação é de 75% a 100%. Trinta e nove (39) informantes não fazem uso de nenhuma das duas formas, o que corresponde a 22% da amostra. Apenas três (3) falantes fazem uso de *teu*, nenhum de forma categórica: i) 38ent.UFS-SaoCristovao2018__desl. III_final_gen.fs.22, do interior de Sergipe que reside na região metropolitana do estado, com uma (1) realização; ii) 29ent.UFS-SaoCristovao2020_desl2_final_vag_pedagogia.fs.36, do interior de Sergipe – que faz o percurso diário para a UFS –, com também uma (1) realização; e iii) 65ent.UFS-SaoCristovao2020_desl4_inicio_bia_biologia.fs.21, de Alagoas, com três (3) realizações para o pronome. Assim, não podemos generalizar que *seu* e *teu* são variáveis em nossos dados, quando apenas três (3) falantes utilizam *teu* em variação, e cento e trinta e nove (139) fazem uso categórico de *seu*, 78% dos falantes. O uso categórico do pronome *seu* é a regra em dados das amostras.

A observação dos fenômenos morfossintáticos por indivíduo nos permite observar que, embora haja variação quando considerado o todo, individualmente, falantes podem apresentar comportamento distinto, ora para a variação entre as formas, com predomínio de uma variante, ora para o uso categórico de uma das formas, como também para o não uso de nenhuma das formas do fenômeno variável.

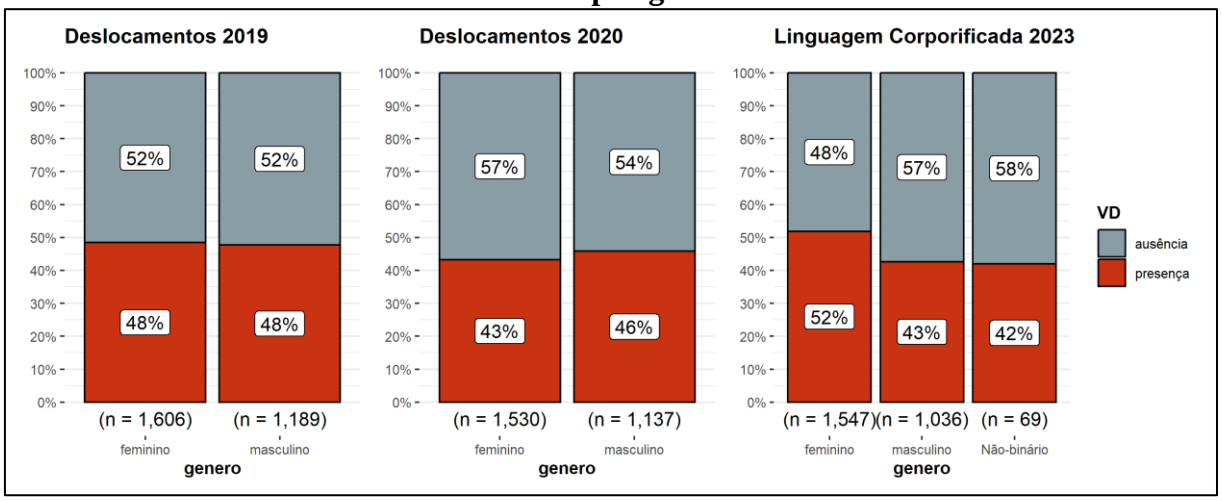
Outros atributos sociais dos indivíduos também interferem em seus usos linguísticos. Na seção que segue, descrevemos o comportamento das variáveis considerando gênero e idade.

5.1.5 Distribuição por gênero e idade

A pesquisa em sociolinguística no Brasil tem se amparado, para a constituição de amostras, em uma estratificação pautada em variáveis sociais macro, como escolaridade, idade e sexo/gênero do falante, cujos estudos descritivos apontam para uma associação entre essas categorias sociais e a variação linguística (Freitag, 2000; Novais; Siqueira, 2020; Siqueira; Novais, 2023). Nesta seção, discorremos sobre a distribuição de nossos dados considerando duas variáveis sociais macro: gênero e idade do informante. Não inserimos, para a variável gênero, a variação nos pronomes possessivos de 2PS, frente ao uso semicategórico de *seu*.

Com base em pesquisas anteriores sobre os fenômenos variáveis, questionamos se fatores extralinguísticos, como gênero e idade, interferem nos usos linguísticos dos falantes quanto às variáveis morfossintáticas descritas. Aventamos a hipótese de que há relação entre essas variáveis sociais e a distribuição dos fenômenos variáveis. As Figura 52, Figura 53 e Figura 54 apresentam a análise a partir da variável gênero.

Figura 52 – Distribuição do uso variável de artigo definido antes de possessivo pré-nominal por gênero

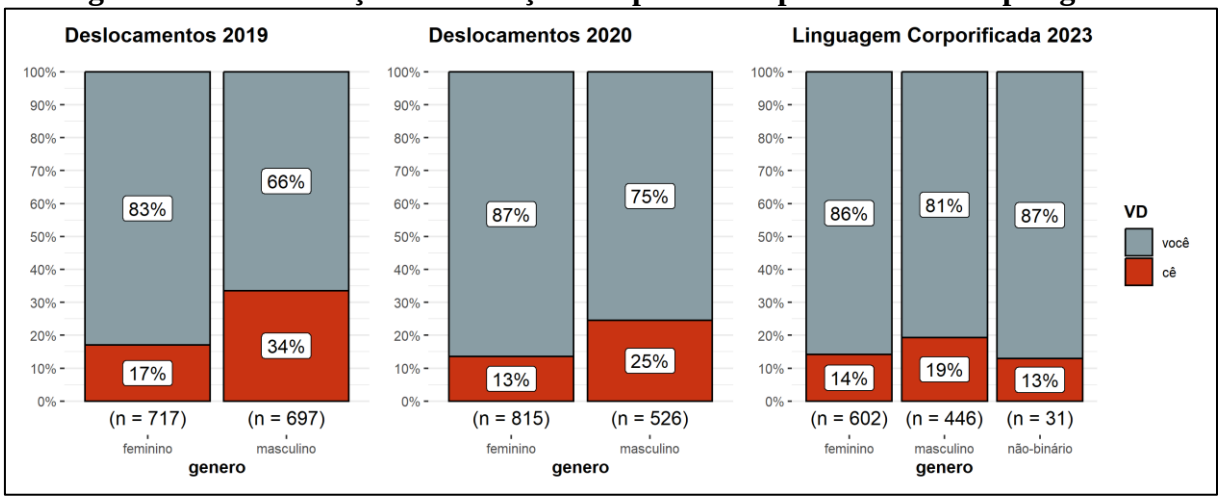


Fonte: elaboração própria.

Na amostra D2019, os percentuais são iguais para ambos os gêneros, o que não nos permite rejeitar a H_0 ($X^2(1, N= 2,795) = 0.16$ $p = 0.69$). Já na amostra D2020, embora a frequência da ausência seja maior com falantes do gênero feminino (57% 869/1530), também não é possível rejeitar a H_0 ($X^2(1, N= 2,667) = 1.81$ $p = 0.18$). Em LC2023, na qual há a inclusão de falantes do gênero não-binário, há relação significante entre gênero e a distribuição da variável ($X^2(2, N= 2,625) = 22.26$ $p < 0.001$): falantes do gênero não-binário

(58% 40/69) e masculino (57% 594/1036) fazem maior uso da ausência de artigo em comparação a falantes do gênero feminino (48% 744/1547), em cujo falar predomina a presença. Em duas amostras, não há relação do gênero do falante com o uso variável de artigo definido antes de possessivos pré-nominais em nossos dados; em LC2023, há relação. Essa diferença de comportamento linguístico pode indicar que o gênero dos falantes interfere na variação em determinados grupos (como no caso de LC2023), mas não é um fator determinante em outros. Passemos para a observação da variação nos pronomes pessoais de 2PS (Figura 53).

Figura 53 – Distribuição da variação dos pronomes pessoais de 2PS por gênero

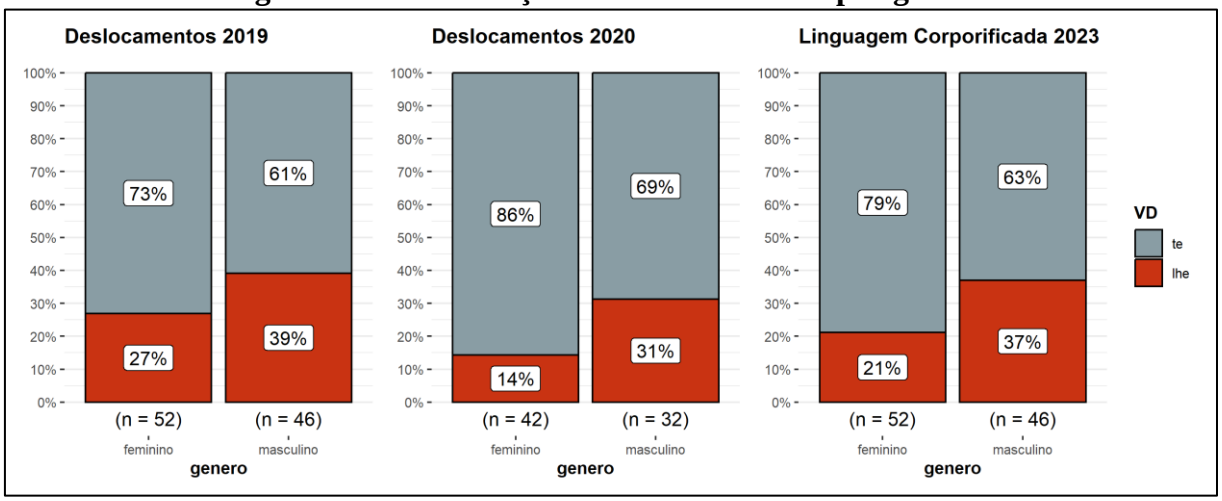


Fonte: elaboração própria.

Nas amostras Deslocamentos, a frequência de *você* é maior com falantes do gênero feminino: na amostra D2019, o percentual é de 83% (595/717), com relação estatisticamente significativa (X^2(1, N= 1,414) = 51.43 p < 0.001) e efeito pequeno (V^2 = 0.19); na amostra D2020, o percentual é de 87% (705/815), com relação também estatisticamente significativa (X^2(1, N= 1,341) = 26.54 p < 0.001) e efeito pequeno (V^2 = 0.14). Falantes do gênero feminino tendem a empregar o pronome *você* em posição de sujeito de 2PS mais do que falantes do gênero masculino, oposto ao observado em Oliveira (2007), na qual são os falantes do gênero masculino que lideram o uso de *você*; mas similar aos dados de Nogueira (2013), nos quais quem lidera são falantes do gênero feminino. Na amostra LC2023, falantes do gênero não-binário fazem o maior uso (87% 27/31), seguidos por falantes do gênero feminino (86% 517/602), mas a relação não é significativa (X^2(2, N= 1,079) = 5.28 p = 0.07).

Os resultados para a variável gênero são indícios de que os usos linguísticos quanto aos pronomes pessoais de 2PS em posição de sujeito são sensíveis ao gênero do falante. A Figura 54 apresenta os resultados com base na variação dos clíticos de 2PS.

Figura 54 – Distribuição dos clíticos de 2PS por gênero



Fonte: elaboração própria.

Os usos de *te* na amostra D2019 são maiores com falantes do gênero feminino (73% 38/52), resultado similar ao da amostra D2020, em que o percentual é de 86% (36/42), e ao da amostra LC2023, com percentual de 79% (41/52). Nos três casos, falhamos em rejeitar a H_0 ($X^2(1, N= 98) = 1.65$ $p = 0.20$; $X^2(1, N= 74) = 3.08$ $p = 0.08$; e $X^2(1, N= 98) = 2.99$ $p = 0.08$, respectivamente). Não há, em nossos dados, efeito do gênero sobre a distribuição dos clíticos de 2PS.

Gênero apresenta efeito estatisticamente significativo apenas no uso variável de artigo antes de possessivo, na amostra LC2023, e na variação dos pronomes pessoais de 2PS, nas amostras D2019 e D2020. Nos fenômenos variáveis com as demais amostras, não há interferência dessa variável social sobre a distribuição dos resultados.

Consideremos a variável idade. Em nossa amostra, idade não é critério de estratificação. Com isso, não temos diferentes faixas etárias, como as pesquisas apresentadas na revisão integrativa. A variável idade, em nossos dados, é uma variável numérica contínua, com distribuição não normal ($p < 0.001$), cuja média é de 21 anos. Frente a isso, recorremos, para nossa análise, a modelos de regressão logística, por meio da função `glm` no RStudio (RStudio Team, 2015), cujos resultados são sumarizados nas Tabela 5,

Tabela 6, Tabela 7 e Tabela 8. Nelas, os valores são apresentados em *log odds* (*log* de razão de chances de a variante de interesse ocorrer). Valores positivos correspondem a *logs* de razão de chances positivos da ocorrência da variante de interesse (favorecimento de ausência de artigo, pronome *você* e clítico *te*), e valores negativos correspondem a *logs* de razão de chances negativos (desfavorecimento), que são lidos em relação ao valor de *intercept*, apresentado na primeira linha da tabela. Todos os modelos tomam, para o *intercept*, idade como 0. As fórmulas dos modelos são apresentadas na última linha das tabelas.

Tabela 5 – Modelo de regressão logística do uso variável de artigo antes de possessivos pré-nominais por idade

	D2019			**D2020**			**LC2023**		
Preditores	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*
(Intercept)	0.98	0.53 – 1.45	**<0.001**	1.48	0.90 – 2.07	**<0.001**	-0.21	-0.88 – 0.47	0.548
idade	-0.04	-0.07 – -0.02	**<0.001**	-0.06	-0.09 – -0.03	**<0.001**	0.01	-0.02 – 0.04	0.403
Observações	2795			2667			2652		
R^2 Tjur	0.006			0.007			0.000		

Fórmula: glm(VD ~ idade, data = data, family = binomial)

Fonte: elaboração própria.

Na amostra D2019, o *intercept* é de 0.98 e significativo ($p < 0.001$). Nesse modelo, a variável independente idade apresenta *log* de razão de chances de -0.04. Uma vez que o número é negativo, entendemos que o *log* de razão de chances de ocorrer a ausência de artigo diminui à medida que a idade aumenta ($p < 0.001$). D2020 segue um caminho similar: dado seu *intercept* de 1.48 ($p < 0.001$) e o *log* de razão de chances da variável idade ser de -0.06 ($p < 0.001$), entendemos que a probabilidade de uso da variante diminui à medida que a idade aumenta. Não há significância no modelo com a amostra LC2023. A idade do falante tende a interferir na ausência de artigo antes de possessivos apenas nas amostras Deslocamentos.

A Tabela 6 apresenta a estatística de regressão com base nos pronomes pessoais de 2PS, excluído o *tu*.

Tabela 6 – Modelo de regressão logística dos pronomes pessoais de 2PS por idade

	D2019			D2020			LC2023		
Preditores	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*
(Intercept)	1.09	0.27 – 1.90	**0.009**	2.32	1.47 – 3.14	**<0.001**	3.98	2.67 – 5.31	**<0.001**
idade	0.00	-0.04 – 0.04	0.999	-0.04	-0.07 – 0.00	0.057	-0.11	-0.16 – -0.05	**<0.001**
Observações	1414			1341			1079		
R^2 Tjur	0.000			0.002			0.012		

Fórmula: glm(VD ~ idade, data = data, family = binomial)

Fonte: elaboração própria.

Para D2019, o *intercept* é de 1.09 e significante (p = 0.009). Contudo, a variável independente idade não apresenta significância (p = 0.999). De forma similar, o *intercept* na amostra D2020 é de 2.32 e significante (p < 0.001), mas a variável não apresentou significância (p = 0.057). Em LC2023, o *intercept* é de 3.98 e significante (p< 0.001), cuja variável preditora idade é significativa (-0.11 p< 0.001): à medida que a idade aumenta em uma unidade, o valor da variável dependente diminui.

A variável social macro idade tende a não interferir na maior parte dos dados da variação dos pronomes pessoais de 2PS em nossos dados, porém apresenta significância na amostra LC2023. Passemos para a observação dos clíticos de 2PS (Tabela 7).

Tabela 7 – Modelo de regressão logística dos clíticos de 2PS por idade

	D2019			**D2020**			**LC2023**		
Preditores	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*
(Intercept)	5.66	2.93 – 8.81	**<0.001**	-3.44	-9.15 – 1.38	0.208	-2.76	-6.88 – 1.12	0.173
idade	-0.23	-0.37 – -0.10	**0.001**	0.23	-0.00 – 0.52	0.088	0.17	-0.01 – 0.37	0.071
Observações	98			74			98		
R^2 Tjur	0.142			0.039			0.032		

Fórmula: glm(VD ~ idade, data = data, family = binomial)

Fonte: elaboração própria.

Observamos significância apenas na amostra D2019, em que o *intercept* é de 5.66 e significante (p < 0.001). Dentro desse modelo, idade apresenta *log* de razões de chances negativo (-0.23) e significante (p < 0.001). À medida que a idade do falante aumenta, há uma diminuição na frequência de uso do clítico *te*. Dados da variável idade nas amostras D2020 e LC2023 não apresentaram efeito sobre a variação nos clíticos de 2PS. É possível, contudo,

que a pouca quantidade de dados tenha interferido no modelo. Essa previsão não se segue nos possessivos de 2PS, conforme Tabela 8, na qual dados de LC2023 foram omitidos dada a inexistência de *teu*.

Tabela 8 – Modelo de regressão logística dos possessivos de 2PS por idade

	D2019			**D2020**		
Variáveis	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*
(Intercept)	8.11	-6.55 – 22.03	0.225	7.74	3.02 – 12.22	**<0.001**
idade	-0.14	-0.68 – 0.61	0.649	-0.18	-0.35 – 0.03	**0.042**
Observações	181			165		
R^2 Tjur	0.001			0.046		

Fórmula: glm(VD ~ idade, data = data, family = binomial)

Fonte: elaboração própria.

Há significância apenas na amostra D2020, em que o *intercept* é de 3.02 ($p < 0.001$). Dentro desse modelo, à medida que a idade do falante aumenta, há uma diminuição (-0.18) na frequência de uso de *seu* ($p = 0.042$). A variável idade se mostrou estatisticamente significante apenas com o uso variável de artigo antes de possessivos, em ambas as amostras, nos clíticos de 2PS, com a amostra D2019, e no uso do possessivo *seu*, na amostra D2020. Nesse sentido, a idade, em nossos dados, não interfere em todas as variáveis, nem de forma totalitária, o que refuta, em partes, nossa hipótese, a de que haveria interferência dessas variáveis sociais sobre a distribuição dos dados.

No que segue, sintetizamos nossos resultados.

5.2 EM SÍNTESE

Neste momento, retomamos as questões lançados ao início deste capítulo: i) qual o padrão de realização das variáveis em nossas amostras?; ii) há diferenças na frequência de uso das quatro variáveis morfossintáticas entre os perfis de deslocamento dos falantes?; iii) o tempo no curso interfere no modo como os falantes fazem uso das variáveis?; iv) como os falantes se comportam individualmente quanto ao uso das variáveis morfossintáticas?; v) fatores extralinguísticos, como gênero e idade, interferem nos usos linguísticos dos falantes quanto às variáveis morfossintáticas descritas? As análises univariadas conduzidas nos permitem responder a essas questões, as quais sintetizamos nesta subseção. Comecemos pela questão em (i):

- A ausência de artigo antes de possessivo é predominante em todas as amostras, com maior frequência em D2020. Todos os falantes apresentam uso variável do fenômeno;
- O pronome pessoal de 2PS predominante é *você* nas amostras, seguido da variante reduzida *cê*. O pronome *tu* ocorre com frequência extremamente baixa. Dois (2) falantes que fazem uso categórico da variante *cê* e trinta e nove (39) fazem categórico de *você*. Para a maior parte dos falantes (N= 96), a variação *você* e *cê* é a regra;
- O clítico de 2PS *te* ocorre com maior frequência. A maior proporção de uso é vista em D2020. Há sessenta e quatro (64) falantes que não utilizam clíticos de 2PS durante a entrevista. Por outro lado, a maior parte dos falantes são categóricos no uso de *te* (N= 76) e vinte e nove (29) são categóricos no uso de *lhe*. De todos os dados, apenas doze (12) falantes fazem uso variável de *te* e *lhe*. O uso categórico de uma das formas é a regra; e
- O possessivo *seu* é o mais recorrente nas amostras, com frequências acima de 97%. Os pronomes *teu* e *seu* variam na fala de apenas três (3) falantes, enquanto cento e trinta e nove (139) falantes fazem uso categórico de *seu*, que é a regra.

Adicionalmente, a observação com base em fatores extralinguísticos evidencia o comportamento da variação quanto ao perfil social do falante e permite a resolução das questões (ii)-(v), conforme sintetizamos no Quadro 11.

Quadro 11 – Síntese dos resultados das variáveis independentes

Fenômeno	**Regra intrafalante**	**Deslocamentos**	**Tempo no curso**	**Gênero**	**Idade**
Uso variável de artigo antes de possessivos *ausência* (53,2% 4313/8114)	Variável.	**D2019**: não há associação significativa. **D2020**: falantes de Sergipe fazem o maior uso da ausência. **LC2023**: falantes externos a Sergipe fazem maior uso da ausência.	**D2019**: falantes ao início do curso fazem maior uso da ausência de artigo antes de possessivos do que falantes ao final do curso. **D2020**: falantes ao início do curso fazem maior uso da ausência de artigo antes de possessivos do que falantes ao final do curso. **LC2023**: não há associação significativa.	**D2019**: não há associação significativa. **D2020**: não há associação significativa. **LC2023**: falantes do gênero não-binário e masculino fazem maior uso da ausência de artigo.	**D2019**: a ausência de artigo diminui à medida que a idade aumenta. **D2020**: a ausência de artigo diminui à medida que a idade aumenta. **LC2023**: não há significância.
Pronomes pessoais de 2PS *você* (79,6% 3064/3848)	Variável, mas restrita a *você* e *cê*.	**D2019**: pronome *você* ocorre mais com falantes dos Deslocamentos 2 e 4. **D2020**: o pronome *você* ocorre mais com falantes do Deslocamento 1, Bahia e 3. **LC2023**: não há associação significativa.	**D2019**: não há associação significativa. **D2020**: falantes ao início do curso fazem maior uso da variante você (do que ao final. **LC2023**: não há associação significativa.	**D2019**: a frequência de *você* é maior com falantes do gênero feminino. **D2020**: a frequência de *você* é maior com falantes do gênero feminino. **LC2023**: não há associação significativa.	**D2019**: não há significância. **D2020**: não há significância. **LC2023**: uso de *você* diminui à medida que a idade aumenta.
Pronomes clíticos de 2PS *te* (71,9% 194/270)	Categórica para *te* ou *lhe*.	**D2019**: não há associação significativa. **D2020**: pronome *te* ocorre mais com falantes do Deslocamento 1 e Bahia. **LC2023**: não há associação significativa.	**D2019**: falantes ao início do curso fazem maior uso de *te* do que ao final. **D2020**: falantes ao final do curso fazem maior uso de *te* do que ao início. **LC2023**: não há associação significativa.	**D2019**: não há associação significativa. **D2020**: não há associação significativa. **LC2023**: não há associação significativa.	**D2019**: à medida que a idade do falante aumenta, há uma diminuição na frequência de uso do clítico *te*. **D2020**: não há significância. **LC2023**: não há significância.
Pronomes possessivos de 2PS *seu* (98,6% 528/533)	Categórica para *seu*.	**D2019**: não há associação significativa. **D2020**: não há associação significativa. **LC2023**: não há associação significativa.	**D2019**: não há associação significativa. **D2020**: não há associação significativa. **LC2023**: não há associação significativa.	**D2019**: não há associação significativa. **D2020**: não há associação significativa. **LC2023**: não há associação significativa.	**D2019**: não há associação significativa. **D2020**: à medida que a idade aumenta, o uso de *seu* diminui. **LC2023**: não há associação significativa.

Fonte: elaboração própria.

A ausência de artigo antes de possessivo é predominante, e todos os falantes apresentam uso variável do fenômeno. Para a variação nos pronomes pessoais de 2PS, há predomínio de *você*, seguido da variante reduzida *cê*, com usos escassos de *tu*, o que o fez ser removido de análises posteriores. A variação *você* e *cê* é a regra para os grupos de falantes. O clítico de 2PS *te* ocorre com maior frequência. De todos os dados, apenas doze (12) falantes fazem uso variável de *te* e *lhe*. O uso categórico de uma das formas é a regra, principalmente de *te*. Para o uso variável de possessivos de 2PS, *seu* é o pronome possessivo mais recorrente, cujo uso categórico é a regra em nossos dados.

Para a observação de possível distinção e saliência dialetal, utilizamos as variáveis deslocamentos e tempo no curso. Distinções dialetais são visíveis em três dos quatro fenômenos, em ao menos uma das amostras: o uso variável de artigo antes de possessivo, a variação nos pronomes pessoais de 2PS, excluído o *tu*, e a variação nos clíticos de 2PS se mostraram dialetalmente distintas, frente as diferenças nos padrões de uso de cada fenômeno. Em relação à força dialetal, vimos que a mudança em relação ao comportamento linguístico de falantes ao final do curso é indício de um reconhecimento, ainda que não consciente, de diferentes usos das variáveis morfossintáticas, apontando evidências indiretas para uma saliência nas formas a partir da exposição, por meio do contato, a diferenças nos usos linguísticos dos fenômenos variáveis.

A observação de variáveis sociais macro, como gênero e idade, apontou, em sua grande maioria, para a não significância, com exceção de (i) gênero no uso variável de artigo antes de possessivos para LC2023, em que falantes dos gêneros não-binário e masculino lideram o uso da ausência, e na variação nos pronomes pessoais de 2PS, em que falantes do gênero feminino tendem a fazer maior uso de *você*; e (ii) idade no uso variável de artigo antes de possessivos, nas amostras D2019 e D2020, na variação dos pronomes pessoais de 2PS na amostra LC2023, na variação dos clíticos de 2PS, na amostra D2019, e variação nos possessivos de 2PS, na amostra D2020.

A análise univariada nos forneceu informações cruciais sobre as variáveis selecionadas nesta pesquisa. A partir da observação do comportamento dessas variáveis e sua (não) relação com os condicionantes sociais (deslocamento, tempo no curso, gênero e idade), obtivemos informações sobre cada variável individualmente. Com isso, a análise univariada nos fornece suporte à interpretação dos resultados da análise de covariação. Por exemplo, sabendo que falantes de certo grupo apresentam comportamento específico, podemos inferir se eles irão apresentar usos conjuntos das formas linguísticas. Além disso, compreendemos que há variáveis

com baixa frequência e/ou que há variáveis cujos falantes fazem uso categórico, podemos prever que isso afetará a estatística de correlação e de agrupamento, por exemplo.

Assim, uma vez que, neste capítulo, compreendemos padrões univariados dos fenômenos variáveis, podemos passar para a observação de padrões de covariação, no qual buscamos observar a relação entre as variáveis morfossintáticas mobilizadas nesta pesquisa.

6. DESCRIÇÃO E ANÁLISE DE VARIÁVEIS CORRELACIONADAS

No capítulo anterior, descrevemos o comportamento das quatro variáveis morfossintáticas de forma isolada, observando sua distribuição e fatores que interferem em seus usos. Os resultados demonstram que esses fenômenos morfossintáticos podem sofrer efeito de fatores externos à língua, como o deslocamento dos falantes, que tendem a apresentar comportamento distinto a depender de sua comunidade de origem, o tempo no curso – dado a visualização de mudança em tempo aparente –, gênero e idade. Este capítulo avança em relação à descrição da variação, com o objetivo de realizar análises de covariação entre os quatro fenômenos morfossintáticos: i) uso variável de artigo antes de possessivo pré-nominal; ii) pronomes pessoais de 2PS; iii) pronomes clíticos de 2PS; e iv) pronomes possessivos de 2PS. Conduzimos as análises de covariação em ordem de responder nossa pergunta de pesquisa, a qual retomamos: a descrição da covariação morfossintática em grupos de falantes de diferentes regiões geográficas possibilita a identificação de sua origem dialetal?

A descrição de covariação se inicia a partir da observação de possíveis relações entre os pares de variáveis, por meio de análise de correlação, conforme Guy (2013) e Oushiro (2015a; 2016a). Após isso, passamos para a observação de padrões de agrupamento social, similar ao feito em Guy (2013) e Oushiro (2015a; 2016a). A análise de *cluster* compõe nossa terceira etapa de análise, de modo a observar agrupamentos dos falantes em grupos naturais, conforme desenvolvido por Freitag (2022).

6.1 A DESCRIÇÃO DA COVARIAÇÃO

Prévias análises de covariação têm evidenciado que alguns conjuntos de variáveis podem apresentar correlação na fala individual dos falantes (Guy, 2013; Oushiro, 2015a; 2016a). As correlações, contudo, não são tão comuns ou fortes quanto deveriam ser para se considerar a existência de uma coesão social/dialetal para a definição de um socioleto/dialeto. Por exemplo, é de se esperar que certas variáveis se correlacionem fruto de pressões externas à língua: marcas zero de concordância são comuns na fala de indivíduos menos escolarizados, o que pode levar a uma correlação no uso de variáveis como concordância verbal e concordância nominal. De forma similar, certas variáveis podem se correlacionar como fruto de pressões internas à língua, como a similaridade estrutural. Por exemplo, a mudança nos pronomes pessoais do PB resultou em outras mudanças no sistema pronominal, o que pode levar à covariação devido a pressões internas.

Se considerarmos as variáveis morfossintáticas descritas neste trabalho, observamos que compartilham de certas similaridades estruturais. Todas as variáveis envolvem pronomes do PB, sejam pessoais, possessivos ou clíticos. Além disso, três das quatro variáveis se relacionam com o paradigma pronominal de 2PS. Uma vez que o estudo de covariação busca observar "se falantes que tendem a empregar a variante x de uma variável A também tendem a empregar a variante y de uma variável B" (Oushiro, 2015b, p. 230), é de se questionar, com base em nossos dados, se falantes que tendem a empregar *você* também tendem a empregar *te*; se a ausência de artigo se correlaciona com *seu*; se as variáveis de 2PS tendem a seguir um mesmo padrão; ou se, apesar da similaridade estrutural entre as variáveis, não há, necessariamente, relação entre seus usos.

Uma vez que nossa questão de pesquisa é voltada para identificação da origem dialetal de falantes por meio de traços morfossintáticos agrupados, compreendemos que, em um ambiente como uma universidade federal, que recebe alunos de diferentes regiões, a presunção de origem seja algo difícil de se observar, frente ao contínuo contato que os falantes mantêm no ambiente universitário e à mudança linguística. Contudo, ainda que falantes compartilhem práticas em comum na universidade, traços linguísticos de sua comunidade de origem podem permanecer em seu vernáculo, possibilitando a identificação de sua origem por meio de seus usos linguísticos.

A ideia de que certas variantes que caracterizam dialetos podem se correlacionar já foi posta em prova, conforme Oushiro (2015a; 2016a). Contudo, diferente da existência de "português paulista", não é possível a observação de um "português universitário da UFS" que seja relativamente homogêneo, mas é possível a observação de padrões mais gerais de uso, que podem se relacionar e apontar para um padrão de uso mais específico para determinada área geográfica. Nas três seções que seguem, buscamos colocar essa ideia em prova por meio de nossos dados.

6.1.1 Estabelecendo correlações

Para a análise de correlação, cada uma das quatro variáveis morfossintáticas foi analisada individualmente de duas formas. Na primeira forma, extraímos a frequência individual de uso da variante de aplicação a partir do total de dados. Por exemplo, o falante 01ent.UFS-SaoCristovao2018__desl.I_final_lui.ms.24 apresentou um total de quarenta e uma (41) ocorrências do uso variável de artigo antes de possessivos. Dessas ocorrências, vinte e três (23) são para a ausência de artigo. Assim, sua taxa de frequência individual é de 56,10%

para a variante ((23/41)*100). Isso foi realizado em cada um dos cento e oitenta e um (181) falantes que compõem nosso *corpus*, em cada um dos fenômenos morfossintáticos variáveis, de forma automatizada no RStudio (RStudio Team, 2015). Falantes que não apresentaram ocorrências da variante ou da variável receberam a taxa 0%. Com base na taxa de frequência, apenas a ausência de artigo antes de possessivos pré-nominais apresenta distribuição normal, conforme o teste de normalidade Shapiro-Wilk (p = 0.21).

Na segunda forma, extraímos os *log odds* de uso da variante de aplicação por falante a partir de modelos de efeitos mistos no RStudio (RStudio Team, 2015), por meio da função `glmer` do pacote `lme4` (Bates *et al*., 2015), com falante como efeito aleatório (fórmula = glmer(VD ~ (1 | informante), family = binomial(link = "logit"), data = data). Neste modelo, (1 | informante) especifica que há a inclusão de um efeito aleatório para o falante. Isso pressupõe a existência de dados de várias observações para cada indivíduo e que essas observações são correlacionadas entre si. O modelo apresenta *log odds* positivos ou negativos para cada falante, representando os *logs* de razão de chances de a variante de interesse acontecer. Essa forma foi mobilizada em ordem de tentar obter dados com distribuição normal. Isso, contudo, não aconteceu de forma total, uma vez que apresentaram distribuição normal apenas a ausência de artigo (p= 0.11) e o pronome pessoal de 2PS *você* (p = 0.34).

Frente à violação do pressuposto de normalidade na maior parte dos dados, as análises de correlação são feitas a partir de teste de correlação não-paramétrico, a correlação de Spearman. O coeficiente da correlação de rho de Spearman (ρ), conforme apresentado no Capítulo 3, gera um coeficiente que varia de -1 a +1: a maior proximidade a um dos extremos (-1 ou 1) evidencia maior força de correlação, à medida que valores próximos a 0 implicam em correlação fraca ou inexistente. Os sinais de mais (+) e menos (-), por sua vez, indicam a direção da correlação: valores positivos indicam que o aumento em uma variável implica no aumento na outra variável, enquanto valores negativos indicam que o aumento de uma variável implica na diminuição de outra.

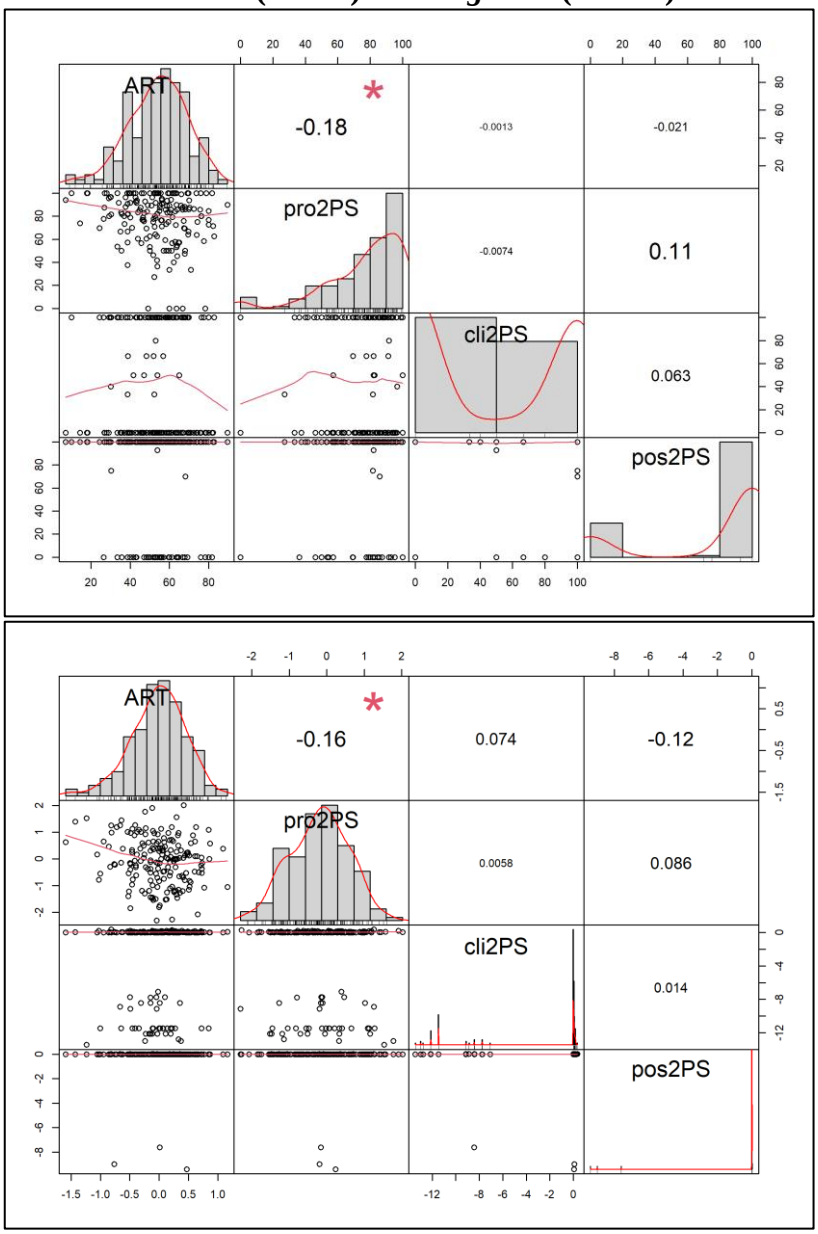
Para a análise de correlação, questionamos se os testes evidenciam relação entre as variáveis e dão suporte à ideia de que variáveis morfossintáticas covariam entre si. Nossa hipótese é que os coeficientes de ρ revelem um panorama da covariação ao evidenciar a existência de relação entre as variáveis, como também a força dessa relação, de modo com que os grupos controlados apresentem correlação significativa entre os pares de variáveis.

No que segue, as Figuras 55-63 são matrizes de correlação que contém os 6 pareamentos possíveis entre as quatro variáveis morfossintáticas. Os quadros na linha diagonal mostram as distribuições das frequências dos falantes por meio de histogramas; os

quadros do canto superior direito mostram os coeficientes de correlação (ρ) (a existência de símbolos representa a significância: **.** p > 0,05; *p < 0,05;**p < 0,01; ***p < 0,001); e os quadros do canto inferior esquerdo mostram os gráficos de dispersão e suas respectivas linhas de regressão. Os fenômenos são representados por siglas: ART (ausência de artigo diante de possessivo), pro2PS (pronome de 2PS *você*), cli2PS (clítico de 2PS *te*) e pos2PS (possessivo de 2PS *seu*).

Na Figura 55, observamos os coeficientes de correlação a partir do conjunto total de dados, tanto com as taxas de frequência de uso individual (acima), quanto com os *log odds* extraídos do modelo de efeitos mistos (abaixo).

Figura 55 – Matriz de correlação entre quatro variáveis com base nas taxas de frequência de uso das variáveis (acima) e em *log odds* (abaixo) em dados das amostras



Fonte: elaboração própria.

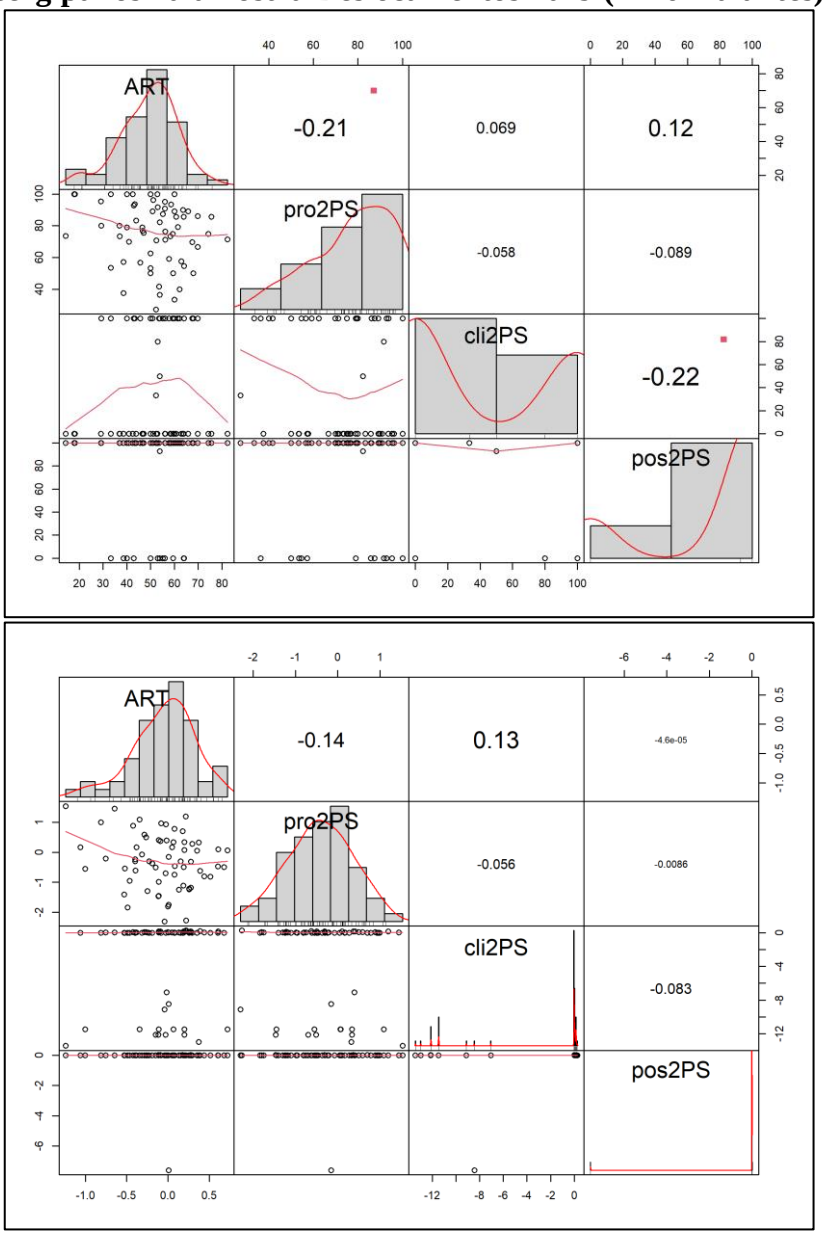
Há correlação negativa e significativa entre ART~pro2PS com as frequências (ρ = -0.18, p = 0.017) e com os *log odds* (ρ = -0,16, p = 0,036). Conforme os valores de ART ou pro2PS aumentam, tanto as frequências quanto os *log odds* da outra tendem a diminuir: há uma tendência contrária na ausência de artigo definido e a variante *você* como sujeito, indicando que os falantes tendem a não utilizar uma forma com a outra. Nos demais pares, o coeficiente de Spearman é baixo, o que resulta na impossibilidade de se rejeitar a hipótese nula: não há correlação entre os pares de variáveis. Embora a covariação entre as variáveis

pronominais tenha sido esperada linguisticamente (paradigma pronominal), isso não ocorreu na análise de correlação.

No conjunto geral de nossos dados, só há correlação significativa entre a ausência de artigo antes de possessivos e o pronome *você* como sujeito de 2PS: os falantes que frequentemente usam a ausência de artigo definido tendem a não usar o *você* como sujeito de 2PS, o que pode refletir escolhas linguísticas específicas de sua região geográfica.

Considerar divisões entre os grupos pode nos fornecer resultados mais específicos para a covariação. Retomamos o argumento de Freitag e Rost-Snichelotto (2004) de que diferenças/similaridades encontradas em análises e descrições podem ser resultantes da metodologia amostral. As amostras que utilizamos lidam com organizações similares, mas ainda apresentam distinções. A análise de correlação por amostra pode permitir a visualização de diferentes relações entre as variáveis (Figura 56).

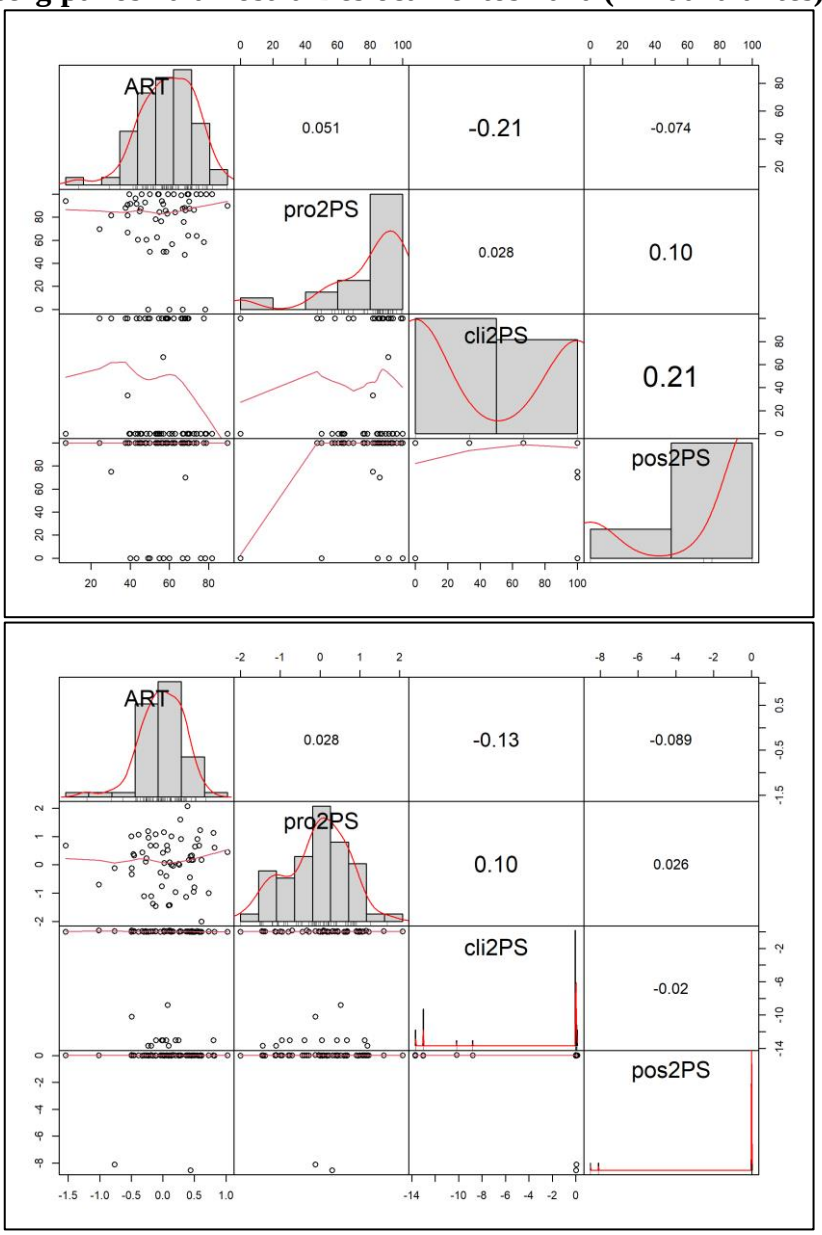
Figura 56 – Matriz de correlação entre quatro variáveis com base nas taxas de frequência (acima) e em *log odds* (abaixo) de uso das variáveis em dados de falantes sergipanos na amostra Deslocamentos 2019 (N= 64 falantes)



Fonte: elaboração própria.

A análise de correlação considerando a amostra D2019 individualmente não altera a significância dos coeficientes de correlação, tanto com a frequência de uso das variantes por indivíduo quanto com os *log odds* individuais. Na amostra D2020, podemos obter outros resultados (Figura 57).

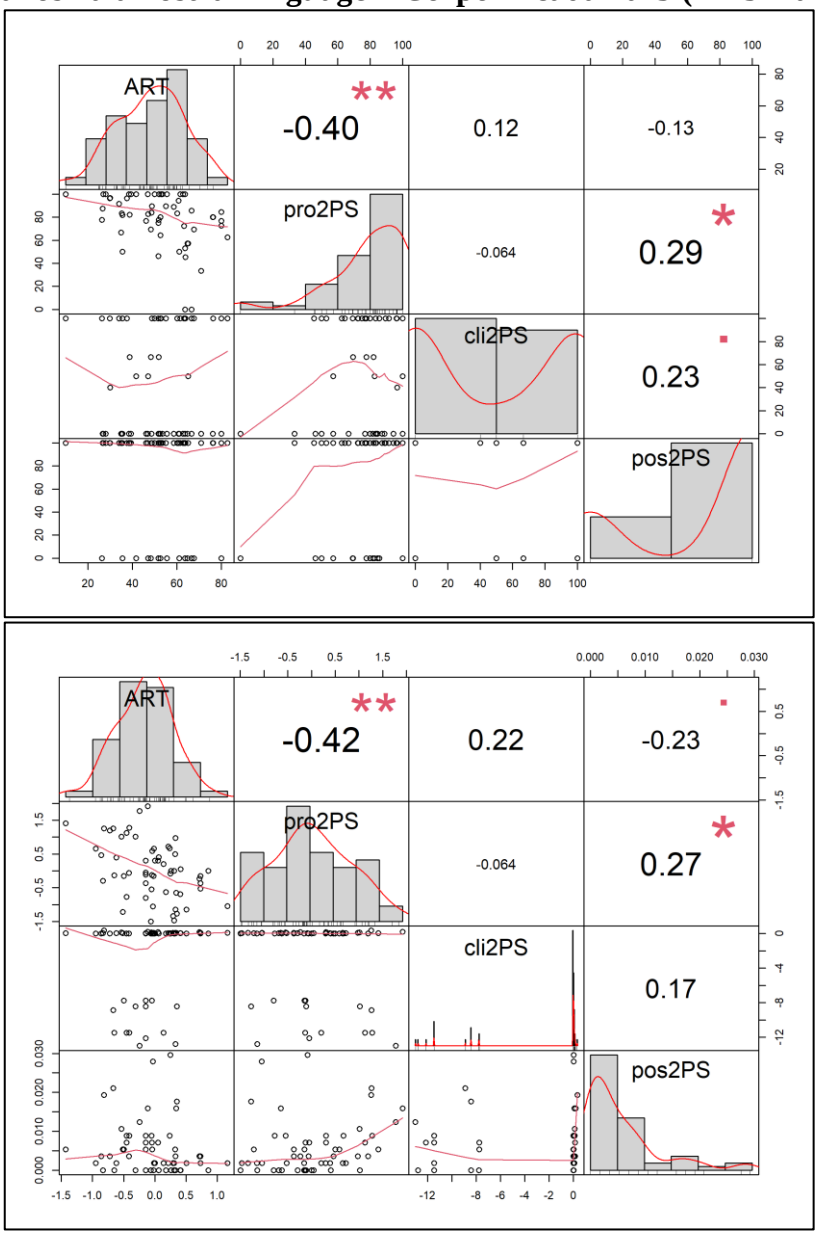
Figura 57 – Matriz de correlação entre quatro variáveis com base nas taxas de frequência (acima) e em *log odds* (abaixo) de uso das variáveis em dados de falantes sergipanos na amostra Deslocamentos 2020 (N= 60 falantes)



Fonte: elaboração própria.

A amostra D2020 segue comportamento similar à amostra D2019 (Figura 57) ao não apresentar relação significativa entre nenhum dos pares de variáveis, com nenhum dos tipos de dados. Resta-nos a observação em LC2023 (Figura 58).

Figura 58 – Matriz de correlação entre quatro variáveis com base nas taxas de frequência (acima) e em *log odds* (abaixo) de uso das variáveis em dados de falantes sergipanos na amostra Linguagem Corporificada 2023 (N= 57 falantes)



Fonte: elaboração própria.

Com dados da amostra LC2023, obtemos correlações significativas: há correlação negativa e significativa entre ART~pro2PS, com a frequência (ρ = -0,40 p= 0.001) e com os *log odds* (ρ = -0,42 p = 0.001), e positiva e significativa entre pro2PS~pos2PS com a frequência (ρ = 0,29 p = 0.03) e com os *log odds* (ρ = 0,27 p = 0.04). Os resultados quanto a ART~pro2PS indicam que, à medida que a ausência de artigo aumenta, a frequência e a probabilidade de uso de *você* como sujeito de 2PS tendem a diminuir. A correlação significativa ($p < 0.05$) e o coeficiente de ρ relativamente forte (-0.40 e -0.42) sugerem que há uma tendência consistente entre os falantes de não combinar essas duas formas. Quanto à

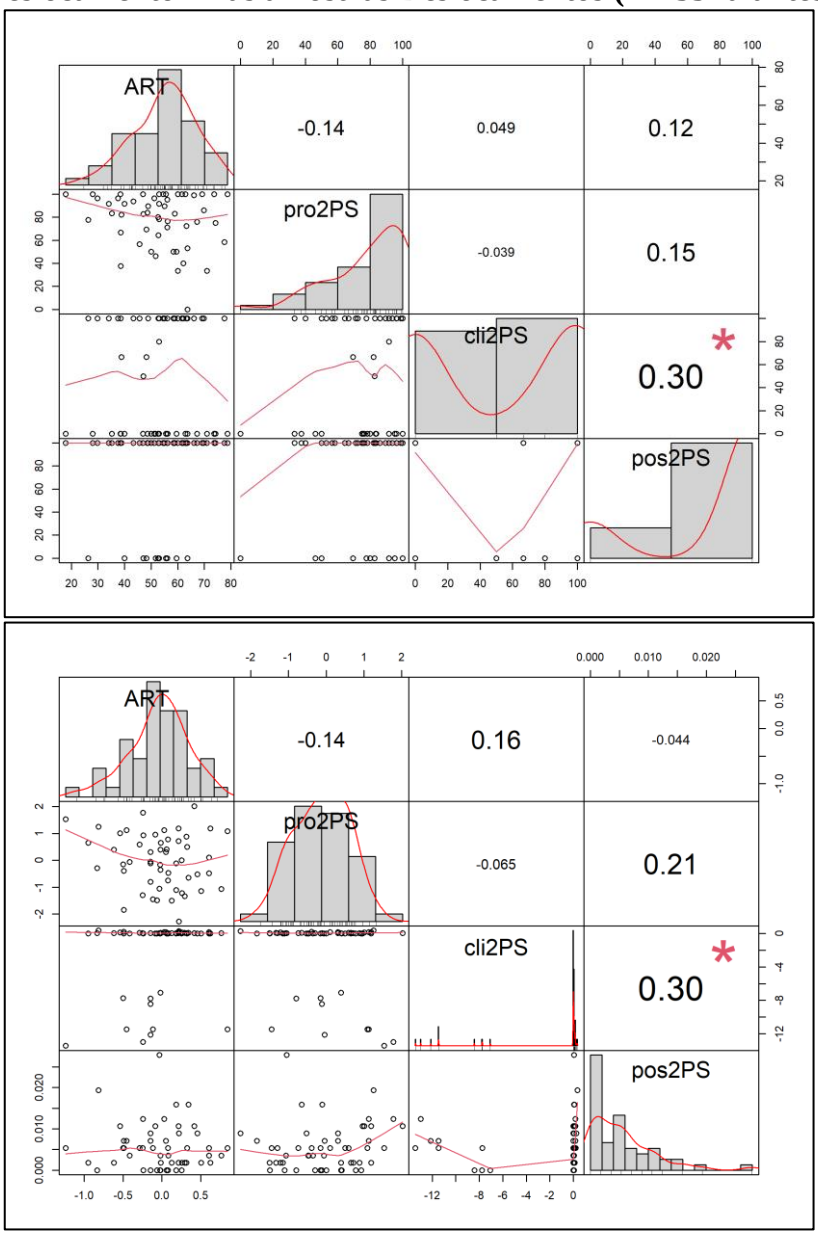
pro2PS~pos2PS, aqui vemos um paralelismo: à medida que o uso de *você* aumenta, a frequência e a probabilidade do uso de *seu* também tendem a aumentar. A correlação, embora moderada, é significativa e reflete um padrão de uso no qual falantes combinam as duas formas em seus usos.

Tais resultados indicam algumas tendências no comportamento dialetal dos falantes, na medida em que utilizam conjuntamente *você* e *seu*, reforçando a ideia de que essas formas estão gramaticalmente relacionadas e tendem a coocorrer, evidenciando um possível efeito de paralelismo, mas não o fazem com a ausência de artigo, privilegiando a presença.

É evidente, contudo, que trabalhamos com grupos muito heterogêneos, frente à diversidade nas origens dos falantes e também em sua integração à universidade, conforme evidenciam Côrrea (2019) e Ribeiro (2019): há falantes da região metropolitana do estado de Sergipe; falantes do interior que fazem o percurso diário para o *campus* da UFS; falantes do interior que migraram para a região metropolitana do estado; e falantes externos a Sergipe que migraram pra região circunvizinha à UFS. Tais falantes podem apresentar comportamentos linguísticos distintos um dos outros, dado o efeito da região geográfica sobre o vernáculo dos grupos de falantes; mas, pertencentes à mesma região, podem apresentar comportamento similar, com relações significativas entre os pares de variáveis.

Dividir a análise de correlação com base nos perfis de deslocamento nos permite observar como os pares de variáveis se comportam quanto a cada um dos grupos, o que nos ajuda a observar se é possível traçar um perfil de comportamento linguístico para falantes de regiões geográficas específicas, possibilitando a identificação de sua origem dialetal. A Figura 59 apresenta, por meio das taxas de frequência de uso dos falantes (à esquerda) e dos *log odds* (à direita), a matriz de correlação considerando dados do Deslocamentos 1 de todas as amostras (N= 53 falantes).

Figura 59 – Matriz de correlação entre quatro variáveis com base nas taxas de frequência (acima) e em *log odds* (abaixo) de uso das variáveis em dados de falantes do Deslocamento 1 nas amostras Deslocamentos (N= 53 falantes)



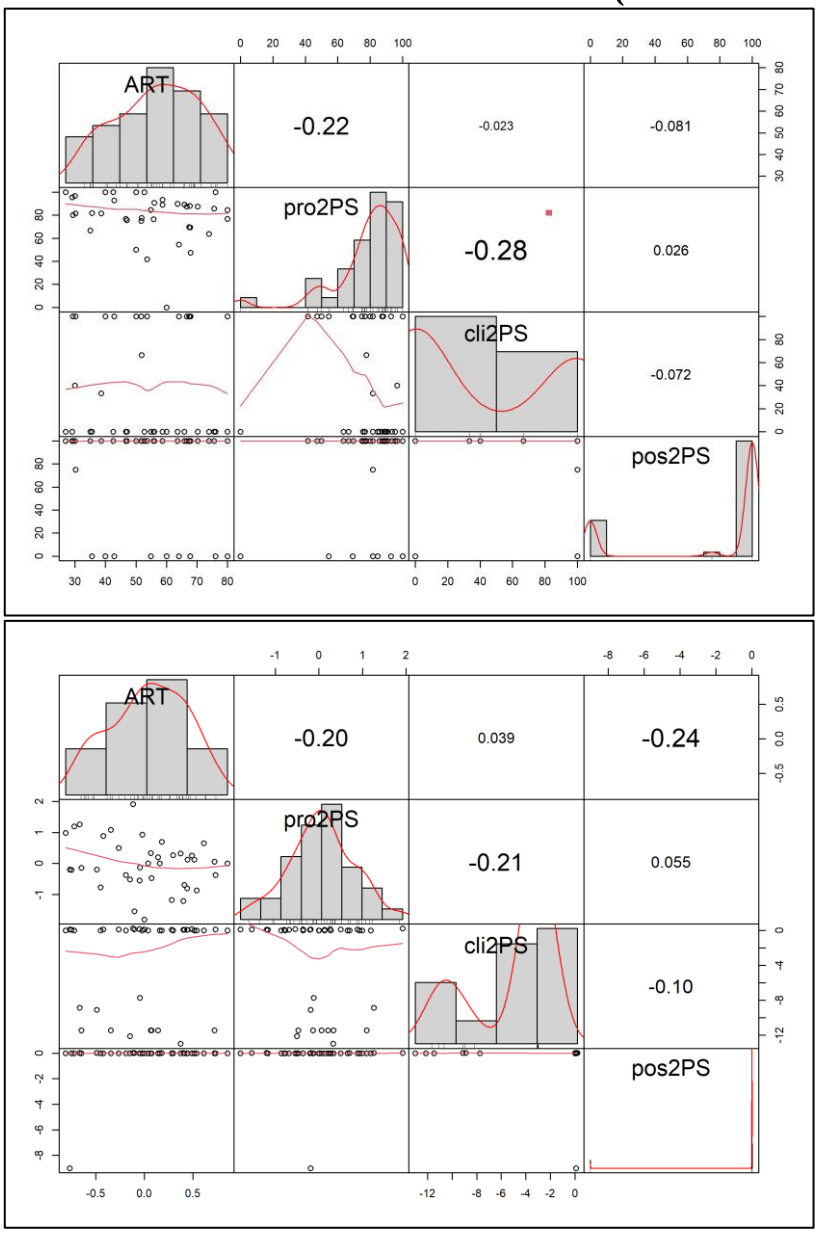
Fonte: elaboração própria.

Ainda que esperássemos uma maior homogeneidade em relação ao comportamento de falantes do Deslocamento 1, dado que são nascidos e residentes na região metropolitana de Sergipe, os coeficientes de correlação nos permitem rejeitar a H_0 apenas em um par: há correlação positiva e significativa entre cli2PS~pos2PS, com as frequências ($\rho = 0{,}30$ p= 0.02) e com os *log odds* ($\rho = 0{,}30$ p= 0.03). Falantes que utilizam *te* como clítico de 2PS tendem a utilizar *seu* como possessivo de 2PS. Similar ao que vimos com pro2PS~pos2PS, a relação

aponta uma coocorrência entre as formas. Falantes do Deslocamento 1 tendem a manter um padrão no qual se usam ambas as formas como referência à 2PS.

A Figura 60 apresenta as matrizes de correlação com base nos dados de falantes do Deslocamento 2 (N= 39 falantes), a partir das taxas de frequência de uso individual (à esquerda) e de *log odds* extraídos do modelo de efeitos mistos (à direita).

Figura 60 – Matriz de correlação entre quatro variáveis com base nas taxas de frequência (acima) e em log odds (abaixo) de uso das variáveis em dados de falantes do Deslocamento 2 nas amostras Deslocamentos (N= 39 falantes)

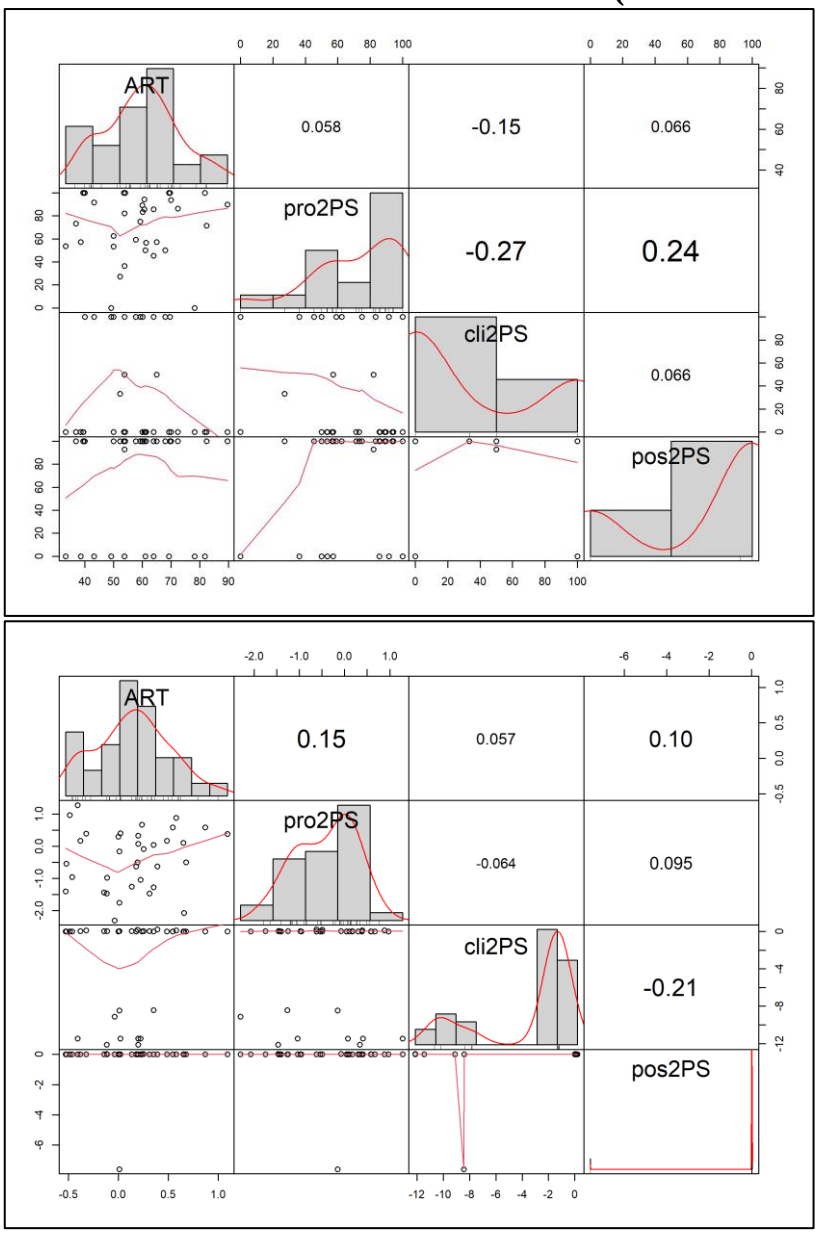


Fonte: elaboração própria.

Os coeficientes de correlação com dados de falantes do Deslocamento 2, falantes do interior do estado de Sergipe que fazem o percurso diário para a UFS, não apresentam significância entre os pares de variáveis, tanto com dados de frequência de uso, quanto com os *log odds* extraídos do modelo de regressão.

Falantes do Deslocamento 3 (N= 35 falantes) podem seguir um caminho similar ao de falantes do Deslocamento 1, já que residem na região metropolitana, ou do Deslocamento 2, já que são do interior (Figura 61).

Figura 61 – Matriz de correlação entre quatro variáveis com base nas taxas de frequência (acima) e em log odds (abaixo) de uso das variáveis em dados de falantes do Deslocamento 3 nas amostras Deslocamentos (N= 35 falantes)

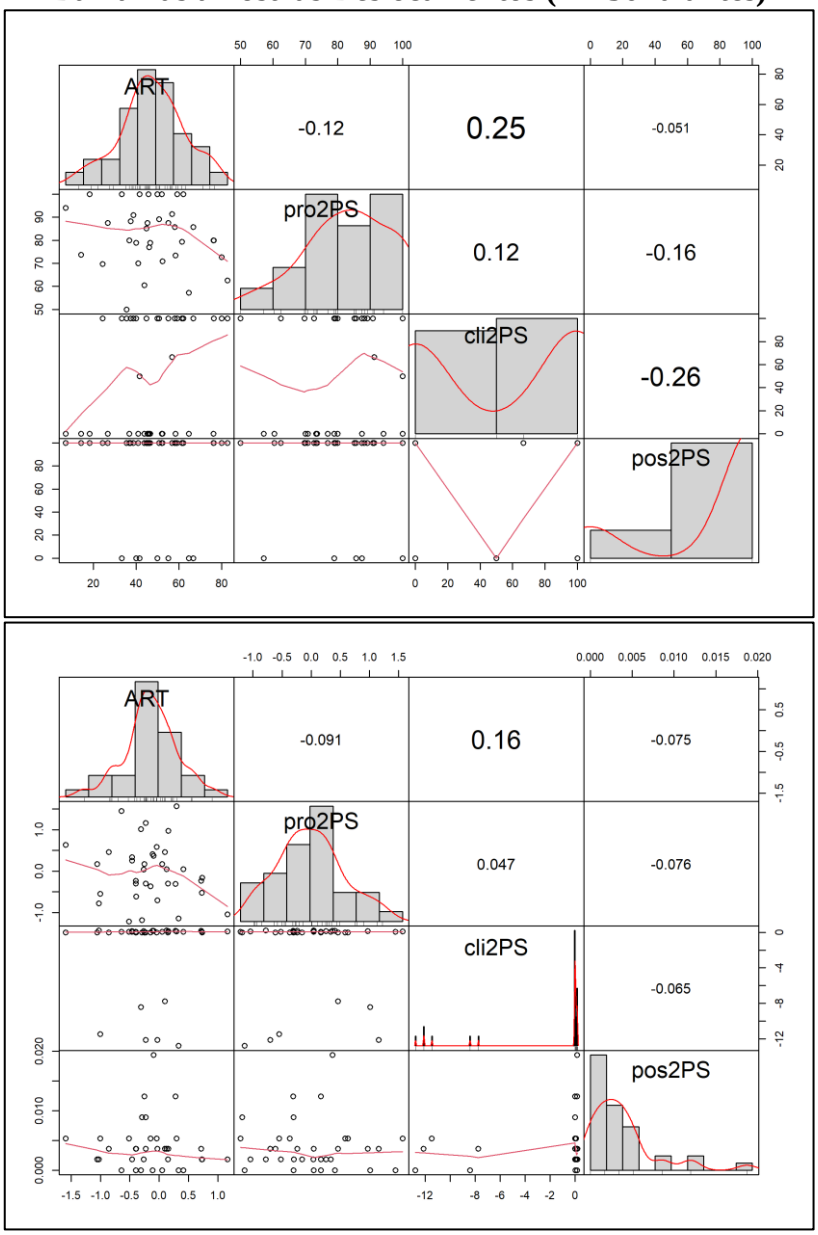


Fonte: elaboração própria.

Na mesma direção que dados do Deslocamento 2, os coeficientes de correlação do Deslocamento 3, falantes do interior do estado que residem na região metropolitana de Sergipe, também não apresentam significância em nenhum dos pares de variáveis, independentemente do tipo de dado utilizado para a condução das análises de correlação. A ausência de correlações significativas pode indicar que os falantes das regiões em questão não apresentam um comportamento linguístico homogêneo. Em outras palavras, a variabilidade individual pode ser mais proeminente do que as tendências regionais.

Temos considerado, contudo, apenas o comportamento de falantes de Sergipe. As amostras utilizadas comportam a fala de universitários oriundos de outros estados, como Alagoas e Bahia, organizadas em áreas dialetais distintas de Sergipe. Conduzimos testes de correlação de Spearman também considerando os usos linguísticos desses falantes. Na Figura 62, visualizamos a matriz de correlação para falantes da Bahia, que organizamos considerando treze (13) da amostra D2019, removendo desta os três (3) falantes oriundos de outros estados (2 de São Paulo e 1 do Mato Grosso do Sul), doze (12) falantes da amostra D2020, e onze (11) da amostra LC2023, removendo três (3) de outros estados (1 de São Paulo, 1 do Maranhão e 1 do Mato Grosso do Sul), totalizando trinta e seis (36) falantes.

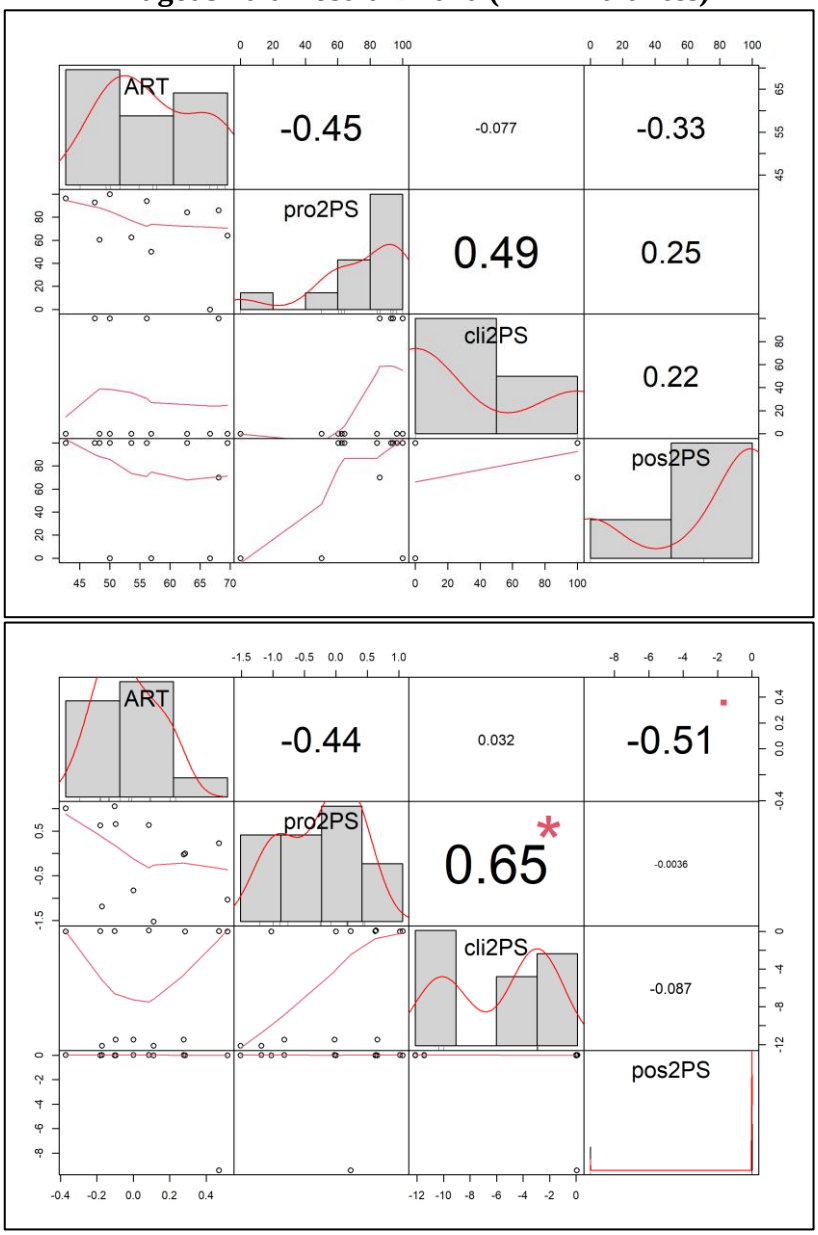
Figura 62 – Matriz de correlação entre quatro variáveis com base nas taxas de frequência (acima) e em *log odds* (abaixo) de uso das variáveis em dados de falantes da Bahia nas amostras Deslocamentos (N= 36 falantes)



Fonte: elaboração própria.

Não há significância entre os pares de variáveis no Deslocamento Bahia. Resta-nos ainda observar o comportamento de falantes de Alagoas, que se encontram apenas na amostra D2020, com 12 falantes, um quantitativo considerado baixo, o que pode interferir nas análises de correlação, já que fenômenos morfossintáticos tendem a ter uma frequência relativamente reduzida. A Figura 63 apresenta a matriz de correlação para esses dados.

Figura 63 – Matriz de correlação entre quatro variáveis com base nas taxas de frequência (acima) e em *log odds* (abaixo) de uso das variáveis em dados de falantes de Alagoas na amostra D2020 (N= 12 falantes)



Fonte: elaboração própria.

Com as taxas de frequência, não há correlação significativa entre os pares de variáveis. Por outro lado, com os *log odds*, há relação significativa e positiva entre o uso de *você* em posição de sujeito e o uso do clítico de 2PS *te* ($\rho = 0{,}65$, $p = 0.008$). À medida que há aumento no uso de *você* na fala de estudantes da UFS oriundos de Alagoas, há também aumento no uso do clítico *te*, sugerindo que há um padrão comportamental para esses falantes, no qual há um uso conjunto de *você*+*te*. Na análise univariada, todavia, falantes de Alagoas apresentam predomínio para o uso de *lhe*, além de terem a segunda menor frequência para o uso de *você*.

É de se observar, contudo, que na visualização da correlação os pontos não aparecem próximos à linha de correlação, o que indica que indica que a relação entre as variáveis não é perfeitamente linear.

As análises de correlação nos permitem algumas respostas. A primeira delas é que a existência de semelhanças estruturais não resulta, necessariamente, em relações significativas entre variáveis morfossintáticas, como ocorreu com outras variáveis morfossintáticas em Oushiro (2015a; 2016a), ou entre variáveis fonológicas e morfossintáticas, como em Guy (2013). Ainda que estejamos lidando com variáveis que se relacionam com o paradigma pronominal do PB, isso não implica que o uso de uma variante se relacione com o de outra.

Contudo, conforme vimos ao longo das análises, há, em conjuntos específicos de dados, correlação significativa entre pares de variáveis (Quadro 12).

Quadro 12 – Correlações significativas

Conjunto	Direção	Dados	Variáveis
Geral	Negativa	Frequência e *log odds*	ART~pro2PS
Amostra Linguagem Corporificada 2023	Negativa	Frequência e *log odds*	ART~pro2PS
Amostra Linguagem Corporificada 2023	Positiva	Frequência e *log odds*	pro2PS~pos2PS
Deslocamento 1	Positiva	Frequência e *log odds*	cli2PS~pos2PS
Alagoas	Positiva	*Log odds*	pro2PS~cli2PS

Fonte: elaboração própria.

A segunda das respostas é que o condicionamento extralinguístico, isto é, o significado dialetal das variáveis, também não resulta diretamente em covariação, dado que pares de variáveis que poderiam ocorrer concomitantes não apresentaram relação significante. Tais resultados levantam a questão da ausência de padrões de uso conjunto que refletem a região dialetal do falante. Isso significa que, embora tenhamos analisado quatro variáveis morfossintáticas dialetalmente distintas, as correlações entre elas não foram estatisticamente significativas na maior parte das observações.

Vimos, por exemplo, que falantes sergipanos tendem a fazer uso da ausência de artigo, do pronome *você* e do clítico *te,* esse exceto o Deslocamento 2 em D2019. Esse resultado, contudo, não resulta em relações positivas significativas entre as variáveis, mas houve relação entre *te* e *seu* com falantes do Deslocamento 1. Igualmente, vimos que falantes externos a Sergipe tendem a fazer menor uso da ausência de artigo, alto uso de *você* e alto uso de *te,* esse

exceto Alagoas em D2020. Contudo, em nossos dados, foram os falantes alagoanos que apresentaram relação significativa e positiva para o uso de *você+te*.

Evidentemente, podemos especular algumas explicações para a falta de significância na maior parte dos coeficientes de correlação. É certo, por exemplo, que a região de origem do falante pode apresentar uma diversidade linguística, não possuindo um comportamento homogêneo, dada à variação ao nível do indivíduo – ainda que haja variação intrafalante em apenas 2/4 dos fenômenos. Também é possível que fatores extralinguísticos que não foram controlados nos testes de correlação, como tempo no curso, idade ou gênero, e que exercem influência nos usos das variáveis morfossintáticas, conforme vimos nos Capítulos 2 e 5, possam ter resultado na diferenciação do padrão linguístico de uso do falante em relação a sua região de origem.

Além disso, cada falante pode possuir seu próprio padrão de uso, conforme vimos nas análises univariadas. A variabilidade individual, como também a categoricidade intrafalante, pode ter interferido nos testes estatísticos, resultando em correlações não significativas. Somam-se a isso o baixo quantitativo de contagens por falante, a não produção de variantes das variáveis por muitos falantes e o número de falantes analisados por grupo, que também podem ter contribuído para não detectar correlações significativas. Um estudo com uma amostra maior, com perguntas que possibilitem maior produção das variáveis focos, e mais representativa da região pode ser necessário para obter resultados mais conclusivos.

Tais explicações, contudo, são apenas conjecturas sobre os resultados obtidos. Como proferiu Dugald Bell (s/d), ausência de evidências não significa evidência da ausência. Os resultados são generalizáveis apenas para nosso conjunto de dados. Apesar de análises de correlação em estudos anteriores (Guy, 2013; Oushiro, 2015a; 2016a) terem apresentado evidências para a covariação nos dados analisados, em nossos dados, a análise de correlação mostrou significância em poucos pares. Nesse sentido, os dados apresentados pelas análises de correlação não nos permitem chegar a conclusões robustas.

Quais outras formas de análise podem acrescentar mais informações para a nossa descrição e contribuir para a identificação de origem dialetal? No que segue, apresentamos resultados obtidos por meio de técnicas de agrupamento de dados.

6.1.2 Padrões de agrupamento social

Nesta seção, descrevemos padrões gerais de agrupamento social, similar ao desenvolvido por Guy (2013) e Oushiro (2015a; 2016a), por meio da classificação de

tendências de usos das variantes pelos falantes. Diferentemente dos trabalhos citados, que utilizaram pesos relativos extraídos de modelos de regressão, utilizamos a taxa de frequência de uso individual das formas, também utilizadas na análise de correlação anterior, com base na variante de aplicação que temos utilizado (ausência de artigo, pronome *você*, clítico *te* e possessivo *seu*).

Nos trabalhos de Guy (2013) e Oushiro (2015a), as classificações das tendências seguiram uma divisão ternária para os pesos relativos – alto (acima de 0,60), médio (entre 0,40 e 0,60) e baixo (abaixo de 0,40) –, enquanto Oushiro (2016a) realizou uma classificação binária, frente ao quantitativo de variáveis descritas (N= 6) e número de falantes (N= 118) – alto (igual ou acima de 0,50) ou baixo (igual ou abaixo de 0,499). Em nosso estudo, considerando a existência de frequências na faixa dos 50% e por termos apenas quatro (4) variáveis morfossintáticas, optamos por uma divisão ternária. Classificamos frequências abaixo de 40% como B (baixas); entre 40% e 60% como M (médias), e acima de 60% como A (altas).

Ao atribuir A, M ou B, há oitenta e um (81) padrões possíveis nos quais um falante pode ser classificado ($3^4 = 81$): AAAA, AAAB, AAMB etc. Considerando a aleatoriedade, teríamos uma expectativa de distribuição equilibrada do número de falantes por padrão – uma média de 2,23 (124 falantes / 81 padrões) para as amostras. Evidentemente, uma vez que grupos de falantes tendem a se comportar de formar similar, não esperamos que todos esses 81 padrões apareçam, e sim que haja um quantitativo de padrões que concentre a maior quantidade de falantes.

Nesta análise, lançamos a seguinte questão: falantes se agrupam em padrões sociais significativos que possibilitem a observação de coesão dialetal para grupos específicos? Hipotetizamos que haverá um alto grau de coesão dialetal para um grupo (perfil de deslocamento) se a maior parte dos falantes (≥ 75%) que compõem esse grupo se encaixar no mesmo padrão (p.ex., AAAA, BBBB etc.).

A categorização das frequências evidencia a ocorrência de trinta e dois (32) dos oitenta e um (81) padrões. A Tabela 9 apresenta os 10 padrões mais frequentes, dado o baixo percentual dos outros padrões. Nela, a ordem dos índices (A, M e B) segue ART, pro2PS, cli2PS e pos2PS.

Tabela 9 – Padrões de agrupamento social das variantes descritas – classificação ternária

Padrão*	Falantes	Frequência
MABA	29	16,0%
MAAA	23	12,7%
AABA	20	11,0%
AAAA	18	9,9%
BABA	16	8,8%
BAAA	11	6,0%
MAAB	8	4,4%
MABB	6	3,3%
MMAA	6	3,3%
AABB	5	2,7%

*B< 40%; M 40%-60%; A> 60%

Fonte: elaboração própria.

O padrão mais recorrente é MABA, com vinte e nove (29) falantes. Esses falantes tendem a empregar uma taxa média (40%-60%) de ausência de artigo antes de possessivo; alta taxa (> 60%) de uso de *você*; baixa taxa (< 40%) do clítico *te*; e alta taxa do possessivo *seu*. Desses falantes, oito (8) do Deslocamento 2, sete (7) são do Deslocamento 4, cinco (5) do Deslocamento 1, três (3) do Deslocamento 3, três (3) da Bahia e três (3) de Alagoas.

O padrão MAAA é o segundo mais frequente, com vinte e três (23) falantes que fazem uso alto de *você*, *te* e *seu*, mas médio uso da ausência de artigo. Desses falantes, nove (9) são do Deslocamento 1, quatro (4) da Bahia, três (3) de Alagoas, do Deslocamento 2 e do Deslocamento 3, e um (1) do Deslocamento 4. Notemos o compartilhamento no comportamento das variáveis de 2PS, em que, anteriormente, hipotetizamos que haveria covariação, frente à semelhança estrutural.

O terceiro padrão mais recorrente é AABA, com vinte (20) falantes que fazem uso alto da ausência de artigo, pronome *você* e possessivo *seu*, mas baixo uso de *te*. Dentro do grupo, seis (6) são do Deslocamento 3, cinco (5) do Deslocamento 2 e do Deslocamento 1, cada, e dois (2) do Deslocamento 4 e Alagoas, cada. Não há falantes do deslocamento Bahia.

O padrão no qual todas as taxas são iguais (AAAA) apresenta dezoito (18) falantes: eles fazem alto uso da ausência de artigo, de *você*, *te* e *seu*. Nesse padrão, há cinco (5) falantes do Deslocamento 1, cinco (5) do Deslocamento 4, quatro (4) falantes do Deslocamento 2,

dois (2) do Deslocamento 3, um (1) do Deslocamento Alagoas e um (1) da Bahia. Por fim, no quinto padrão, BABA, há dezesseis (16) falantes que apresentam frequências baixas de ausência de artigo e de *te*, e frequências altas de *você* e de *seu*. Os falantes que compõem esse grupo são do Deslocamento 4 (N= 4), Deslocamento 2 (N= 4), Deslocamento 1 (N= 4), Deslocamento 3 (N=3) e Bahia (N= 1).

Dos seis (6) padrões mais frequentes, todos eles apresentam frequência alta de *você* e *seu*, o que representa 64,4% dos falantes, indício de que há uma certa relação entre os usos dessas variantes no comportamento linguístico dos falantes. Além disso, em três (3) desses padrões, os falantes fazem baixo uso de *te*. A ausência de artigo foi a variável que mais apresentou variância nos cinco grupos mais frequentes. A Tabela 10 apresenta os padrões obtidos por perfil de deslocamento com base na classificação ternária e o percentual que o padrão mais frequente representa para o deslocamento (%).

Tabela 10 – Padrões mais frequentes por perfil de deslocamento - classificação ternária – e o percentual que o padrão mais frequente representa para o deslocamento (%)

Padrão*	**D1**	**D2**	**D3**	**D4**	**BA**	**AL**
MABA	5	8	3	7	3	3
MAAA	9	3	3	1	4	3
AABA	5	5	6	2	0	2
AAAA	5	4	3	5	1	1
BABA	4	4	3	4	1	0
BAAA	5	2	0	0	3	0
MAAB	2	2	1	3	0	0
MABB	4	1	0	0	0	1
MMAA	3	2	0	0	0	0
AABB	0	2	3	0	0	0
%	16,9%	20,5%	17,1%	23,3%	33,3%	25%

*B< 40%; M 40%-60%; A> 60%

Fonte: elaboração própria.

Falantes do Deslocamento 1 se distribuem em dezoito (18) padrões, com maior quantitativo de falantes (N = 9) aquele no qual há taxa média (40%-60%) da ausência de

artigo e alta taxa (>60%) dos demais fenômenos (MAAA). Os falantes do Deslocamento 2 se distribuem em dezesseis (16) padrões, com maior quantitativo de falantes (N= 8) no padrão no qual há médio uso da ausência de artigo, alto uso de *você* e *seu* e baixo uso (<40%) de *te* (MABA). No Deslocamento 3, no qual falantes se distribuem em dezoito (18) padrões, o maior quantitativo de falantes é observado no padrão AABA (N= 6), em que falantes fazem alto uso da ausência de artigo, de *você* e de *seu*, e baixo uso de *te*.

Falantes do Deslocamento 4 se distribuem em quatorze (14) padrões, sendo o com maior quantitativo o MABA (7), similar a falantes do Deslocamento 2. Com falantes da Bahia, há cinco (5) padrões, sendo o mais frequente o padrão MAAA (N= 4), no qual falantes fazem médio uso da ausência de artigo e alto uso das demais variáveis, similar a falantes do Deslocamento 1. Falantes de Alagoas seguem um caminho parecido, no qual, dos seis (6) padrões, o com maior quantitativo de falantes é o MAAA (N= 4), empatado com o padrão MABA (N= 3).

Retomemos nossa hipótese, a de que haveria alto grau de coesão dialetal para um grupo se a maior parte dos falantes que compõem esse grupo se encaixasse no mesmo padrão. Essa hipótese, contudo, não pôde ser confirmada em nossos dados, uma vez que, em nenhum dos perfis de deslocamento, a maior parte dos falantes se encaixou na mesma categoria. Os falantes em nossa amostra não tendem a se agrupar em conjuntos coerentes.

Similar a Guy (2013) e Oushiro (2015a; 2015b), também fizemos uma classificação binária para nossos dados, na qual adotamos A para taxas iguais ou superiores a 50%, e B para taxas inferiores a 50%. Ao atribuir A ou B, há dezesseis (16) padrões possíveis nos quais um falante pode ser classificado ($2^4 = 16$): AAAA, AAAB, AABB etc. Considerando a aleatoriedade, teríamos uma expectativa de distribuição equilibrada do número de falantes por padrão – uma média de 11,31 (181 falantes / 16 padrões). A Tabela 11 apresenta os dez (10) dos quatorze (14) padrões obtidos. A ordem dos índices (A e B) segue ART, pro2PS, cli2PS e pos2PS.

Tabela 11 – Padrões de agrupamento social das variantes descritas – classificação binária

Padrão*	Falantes	Frequência
AAAA	43	23,75%
AABA	36	19,88%
BABA	30	16,57%
BAAA	18	9,94%
AABB	17	9,39%
BAAB	9	4,97%
AAAB	7	3,86%
ABBB	6	3,31%
ABAA	5	2,76%
BABB	5	2,76%

*B < 50%; A ≥ 50%

Fonte: elaboração própria.

Os dois padrões mais frequentes, correspondendo a 43,64%, implicam no alto uso (≥ 50%) de ausência de artigo, pronome *você* e possessivo *seu*. O que destoa entre o padrão mais frequente (AABA) e o segundo mais frequente (AAAA) é a frequência de *te*. Considerando que as três formas se enquadram no padrão atual para a 2PS, era esperada uma relação entre as formas *você, te* e *seu*. A existência de muitos falantes que não fazem uso de *te* ou de nenhum clítico de 2PS certamente interfere nessa distribuição.

O padrão mais frequente, AAAA, composto por quarenta e três (43) falantes, organiza-se com os Deslocamentos 1 (N= 15), Deslocamento 2 (N= 8), Deslocamento 3 (N= 7), Deslocamento 4 (N= 6), Alagoas (N= 3) e Bahia (N= 4). O segundo padrão mais frequente, AABA, formado por trinta e sete (37) falantes, organiza-se com os Deslocamentos 1 (N= 8), Deslocamento 2 (N= 10), Deslocamento 3 (N= 11), Deslocamento 4 (N= 6) e Alagoas (N= 1), sem a presença de falantes do deslocamento Bahia.

No terceiro padrão mais frequente, BABA, trinta (30) falantes fazem baixo uso (< 50%) da ausência de artigo e do clítico *te*, mas alto uso de *você* e *seu*. O padrão organiza-se com os Deslocamentos 1 (N= 5), Deslocamento 2 (N= 8), Deslocamento 3 (N= 3), Deslocamento 4 (N= 8), Alagoas (N= 2) e Bahia (N= 4). O padrão BAAA, o quarto mais frequente, apresenta dezoito (18) falantes que fazem baixo uso da ausência de artigo, mas baixo uso de *você, te* e *seu*, formas de 2PS, e organiza-se com os Deslocamentos 1 (N= 8),

Deslocamento 2 (N= 2), Deslocamento 3 (N= 1), Deslocamento 4 (N= 2), Alagoas (N= 1) e Bahia (N= 4). No quinto padrão mais frequente, AABB, dezessete (17) falantes fazem alto uso da ausência de artigo e de *você*, mas baixo uso de *te* e *seu*. São organizados nos Deslocamentos 1 (N= 5), Deslocamento 2 (N= 3), Deslocamento 3 (N= 4), Deslocamento 4 (N= 1), Alagoas (N= 4), sem falantes do deslocamento Bahia.

A variante *você* apresenta o comportamento mais estável entre os cinco padrões mais frequentes – também entre os dez (10) –, com altas taxas de uso. O pronome *seu* tende a acompanhar a tendência de *você* em cinco entre os dez padrões. A ausência de artigo e o clítico *te* são os fenômenos que apresentam a classificação mais variável dentre os dez padrões mais frequentes.

A Tabela 12 apresenta os padrões obtidos por perfil de deslocamento com base na classificação binária.

Tabela 12 – Padrões mais frequentes por perfil de deslocamento - classificação binária – e o percentual que o padrão mais frequente representa para o deslocamento (%)

Padrão*	**D1**	**D2**	**D3**	**D4**	**BA**	**AL**
AAAA	15	8	7	6	4	3
AABA	8	10	11	6	0	1
BABA	5	8	3	8	4	2
BAAA	8	2	1	2	4	1
AABB	5	3	4	1	0	4
BAAB	3	2	1	3	0	0
AAAB	1	2	1	3	0	0
ABBB	2	1	1	1	0	1
ABAA	2	2	1	0	0	0
BABB	2	1	2	0	0	0
%	26,7%%	25,6%	31,4%	26,6%	33,3%	33,3%

*B < 50%; A ≥ 50%

Fonte: elaboração própria.

Falantes do Deslocamento 1 se agrupam em doze (12) padrões, sendo o mais frequente o padrão AAAA (N= 15), com altas taxas (≥ 50%) de todas as variantes. No Deslocamento 2,

no qual há dez (10) padrões, o com maior quantitativo de falantes é o AABA (N= 10), com altas taxas de uso da ausência de artigo, *você* e *seu*, mas baixa taxa (< 50%) de *te*. Falantes do Deslocamento 3 seguem tendência similar à do Deslocamento 2: ainda que haja treze (13) padrões, falantes se agrupam mais no padrão AABA (N= 11).

No Deslocamento 4, observamos oito (8) padrões, sendo o com maior quantitativo de falantes o padrão BABA (N= 8), no qual falantes fazem baixo uso da ausência de artigo e de *te*, mas alto uso de *você* e de *seu*. Falantes da Bahia só se agrupam em três (3) padrões, todos com quatro (4) falantes cada, com o alto uso de *você* e *seu* como constante. Com falantes de Alagoas, observamos seis (6) padrões, sendo o mais frequente o padrão AABB (N= 4), com alto uso da ausência de artigo e de *você*, mas baixo uso de *te* e de *lhe*.

Em todos os perfis de deslocamento, há uma constante para o padrão mais frequente: os falantes sempre fazem alto uso de *você*. Além disso, em cinco (5) dos seis (6) perfis de Deslocamento, falantes se agrupam mais em padrões nos quais se faz alto uso da ausência de artigo ou alto uso de *seu*. O uso do clítico é o que mais varia. Contudo, similar à classificação ternária, uma vez que a maior parte dos falantes nos perfis de deslocamento não se encaixa no mesmo padrão, não podemos pontuar que há uma coesão dialetal, refutando nossa hipótese.

A descrição por vias de agrupamento social, similar aos resultados Guy (2013) e Oushiro (2015a; 2015b; 2016a), não nos permite observar uma coesão dialetal entre os falantes que compõem nossa amostra com base nos quatro fenômenos morfossintáticos. Além disso, em relação ao problema de nossa tese, compreendemos que, uma vez que os falantes não se agrupam em padrões coerentes de uso por perfil, não é possível, através dos dados visualizados, identificar a origem dialetal de nossos falantes.

Resta-nos, ainda, considerar uma terceira técnica de análise para a covariação: a análise de *cluster*.

6.1.3 Análise de *cluster*

De modo a observar padrões conjuntos de uso dos indivíduos em relação aos quatro fenômenos morfossintáticos variáveis e à sua origem dialetal, conduzimos, conforme feito por Freitag (2022), análises de *cluster* usando a técnica de k-medoids, por meio da função `pam` (*partitioning around medoids*, particionamento/agrupamento em torno de medóides) do pacote `cluster` (Maechler *et al*., 2023) para agrupar o conjunto de dados obtidos por meio das taxas médias de uso inidivual das quatro variáveis.

Uma vez que as análises de Horvath e Sankoff (1987) e Beaman (2021), por meio da técnica de PCA – que também utiliza, similar à análise de *cluster*, a redução de massa de dados, com perda mínima possível da informação –, permitiram a observação de variedades linguísticas distintas, questionamos se o agrupamento de falantes com base em seu comportamento linguístico, por meio da técnica de *clustering*, possibilita a identificação da região dialetal dos falantes. Nossa hipótese é a de que teremos a identificação da região dialetal dos falantes se a maior parte dos falantes (≥ 75%) do mesmo perfil de deslocamento se agruparem no mesmo *cluster*, isto é, esperamos que os agrupamentos gerados pela técnica de análise de agrupamento (*clustering*) dos dados reflitam à origem dialetal do falante, com falantes pertencentes à mesma região dialetal tendendo a serem agrupados de forma relativamente homogênea.

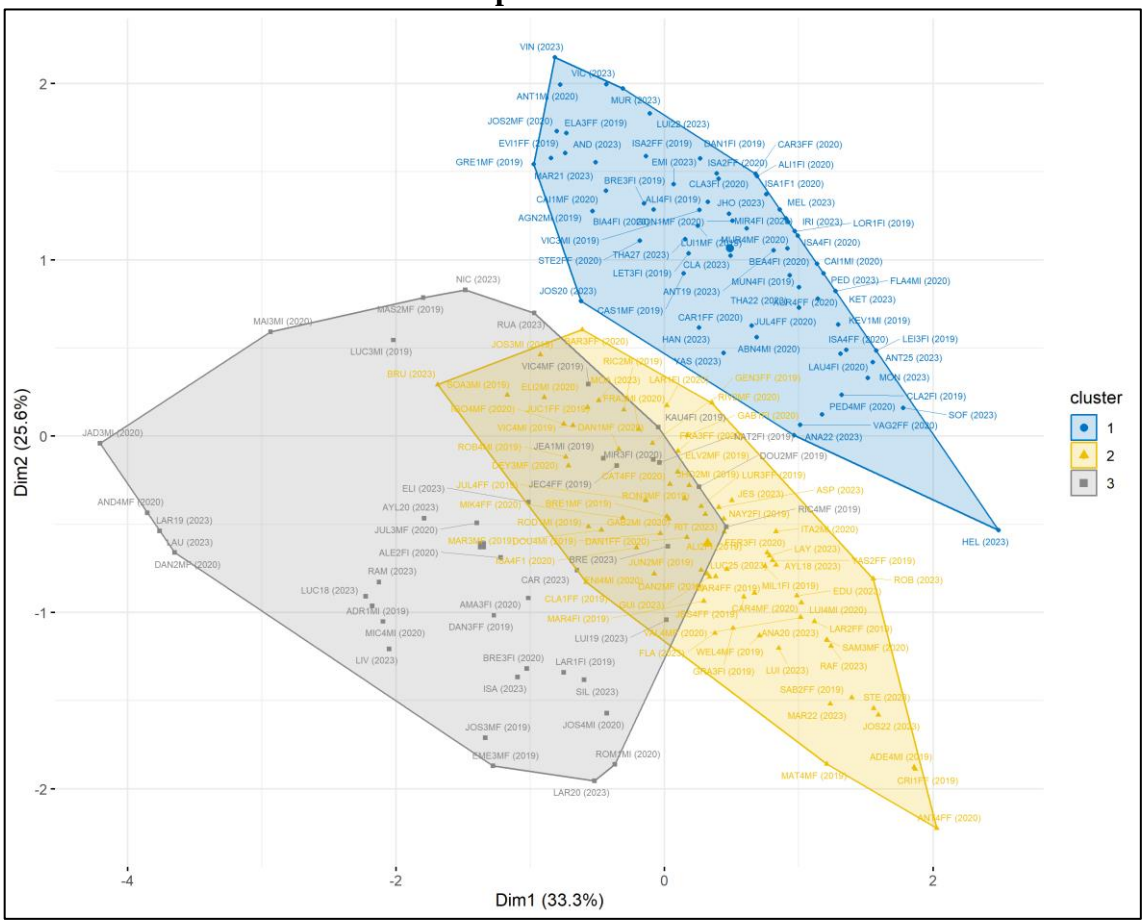
Para a aplicação da técninca de análise de *cluster*, utilizamos dois tipos de dados. O primeiro tipo não considera informações sociais dos falantes (deslocamento, tempo no curso, gênero e idade) para o cálculo utilizado na análise de agrupamento, os quais chamamos de Dados Básicos (DB). O segundo tipo integra essas informações, de modo a observar se há efeitos sociais sobre o agrupamento dos dados, o qual chamamos de Dados Sociais (DS). Em ambos os casos, os dados foram dimensionados por meio da função base do R `scale`, que calcula a média e o desvio padrão de todo o vetor, depois "dimensiona" cada elemento por esses valores subtraindo a média e dividindo pelo desvio padrão.

Após a inspeção dos dados e seu dimensionamento, utilizamos a função `fviz_nbclust` do pacote `factoextra` (Kassambara; Mundt, 2020) para observar o número ideal de *clusters* em ambos os conjunto de dados. A utilização da função apontou um número ideal de três (3) *clusters* para ambos os conjuntos de dados. Em um contexto ideal, uma vez que possuímos seis (6) perfis de deslocamentos (Deslocamento 1, Deslocamento 2, Deslocamento 3, Deslicamento 4, Bahia e Alagoas), poderíamos esperar a obtenção de um quantitativo que representasse o número de perfis, de modo com que houvesse um comportamento próprio para cada grupo. Contudo, dado que regiões dialetais distintas podem apresentar comportamento similar, que falantes de regiões próximas podem se comportar linguisticamente de forma semelhante, que falantes individualmente apresentam comportamentos específicos e que o contato linguístico pode resultar em mudanças ao nível do indivíduo, a obtenção de apenas três grupos não nos é surpresa.

A Figura 64 apresenta os agrupamentos obtidos por meio do conjunto DB. Nela, diferentes cores representam grupos diferentes: os pontos são coloridos de acordo com o grupo ao qual pertencem. Pontos representam observações: cada ponto na figura representa

uma observação em seus dados (um falante). Se os pontos estão agrupados em áreas específicas, isso sugere que as observações nesses grupos são mais similares entre si do que com observações de outros grupos. Os medóides são representados por formas geométircas (triângulo, círculo e quadrado) maiores que as demais na figura. As observações em torno dessas formas são os pontos mais representativos (medóides) de cada grupo. A distância entre as formas dos medóides pode ser interpretada como a dissimilaridade média entre os grupos. Quanto maior a distância, mais separados estão os grupos. Pontos que estão isolados ou longe dos *clusters* principais podem ser considerados *outliers*.

Figura 64 – Análise de *cluster* do conjunto DB com base na taxa de uso das variáveis por falante



Fonte: elaboração própria.

Os termos “Dim1” e “Dim2” na figura se referem às duas primeiras dimensões principais após a realização de uma análise de componentes principais (PCA). Dim1 representa a primeira dimensão principal após a redução de dimensionalidade. A primeira dimensão principal é aquela que captura a maior variabilidade nos dados. Dim2 representa a segunda dimensão principal após a redução de dimensionalidade. Dim2 é ortogonal à Dim1

e captura a maior variabilidade restante nos dados que não foi explicada pela Dim1. Na análise com DB, a Dim1 representa 33,3% da variabilidade, enquanto Dim2 representa 25,6%. Juntas, explicam 58,9% da variabilidade no conjunto dos dados.

O Grupo 2 é o maior dos *clusters*: setenta e quatro (74) falantes compõem esse grupo, agrupados por apresentarem a menor mediana de uso da ausência de artigo antes de possessivo (*Md*= 52,3), a maior mediana de uso do pronome *você* em posição de sujeito (*Md*= 85), mediana do clítico *te* baixa (*Md*= 0) e mediana de uso do possessivo *seu* alta (*Md*= 100). Dentro desse grupo, não há uma clara divisão quanto às características sociais dos falantes. Considerando deslocamento, há dezessete (17) falantes do Deslocamento 1, dezoito (18) do Deslocamento 2, dezesseis (16) do Deslocamento 3, quatorze (14) do Deslocamento 4, cinco (5) de Alagoas e quatro (4) da Bahia. Para tempo no curso, há trinta e seis (36) no início do curso e trinta e oito (38) do final. Gênero se divide em trinta e sete (37) falantes do gênero feminino (27), trinta e cinco (35) do masculino e dois (2) do gênero não-binário. A média de idade é de 20,98 anos.

O grupo apresenta uma distribuição relativamente homogênea em relação ao deslocamento, com falantes de todos os deslocamentos presentes, indício de que a variável deslocamento não interfere na formação do *cluster*. O mesmo pode ser dito em relação ao tempo no curso e ao gênero, dada a distribuição similar para os grupos.

O Grupo 1 é o segundo maior grupo, com sessenta e seis (66) falantes que apresentam a maior mediana de ausência de artigo antes de possessivo (*Md*= 57,3), segunda maior mediana de uso de *você* (*Md*= 83,1) e altas medianas de uso de *te* (*Md*= 100) e *seu* (*Md*= 100). A frequência no uso de *te* é o que mais diferencia este grupo do anterior. Compõem esse grupo vinte e cinco (25) falantes do Deslocamento 1, doze (12) do Deslocamento 2, nove (9) do Deslocamento 3, oito (8) do Deslocamento 4, quatro (4) de Alagoas e oito (8) da Bahia, divididos, quanto ao tempo, há vinte e oito (28) do início do curso e trinta e cinco (35) do final. Desses sessenta e seis falantes, quarenta e um (41) são do gênero feminino e vinte e cinco (25) do masculino, cuja média de idade é 21,25 anos, mais velhos comparados ao Grupo 2.

Similar ao Grupo 2, o Grupo 1 é heterogêneo, indicando que a variável deslocamento não é um fator determinante para as características linguísticas observadas. No entanto, o alto quantitativo de falantes do Deslocamento 1 e da Bahia pode indicar que a ausência de artigo antes de possessivo e o alto uso de *você*, *te* e *seu* são características desses falantes.

A divisão próxima entre falantes do início e do final do curso indica que o tempo no curso não é fator determinante para as características do grupo quanto aos usos linguísticos. A diferença quanto ao gênero, contudo, pode sugerir que falantes do gênero feminino

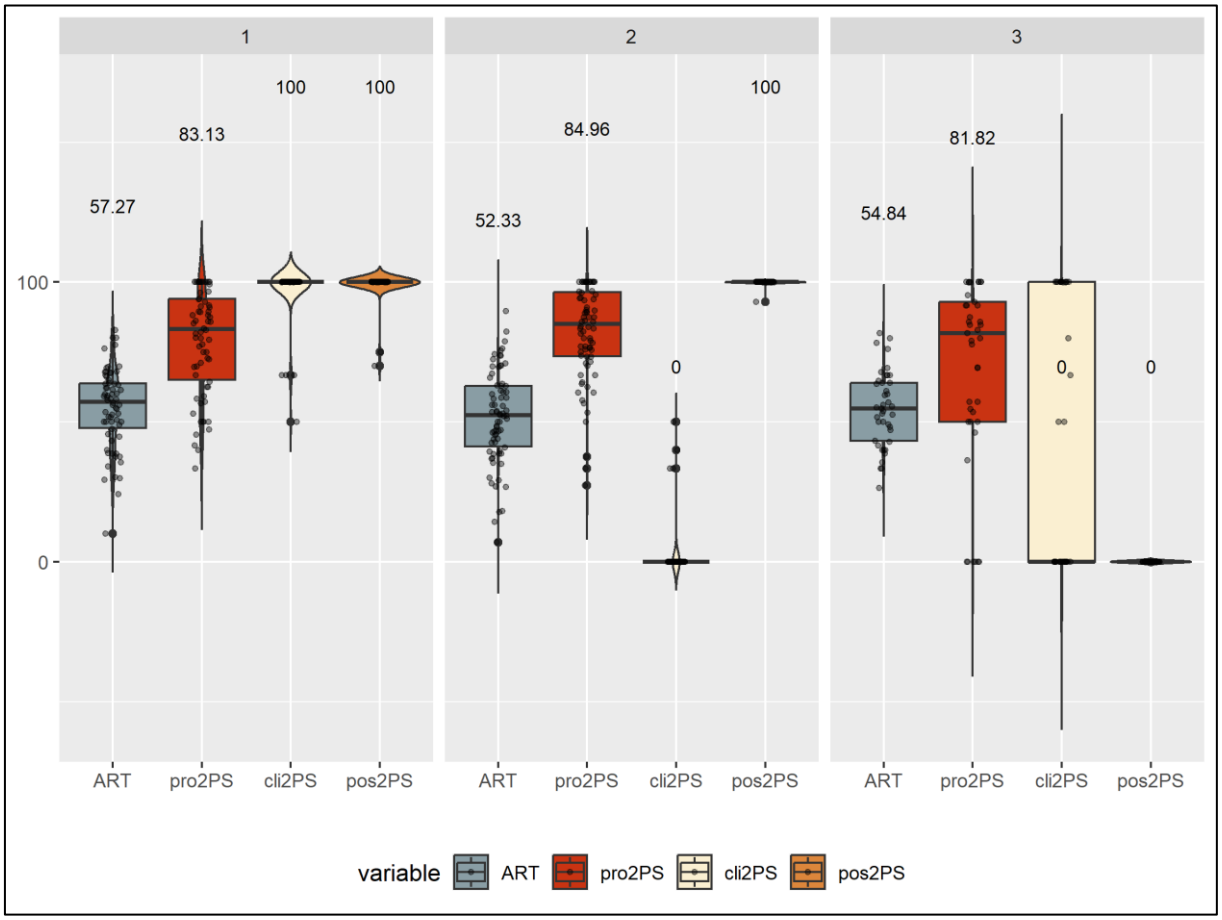
apresentam uma maior frequência das características linguísticas que definem o grupo, especialmente quando ao pronome pessoal de 2PS, no qual falantes do gênero feminino fazem maior uso de *você*, conforme discutido na análise univariada.

O Grupo 3 é o menor, com quarenta e um (41) falantes. O que une esses falantes são a segunda maior mediana de uso de artigo antes de possessivo (*Md*= 54,8), menor mediana de uso de *você* (*Md*= 81,8) e mediana de uso de *te* e *seu* de 0. O grupo é composto por onze (11) falantes do Deslocamento 1, nove (9) do Deslocamento 2, dez (10) do Deslocamento 3, oito (8) do Deslocamento 4 e três (3) de Alagoas, sem falantes da Bahia. Desses falantes, vinte e seis (26) são do início do curso e quinze (15) do final; vinte (20) são do gênero feminino e vinte e um (21) do masculino, cuja idade média é 20,63 anos, a menor entre os três grupos.

A ausência de falantes da Bahia no Grupo 3 sugere que o uso de *você* e a omissão de *te* e *seu* são mais frequentes em outros perfis de deslocamento. A composição equilibrada de gêneros reforça a ideia de que essas características linguísticas não são influenciadas por questões de gênero.

A Figura 65 apresenta as distribuições considerando as taxas de uso dos três grupos, enquanto o Quadro 13 apresenta os resultados da análise de *cluster* a partir do conjunto DB.

Figura 65 – Distribuição das taxas de uso das variáveis morfossintáticas no conjunto DB



Fonte: elaboração própria.

Quadro 13 – Síntese da análise de *cluster* com o conjunto DB

Característica	**Grupo 1**	**Grupo 2**	**Grupo 3**
Quantidade	66 falantes	74 falantes (maior)	41 falantes (menor)
Ausência de Artigo (*Md*)	57,3 (maior)	52,3 (menor)	54,8
Pronome *você* (*Md*)	83,1	85 (maior)	81,8 (menor)
Clítico *te* (*Md*)	100	0	0
Possessivo *seu* (*Md*)	100	100	0
Deslocamento	Deslocamento 1 e Bahia (predominante)	Distribuição igual	Deslocamento Bahia (ausente)
Tempo no curso	Distribuição igual	Distribuição igual	Mais falantes do início
Gênero	Feminino (predominante)	Masculino (predominante)	Distribuição igual
Idade (*M*)	21,25 anos (maior)	20,98 anos	20,63 anos (menor)

Fonte: elaboração própria.

Os resultados obtidos por meio da técnica de *cluster* com o conjunto de Dados Básicos não nos permite confirmar nossa hipótese, isto é, nos resultados obtidos, a maior parte dos falantes do mesmo perfil de deslocamento não se agrupa no mesmo *cluster*, não apresentando comportamento relativamente homogêneo para as taxas de uso das variantes. Os falantes do deslocamento Bahia seriam uma possível exceção, dado que oito (8) se agrupam no *cluster* 1, mas representam apenas 66,6% do grupo (8/12).

A técnica de *cluster* apresentada, contudo, foi aplicada sem a inserção das características sociais dos falantes. Como pontuamos na análise univariada, fatores extralinguísticos, como deslocamento, tempo no curso, gênero e idade, tendem a interferir nos usos linguísticos das variáveis linguísticas. A Figura 66 apresenta a aplicação da técnica de *cluster* considerando atributos sociais dos falantes controlados nas amostras Deslocamentos, com base no conjunto de Dados Sociais (DS).

Figura 66 – Análise de *cluster* do conjunto DS com base na taxa de uso das variáveis por falante

Fonte: elaboração própria.

Na análise com o conjunto DS, visualizamos uma diminuição nos valores de Dim1 e Dim2. Com DS, a Dim1 representa 21,3% da variabilidade, enquanto Dim2 representa 18,1%. Juntos, explicam 39,4% da variabilidade. Em comparação com a análise com DB, a análise sem informações sociais do falante captura melhor a maior variabilidade nos dados.

O Grupo 2 é o maior dos *clusters*, composto por noventa (90) falantes que apresentam a maior mediana da ausência de artigo (*Md*= 56,1), menor mediana de uso de *você* (*Md*= 76.9), baixa mediana de *te* (*Md*= 16,7) e alta mediana de uso de *seu* (*Md*= 100). Dentro desse grupo, vinte e quatro (24) falantes são do Deslocamento 1, vinte e dois (22) do Deslocamento 2, dezoito (18) do Deslocamento 3, treze (13) do Deslocamento 4, sete (7) de Alagoas e seis (6) da Bahia, divididos entre setenta e sete (77) do início do curso e treze (13) do final, com distribuição de gênero representado por cinquenta e seis (56) do gênero feminino, trinta e dois (32) do masculino e dois (2) do não-binário. A média de idade é de 18,81 anos.

O perfil do deslocamento, neste grupo, não é um fator que interfere nos usos dos grupos, dada à distribuição entre os deslocamentos, mas observemos o alto quantitativo de falantes alagoanos (7/12). O tempo no curso, por outro lado, parece interferir nos usos dos falantes, uma vez que o maior quantitativo no grupo é de falantes que estão ao início do curso. Nesse sentido, falantes ao início do curso compartilham um comportamento linguístico que os enquadre no Grupo 2. Gênero, por sua vez, não apresenta uma diferença dado que há forte predomínio de falantes do gênero feminino. A média de idade, comparada com os demais grupos, é a menor.

O Grupo 3 é o segundo maior grupo, com cinquenta e quatro (54) falantes. Esses falantes fazem o segundo maior uso da ausência de artigo (*Md*= 53), segundo maior uso de *você* (*Md*= 82,8), baixo uso de *te* (*Md*= 0) e alto uso de *seu* (*Md*= 100). O grupo é composto por dez (10) falantes do Deslocamento 1, onze (11) do Deslocamento 2, dezessete (17) do Deslocamento 3, quinze (15) do Deslocamento 4 e um (1) da Bahia, sem falantes alagoanos, que se dividem entre dez (10) do início do curso e quarenta e quatro (44) do final. Gênero apresenta uma distribuição homogênea, sendo vinte e seis (26) falantes do gênero feminino e vinte e oito (28) do masculino. A média de idade é a segunda maior entre os três grupos (*M*= 22,09).

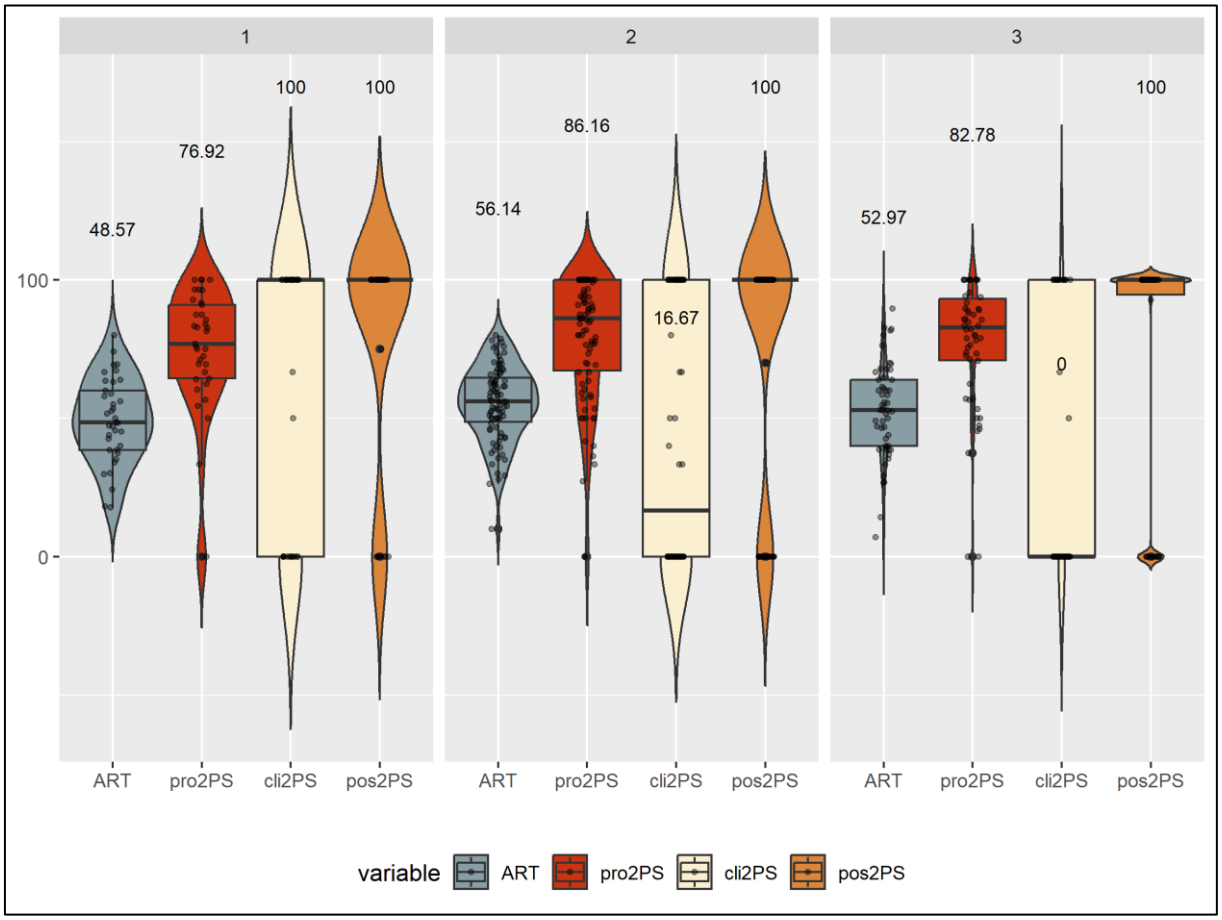
Mais uma vez, vemos uma distribuição homogênea em relação aos perfis de deslocamento, mas há ausência de falantes alagoanos, presentes fortemente no Grupo 2 (8/12). O predomínio de falantes do final do curso (44), com apenas dez do início, pode indicar uma relação entre tempo no curso e formação do grupo, oposto ao observado no Grupo 1.

Por fim, o Grupo 1 é o menor deles, contendo trinta e sete (37) falantes, cujas frequências apresentam as seguintes medianas: ausência de artigo (*Md*= 48,6), pronome *você*

(*Md*= 76,9), clítico *te* (*Md*= 100) e possessivo *seu* (*Md*= 100). Desses falantes, dezenove (19) são do Deslocamento 1, seis (6) do Deslocamento 2, dois (2) do Deslocamento 4, cinco (5) de Alagoas e cinco (5) da Bahia, divididos em seis (6) do início do curso e trinta e um (31) do final, sendo dezesseis (16) do gênero feminino e vinte e um (21) do gênero masculino. A média de idade é a maior entre os três grupos (*M*= 24,75).

Este grupo apresenta a menor frequência de ausência de artigo e do pronome *você* entre os três grupos, maior frequência do clítico *te*, e alta frequência do pronome possessivo *seu*, igual aos Grupos 1 e 2. A predominância de falantes do final do curso (31) em relação ao início (6) pode indicar que o tempo de exposição no *campus* está associado a uma maior adoção das variantes observadas, e a média de idade mais alta (*M*= 24,75) em comparação aos outros grupos pode reforça a ideia de que esses falantes são, em sua maioria, falantes com maior exposição às variantes. A divisão quase equilibrada entre gêneros sugere que as características observadas não estão fortemente vinculadas a essa variável. A Figura 67 apresenta a distribuição das taxas por grupo, enquanto o Quadro 14 apresenta os resultados da análise com os Dados Sociais.

Figura 67 – Distribuição das taxas de uso das variáveis morfossintáticas no conjunto DS



Fonte: elaboração própria.

Quadro 14 – Síntese da análise de *cluster* com o conjunto DS

Característica	**Grupo 1**	**Grupo 2**	**Grupo 3**
Quantidade	37 falantes (menor)	90 falantes (maior)	54 falantes
Ausência de Artigo (*Md*)	48,6 (menor)	56,1 (maior)	53
Pronome *Você* (*Md*)	76,9 (menor)	86,2 (maior)	82,8
Clítico *te* (*Md*)	100 (maior)	16,7	0 (menor)
Possessivo *seu* (*Md*)	100	100	100
Deslocamento	Deslocamento 1 (predominante)	Distribuição igual	Deslocamento Alagoas (ausente)
Tempo no curso	Final (predominante)	Início (predominante)	Final (predominante)
Gênero	Distribuição igual	feminino (predominante)	Distribuição igual
Idade (*M*)	24,75 anos (maior)	18,81 anos (menor)	22,09 anos

Fonte: elaboração própria.

Em relação à nossa hipótese para a técnica de *cluster*, semelhante ao visto com o conjunto DB, não é possível a sua confirmação, dado que a maior parte dos falantes do mesmo perfil de deslocamento não se agrupam no mesmo *cluster*.

Vale salientar que a ausência de falantes de determinado deslocamento em um grupo aponta para a divergência no seu comportamento linguístico. É o caso da ausência de baianos no Grupo 3 na análise DB. Isso implica que os falantes da Bahia podem apresentar padrões linguísticos que os diferenciam significativamente dos falantes do Grupo 3. Como resultado, eles foram agrupados em outros *clusters* (Grupo 1 e Grupo 2) que melhor representam suas características linguísticas. Nesse sentido, o Grupo 3 pode ser composto por falantes que compartilham características linguísticas muito específicas, que não são comuns entre os falantes da Bahia. Isso pode indicar que esses falantes têm usos linguísticos que os afastam desse grupo, por uma possível influência regional. De forma similar, falantes de Alagoas não estão presentes no Grupo 3 na análise DS. Contudo, dado que características sociais – deslocamento, tempo no curso, gênero e idade – são calculadas na análise, explicações pautadas somente nos usos linguísticos não são viáveis.

Quanto à performance de ambas as análises, cada uma delas apresenta sua contribuição para a descrição de padrões conjuntos de usos linguísticos. A análise DB possui uma distribuição mais equilibrada por *cluster*. Além disso, o agrupamento considera apenas as taxas de uso das variantes. Os falantes são agrupados puramente por seus usos linguísticos. A análise DS, por sua vez, considera informações extralinguísticas dos falantes, o que acaba interferindo no agrupamento dos falantes, mas insere a informação social, o que é importante, dado que a variação linguística é socialmente motivada.

A motivação para o agrupamento considerando dados sociais, contudo, não nos é clara, o que nos permite apenas tecer inferências. Por exemplo, poderíamos dizer que falantes do Grupo 2 tenham sido agrupados por conta de sua idade; como poderíamos dizer que o agrupamento seja resultado da diferenciação na mediana da ausência de artigo. Evidentemente, essas inferências não se seguem, dado que todos as variáveis inseridas na análise são computadas no cálculo de agrupamento. Assim, em vias de compreensão das motivações para agrupamento, a análise DB permite uma melhor observação das diferenças entre os grupos, dado que não mobiliza informações sociais.

Todavia, nenhuma das duas análises nos permitiu confirmar nossa hipótese. Sobre isso, apresentamos algumas possíveis motivações, similares às motivações levantadas na análise de correlação. A primeira delas é a de que nem todos os falantes fazem uso das variantes/variáveis, o que resulta em muitas taxas em 0%, como também em muitas taxas em

100%, decorrente do uso exclusivo de apenas uma forma. A não existência de variação ao nível do indivíduo, em 2/4 dos fenômenos, interfere diretamente na análise estatística. A segunda delas é que temos limitações em nossas amostras, dado o pouco quantitativo de indivíduos (N= 181). A utilização de mais informantes por perfil de deslocamento poderia contribuir para uma descrição mais satisfatória de usos linguísticos conjuntos por falantes da mesma região dialetal.

Nesse sentido, diferentemente da análise de Horvath e Sankoff (1987), cujos resultados sugerem a existência de pelo menos duas variedades de inglês falado em Sydney, que são relativamente distintas, e da análise de Beaman (2021), que apresenta a existência de dois letos distintos, nossos dados não permitem a observação de dialetos específicos, mas revelam indícios que apontam para a possibilidade de identificar padrões dialetais específicos, particularmente na distinção entre falantes da Bahia e não Bahia. A ausência de falantes baianos no Grupo 3 na análise DB é um indicativo de que há diferenças linguísticas significativas associadas à origem regional desses falantes. Isso quer dizer que a presença de falantes da Bahia nos Grupos 1 e 2, mas não no Grupo 3, reforça a ideia de que sua origem regional está associada a usos linguísticos específicos.

6.2 EM SÍNTESE

Este capítulo objetivou realizar análises de covariação entre quatro fenômenos morfossintáticos: i) uso variável de artigo antes de possessivo pré-nominal; ii) pronomes pessoais de 2PS; iii) pronomes clíticos de 2PS; e iv) pronomes possessivos de 2PS. Para tanto, empregamos diferentes técnicas estatísticas em ordem de observar a existência de covariação entre os quatro fenômenos selecionados, com vistas a responder se a análise de covariação entre variáveis morfossintáticas em grupos de falantes de diferentes regiões geográficas possibilita a identificação de sua origem dialetal. O Quadro 15 apresenta nossos principais achados para essas análises, em diálogo com as questões levantadas em cada análise.

Quadro 15 – Síntese das técnicas estatísticas para covariação

Técnica	**Evidenciou covariação?**	**Identifica a origem do falante?**
Análise de correlação	Sim, mas apenas com uma pequena parcela dos pares.	Não, dado que a maior parte dos pares de variáveis, na maior parte dos grupos, não se correlacionam significativamente.

Padrões de agrupamento	Sim, dado que os cinco padrões mais frequentes apresentam frequência alta de *você* e *seu*.	Não, dado que a maior parte dos falantes que compõem o mesmo perfil de deslocamento não se encaixou no mesmo padrão.
Análise de *cluster*	Sim, uma vez que o agrupamento dos falantes em grupos específicos reflete semelhanças em seus usos linguísticos.	Em partes, dados os indícios de origem dialetal para falantes baianos, cujo comportamento linguístico na análise DB apresentou usos específicos.

Fonte: elaboração própria.

Tais resultados são, possivelmente, efeito de alguns problemas linguísticos e extralinguísticos. De ordem linguística, podemos citar a frequência de ocorrência de variáveis morfossintáticas. Como discutem Milroy e Gordon (2003, p. 172),

> uma vez que os falantes fazem uso de um inventário nitidamente limitado e, portanto, frequentemente recorrente de contrastes fonológicos, as realizações de qualquer variável tendem a aparecer com bastante frequência, mesmo em uma amostra curta de fala. Este não é o caso para variáveis morfológicas e, mais especialmente, de variáveis sintáticas mais altas, uma vez que uma quantidade suficiente de *tokens* de um determinado tipo de construção geralmente não pode ser garantida para aparecer em um trecho de discurso espontâneo.[27]

Fenômenos morfossintáticos, comparados a fenômenos fonológicos, tendem a ter baixa frequência no discurso espontâneo. Adicionalmente, fenômenos morfossintáticos específicos tendem a ter frequência ainda menor a depender do tipo de discurso dos quais os dados são extraídos, como nossas entrevistas, que não lidam com interações diretas entre dois falantes, mas apenas diálogo documentador e informante. O resultado é o que vimos com os clíticos de 2PS e os possessivos de 2PS, cujas frequências, comparadas às frequências do uso variável de artigo antes de possessivos e de pronomes pessoais de 2PS, são relativamente baixas.

Isso, evidentemente, afetou as análises estatísticas. Testes de correlação, por exemplo, lidam com tabelas de contingência que necessitam possuir o mesmo quantitativo (n x n). Nesse sentido, a não existência de dados de *te* não poderia ser diretamente relacionada com a existência da ausência de artigo. Para contornar isso, foi necessária a inserção de zeros (0), o que interferiu na distribuição dos dados.

[27] No original: "Since speakers make use of a sharply limited and therefore frequently recurring inventory of phonological contrasts, realizations of any given variable are likely to show up quite frequently in even a short sample of speech. This is not the case for morphological and more especially higher level syntactic variables, since a sufficient quantity of tokens of a given type of construction cannot usually be guaranteed to appear in a piece of spontaneous discourse".

De ordem extralinguística, os resultados também evidenciam limitações da amostra, como a existência de poucos indivíduos, especialmente em perfis de deslocamento específicos, como falantes de Alagoas e da Bahia, que possuem apenas doze (12) falantes cada.

Quanto à técnica aplicada, no Quadro 16 listamos os prós e os contras de cada técnica observados nesta pesquisa.

Quadro 16 – Avaliando as técnicas empregadas

Técnica	**Prós**	**Contras**
Análise de correlação	1. Os coeficientes de correlação fornecem material para a observação da força e da direção da relação entre duas variáveis linguísticas; 2. Havendo relação significativa, há evidencia de covariação para as formas, o que auxilia na identificação da região dialetal dos grupos de falantes; 3. Diferentes tipos de dados numéricos podem ser utilizados, a depender das escolhas metodológicas do pesquisador.	1. Ainda que seja apresentado o coeficiente de correlação, depende de significância para atestar relação; 2. Amostras pequenas resultam em erros no cálculo do coeficiente de correlação e no valor de p; 3. Não lida com células vazias. Fenômenos pouco frequentes geram erros na análise; 4. Não possibilita uma observação detalhada do papel do indivíduo.
Padrões de agrupamento	1. Possibilita a observação de padrões gerais de uso, tornando valores numéricos em valores categóricos; 2. Permite a observação da coesão dialetal para um grupo se houver maior concentração de falantes em um padrão.	1. Não comporta gradientes, uma vez que a classificação lida com valores discretos; 2. Diferentes tipos de dados resultam em resultados distintos para o agrupamento; 3. Pesquisadores diferentes podem atribuir classificações distintas; 4. Os intervalos das classificações são muito grandes.
Análise de *cluster*	1. Possibilita a identificação de agrupamentos não lineares de dados, o que é útil quando as relações entre as variáveis não são lineares; 2. Grupos são formados a partir da similaridade existente entre cada observação (cada falante), o que permite a consideração de que há um compartilhamento no uso linguístico dos falantes que compõem cada grupo; 3. Permite a inserção de informações sociais dos falantes;	1. A análise de *cluster* com k-medoids requer a escolha do número de *clusters*, que pode ser subjetiva e influenciar significativamente os resultados da análise, dado que métodos para determinar o número ideal de *clusters* nem sempre são conclusivos; 2. É necessária a utilização de amostra robusta, dado que poucas observações podem resultar em erros no agrupamento; 3. Necessita de um conhecimento mais aprofundado

	4. O falante toma maior espaço, podendo ser observado individualmente.	em estatística para interpretação.

Fonte: elaboração própria.

Com o teste de correlação de Spearman, pudemos observar a relação entre as quatro variáveis morfossintáticas e identificar se haveria uma associação significativa entre elas. Com a análise de padrões de agrupamento social, pudemos descrever se falantes se organizam em padrões relativamente homogêneos. A análise de *cluster* por k-medoids permitiu o grupamento de falantes com base nas características morfossintáticas, identificando padrões distintos de uso linguístico entre diferentes grupos de falantes. Nesse sentido, as técnicas isoladamente apresentam suas contribuições para a análise de covariação na língua.

Advogamos, contudo, para a maior adequação da técnica de *cluster* com k-medoids, a qual argumentamos que apresenta um melhor resultado para o nosso objetivo, dado que permite agrupar falantes de forma mais precisa com base nas semelhanças em seus traços morfossintáticos ao capturar padrões distintos de uso linguístico e facilitar a identificação de origens dialetais de maneira mais eficaz e intuitiva (ainda que isto não tenha sido possível), uma vez que fornece a delimitação de grupos, possibilitando observar onde cada falante se insere. Isso possibilitou a observação de uma separação dos falantes baianos em grupos específicos, o que aponta indícios para uma influência dialetal.

De forma mais específica, a utilização de k-medoids possibilita a inserção de *outliers* e se baseia na minimização das distâncias médias em relação aos medóides dos *clusters*, o que resulta em um agrupamento mais representativo e confiável. Assim, ao contrário de análises de correlação, que medem a força e direção das relações entre variáveis, a técnica de *cluster* separa os falantes em grupos distintos com base nas semelhanças em seus traços linguísticos, o que facilita a visualização e interpretação de padrões de uso linguístico que podem estar associados à região dialetal dos falantes. É de se considerar, contudo, que os objetivos de ambas as análises são diferentes, conforme discussões no Capítulo 3.

Além disso, a técnica de *cluster* também permite explorar interações complexas entre múltiplas variáveis morfossintáticas associadas a características sociais dos falantes, algo que seria difícil de capturar apenas com correlações ou descrições como padrões de agrupamento social. Assim, sua capacidade de sintetizar grandes volumes de dados linguísticos torna a técnica mais adequada para a identificação da origem dialetal de falantes.

É evidente, contudo, que combinar essas técnicas fornece resultados mais satisfatórios para a condução de descrição de covariação, possibilitando descrições mais aprofundadas

sobre usos linguísticos conjuntos, dado que considera diferentes parâmetros de análise. Em nossa pesquisa, a combinação possibilitou obter uma compreensão mais completa da covariação entre as variáveis morfossintáticas e sua relação com as diferenças dialetais existentes no vernáculo dos falantes.

7. CONSIDERAÇÕES FINAIS

A descrição de traços linguísticos conjuntos para a observação de variedades dialetais da língua é um passo importante para a compreensão de como grupos de falantes se comportam linguisticamente e se há uma coesão nesse comportamento. Ao longo deste trabalho, discutimos a necessidade de se considerar mais de uma variável linguística para a compreensão de um dialeto e, mais precisamente, a descrição de traços morfossintáticos, que também podem caracterizar variedades dialetais de uma língua.

Isso resultou no cerne de nossa pesquisa, na qual questionamos se a descrição da covariação morfossintática em grupos de falantes de diferentes regiões geográficas identifica sua origem dialetal. Investigamos a hipótese, diante da questão, de que a identificação da origem dialetal do falante a partir de variáveis morfossintáticas pode ser realizada utilizando técnicas de descrição de covariação, dado que a descrição de padrões de variação na morfossintaxe por meio da fala de indivíduos de diferentes origens possibilita a identificação de características distintivas de determinadas regiões geográficas.

Com vistas a responder à pergunta de pesquisa, e confirmar ou refutar nossa hipótese, objetivamos, ao longo deste trabalho, descrever a covariação entre variáveis morfossintáticas geograficamente distintas (i)-(iv), buscando identificar padrões que permitam a caracterização das variedades dialetais do PB e a associação entre falantes e suas respectivas regiões de origem.

i) uso variável de artigo definido antes de possessivo pré-nominal;
ii) pronomes pessoais de segunda pessoa do singular (2PS) em posição de sujeito;
iii) pronomes clíticos de 2PS; e
iv) pronomes possessivos de 2PS;

Utilizamos, para tal, técnicas de covariação para identificar padrões que possam indicar a origem dialetal dos falantes. Também lançamos como objetivos específicos (i) descrever fenômenos morfossintáticos variáveis do PB com base em fatores sociais; (ii) investigar covariação entre quatro fenômenos morfossintáticos variáveis, de modo a sistematizar padrões de uso conjuntos; e (iii) identificar as diferenças e semelhanças nos usos de traços sociolinguísticos morfossintáticos por estudantes universitários da UFS.

Assim, no Capítulo 2, discutimos sobre a existência de significado dialetal na língua e, mais especificamente, na morfossintaxe, observando, por meio de revisão da literatura, o

comportamento dos fenômenos que selecionamos no português brasileiro (PB), o que nos permitiu reforçar a constatação de que são variáveis que se distinguem dialetalmente na língua, podendo, então, serem distintas em nossas amostras. O Capítulo 3, por sua vez, nos permitiu observar que, para compreender como variedades se comportam, é necessária a consideração de que formas linguísticas podem levar ao uso de outras formas, como também de que formas linguísticas ocorrem conjuntamente. Desse modo, o estudo de covariação contribui para a descrição de como uma variedade se comporta em termos de usos linguísticos conjuntos. Caminhamos, então, para nosso capítulo metodológico, no Capítulo 4, no qual apresentamos as amostras Deslocamentos (2019), Deslocamentos (2020) e Linguagem Corporificada (2023), as quais deram suporte para nossas análises, por meio da utilização de técnicas estatísticas univariadas e de covariação, e, dessa forma, podemos responder nossa questão de pesquisa e cumprir com nossos objetivos (geral e específicos).

Na descrição dos fenômenos morfossintáticos variáveis do PB, no Capítulo 5, para a distribuição geral dos fenômenos, observamos o predomínio, em todas as amostras, da ausência de artigo antes de possessivos, do pronome *você* em posição de sujeito de 2PS, do clítico de 2PS *te* e do possessivo de 2PS *seu*. Com base em fatores sociais, conforme o objetivo especifico de descrever os fenômenos correlacionando-os às variáveis sociais que as amostras abordam, pudemos obter alguns resultados:

1. **Indivíduo**: falantes individuais são variáveis na ausência/presença de artigo e no uso de *você/cê*, mas são categóricos com os clíticos e possessivos de 2PS;
2. **Deslocamento**: há associação com o deslocamento do falante no uso variável de artigo, nos pronomes pessoais e clíticos de 2PS, mas não há com possessivos de 2PS;
3. **Tempo**: há associação com o tempo no curso do falante no uso variável de artigo, nos pronomes pessoais e clíticos de 2PS, mas não há com possessivos de 2PS;
4. **Gênero**: há associação do gênero do falante com o uso variável de artigo e nos pronomes pessoais de 2PS, mas não há com clíticos e possessivos de 2PS;
5. **Idade**: há associação da idade do falante com o uso variável de artigo e nos pronomes pessoais, clíticos e possessivos de 2PS.

Os dados possibilitam observar que a variação morfossintática se correlaciona com variáveis sociais, na medida em que fatores controlados neste estudo mostraram associação com os usos linguísticos dos falantes. As informações relativas à variável deslocamento permitem a inferência de que há distinção dialetal para ao menos três dos quatro fenômenos, dada a

distribuição dos possessivos de 2PS, o que nos ajuda na resolução de nossa questão, uma vez que usos específicos das variáveis podem apontar para regiões dialetais dos falantes que compõem a amostra.

A partir disso, no Capítulo 6, pudemos passar para a investigação da possível covariação entre os quatro fenômenos morfossintáticos variáveis, de modo a sistematizar padrões de uso conjuntos. Para tanto, três técnicas de análise foram mobilizadas: correlação não-paramétrica, padrões de agrupamento social e análise de *cluster*. As técnicas possibilitaram a obtenção de padrões covariáveis.

1. **Correlação**: relação significativa nos pares ART~pro2PS, pro2PS~pos2PS, cli2PS~pos2PS e pro2PS~cli2PS, mas as correlações não apareceram todas no mesmo conjunto de dados.
2. **Padrão de agrupamento social**: padrões nos quais > 30% dos falantes do grupo fazem parte e compartilham de frequências similares;
3. **Análise de *cluster***: agrupamento natural dos falantes em três grupos, a partir de similaridades e dissimilaridades nos usos linguísticos.

As análises realizadas também permitiram a observação de diferenças e semelhanças nos usos de traços sociolinguísticos morfossintáticos por estudantes universitários da UFS, conforme sintetizamos no Quadro 11. Por exemplo, nas amostras Deslocamentos (2020) e Linguagem Corporificada (2023), falantes externos a Sergipe fazem maior uso da ausência de artigo, enquanto na amostra Deslocamentos (2020), o pronome *você* ocorre mais com falantes do Deslocamento 1, Bahia e 3. As sínteses nos Quadro 12, Quadro 13, Quadro 14, Tabela 10 e Tabela 12 são outros exemplos de similaridades e diferenças entre os usos linguísticos feitos pelos estudantes da UFS. Nas análises de *cluster*, é interessante observar a mudança no percentual de captura de variabilidade nos dados entre o conjunto Dados Básicos e Dados Sociais: dados não estruturados permitiram agrupamentos mais consistentes do que dados estruturados. Conforme defendem Freitag e Gois (2024), Modelos de Língua em Larga Escala podem ser eficientes para observar esses padrões.

A partir de todas as análises realizadas, pudemos cumprir nosso objetivo geral ao descrever o comportamento conjunto de quatro variáveis morfossintáticas geograficamente distintas do PB utilizando técnicas de covariação para identificar padrões que possam indicar a variedade dialetal dos falantes. E é a partir de todas as análises, em especial às três técnicas de análise de covariação, que podemos responder à questão de pesquisa lançada ao início da

pesquisa, isto é, se a descrição da covariação morfossintática em grupos de falantes de diferentes regiões geográficas identifica sua origem dialetal. A utilização de diferentes técnicas aponta que a descrição da covariação morfossintática em grupos de falantes de diferentes regiões geográficas pode apontar indícios de origem dialetal, mas não identificar, especialmente quando há padrões consistentes associados a uma região. É o que obtivemos com a análise de *cluster* no conjunto Dados Básicos, na qual os falantes da Bahia apresentam um comportamento linguístico específico, sendo agrupados nos Grupos 1 e 2, mas ausentes no Grupo 3. Essa distinção sugere que os baianos compartilham características morfossintáticas distintas. As outras técnicas, contudo, não possibilitaram esse tipo de observação.

Em termos práticos, isso implica que a identificação de padrões dialetais a partir da covariação morfossintática depende tanto da escolha das técnicas de análise quanto da natureza dos dados. A análise de *cluster*, no caso dos Dados Básicos, mostrou-se eficaz em revelar indícios de variação dialetal, ao segregar os falantes da Bahia em grupos específicos com base em suas características linguísticas. No entanto, as outras técnicas utilizadas não permitiram observações semelhantes, o que sugere que a detecção de padrões dialetais pode ser sensível a outras questões.

Por exemplo, a baixa produtividade de certos fenômenos variáveis aqui descritos, como é o caso dos clíticos e possessivos de 2PS. A presença baixa das formas, em especial dos clíticos, que não ocorrem em grande parte dos falantes, pode ter interferido nos resultados estatísticos. A construção de amostras que possibilitem um maior uso de formas de 2PS – interações, por exemplo – pode ser um elemento importante para a maior produtividade de formas de 2PS, em especial clíticos e possessivos. O trabalho de covariação com fenômenos mais produtivos também pode ser uma saída.

Além disso, é necessário reconhecer o tamanho da amostra. Embora superior, quanto ao quantitativo de falantes (N= 181), às pesquisas revisadas (Labov, 2006[1966]; Horvath; Sankoff, 1987; Guy, 2013; Oushiro, 2015a; 2016a; Tamminga, 2019; Beaman, 2021; Freitag, 2022), a análise não se sustenta em cálculos amostrais, tanto pelo baixo quantitativo, quanto pela existência de três recortes temporais (Freitag, 2018). Consideremos que a UFS, no semestre 2024.2, possuía 16.748 estudantes de graduação matriculados, público-alvo das amostras: ainda que 181 participantes possam ser considerados uma amostra razoável em muitos contextos, a amostra representa apenas uma pequena fração da população total, 1,08% dos falantes da universidade. Embora a análise tenha revelado correlações significativas e

agrupamentos, a impossibilidade de identificar a origem dialetal da maior parte dos grupos de falantes pode ser reflexo das limitações do tamanho amostral.

Para identificar padrões dialetais mais precisos e com maior robustez estatística, seria necessário um número maior de participantes que refletissem a diversidade linguística e geográfica da população total da UFS. Além disso, diferenças dialetais são oriundas de diferentes fatores: uma amostra de 181 falantes pode não ter sido grande o suficiente para capturar as nuances de todos esses fatores, especialmente quando se considera que a universidade pode englobar uma ampla gama de dialetos e variáveis linguísticas. Como argumentam Campbell-Kibler *et al.* (2014, p. 21, tradução nossa), "a universidade é uma experiência de realocação comum, embora não universal, para jovens adultos, muitas vezes envolvendo movimentos para novas regiões dialetais, ou interação com outros que vieram de outras regiões",[28] o que possibilita o contato e consequente mudança. Assim, para uma generalização mais confiável sobre a origem dialetal dos falantes, uma amostra maior seria necessária. Com isso, embora tenhamos feito uso de técnicas corriqueiras para identificar covariação, a amostra de 181 falantes não foi suficiente para fornecer uma representação completa e precisa da variação dialetal na UFS.

Também é importante considerar o papel do contato linguístico sobre os usos dos falantes que, efetivamente, pode ter interferido na produção das variáveis descritas. Conforme discorremos ao longo deste trabalho, lidamos com dados de universitários – grande parte migrante –, que estão constantemente em processo de interação com pessoas que podem ser falantes de outros dialetos. O contato, conforme defendido por Trudgill (1986), pode resultar em mudanças linguísticas. Assim, é possível que os falantes não mais falem de forma similar àqueles de sua comunidade de origem, mas similar ao padrão de seu novo ambiente, que é heterogêneo, o que não nos permitiu, a partir da descrição de seus usos linguísticos, presumir sua região de origem.

Os resultados deste trabalho apresentam algumas contribuições com a agenda de pesquisa em Sociolinguística Variacionista, em especial à descrição morfossintática e à covariação. A primeira contribuição é quanto ao reforço da visão de que variáveis morfossintáticas também carregam significado dialetal e podem ser estratificadas quanto à região dialetal dos grupos de falantes. Ainda que seja uma discussão antiga, pouco a pesquisa em Sociolinguística tem se detido para a observação desses aspectos na morfossintaxe da língua, frente ao privilégio a descrições fonético-fonológicas para a compreensão de dialetos. Com isso, os resultados do trabalho

[28] No original: "college is a common, though by no means universal, relocation experience for young adults, often involving moves to new dialect regions, or interaction with others who have come from other regions.

enfatizam a importância de se observar a variação morfossintática como nível importante na caracterização de dialetos das línguas, em espacial do PB.

Quanto à contribuição à agenda de descrição da covariação, os resultados apontam para a aplicabilidade das três técnicas no entendimento de uso conjunto de variáveis linguísticas, conforme já vêm sendo empregadas por Guy (2013), Oushiro (2015a; 2016a) e Freitag (2022). O emprego de técnicas de correlação, de padrões de agrupamento social e de análise de *cluster* possibilita a observação de como variáveis linguísticas se correlacionam nos usos individuais dos falantes, auxiliando na compreensão de como dialetos da língua são organizados, com especial foco a variáveis morfossintáticas. Assim, a pesquisa evidencia que essas técnicas de análise são ferramentas eficazes para mapear a interação entre variáveis linguísticas nos usos reais dos grupos de falantes. O estudo oferece uma metodologia para a descrição da covariação que pode ser replicado por pesquisas anteriores, dado que também disponibilizados os *scripts* no OSF e GitHub.

Por fim, consideramos que este estudo abre possibilidades de investigação para pesquisas futuras, especialmente no que tange à relação entre variáveis morfossintáticas e a identificação de dialetos no PB. Embora haja um potencial de estratificação dialetal para variáveis morfossintáticas, a abordagem da covariação não nos permitiu identificar a origem dialetal dos falantes. Isso não implica, necessariamente, no mesmo comportamento para todo o PB ou para outros fenômenos na morfossintaxe, o que torna imperativo novos estudos.

REFERÊNCIAS

ALMEIDA, G. S. *Quem Te Viu Quem Lhe Vê*: a expressão do objeto acusativo de referência à segunda pessoa na fala de Salvador. Dissertação (Mestrado em Língua e Cultura) – Universidade Federal da Bahia, Instituto de Letras, Salvador, 2009.

ALMEIDA, G. S. *Uso variável dos pronomes-objeto na expressão do dativo e do acusativo de segunda pessoa em Santo Antônio de Jesus–BA*. 2014. 256f. Tese (Doutorado em Língua e Cultura) – Universidade Federal da Bahia, Instituto de Letras, Salvador, 2014.

ALVES, C. C. B. *Pronomes de segunda pessoa no espaço maranhense*. 2015. 153f. Tese (Doutorado em Linguística) – Universidade de Brasília, Instituto de Letras, Departamento de Linguística, Português e Línguas Clássicas, Programa de Pós-Graduação em Linguística, 2015.

AMARAL, A. *O dialeto caipira*: gramática, vocabulário. Hucitec, 1976.

ARAUJO, A. S. *O uso variável dos pronomes tu, você e cê na função de sujeito: um estudo do padrão de comportamento referencial*. 2022. 198 f. Tese (Doutorado em Letras) – Universidade Federal de Sergipe, São Cristóvão, 2022.

ARAUJO, A. S.; BORGES, D. K. V. Variação no uso de pronomes-objeto de segunda pessoa na fala de estudantes Itabaianenses. *Paraguaçu*, v. 1, n. 1, p. 146-167, 2021.

ARDUIN, J. *A variação dos pronomes possessivos de segunda pessoa do singular teu/seu na Região Sul do Brasil*. 2005. 124f. Dissertação (Mestrado em Linguística) - Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão, Programa de Pós-Graduação em Linguística, 2005.

ARDUIN, J. A descrição do sistema possessivo de segunda pessoa na fala catarinense. In: *Anais da XX Jornada – GELNE*, João Pessoa, Paraíba, p. 1163-1172, 2004.

ASSIS, V. A. *Alegro-me em ver o outro sofrer? Uma descrição do vocabulário emocional e das expressões faciais da Schadenfreude*. 2025. 108 f. Dissertação (Mestrado em Psicologia) – Programa de Pós-Graduação em Psicologia, Universidade Federal de Sergipe, São Cristóvão, Sergipe, 2025.

BARBADINHO NETO, R. *Estudos filológicos*: volume dedicado à memória de Antenor Nascentes. Academia Brasileira de Letras, 2003.

BARRETO, E. A. *A expressão do aspecto habitual: um estudo na fala e na escrita de Itabaiana/SE*. 2014. 94f. Dissertação (Pós-Graduação em Letras) - Universidade Federal de Sergipe, São Cristóvão, 2014.

BATES, D.; MÄCHLER, M.; BOLKER, B.; WALKER, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, v. 67, n. 1, p. 1-48, 2015. DOI: 10.18637/jss.v067.i01.

BAILEY, G.; TILLERY, J. Some sources of divergent data in sociolinguistics. In: FOUGHT, C. *Sociolinguistic variation: critical reflections*. New York: Oxford University, 2004. p.11–30

BEAMAN, K. V. Exploring an approach for modelling lectal coherence. In: VELDE, H. V.; HILTON, N. H.; KNOOIHUIZEN, R. (Eds.). *Language Variation–European Perspectives VIII*, Leeuwarden, 2021, p. 135-160.

BORTOLETTO, F. F.; ANTONELLI, A. O uso dos pronomes de segunda pessoa em duas sincronias do português brasileiro. In: *Anais do I SIELLI*, p. 1-9, 2020.

BRITAIN, D. Dialect and Accent. In: AMMON, U.; DITTMAR, N.; MATTHEIER, K.; TRUDGILL, P. (Eds.). *Volume 1: An International Handbook of the Science of Language and Society*. Berlin-New York: De Gruyter Mouton, 2004, p. 267-273. https://doi.org/10.1515/9783110141894.1.2.267

BRITAIN, D. Dialect contact and new dialect formation. In: BOBERG, C.; NERBONNE, J.; WATT; D. (eds.). *The Handbook of Dialectology.* Wiley Blackwell, 2018, p. 143-158.

BRITAIN, D. Space, diffusion and mobility. In: CHAMBERS, J. K.; TRUDGILL, P.; SCHILLING-ESTES, N. (eds.). *The handbook of language variation and change*. Blackwell publishing, 2008, p. 604-637.

BUCHSTALLER, I.; KHATTAB, G. Population samples. In: PODESVA, R. J.; SHARMA, D. *Research methods in linguistics*. New York: Cambridge University Press, 2013, p. 74-95.

CALLOU, D.; MORAES, J. A norma de pronúncia do S e R pós-vocálicos: distribuição por áreas regionais. In: CARDOSO, S. (Org.). *Diversidade lingüística e ensino*. Salvador: EDUFBA, 1996, p. 133-147.

CALLOU, D.; MORAES, J.; LEITE, Y. Variação e diferenciação dialetal: a pronúncia do /r/ no português do Brasil. In: *Gramática do português falado, vol. VI*. Campinas: Editora da Unicamp, 1996, p. 465-493.

CALLOU, D.; SILVA, G. M. O. O uso do artigo definido em contextos específicos. In: HORA, D. (Org.). *Diversidade Lingüística no Brasil***.** João Pessoa: Idéia, 1997, p. 11-27.

CAMPBELL-KIBLER, K.; WALKER, A.; ELWARD, S.; CARMICHAEL, K. Apparent time and network effects on long-term cross-dialect accommodation among college students. U. Penn *Working Papers in Linguistics***,** v. 20, n. 2, 2014.

CAMPOS, H. A variação morfossintática do artigo definido na capital capixaba. *PERcursos Linguísticos*, v. 2, n. 5, p. 21–39, 2012.

CARDOSO, D. P. *Atitudes linguísticas e avaliações subjetivas de alguns dialetos brasileiros.* São Paulo: Editora Blucher, 2015.

CARDOSO, P. B. *Efeitos linguísticos e paralinguísticos na inferência dos sentidos indicados por (eu) acho que em entrevistas sociolinguísticas*. 2021. 99 f. Dissertação (Mestrado em Letras) - Universidade Federal de Sergipe, São Cristóvão, 2021.

CARDOSO, P. B. *Entre palavras e gestos manuais: uma abordagem multimodal da negação no português brasileiro*. 2025. Tese (Doutorado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal de Sergipe, São Cristóvão, Sergipe, 2025.

CARDOSO, P. B.; FREITAG, R. M. K.; ASSIS, V. A.; SIQUEIRA, J. M.; MENEZES, K. V.; GREGORIO, L. L.; DORIA, Y. A. A. T. M.; TAVARES, B. N.; CONCEICAO, N. S.; MATOS, A. W. O.; FREITAS, F. O. *Linguagem Corporificada 2023*. 2024. DOI: doi.org/10.17605/OSF.IO/X5S68. Disponível em: < https://osf.io/x5s68/>. Acesso em: 15 out. 2024.

CHAMBERS, J. K.; TRUDGILL, P. *Dialectology*. 2. ed. Cambridge: Cambridge University Press, 2004.

CHESHIRE, J. Taming the vernacular: Some repercussions for the study of syntactic variation and spoken grammar. *Cuadernos de Filologia Inglesa*, v.8n n. 1, p. 59–80, 1999.

COMITÊ NACIONAL DO PROJETO ALIB. *Atlas Linguístico do Brasil*: questionário 2001. Londrina: Ed. UEL, 2001.

CORREA, T. R. A. *A variação na realização de /t/ e /d/ na comunidade de práticas da UFS: mobilidade e integração.* 121f. 2019. Dissertação (Mestrado em Estudos Linguísticos) – Universidade Federal de Sergipe, 2019.

CRISTÓFARO-SILVA, T. *Fonética e fonologia do português*. Roteiro de estudos e guia de exercícios. São Paulo: Editora Contexto, 9 ed., 2007.

DALTO, C. D. L. *Estudo sociolingüístico dos pronomes-objeto de primeira e de segunda pessoas nas três capitais do Sul do Brasil*. 2002. 132f. Dissertação (Mestrado em Letras) – Universidade Federal do Paraná, Curitiba, 2002.

DANCEY, C.; REIDY, J. *Estatística Sem Matemática para Psicologia.* Porto Alegre: Artmed, 2006.

DRAGER, K.; KIRTLEY, M. Awareness, Salience, and Stereotypes in Exemplar-Based Models of Speech Production and Perception. In: BABEL, A. (Ed.). *Awareness and Control in Sociolinguistic Research*. Cambridge: Cambridge University Press, 2016, p. 1-24. DOI: https://doi.org/10.1017/CBO9781139680448.003.

DUARTE, M. E. L.; VAREJÃO, F. Null subjects and agreement marks in European and Brazilian Portuguese. *Journal of Portuguese Linguistics*, Lisboa, v. 12, n. 2, p. 101-124, 2013.

FIGUEIREDO FILHO, D. B.; SILVA JÚNIOR, J. A. Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). *Revista Política Hoje*, v. 18, n. 1, p. 115-146, 2009.

FLECK, L.; SIMIONI, T. O uso real dos pronomes pessoais em Itaqui-RS. In: *Anais do Salão Internacional de Ensino, Pesquisa e Extensão*, v. 8, n. 4, 2016.

FRANCESCHINI, L. T. Variação pronominal tu/você em Concórdia/SC: o papel dos fatores sociais. *Signótica*, v. 27, n. 2, p. 265-286, 2015.

FREITAG, R. M. K. Idade: uma variável sociolingüística complexa. *Línguas & Letras*, v. 6, n. 11, p. 105–121, 2000. DOI: 10.5935/rl&l.v6i11.875.

FREITAG, R. M. K. Banco de dados Falares Sergipanos. *Working Papers em Linguística*, v. 14, n. 1, p. 156-164, 2013.

FREITAG, R. M. K. Amostras sociolinguísticas: probabilísticas ou por conveniência?. *Revista de Estudos da Linguagem*, v. 26, n. 2, p. 667-686, 2018.

FREITAG, R. M. K. *Projeto de pesquisa: A língua do universitário: fala, leitura e escrita para o letramento acadêmico.* Universidade Federal de Sergipe, Programa de Apoio Pedagógico, 2018.

FREITAG, R. M. K. Reparos na leitura em voz alta como pistas de consciência sociolinguística. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, v. 36, n. 2, 1-22, 2020. DOI: https://doi.org/10.1590/1678-460x2020360206.

FREITAG, R. M. K. Mobility and higher education in grammatical patterns of Brazilian Portuguese. In: MUHR, R.; DUARTE, E.; RODRIGUES, C.; THOMAS, J. *Pluricentric Languages in the Americas.* Graz/Berlin: PCL-PRESS, 2022, p. 201-218.

FREITAG, R. M. K. O tratamento da variação linguística na escola para o combate à discriminação e ao preconceito. In: SILVA, N. I.; MINUSSI, R. D. *Linguística na educação básica*. Campinas: Mercado de Letras, 2023, p. 209-222.

FREITAG, R. M. K. *Não existe linguagem neutra!: gênero na sociedade e na gramática do português brasileiro.* São Paulo: Contexto, 2024.

FREITAG, R. M. K.; GOIS, T. S. Performance in a dialectal profiling task of LLMs for varieties of Brazilian Portuguese. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 2024. **Anais...** [S.l.]: SBC, 2024. p. 317-326. DOI: https://doi.org/10.5753/stil.2024.241891.

FREITAG, R. M. K.; ROST-SNICHELOTTO, C. A. Análises contrastivas: estabilidade, variedade ou metodologia?. *Working Papers em Linguística*, v. 16, n. 1, p. 157-169, 2015.

FREITAG, R. M. K.; SEVERO, C. G.; ROST-SNICHELOTTO, C. A.; TAVARES, M. A. Como os brasileiros acham que falam? Percepções sociolinguísticas de universitários do Sul e do Nordeste. *Todas as Letras-Revista de Língua e Literatura*, v. 18, n. 2, p. 64-84, 2016.

GAMA, D. E. R. S. O Uso Variável dos Clíticos para Referenciar o Interlocutor. *Revista A Cor das Letras*, v. 19, n. 2, p. 102-115, 2019.

GARSON, G. D. *Statnotes: Topics in Multivariate Analysis*. 2009. Disponível em: http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm. Acesso em: 09 de fev. 2024.

GUEDES, S. Emprego do artigo definido em situação de contato dialetal. *Domínios de Lingu@gem*, v. 13, n. 4, p. 1401-1432, 2019.

GUIMARÃES, T. A. A. S. *Tu e você no falar de Fortaleza-CE: variação e avaliações linguísticas*. 2019. 219f. Tese (Doutorado em Linguística Aplicada) – Universidade Estadual do Ceará, 2019.

GUY, G. R. The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables?. *Journal of pragmatics*, v. 52, p. 63-71, 2013.

GUY, G. R.; HINSKENS, F. Linguistic coherence: Systems, repertoires and speech communities. *Lingua*, v. 172, n. 173, p. 1-9, 2016.

GUY, G. R.; OUSHIRO, L.; MENDES, R. B. Indexicality and coherence. In: BEAMAN, K. V.; GUY, G. *The Coherence of Linguistic Communities*. Routledge, 2022, p. 53-68.

HARZING, A. W. *Publish or Perish.* Disponível em: https://harzing.com/resources/publish-or-perish. 2007. Acesso em: 10 ago. 2022.

HELLWIG, B.; GEERTS, J. *ELAN: Linguistic Annotator*. Versão 4.4.0. 2013. Disponível em: mpi.nl/corpus/manuals/manual-elan.pdf. Acesso em: 11 ago. 2023 .

HONNIBAL, M.; MONTANI, I.; VAN LANDEGHEM, S.; BOYD, A. *spaCy: Industrial-strength Natural Language Processing in Python*. 2020. doi.org/10.5281/zenodo.1212303

HORVATH, B. M.; SANKOFF, D. Delimiting the Sydney Speech Community. *Language in Society*, v. 16, n. 2, p. 179–204, 1987.

JOHNSON, D. E. Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and Linguistics Compass*, v. 3, n. 1, p. 359–383, 2009.

KASSAMBARA, A.; MUNDT, F. *Factoextra:* Extract and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0.7. 2020
Disponível em: <https://CRAN.R-project.org/package=factoextra>. Acesso em: 01 abr. 2023.

KLUYVER, T.; RAGAN-KELLEY, B.; FERNANDO, P.; GRANGER, B.; BUSSONNIER, M.; FREDERIC, J.; WILLING, C. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: LOIZIDES, F.; SCHMIDT, B. (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 2016, p. 87–90.

LABOV, W. The social motivation of a sound change. *Word*, v. 19, p. 273–309, 1963.

LABOV, W. *Language in the inner city: Studies in the Black English vernacular*. University of Pennsylvania Press, 1972.

LABOV, W. Where Does the Linguistic Variable Stop? A Response to Beatriz Lavandera. *Working Papers in Sociolinguistics*, n. 44, 1978.

LABOV, W. *The unobservability of structure and its linguistic consequences*. Paper presented at New Ways in Analyzing Variation (NWAV) 22. Ottawa, ON: University of Ottawa, 1993.

LABOV, W. *Principles of linguistic change, volume 1: Internal factors*. Oxford: Blackwell, 1994.

LABOV, W. *Principles of linguistic change, volume 2: Social factors*. Oxford: Blackwell, 2001.

LABOV, W. *The social stratification of English in New York city*. Cambridge University Press, 2006[1966].

LABOV, W. *Padrões Sociolinguísticos*. Trad. de M. Bagno; M. M. P. Scherre; C. R. Cardoso. São Paulo: Parábola Editorial, 2008[1972].

LAVANDERA, B. R. Where does the sociolinguistic variable stop?. *Language in society,* v. 7, n. 2, p. 171-182, 1978.

LOPES, C. R. S. Retratos da variação entre "você" e "tu" no português do Brasil: sincronia e diacronia. In: RONCARATI, C.; ABRAÇADO, J. (Org.). *Português Brasileiro II - contato lingüístico, heterogeneidade e história*. 1 ed. Niterói: EDUFF, 2008, p. 55-71.

LUCCHESI, D. The article systems of Cape Verde and São Tomé creole Portuguese**:** general principles and specific factors. *Journal of Pidgin and Creole Languages,* v. 8, n. 1, p. 81-108, 1993.

MAECHLER, M; ROUSSEEUW, P; STRUYF, A; HUBERT, M; HORNIK, K. *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.6. 2023. Disponível em: <https://CRAN.R-project.org/package=cluster>. Acesso em: 01 abr. 2023.

MANSFIELD, J.; LESLIE-O'NEILL, H.; LI, H. Dialect differences and linguistic divergence: A crosslinguistic survey of grammatical variation. *Language Dynamics and Change*, v. 1, n. aop, p. 1-45, 2023.

MATTHEWS, P. *The Concise Oxford Dictionary of Linguistics*. Oxford: Oxford Press, 1997.

MENDES, F. Variação estilística e genericidade: a variação de pronomes possessivos de segunda e terceira pessoa do singular. In: *Anais do CELSUL,* p. 1-16, 2008.

MENDONÇA, J. J. *Traços semânticos da referência à primeira pessoa do plural do português brasileiro*: um estudo em tempo real. 2022. 121f. Tese (Doutorado em Letras) – Universidade Federal de Sergipe, 2022.

MENDONÇA, J. J. *Variação na expressão da 1ª pessoa do plural: indeterminação do sujeito e polidez*. 2016. 102 f. Dissertação (Mestrado em Letras) - Universidade Federal de Sergipe, São Cristóvão, 2016.

MEYERHOFF, M.; WALKER, J. A. An existential problem: the sociolinguistic monitor and variation in existential constructions on Bequia (St. Vincent and the Grenadines). *Language in Society*, v. 42, n. 4, p. 407–428, 2013.

MILROY, L.; GORDON, M. *Sociolinguistics: Method and interpretation*. John Wiley & Sons, 2003.

MOORE, D. S.; McCABE, G. *Introduction to the practice of statistics*. New York: Freeman, 2004.

MOORE, D; S. *The Basic Practice of Statistics.* New York: Freeman, 2003.

MOORE, E. The Social Meaning of Syntax. In: HALL-LEW, L.; MOORE, E.; PODESVA, R. J. *Social Meaning and Linguistic Variation: Theorizing the Third Wave*, p. 54-79, 2021.

NASCENTES, A. *O linguajar carioca*. 2.ed. Completamente refundida. Rio de Janeiro: Organização Simões, 1953.

NASCIMENTO, L. C. R.; PAIM, M. M. T. A variação tu/você no português popular falado de Salvador e Amargosa, na Bahia. In: *Anais do VI Encontro de Sociolinguística*: Estudos sobre a relação entre língua e sociedade, p. 31-45, 2016.

NOGUEIRA, F. M. S. B. *Como os falantes de Feira de Santana e Salvador tratam o seu interlocutor?*. 2013. 138f. Dissertação (Mestrado em Língua e Cultura) – Universidade Federal da Bahia, 2013.

NOVAIS, V. S. *Variação na concordância verbal de terceira pessoa do plural Na fala de universitários sergipanos*. Dissertação (Mestrado em Letras) – Universidade Federal de Sergipe, São Cristóvão, SE, 2021.

NOVAIS, V.; SIQUEIRA, M. A variável sexo/gênero no português falado no sertão alagoano. *Leitur*a, n. 66, p. 35–50, 2020. DOI: 10.28998/2317-9945.202066.35-50.

OLIVEIRA, L. A. F. *Tu* e *você* no português rural do estado da Bahia. In: *Anais do SEMOC-Semana de Mobilização Científica-Meio Ambiente e Desenvolvimento Sustentável*, Salvador, p. 1-12, 2007.

OUSHIRO, L. A coesão dialetal nas concordâncias nominal e verbal no português paulistano. *Cuadernos de la ALFAL*, n. 7, p. 68-89, 2015a.

OUSHIRO, L. *Identidade na pluralidade: avaliação, produção e percepção linguística na cidade de São Paulo*. 2015. Tese (Doutorado em Semiótica e Linguística Geral) – Universidade de São Paulo, 2015b.

OUSHIRO, L. Social and structural constraints in lectal cohesion. *Lingua*, v. 172-172, p. 116-130, 2016a.

OUSHIRO, L. *A acomodação dialetal e a estabilidade de padrões sociolinguísticos na fala adulta.* Relatório Científico (Pós-doutorado em Linguística) – Faculdade de Letras, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016b.

PATIL, I. Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software*, v. 6, n. 61, p. 3167, 2021. DOI: https://doi.org/10.21105/joss.03167

PEREIRA, D. K. F. *A realização de artigo definido no português falado na região do sertão do Pajeú - PE***.** Dissertação (Mestrado em Linguística) – Centro de Artes e Comunicação, Universidade Federal de Pernambuco, 2017.

PINHEIRO, B. F. M. *Pistas linguísticas e paralinguísticas para os sentidos diminutivos*. 2021. 97 f. Dissertação (Mestrado em Letras) - Universidade Federal de Sergipe, São Cristóvão, 2021.

QIAO, S.; GENG, X.; WU, M. An Improved Method for K-Medoids Algorithm. In: *Business Computing and Global Informatization (BCGIN), 2011 International Conference*, p. 440,444, 2011.

R CORE TEAM. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria, 2018. Disponível em: <https://www.r-project.org/>. Acesso em: 06 jun. 2022.

RAMOS, C. M. A. *O clítico de 3 pessoa*: um estudo comparativo português brasileiro / espanhol peninsular. 1999. Tese (Doutorado em Linguística) – Universidade Federal de Alagoas, Maceió, 1999.

RIBEIRO, C. C. S. *Deslocamento geográfico e padrões de uso linguístico: a variação entre as preposições em ~ ni na comunidade de práticas da Universidade Federal de Sergipe*. 84f. 2019. Dissertação (Mestrado em Estudos Linguísticos) – Universidade Federal de Sergipe, 2019.

RODRIGUES, F. G. C. *Variação na regência de complementos locativos de verbos de movimento na fala de universitários da UFS*. 2021. Dissertação (Mestrado em Letras) – Universidade Federal de Sergipe, São Cristóvão, SE, 2021.

ROMAINE, S. On the problem of syntactic variation and pragmatic meaning in sociolinguistic theory. *Folia linguistica*, v. 51, n. s1000, p. 1-29, 2017[1981].

RSTUDIO TEAM. *RStudio: Integrated Development Environment for R*. Boston: MA, 2015. Disponível em: http://www.rstudio.com. Acesso em: 06 fev. 2023.

SANKOFF, D.; HENRY, R. 1975. A probabilistic model of variation in language. *Language*, v. 51, n. 2, p. 374–398, 1975.

SANKOFF, D.; LABOV, W. On the uses of variable rules. *Language in society*, v. 8, n. 2-3, p. 189-222, 1979.

SANKOFF, D.; TAGLIAMONTE, S.; SMITH, E. *Goldvarb X*: A variable rule application for Macintosh and Windows. Department of Linguistics, University of Toronto, 2005.

SANKOFF, G. Above and beyond phonology in variable rules. In: BAILEY, C-J.; SHUY, R. (eds.). *New Ways of Analyzing Variation in English.* Washington, D. C.: Georgetown Univ. Press, p. 42-62, 1973.

SANTANA, R. R. *Tipos de tipo em uma comunidade de prática universitária*. 2019. Dissertação (Mestrado em Letras) - Universidade Federal de Sergipe, 2019.

SANTOS, K. C. *Estratégias de polidez e a variação de nós x a gente na fala de discentes da Universidade Federal de Sergipe*. 2014. 88 f. Dissertação (Pós-Graduação em Letras) - Universidade Federal de Sergipe, São Cristóvão, SE, 2014.

SARDINHA, T. B. Lingüística de Corpus: histórico e problemática. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, v. 16, n. 2, p. 323-367, 2000. DOI: 10.1590/S0102-44502000000200005.

SCHERRE, M. M. P.; ANDRADE, C. Q.; CATÃO, R. C. Por onde transitam o tu e o você no Nordeste?. *Revista de Letras*, v. 1, n. 40, p. 164-197, 2021.

SCHERRE, M. M. P.; DIAS, E. P.; ANDRADE, C. Q.; MARTINS, G. F. Variação dos pronomes "tu" e "você". In: MARTINS, M. A.; ABRAÇADO, J. *Mapeamento sociolinguístico do português brasileiro*. São Paulo: Contexto, 2015, p. 133-172.

SCHERRE, M. M. P.; NARO, A. J. Sobre a concordância de número no português falado do Brasil. In.: RUFFINO, G. (org.) *Dialettologia, geolinguistica, sociolinguistica*. (Atti del XXI Congresso Internazionale di Linguistica e Filologia Romanza) Centro di Studi Filologici e Linguistici Siciliani, Universitá di Palermo. Tübingen: Max Niemeyer Verlag. p. 509- 523, 1998.

SCHERRE; M. M.; DUARTE, M. E. L. Main current processes of morphosyntactic variation. *In*: WETZELS, L; COSTA, J.; MENUZZI, S. (Eds.) *The Handbook of Portuguese Linguistics*. John Wiley & Sons, Inc., 2016, p. 526-544. DOI: https://doi.org/10.1002/9781118791844.ch29.

SEDRINS, A. P.; PEREIRA, D. K. F.; SILVA, C. R. T. O uso do artigo definido diante de antropônimos e pronomes possessivos em duas cidades do sertão pernambucano. *Caletroscópio*, v. 5, n. 8, p. 12-33, 2017.

SILVA, G. M. O. Emprego do artigo diante de possessivo e de patronímico: resultados sociais. In: SILVA, G. M. O; SCHERRE, M. M. P. (Orgs.). *Padrões sociolingüísticos:* análise de fenômenos variáveis do português falado na cidade do Rio de Janeiro. Rio de Janeiro: Tempo Brasileiro, 1998b, p. 265-281.

SILVA, G. M. O. *Estudo da Regularidade na Variação dos Possessivos no Português do Rio de Janeiro*. Tese (Doutorado) - Universidade Federal do Rio de Janeiro, 1982.

SILVA, G. M. O. Realização facultativa do artigo definido diante de possessivo e de patronímico. In: SILVA, G. M. O.; SCHERRE, M. M. P. (Orgs.). *Padrões sociolingüísticos:* análise de fenômenos variáveis do português falado na cidade do Rio de Janeiro. Rio de Janeiro: Tempo Brasileiro, 1998a, p. 120-145.

SILVA, J. M. S. *Variação no preenchimento da posição determinante antes de possessivos pré-nominais:* padrões dialetais e contatos. 2020. 140f. Dissertação (Mestrado em Letras) - Universidade Federal de Sergipe, São Cristóvão, 2020.

SILVA, J. M. S.; VITÓRIO, E. G. S. L. A. A implementação de a gente não sujeito no sertão alagoano. *A Cor das Letras*, v. 19, n. 2, p. 183-198, 2018.

SILVA, L. S. *Análise acústica ou de oitiva? Contribuições para o estudo da palatização em Sergipe*. 2021. 117 f. Dissertação (Mestrado em Letras) - Universidade Federal de Sergipe, São Cristóvão, 2021.

SILVA, S. O. P.; VITÓRIO, E. G. S. L. A. A variação *você* e *cê* no sertão alagoano. *Leitura*, n. 59, p. 122-142, 2017.

SIMPSON, J. Accent. In: ASHER, R (Ed.). *Encyclopaedia of Language and Linguistics*. Oxford: Oxford Press, 1994, p. 8-12.

SIQUEIRA, A. L. S. Análise da ocorrência de artigos definidos diante de possessivos pré-nominais e antropônimos em dados de fala. In: *Anais da 25ª Jornada Nacional do GELNE*, 2014.

SIQUEIRA, M. Efeitos do contato entre normas na variação linguística: a presença de artigo definido antecedendo possessivos no falar universitário da UFS. *Porto das Letras*, v. 6, n. 1, p. 8-33, 2020.

SIQUEIRA, M. O tratamento da mobilidade em variáveis morfossintáticas. In: *Caderno de resumos do V Congresso Internacional de Linguística Histórica - V CILH*, 2021.

SIQUEIRA, M.; FREITAG, R. M. K. Can mobility affect grammar at the morphosyntactic level? A study in Brazilian Portuguese. *Organon*, v. 37, n. 73, p. 14-35, 2022.

SIQUEIRA, M.; NOVAIS, V. Controle da escolarização no português falado do sertão alagoano. *R. Letras*, v. 25, n. 46, p. 81-100, 2023.

SIQUEIRA, M.; SOUSA, M. D. A. F.; RODRIGUES, F. G. C. Sistematizando Padrões Dialetais Morfossintáticos: Mobilidade e Contato. In: FREITAG, R. M. K.; SAVEDRA, M. M. G. *Mobilidades e Contatos Linguísticos no Brasil*. São Paulo: Blucher, 2023, p. 165 - 188. DOI: https://doi.org/10.5151/9786555502121-08

SOUSA, M. D. A. F. *Protocolo para anotação linguística e gerenciamento de amostras sociolinguísticas: o caso da amostra Deslocamentos 2019*. 2023. Tese (Doutorado em Letras) Programa de Pós-graduação em Letras, Universidade Federal de Sergipe, 2023. Disponível em: https://ri.ufs.br/jspui/handle/riufs/18363. Acesso em: 04 jan. 2024.

SOUSA, M. D. A. F.; SOUZA, V. R. A. Transcrição e anotação de dados linguísticos usando as ferramentas ELAN e LancsBox. *Domínios de Lingu@gem*, Uberlândia, v. 16, n. 3, p. 1173–1202, 2022. DOI: 10.14393/DL51-v16n3a2022-10.

SOUZA, E. S. *A preposição 'ni' no continuum rural-urbano de comunidades baianas.* Dissertação (Mestrado em Estudos Linguísticos) – Departamento de Letras e Artes da Universidade Estadual de Feira de Santana, Feira de Santana, 2015.

SOUZA, G. G. A. *Palatalização de oclusivas alveolares em Sergipe*. 2016. 76 f. Dissertação (Mestrado em Letras) - Universidade Federal de Sergipe, São Cristóvão, 2016.

SOUZA, V. R. A. *Monotongação dos ditongos decrescentes orais [oʊ̯], [eɪ̯], [aɪ̯] e [oɪ̯] na fala e na leitura em voz alta de universitários sergipanos*. 2022. 165 f. Dissertação (Mestrado em Letras) - Universidade Federal de Sergipe, São Cristóvão, 2022.

SQUIRES, L. Processing Grammatical Differences: Perceiving versus Noticing. In: BABEL, A. (Ed.). *Awareness and Control in Sociolinguistic Research*. Cambridge: Cambridge University Press, 2016, p. 80-103. DOI: https://doi.org/10.1017/CBO9781139680448.006.

TAMMINGA, Meredith. Interspeaker covariation in Philadelphia vowel changes. *Language Variation and Change*, v. 31, n. 2, p. 119-133, 2019.

TRUDGILL, P. *Dialects in contact.* Oxford: Basil Blackwell, 1986.

VALLI, M. Análise de cluster. *Augusto Guzzo Revista Acadêmica*, v. 4, p. 77-87, 2012.

VARELLA, C. A. A. Análise de componentes principais. *Seropédica*, Universidade Federal Rural do Rio de Janeiro, p. 38, 2008.

WEINREICH, U.; LABOV, W.; HERZOG, M. *Fundamentos empíricos para uma teoria da mudança linguística*. Trad. de Marcos Bagno. São Paulo: Parábola Editorial, 2006[1968].

WICKHAM, H. *ggplot2*: Elegant Graphics for Data Analysis. New York: Springer-Verlag, 2016. Disponível em: https://ggplot2.tidyverse.org. Acesso em: 10 mar. 2022.

ZAR, J. H. Spearman rank correlation. *Encyclopedia of biostatistics*, v. 7, 2005.