DOI: 10.14393/DLv18a2024-44
e1844

Artigo

Covariação entre os verbos *botar* e *colocar* nas capitais do Brasil: uma pesquisa sociolinguística com dados do ALiB

Covariation between the verbs *botar* and *colocar* in the capitals of Brazil: a sociolinguistic research with data from ALiB

Cassio Murilio Alves de LAVOR[*]
Rakel Beserra de Macêdo VIANA[**]
Aluiza Alves de ARAÚJO[***]

RESUMO: Este artigo discute o uso dos verbos concorrentes *botar* e *colocar* no Português brasileiro em dados das capitais das cinco regiões do Brasil (Centro-Oeste, Nordeste, Norte, Sudeste e Sul) extraídos dos registros de fala do Questionário Fonético-Fonológico (QFF), Questionário Semântico-Lexical (QSL) e o Questionário Morfossintático (QMS) do Atlas Linguístico do Brasil – ALiB, totalizando 200 informantes. Para essa empreitada, objetivamos verificar quais fatores sociais como *sexo*, *faixa etária*, *escolaridade* e *localidade*, favorecem o uso do verbo *botar*, tido como de menor prestígio social, na amostra analisada e se a concorrência entre os verbos *botar* e *colocar* representa um caso de variação estável ou um caso de mudança em progresso. Nosso modelo metodológico parte da coleta, codificação e discussão dos resultados, desenvolvido a partir dos pressupostos teórico-metodológicos da Sociolinguística Variacionista (Labov, 2008; Weinreich; Labov; Herzog, 2006 [1968]), usando, para coleta das ocorrências, a ferramenta tecnológica definida como linguagem R (R CORE TEAM, 2021). Os resultados estatísticos permitiram inferir que 48,5% dos entrevistados usam o verbo *botar* (1.473 dados) e 51,5% o verbo *colocar* (1.567 dados). Além disso, foi possível inferir que a variável sexo não apresenta correlação com a variação entre os verbos estudados, se contrapondo às variáveis *faixa etária*, *escolaridade*, *região* e *localidade*, que estão correlacionadas à variação entre os verbos estudados.
PALAVRAS-CHAVE: ALiB. Botar. Colocar. Variação linguística.

ABSTRACT: This article discusses the use of the competing verbs *botar* and *colocar* in Brazilian Portuguese in data from the capitals of the five regions of Brazil (Centre West, North East, North, South East and South) extracted from the speech records of the Phonetic-Phonological Questionnaire (QFF), the Semantic-Lexical Questionnaire (QSL) and the Morphosyntactic Questionnaire (QMS) of the Linguistic Atlas of Brazil - ALiB, totalling 200 informants. Thus, our general objective is to verify which social factors, such as *gender*, *age group*, *schooling* and

[*] Mestre em Linguística Aplicada, Secretaria da Educação do Ceará – SEDUC-CE. murilolavor.rh@gmail.com
[**] Doutora em Linguística Aplicada, Secretaria da Educação do Ceará – SEDUC-CE. rakelbeserra@gmail.com
[***] Doutora em Linguística, Universidade Estadual do Ceará – UECE. aluizazinha@hotmail.com

location, favour the use of the verb *botar*, considered to have less social prestige, in the sample analysed. Our methodological model starts with the collection, coding and discussion of the results, developed based on the theoretical-methodological assumptions of Variationist Sociolinguistics (Labov, 2008; Weinreich; Labov; Herzog, 2006 [1968]), using the technological tool defined as R language (R CORE TEAM, 2021) to collect the occurrences. The statistical results allowed us to infer that 48.5 per cent of interviewers use the verb *botar* (1,473 data points) and 51.5 per cent the verb *colocar* (1,567 data points). In addition, it was possible to infer that the gender variable does not correlate with the variation between the verbs studied, as opposed to the *age*, *schooling*, *region* and *location* variables, which are correlated with the variation between the verbs studied.
KEYWORDS: ALiB. Botar. Colocar. Linguistic variation.

Artigo recebido em: 01.03.2024
Artigo aprovado em: 16.09.2024

1 Introdução

Não há discordância no tocante ao fenômeno variacionista entre os verbos **botar** e **colocar** no Português do Brasil (doravante PB), diferentemente do Português de Portugal, (PE), em que, conforme Batoréo e Casadinho (2009), não se verificou a covariação entre esses verbos em estudo, pois, em vez disso, vislumbramos dois verbos com sentidos semânticos já bem estabelecidos e consolidados na comunidade lusitana.

Dessarte, quando os pesquisadores variacionistas direcionam sua atenção para esse fenômeno entre os verbos **botar** e **colocar** no PB, inferem que a concorrência entre os citados verbos se realiza de forma harmônica, logo, ela se apresenta em diferentes espaços geográficos e sociais, independente do sexo, da faixa etária, da escolaridade e da idade. Isso não quer dizer que não haja grupos sociais mais propensos à realização da variante[1] **botar** em detrimento da variante **colocar**, haja vista existir uma tendência a associar o uso do verbo **botar** às pessoas com menos escolaridade e pertencentes a uma comunidade de fala entendida como popular em comparação com outra comunidade defendida como culta. Todavia, isso só pode ser confirmado a partir de distintas pesquisas e bancos de dados.

[1] O termo variante é usado para referir as diferentes maneiras de dizer a mesma coisa do ponto de vista da língua (Labov, 2008).

Ademais, observa-se, principalmente, pelo senso comum, que há uma atribuição de valor ao verbo **colocar**, como sendo o correto e pertencente à norma culta, portanto, a variante padrão; já, ao verbo **botar**, foi atribuído um valor de não correto, errado, pertencente às classes populares, à fala popular, portanto, é uma variante não padrão, que, para efeito de investigação para este artigo, esse verbo é considerado inovador e, portanto, usado como valor de aplicação da regra variável[2]. Porém, buscando em algumas gramáticas (Bechara, 2005, 2015; Ferreira, 2003; Rocha Lima, 1992), não há nenhuma valoração atribuída aos verbos em concorrência.

Dessarte, a pesquisa cujos resultados são discutidos neste artigo é delineada a partir de um questionamento: como se comportam os verbos **botar** e **colocar** no PB em dados do ALiB? A partir desse, definiu-se como objetivo geral, então, verificar quais fatores sociais[3], como **sexo**, **faixa etária**, **escolaridade** e **localidade**, favorecem o uso do verbo **botar**, tido como de menor prestígio social, na amostra analisada. Ademais, definiu-se como objetivos específicos verificar: i) qual variante em concorrência é mais frequente na amostra analisada; ii) em qual localidade regional a variante **botar** é mais frequente; iii) em qual capital do Brasil a variante **botar** é mais frequente; iv) qual ou quais grupos de fatores controlados favorece o uso do verbo **botar**; v) e ainda, se a concorrência entre os verbos **botar** e **colocar** representa um caso de variação estável ou um caso de mudança em progresso.

Para esta tarefa, optamos por recorrer ao banco de dados do ALiB (Atlas Linguístico do Brasil), utilizando os dados de todas as capitais do Brasil para verificar esse fenômeno de coocorrência entre os verbos **botar** e **colocar**. Com este trabalho, possibilitaremos a construção de uma fotografia sociolinguística sobre a variação dos verbos **botar** e **colocar** nas capitais brasileiras, pois se trata de um trabalho pioneiro

[2] Em uma análise feita pelo programa Varbrul, "o pesquisador deve escolher qual das variantes será tratada como **aplicação da regra** e, ao realizar a rodada dos dados, deve informar ao programa o respectivo código dessa variante" (Guy; Zilles, 2007, p. 229).

[3] Os fatores sociais controlados pela pesquisa que deram origem a este artigo estão disponíveis na seção 4, a metodologia aplicada.

em razão do número de localidades pesquisadas em um mesmo estudo, situação que justifica esta pesquisa.

Considerando, principalmente, pesquisas de cunho variacionista a partir dos pressupostos teórico-metodológicos desenvolvidos por Weinreich, Labov e Herzog (2006) e Labov (2008), localizamos uma quantidade considerável de trabalhos publicados nos últimos anos, todavia, fizemos um recorte, usando, como critério de inserção, pesquisas sobre o fenômeno em pauta realizadas com dados do banco de fala do ALiB, o mesmo usado por este estudo. À vista disso, trazemos, então, as pesquisas de Lavor, Araújo e Viana (2018), Lavor, Viana e Araújo (2019), Lavor, Vieira e Araújo (2019) e Lavor, Araújo e Pereira (2020), como norteadoras deste trabalho. Portanto, as pesquisas elencadas deram sustentação à construção das hipóteses[4] iniciais a serem confirmadas ou refutadas, a partir dos resultados estatísticos alcançados, e à definição das variáveis controladas[5].

Acreditando que nada na língua acontece por acaso, que a variação não ocorre aleatoriamente e que é possível fornecer uma descrição sobre os fenômenos variacionistas presentes na língua portuguesa, buscamos elaborar reflexões e encontrar explicações para o fenômeno apresentado, procurando entender o funcionamento dessa alternância.

Diante do exposto, este texto está estruturado da seguinte forma: parte introdutória, na qual aqui apresentamos a problemática; uma seção que traz uma resenha sobre os trabalhos norteadores da pesquisa, uma revisão bibliográfica; a seção que apresenta os pressupostos teóricos que possibilitaram este estudo; a seção que traz a metodologia investida na realização desta pesquisa, dividida em subseções apresentando o *corpus*, a ferramenta digital e o caráter da pesquisa; a seção de apresentação dos resultados estatísticos e a discussão desse; as conclusões alcançadas ou considerações finais; e, por fim, as referências aqui utilizadas.

[4] As hipóteses estão apresentadas na seção 4, na metodologia aplicada.
[5] As variáveis ou grupos de fatores controlados estão apresentados na seção 4, na metodologia.

Na próxima seção, temos a revisão bibliográfica dos trabalhos selecionados como norteadores deste estudo.

2 *Botar* e *colocar* no ALiB

Nesta seção, optamos por realizar uma pequena resenha dos trabalhos — por ordem de publicação — que norteiam a construção das hipóteses e a definição das variáveis controladas nesta pesquisa, a saber: Lavor, Araújo e Viana (2018), Lavor, Viana e Araújo (2019), Lavor, Vieira e Araújo (2019), e Lavor, Araújo e Pereira (2020).

Os trabalhos elencados foram construídos a partir de recortes de fala de dados do ALiB[6]. Ademais, todos eles estão em consonância com os pressupostos teóricos e metodológicos da Sociolinguística Variacionista (Labov, 2008; Weinreich; Labov; Herzog, 2006). E, ainda, essas pesquisas trabalharam com a ferramenta computacional GoldVarb X[7] (Sankoff; Tagliamonte; Smith, 2005), diferentemente deste trabalho que usa o programa *RStudio*[8] na coleta das ocorrências localizadas.

Logo, apresentaremos, aqui nesta seção, apenas as variáveis selecionadas como relevantes e seus resultados estatísticos[9] (Frequência de uso das variantes controladas e os Pesos Relativos[10]) obtidos com as rodadas livres, após a retirada de possíveis nocautes[11], para cada um dos trabalhos resenhados. Essa escolha se justifica por já ter

---

[6] O ALiB está apresentado na subseção 4.1 na metodologia.

[7] O Goldvarb X, programa estatístico computacional desenvolvido por Sankoff, Tagliamonte e Smith (2005) (http://individual.utoronto.ca/tagliamonte/goldvarb.html), é uma das ferramentas-chave da Sociolinguística Variacionista, em termos metodológicos. Cabe a esse programa processar um grande volume de dados linguísticos, com o objetivo de definir uma regra variável que ajude a explicar determinado fenômeno sociolinguístico.

[8] O programa computacional RStudio está apresentado na subseção 4.2.

[9] Os resultados obtidos se apresentam como pesos relativos. A tabulação cruzada, por sua vez, mostra as relações – ou a falta delas – entre as variáveis independentes (Guy; Zilles, 2007).

[10] O Peso Relativo das variáveis, em que PR<0,5, PR=0,5 e PR>0,5, indicam, respectivamente, desfavorecimento, neutralidade e favorecimento de uma variável independente em relação à variante escolhida como aplicação da regra.

[11] Nocaute ou knockOut é uma terminologia de análise do GoldVarb X usada em todos os programas da série Varbrul, "que, num dado momento da análise, corresponde a uma frequência de 0% ou 100% para um dos valores da variável dependente" (Guy; Zilles, 2007, p. 158). Em outras palavras, os nocautes são as indicações de fatores ou grupos de fatores que se configuram como categóricos, ou seja, as variantes não competem com base nesses fatores ou grupos de fatores indicados.

sido apresentada, nesse introdutório, outras informações que são comuns a todos os trabalhos, já que esses usam dados de um mesmo banco de dados e trabalham com os mesmos pressupostos teórico-metodológicos.

Feitas essas considerações iniciais, pontuamos que a pesquisa de Lavor, Araújo e Viana (2018) usou amostras de fala de 84 informantes, 42 do sexo feminino e 42 do sexo masculino, provenientes de três estados da Região Nordeste do Brasil — Alagoas, Ceará e Piauí — selecionados a partir da audição, na íntegra, de todos os inquéritos de cada Estado pesquisado.

Os pesquisadores analisaram a atuação da **forma verbal** (presente, pretérito e demais formas encontradas); do **sexo** (masculino e feminino); da **faixa etária** (faixa I: 18 a 30 anos, e faixa II: 50 a 65 anos); do **tipo de questionário do** *ALiB* (Questionário Fonético-Fonológico (QFF); Questionário Semântico-Lexical (QSL); Questionário Morfossintático (QMS); Questões de Prosódia, Discursos Semidirigidos e Perguntas Metalinguísticas) e da **localidade** (Alagoas: Arapiraca, Santana do Ipanema e Maceió; Ceará: Camocim, Canindé, Crateús, Crato, Iguatu, Ipu, Limoeiro do Norte, Quixeramobim, Russas, Sobral, Tauá e Fortaleza; Piauí: Canto do Buriti, Corrente, Picos, Piripiri e Teresina), usando o verbo **botar** como valor de aplicação da regra, sobre a concorrência das variantes **botar** e **colocar** na amostra analisada[12].

O programa computacional utilizado, o Goldvarb X, apresentou um total de 706 ocorrências, sendo 353 (50,1%) para o verbo **botar** e 351 (49,9%) para o verbo *colocar*. Esses resultados demonstraram que, mesmo o verbo **botar** sendo o mais frequente, não há uma diferença significativa de uso desses verbos na amostra analisada.

A ferramenta estatística definiu, também, que a variável *sexo* (no fator masculino com PR[13] 0,624), a variável **faixa etária** (no **fator 50 a 65 anos**, com PR 0,650*)* e a variável **localidade** (nas cidades de Camocim/CE, com PR 0,819, e Teresina/PI com PR 0,710) são favorecedoras do verbo botar.

[12] Até a submissão do artigo, os autores eram membros do Grupo de Pesquisa ALiB.
[13] Abreviatura para Peso Relativo.

Os resultados apontados pela pesquisa de Lavor, Araújo e Viana (2018) indicam que a concorrência entre **botar** e **colocar** na amostra analisada. Trata-se, portanto, de uma variação estável, ou seja, sem indícios de que uma variante esteja tomando o lugar da outra nas localidades pesquisadas. De igual modo, os estudiosos verificaram que o emprego do verbo **botar**, em coocorrência com **colocar**, ocorre de modo sistemático por meio da influência, especificamente nessa amostra analisada, de fatores essencialmente externos ao sistema linguístico (**sexo**, **faixa etária** e **localidade**).

Na segunda pesquisa, Lavor, Viana e Araújo (2019), selecionou um total de 48 informantes estratificados em: **sexo** (24 mulheres e 24 homens), **localidade** (Camocim, Canindé, Crato, Crateús, Fortaleza, Iguatu, Ipu, Limoeiro do Norte, Quixeramobim, Russas, Sobral e Tauá) e **faixa etária** (18 a 30 anos, e 50 a 65 anos). Além dessas variáveis extralinguísticas, foram controladas as variáveis linguísticas **tipo de questionário** (QFF, QLS e QMS) e **tempo e forma verbal** (presente, pretérito, futuro, infinitivo, gerúndio e particípio).

O programa computacional utilizado selecionou 195 ocorrências totais, 105 (53,8%) para a variante **botar** e 90 (46,20%) para a variante **colocar**. Após a retirada de alguns nocautes[14], o programa definiu como estatisticamente relevantes, para o uso do verbo **botar**, dois grupos de fatores, a saber: a **faixa etária** (18 a 30 anos com PR 0,680) e a **forma verbal** (fator tempo **presente** com PR 0,605). Em contrapartida, os informantes da faixa etária de 50 a 65 anos, com PR 0,338, se comportaram de modo a inibir a realização dessa mesma forma variante.

Como forma de ampliar a pesquisa, os autores decidiram fazer uma nova rodada de dados, utilizando o grupo de fatores **faixa etária** vs. o grupo de fatores **localidade**. Dessarte, após a retirada dos nocautes apresentados pelo programa computacional, constatou-se, em seu melhor nível de análise, que a variável **forma**

[14] Os nocautes são entendidos como um problema para as análises estatísticas fornecidas pelo GoldVarb X, pois implicam dizer que, em um dado contexto, o uso de uma determinada variante foi categórico, ou seja, não houve variação (Guy; Zilles, 2007).

verbal e a variável **localidade** vs. **faixa etária** são relevantes para a aplicação da regra. Assim como nas rodadas anteriores, a variável **forma verbal**, mais especificamente o fator **presente**, com PR 0,640, revelou-se favorável à realização do verbo **botar**, enquanto os demais fatores comportaram-se como inibidores da regra variável.

No tocante à variável **localidade**, os resultados estatísticos demonstraram que, no *Ceará*, o verbo *botar* é favorecido pelas cidades de Camocim (PR 0,819), Quixeramobim (PR 0,815), Limoeiro do Norte (PR 0,819) e Ipu (PR 0,598). Em contrapartida, as cidades de Iguatu (PR 0,276) e Crateús (PR 0,308) se comportaram como inibidoras do verbo **botar**, usado como valor de aplicação da regra variável.

A terceira pesquisa resenhada, Lavor, Viera e Araújo (2019), apresenta um recorte de fala da Região Sul e Nordeste composto por uma amostra contendo 16 informantes estratificados em **localidade** (metade de Salvador e metade de Porto Alegre), **sexo** (masculinos e femininos), **faixa etária** (de 18 a 30 anos de idade e 50 a 65 anos de idade) e **escolaridade** (com escolaridade até a 4ª série do ensino fundamental e com nível universitário).

Logo, essa estratificação, provinda do próprio banco de dados, permitiu o controle das variáveis extralinguísticas, a saber: **sexo**, **faixa etária, escolaridade** e **localidade**. Além dessas, optou-se, ainda, por controlar as variáveis linguísticas: **forma verbal** (presente, pretérito, futuro, infinitivo, gerúndio e particípio); **tipo de questionário** (QFF, QSL e QMS) e **tópico discursivo** (trabalho, religião, relacionamento, lazer, vestuário e acessórios, alimentação/cozinhar e outros).

Tal como as duas pesquisas referenciadas anteriormente, a pesquisa de Lavor, Vieira e Araújo (2019) elegeu o verbo **botar** como fator de aplicação da regra e os dados foram submetidos ao programa computacional que, após a retirada dos nocautes registrados, selecionou um total geral de 247 ocorrências: 185 (66,5%) para a variante **botar** e 62 (33,5%) para a variante **colocar**. Ademais, o programa computacional, em seu melhor nível de análise, selecionou a variável **tipo de questionário** e a variável

escolaridade, nessa ordem de importância, como favorecedoras da ocorrência do verbo **botar**.

Em linhas gerais, os resultados estatísticos revelaram que, no grupo de fatores **tipo de questionário**, o fator *QSL* favoreceu o verbo **botar** (PR 0,767), enquanto os demais fatores inibiram a sua aplicação. Para a segunda variável selecionada, a **escolaridade**, o fator **nível fundamental** comportou-se como aliado do verbo **botar** (PR 0,569).

Considerando os resultados separadamente, ou seja, resultados obtidos na capital Salvador e resultados obtidos na capital Porto Alegre, o programa selecionou um total de 111 ocorrências para Salvador: 79 (71,20%) para **botar** e 32 (28,80%) para **colocar**. Ainda na capital Salvador, no melhor nível de análise, o programa computacional selecionou, como estatisticamente relevante para a ocorrência da variante **botar**, os grupos de fatores **tipo de questionário**, no fator *QSL* (PR 0,741), **forma verbal**, nos fatores **tempo presente** (PR 0,656) e o fator ***forma no infinitivo*** (PR 0,592) e **sexo**, no fator **masculino** (PR 0,639).

Concernente aos resultados apresentados na capital Porto Alegre, o programa selecionou um total geral de 74 ocorrências: 44 (59,5%) para **botar** e 30 (40,5%) para **colocar**. Ainda com dados da capital Porto Alegre, após a retirada dos nocautes apresentados, o programa computacional, no melhor nível de análise, selecionou a variável **escolaridade**, no fator nível **fundamental** (PR 0,682) como favorecedor da variante *botar*.

A última pesquisa resenhada é a de Lavor, Araújo e Pereira (2020), que investiga a concorrência entre os verbos **botar** e **colocar** na capital maranhense com o objetivo de analisar os fatores linguísticos e extralinguísticos que condicionam a realização do verbo *botar*.

A pesquisa em foco foi estratificada em **sexo** (masculino e feminino), **faixa etária** (de 18 a 30 anos e 50 a 65 anos de idade), **escolaridade** (até a 4ª série do ensino fundamental e nível universitário), e **localidade** (São Luís do Maranhão; Alto

Paranaíba; Bacabal; Balsas; Brejo; Imperatriz; São João de Patos; Tuntum e Turiaçu) e *tipo de* questionário (QFF, QSL, QMS, Questões de Prosódia, Discurso Semidirigido e Perguntas Metalinguísticas).

A partir dessa estratificação, provenientes do próprio banco de dados, os autores definiram os grupos de fatores extralinguísticos controlados. Acrescentaram a esses, ainda, o grupo de fator linguístico **forma verbal** (presente, pretérito e demais formas).

Após audição, na íntegra, dos inquéritos, as ocorrências obtidas foram curadas no programa computacional escolhido que, após a retirada dos nocautes localizados, apresentou um total de 711 ocorrências, sendo dessas, 507 ocorrências (71,3%) para **botar** e 204 ocorrências (28,7%) para **colocar**. Ademais, no melhor nível de análise, o programa definiu **localidade** e **faixa etária**, nesta ordem de importância, como os únicos a favorecer o uso do verbo **botar**.

O primeiro grupo de fatores selecionado como estatisticamente pertinente para a variação entre **botar** e **colocar**, a **localidade**, revelou que os municípios de São Luís do Maranhão (PR 0,882), Brejo (PR 0,864) e Bacabal (PR 0,658) favorecem a realização do verbo **botar** nos dados analisados. As cidades de Turiaçu (PR 0,482), Tuntun (PR 0,376), Imperatriz (PR 0,338), Balsas (PR 0,241), Alto Paranaíba (PR 0,197) e São João dos Patos (PR 0,188), por sua vez, mostraram-se desfavoráveis à realização do verbo *botar*.

O segundo grupo de fatores apontado como pertinente na seleção estatística feita pelo GoldVarb X, a faixa **etária**, demonstrou que, na amostra desta pesquisa, a **faixa etária de 50 a 65 anos** (PR 0,604) indica que os informantes idosos favorecem o uso de **botar** - considerada por nós, a variante inovadora. Em contrapartida, os falantes mais jovens, isto é, de 18 a 30 anos (PR 0,335), inibiram a realização do verbo *botar*.

Além desses resultados obtidos com rodadas usando o verbo **botar** como valor de aplicação, os autores optaram por realizar uma rodada cruzando dados da variável *sexo* vs. dados da variável **faixa etária**. Assim, nesta rodada, o GoldVarb X selecionou,

em seu melhor nível (*input*[15] 0,771 e *significance*[16] 0,000), as variáveis **localidade** e **sexo** (masculino de 50 a 65 anos e feminino de 50 a 65 anos) como relevantes para a variação entre **botar** e **colocar**. Por terem sido esses os únicos grupos de fatores selecionados nessa rodada, por conseguinte, os demais fatores controlados, mas não selecionados, podem ser considerados como irrelevantes estatisticamente, ou não condicionadores da aplicação da regra variável.

Ademais, essa rodada demonstrou que o município de São Luís do Maranhão (PR 0,886) favorece a realização do verbo **botar**, seguido da cidade de *Brejo* (PR 0,882) e *Bacabal* (0,643). E, ainda, pode-se perceber que o município de *Turiaçu* (PR 0,471), mais uma vez, atua de modo a inibir o uso de ***botar*** *(PR < 0,5)*, logo, pode-se considerar, então que esse resultado estatístico favorece o verbo **colocar**.

Consoante os resultados para a segunda variável selecionada como favorecedora do verbo **botar**, **sexo** vs. **faixa etária**, os resultados estatísticos indicam que os **homens** na faixa dos **50 a 65 anos** (PR 0,608) favorecem o verbo **botar**; enquanto os **homens** com idade de **18 a 30** (PR 0,416) inibem a realização dessa variante. Além disso, constatou-se que o fator **mulher** com **50 a 65 anos** (PR 0,603) favorece o verbo **botar**, ao contrário das **mulheres** com **18 a 30** (PR 0,246) que atuaram de modo a inibir o uso dessa forma variante.

Os resultados estatísticos apresentados nesta seção, consoante às pesquisas resenhadas — Lavor, Araújo e Viana (2018), Lavor, Viana e Araújo (2019), Lavor, Vieira e Araújo (2019) e Lavor, Araújo e Pereira (2020) — permitem, então, observar o quão sistemática tem sido a variação dos verbos **botar** e **colocar** em diferentes localidades do Brasil. Permitem, ainda, inferir-se que, ao contrário do que tenta fazer crer o senso comum, a variação entre **botar** e **colocar** não ocorre, em instância alguma, de modo

[15] O input consiste no "nível geral de uso de um determinado valor da variável dependente" (Guy; Zilles, 2007, p. 238).

[16] O nível de *significance* (significância) pode ser considerado a margem de erro de uma pesquisa. A margem utilizada pelo Varbrul é de 5% (*threshold*, 05), o que representa o grau de confiabilidade dos resultados: "se o nível de significância for acima deste valor, previamente arbitrado, os resultados não são considerados estatisticamente significativos" (Scherre, 1993, p. 27).

aleatório, mas, sim, por meio de um delicado jogo de interação e atuação de fatores linguísticos e extralinguísticos à língua.

Na seção seguinte, fazemos um breve resumo histórico da Sociolinguística Variacionista, quantitativa ou laboviana.

3 A Sociolinguística Variacionista

O estudo da língua se baseia no entendimento de que ela é um conjunto estruturado de normas sociais e que "é usada por seres humanos num contexto social, comunicando suas necessidades, ideias e emoções uns aos outros" (Labov, 2008, p. 215). Logo, entendemos que a língua é elaborada a partir da interação social entre indivíduos e, consequentemente, suscetível às particularidades que envolvem os sujeitos que dela faz uso e sua historicidade. Dessa forma, a língua muda, se reconstrói e se adapta em função de contextos sociais e históricos, permitindo, então, vislumbrar uma língua heterogênea, tanto dentro de um mesmo idioma quanto entre línguas diferentes.

Partindo desse entendimento, reagindo a uma concepção de língua homogênea e, consequentemente, ao estruturalismo, Saussure, e gerativismo, Chomsky, na década de 1960, surgiu a Sociolinguística, defendida, principalmente, por Labov (2008). Todavia, Labov não foi o único a defender as ideias sociolinguísticas, pois, segundo Coelho *et al.*, (2015), outros linguistas já defendiam o caráter social da língua desde o início do século XX, como Meillet (1866-1936), Marr (1865-1934) e Bakhtin (1895-1975), que se posicionavam contrários às ideias de Saussure (1916), ao conceber a língua em seu caráter social (Coelho *et al*., 2015).

Se, anteriormente, às teorias de Labov, a linguística desconsiderava as regras variáveis, entendendo que as variantes linguísticas não possuíam "valor' ou não cumpriam funções no processo comunicativo, o que poderia refletir uma espécie de caos (Camacho, 2006), na atualidade, a Sociolinguística apresenta a heterogeneidade linguística como um fato inquestionável, situação, essa, que permitiu a confirmação

daquela como ciência, e, como tal, possibilita várias discussões quantitativas sobre incontáveis fenômenos linguísticos em diferentes níveis, como lexical, fonológico, sintático, discursivo, e até de interface entre esses.

Dessa forma, considera-se hoje, consoante as fundamentações teóricas de Weinreich, Labov e Herzog (2006 [1968]), haver dois princípios básicos para se estudar a língua: 1) deixar de identificar estrutura linguística com homogeneidade e conceber como opção racional a possibilidade de descrever ordenadamente a diferenciação numa língua que serve à comunidade; 2) compreender que as gramáticas nas quais uma mudança linguística ocorre representam gramáticas de comunidade de fala. Entende-se, ainda, que tais princípios estão concatenados à heterogeneidade sistemática da língua.

Portanto, estudar a variação linguística, teorizada como inerente a todas as línguas, é contemplá-la como instrumento de comunicação usado por falantes de uma comunidade, num sistema de associações comumente aceito entre formas arbitrárias e seus significados. Logo, este texto descreve a regularidade da variação em termos de frequência das variantes botar e colocar, descrevendo os ambientes extralinguísticos ou sociais que se apresentam como mais significativos para a atuação do fenômeno linguístico estudado. A seguir, apresentamos nosso percurso metodológico.

4 Procedimentos Metodológicos

Nesta seção, apresentamos o *corpus* da pesquisa, a ferramenta computacional usada para curar os dados coletados, o caráter da pesquisa, as variáveis controladas e as hipóteses iniciais.

4.1 O ALiB

Hodiernamente, o Projeto ALiB é considerado o maior banco de amostra do PB falado no Brasil, principalmente, por esse conter dados de fala de pessoas em dois níveis de escolaridade, das cinco regiões do país a partir de 250 pontos, contabilizando

1.100 informantes espalhados pelo território brasileiro (Mota; Cardoso, 2009; Cardoso, 2012). Esse banco de dados originou-se no campo da variação linguística e da dialetologia, baseado na geolinguística que, com a ajuda de dados cartográficos, apresenta como finalidade a descrição de fenômenos linguísticos variáveis e dialetais do PB (Cardoso, 2010).

Iniciado em Salvador - BA, mais precisamente no ano de 1996, durante o Seminário Nacional Caminhos e Perspectivas para a Geolinguística no Brasil, sediado no Instituto de Letras da Universidade Federal da Bahia (UFPB), o Projeto ALiB possui uma estratificação homogênea na coleta de seus dados, haja vista os informantes estarem distribuídos igualmente em dois gêneros (masculino e feminino),[17] em duas faixas etárias (de 18 a 30 anos, e de 50 a 65 anos) e em dois níveis de escolaridade (ensino fundamental incompleto e universitário completo), contabilizando um total de oito informantes em cada uma das vinte e cinco[18] capitais e quatro informantes nas demais cidades, que compõem os Estados da Federação.

Assim, a disposição dos informantes do ALiB se apresenta de duas formas distintas, a saber: para as capitais, o projeto possui oito informantes, sendo quatro mulheres: duas com nível médio e duas com ensino superior; e quatro homens: dois com nível médio e dois com nível superior. Para as cidades do interior dos estados brasileiros, o projeto possui dados de fala de quatro informantes: dois homens e duas mulheres, todos somente com nível fundamental de escolaridade.

Esses informantes forneceram dados que podem ser utilizados para análises linguísticas a partir das dimensões diagenérica, diageracional e diastrática, consoante Mota e Cardoso (2009).

Além disso, o ALiB é constituído de questionários que contemplam diferentes níveis da língua: um questionário "fonético-fonológico (QFF), semântico-lexical (QSL)

[17] Para o ALiB, o gênero está restrito apenas ao sexo biológico.
[18] Relatamos somente 25 capitais com a exceção do Distrito Federal e de Palmas-TO, por serem estas, à época da constituição do projeto, capitais com poucos anos de fundação, sem possuírem, por isso, indivíduos nativos da capital.

e morfossintático (QMS), e contendo, além desses, quatro questões de pragmática, seis perguntas de natureza metalinguística, quatro temas para a documentação de discursos semidirigidos e um texto para leitura" (Mota; Cardoso, 2009, p. 249).

4.2 A ferramenta computacional

A pesquisa a que se vincula este artigo conta com uma tecnologia na cura das ocorrências coletadas que se pode considerar inovadora para os trabalhos sociolinguísticos. Essa é definida como linguagem R (R CORE TEAM, 2021), pois, segundo Oushiro (2014, p. 134), "O R é uma linguagem de programação voltada à análise de dados, que pode ser utilizada para realizar computações estatísticas e gráficas, compilar e anotar *corpora*, produzir listas de frequências, entre diversas outras tarefas", assim, a autora destaca, ainda, que "[...] Uma de suas principais vantagens é o fato de ser gratuito e estar disponível para uma variedade de plataformas (UNIX, Windows e MacOS)".

Essa linguagem, conhecida como R, permite, facilmente, desenvolver e compartilhar inúmeras tarefas que podem agilizar o desenvolvimento de diversos estudos, pois conta com uma grande comunidade de usuários que se encontram em fóruns online em regime de colaboração mútua. Logo, é patente que esta pesquisa se aproprie dessa ferramenta na inovação dos resultados aqui apresentados.

Portanto, nos utilizamos da interface *RStudio* (RSTUDIO TEAM, 2020), um similar do R detalhado. Com o *RStudio*, realizamos a descrição, elaboração de tabelas e de gráficos com a finalidade de compreender o condicionamento do fenômeno variável em estudo, a partir de um método estatístico já comprovada sua consistência, pois conforme Gil (2008, p. 17), "[...] este método se fundamenta na aplicação da teoria estatística da probabilidade e constitui importante auxílio para a investigação em ciências sociais […]", o método estatístico passa a se caracterizar por razoável grau de precisão, o que o torna bastante aceito por parte dos pesquisadores com preocupações de ordem quantitativa.

4.3 A definição da metodologia

A pesquisa que gerou este texto foi construída a partir dos seguintes procedimentos: i) delimitação da amostra; ii) audição na íntegra de todos os inquéritos selecionados para compor a amostra de linguagem real usada na pesquisa; iii) definição dos grupos de fatores que podem exercer pressão sobre o uso das variantes investigadas em uma determinada regra variável; iv) codificação de todos os fatores considerados na pesquisa; v) quantificação das informações resultantes da análise, com o auxílio do programa *RStudio*; vi) interpretação dos resultados estatísticos à luz da teoria da Sociolinguística Variacionista.

Possuímos, então, um texto de caráter descritivo e quantiqualitativo e foi desenvolvido a partir dos pressupostos teórico-metodológicos da Sociolinguística Variacionista (Labov, 2008; Weinreich; Labov; Herzog, 2006 [1968]). O caráter quantitativo foi defendido em razão de considerar-se que os resultados dessa pesquisa "pode ser quantificável, o que significa traduzir em números opiniões e informações para classificá-las e analisá-las" (Prodanov; Freitas, 2013, p. 69). Já no que concerne ao caráter descritivo, nosso texto "visa descrever as características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis" (Prodanov; Freitas, 2013, p. 52).

Além disso, a pesquisa que aqui trazemos pode, também, ser defendida como qualitativa, em conformidade com Silveira e Córdova (2009), pois, nesse tipo de abordagem, o pesquisador preocupa-se com aspectos da realidade que não podem ser quantificados e, então, passam a ser explicados a partir de uma dinâmica subjetiva que envolve as relações sociais e a compreensão do pesquisador.

Para este recorte, utilizamo-nos de um *corpus* formado pelas 25 capitais das cinco regiões do Brasil, a saber: Região Nordeste (Aracaju/SE, Fortaleza/CE, João Pessoa/PB, Maceió/AL, Natal/RN, Recife/PE, Salvador/BA, Teresina/PI, São Luís/MA); Região Norte (Belém/PA, Manaus/AM, Boa Vista/RR, Porto Velho/RO, Rio Branco/AC, Macapá/AP); Região Sul (Curitiba/PR, Florianópolis/SC, Porto Alegre/RS); Região

Sudeste (São Paulo/SP, Rio de Janeiro/RJ, Belo Horizonte/MG, Vitória/ES); Região Centro-Oeste (Goiás/GO, Cuiabá/MT, Campo Grande/MS). Dessarte, o estudo totalizou 200 informantes.

Com base na literatura pertinente, levantamos algumas hipóteses iniciais, tanto para o comportamento das variantes **botar** e **colocar**, bem como para as variáveis sociais controladas na amostra deste trabalho e foram elaboradas a partir dos trabalhos que nos serviram como norte: 1) o uso do verbo **botar** é mais frequente que a realização do verbo **colocar** na amostra analisada; 2) os informantes do sexo masculino, por serem menos conservadores que as **mulheres**, favorecem o uso de **botar**, enquanto as informantes do sexo feminino, que tendem a ser mais conservadoras que os **homens**, favorecem o uso do verbo **colocar**; 3) os falantes mais velhos (**faixa etária – 45 a 60 anos**) tendem a favorecer o verbo **botar**, enquanto os mais jovens (**faixa etária** – 18 a 30 anos) beneficiam a variante **colocar**; 4) os informantes menos escolarizados, nível fundamental, favorecem o uso do verbo **botar**, contrapondo-se aos informantes com maior escolaridade, nível universitário, que inibem o uso do verbo **botar**; 5) a região Nordeste favorece o uso do verbo **botar**; 6) a região Norte favorece o uso do verbo **botar**; 7) a região Sul inibe o uso do verbo **botar**; 8) as regiões Sudeste e Centro-Oeste inibem o uso do verbo **botar** e favorecem o uso do **colocar**; 9) a variação na comunidade estudada trata-se de um caso de variação estável.

O *script*[19] que seguiremos está apoiado em Oushiro (2022) que propõe as melhores etapas e testes estatísticos para a análise variável de uma variável nominal dos efeitos aleatórios, ou seja, os informantes.

5 Apresentação e discussão dos resultados

Os dados coletados somam um total de 3.040 tokens/dados, sendo 48,5% de *botar* (1.473 dados) e 51,5% de **colocar** (1.567 dados). Esses resultados, a partir de

[19] Grosso modo, *script* é um texto que possui uma sequência de passos que o programa computacional vai interpretar para executar e apresentar os resultados obtidos.

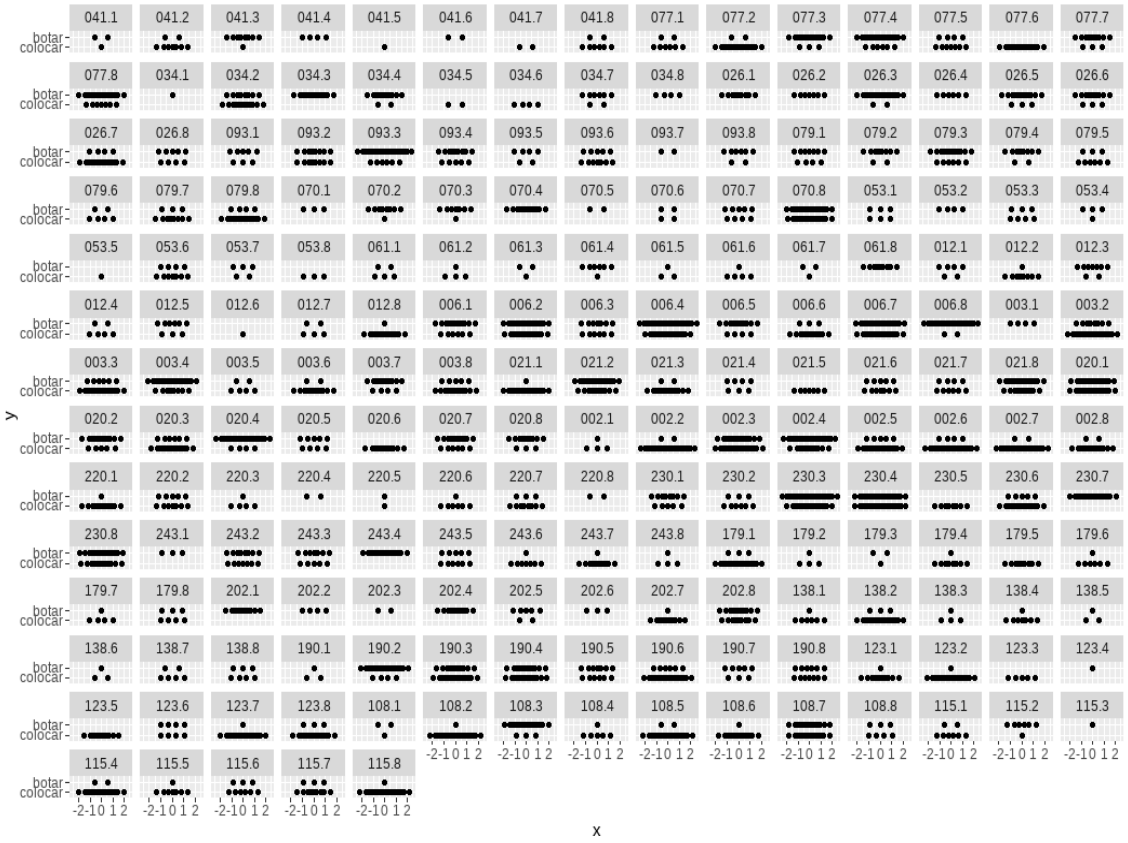
amostras de fala das capitais das cinco regiões juntas, se contrapõem aos resultados alcançados pelas pesquisas apresentados na revisão bibliográfica, realizadas apenas em algumas capitais das regiões do país, como em Lavor, Araújo e Viana (2018), que concluíram que o verbo **botar** apresenta frequência de 50,1% e **colocar** 49,9%; Lavor, Viana e Araújo (2019), que trouxeram **botar** com frequência de 53,80% e **colocar** 46,20%; Lavor, Vieira e Araújo (2019), com frequência de uso de 66,50% de **botar** e 33,50% de **colocar**; e, Lavor, Araújo e Pereira (2020) com uma frequência de 71,30% para **botar** e com 28,70% para **colocar**. Logo, esses trabalhos apresentam uma média de 60,55% para **botar** e 34,45% para **colocar**. E, ainda, adicionando os resultados desta pesquisa, com todas as capitais juntas, aos das pesquisas norteadoras deste trabalho, elencadas anteriormente, tem-se uma média de 59,36% para **botar** e 40,47% para **colocar**.

Assim, consideramos que há uma média de 69,36% de frequência para o verbo **botar**, contrapondo-se uma média de frequência de 40,47% para o verbo **colocar**, em dados provenientes do ALiB. No entanto, quando se trabalha com dados de todas as capitais juntas, os resultados estatísticos alcançados não apresentam o verbo **botar** como o mais frequente, havendo apenas uma diferença de 3%, o que nos surpreende, haja vista ser a primeira vez em que se apresenta a variante **colocar** como a mais frequente na fala do PB.

Após essas primeiras considerações, avançamos para a visualização da distribuição dos dados por informantes desta pesquisa, a partir do Gráfico 1.

O Gráfico 1 apresenta a distribuição das duas variantes por cada um dos 200 informantes da amostra, ou seja, o Gráfico 1 é composto por 200 minigráficos representando, cada um, a produção de fala de cada informante da amostra.

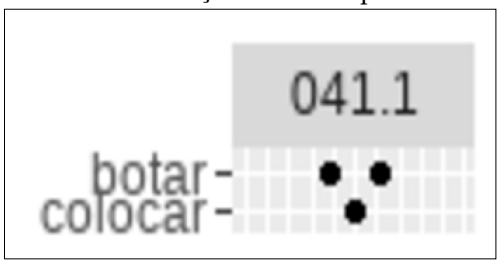
Gráfico 1 — Distribuição de dados por informante.



Fonte: elaborado pelos autores.

Por exemplo, na parte superior esquerda, temos o seguinte minigráfico, como apresentado pela Figura 1:

Figura 1 — Distribuição de dados por informante.



Fonte: elaborado pelos autores.

Na Figura 1, temos o informante 041.1 e os dados referentes à produção dos verbos em estudo. Lembramos que os números que vão de 001. a 200. correspondem a cada uma das localidades da amostra, enquanto os números 1 a 8, após o ponto, representam o tipo de informante (estratificado em idade, sexo e escolaridade), como podemos visualizar nos Anexos 1 e 2, neste trabalho. Na figura, visualizamos um

informante da cidade de Fortaleza (041), do sexo masculino, de faixa 1, com escolaridade fundamental.

Após esses esclarecimentos, vamos às análises propriamente ditas.

Ao realizarmos as primeiras análises descritivas, verificamos, com o χ^2 de Pearson[20], que somente a variável **sexo** apresentou um p-valor acima de 0,05, o que significa dizer que o **sexo** não tem correlação significativa com a variação de **botar** e **colocar** nesse banco de dados nas capitais brasileiras.

Esse resultado está em consonância com a pesquisa de Lavor, Viana e Araújo (2019), onde a variável não é relevante para a amostra com dados de diferentes localidades do ALiB. Todavia, esse resultado diverge da pesquisa de Lavor, Araújo e Viana (2018), Lavor, Vieira e Araújo (2019) e Lavor, Araújo e Pereira (2020), que consideraram a variável *sexo* como relevante para as amostras de fala analisadas em outras localidades do Brasil, com dados do ALiB.

Como Fischer (1958), em seu estudo sobre a pronúncia do sufixo inglês *-ing* formador de gerúndio, o fenômeno variável pode estar associado mais à **localidade** do falante e não propriamente ao fato de ser, o indivíduo, do *sexo* masculino ou feminino.

Continuando a análise estatística, com o χ^2 de Pearson abaixo de 0,05, tivemos as demais variáveis (*f*aixa etária, escolaridade, região e **localidade**) correlacionadas à variação. Assim, aplicamos a função vif() (pacote *car*) para calcular multicolinearidade[21] e não encontramos, embora saibamos que a variável região compreende a variável localidade, pois aquelas estão dentro desta, assim, ao calcularmos uma possível correlação entre estas, usamos a função pairs.panels()

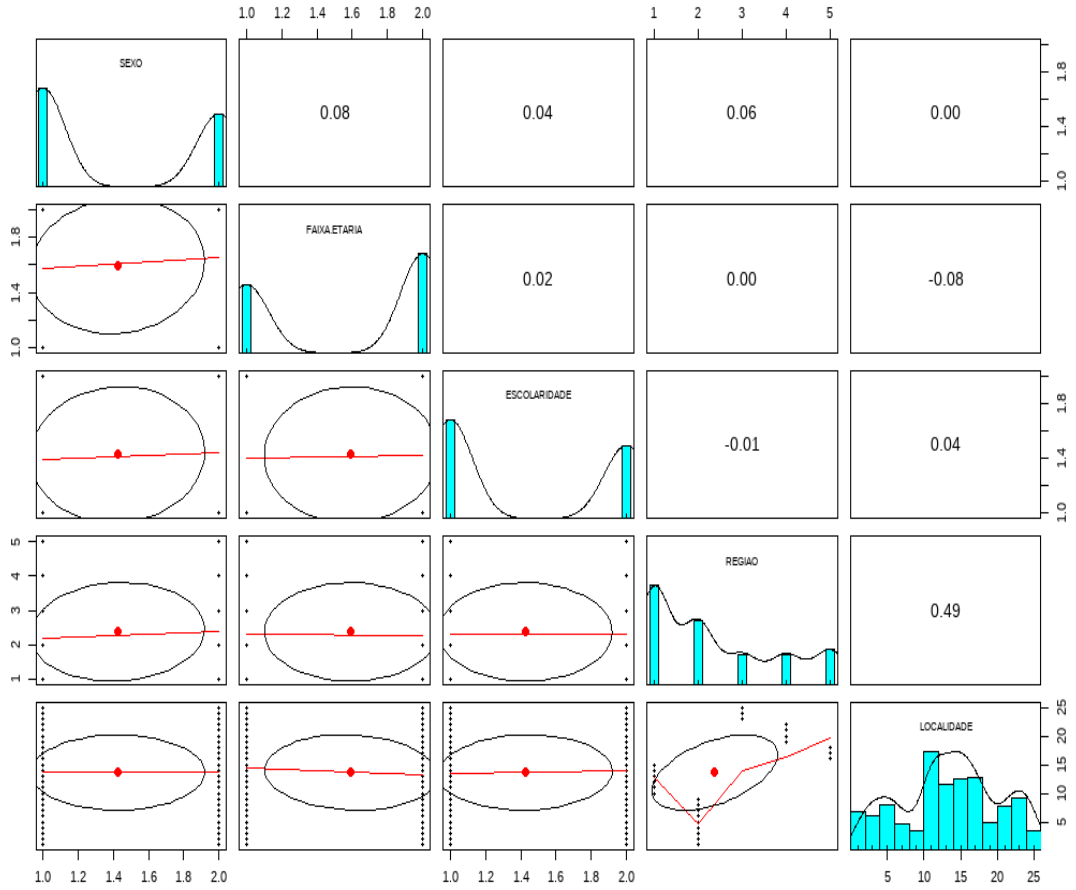
[20] O teste χ2 (qui-quadrado) de Pearson é um teste estatístico realizado em dados categóricos (1 ou 0, sim ou não) para avaliar o quão provável é que qualquer diferença observada entre os dados possa acontecer ao acaso. Esse teste é um dos mais conhecidos e é utilizado para analisar variáveis nominais (ou qualitativas), para determinar a existência ou não de independência entre essas variáveis.

[21] Chamamos de multicolinearidade a situação em que duas ou mais variáveis independentes em um determinado modelo de regressão encontram-se altamente correlacionadas, como por exemplo, *idade* é altamente correlacionada com *faixa etária*, pois ambas calculam a mesma medida.

(pacote *psych*) que tem como objetivo, gerar uma matriz de gráficos de dispersão que mostra as relações entre pares de variáveis em seus dados.

Nesse caso, encontramos um valor de 0.49 que já pode nos apresentar colinearidade, como pode ser verificado através do Gráfico 2.

Gráfico 2 — Colinearidade entre variáveis.



Fonte: elaborado pelos autores.

O Gráfico 2 representa a plotagem de uma matriz de correlação, que traz, na diagonal inferior, os gráficos de dispersão; no centro (quadros diagonais em azul), os histogramas dos dados da variável plotada e sua distribuição; e, na diagonal superior, os resultados da correlação de Pearson entre as variáveis selecionadas, que são os números que aqui nos interessam.

Diante disso, consideramos que há multicolinearidade quando o resultado numérico da correlação de Pearson entre variáveis for maior que 0,5 (os números no

Gráfico 2). Se encontramos uma alta multicolinearidade, isso pode levar a um erro de estimativa nos coeficientes de regressão. Esse fato nos chama atenção para as variáveis **localidade** e **região**.

Para as análises de regressão, atendendo, dessa maneira, ao pressuposto (*c*), realizamos regressões logísticas univariadas entre a variável predita (ou dependente) e as variáveis previsoras (ou independentes), para que pudéssemos já conhecer melhor os dados. Confirmamos a falta de correlação entre **sexo** e a variação entre **botar** e **colocar**. Contudo, continuamos deixando a variável para analisar uma possível interação com outras variáveis, como a variável faixa etária, por exemplo.

Dando prosseguimento, checamos a ortogonalidade[22] com a função with() (função base do R)[23] entre os dados, quando há dados em todas as variáveis, não deixando células de dados vazias, pudemos verificar, como era de esperar, grande falta de ortogonalidade entre as variáveis **localidade e região**, obviamente, porque há localidades que não estão inseridas somente em uma região. Nosso objetivo é verificar se os dados agrupados por **região** (Norte, Sul, Nordeste, Centro-Oeste, Sudeste) podem trazer melhores resultados estatísticos que os mesmos dados agrupados por **localidade** (Aracaju, Fortaleza, João Pessoa, Maceió, Natal, Recife, Salvador, São Luís, Teresina, Belém, Manaus, Boa Vista, Porto Velho, Rio Branco, Macapá, Curitiba, Florianópolis, Porto Alegre, São Paulo, Rio de Janeiro, Belo Horizonte, Vitória, Goiânia, Cuiabá, Campo Grande).

Passando a realizar modelos de regressão logística[24] multivariada, verificamos que modelos com as variáveis **localidade** e **região** não convergiam, e que modelos com a localidade (sem a região) possuíam os melhores índices de AIC (*Akaike Information*

[22] Ortogonalidade é um conceito que se refere à propriedade de dois vetores serem perpendiculares entre si, o que implica que o produto interno entre eles é igual a zero. De outro modo, em um contexto de banco de dados, dizemos que um conjunto de dados é ortogonal quando é possível realizar operações de forma independente, pois há dados em todas as células.

[23] Função, no programa estatístico, é um comando de cálculo.

[24] A regressão logística é um cálculo estatístico que calcula a probabilidade de ocorrência ou não de um determinado evento, como, por exemplo, o voto ou não voto, baseado em um conjunto de dados específicos com variáveis independentes.

Criterion, uma medida que permite comparar modelos de regressão), em comparação com modelos com a variável região (sem a localidade).

Após gerar esses modelos, utilizamos a função vif() (pacote *car*) para avaliar colinearidade entre as variáveis e pudemos checar que o modelo com a variável localidade trazia uma alta falta de ortogonalidade com o índice 24 (índices acima de 10 mostram que há colinearidade entre as variáveis do modelo, assumindo, então, que os coeficientes de regressão estão mal estimados devido à multicolinearidade), portanto, a variável **localidade** deve ser retirada dos modelos seguintes.

A partir da função lrm() (pacote *rms*), verificamos índices como o R^2 e o índice C (o índice de Concordância 'C'), que, conforme Hosmer Jr. e Lemeshow (2000), é a estatística mais reportada para regressões logísticas. Nessa regressão, o mod3 (**sexo**, **faixa etária**, **escolaridade** e **localidade**) apresentou os seguintes índices: R^2 de 0.169 e índice C de 0.707, dessa forma, um mod4 seria o melhor modelo inicial (**faixa etária**, **escolaridade e região**) para analisar as variáveis sem interação, embora os índices de R^2 e índice C permaneceram os mesmos, seguindo o Princípio da Navalha de Occam[25] selecionamos o modelo mais simples e explicativo possível: o mod4.

Partindo desse modelo, checamos possíveis interações entre as variáveis em modelos de regressão logística, entre as variáveis do mod3 (**sexo, faixa etária, escolaridade e região**), aplicando a função drop1() (pacote *stats*) que, verifica, em modelos de regressão logística, a significância de cada variável previsora no modelo e se alguma dessas variáveis deve ser descartada, o que nos fez descartar **sexo** e as interações não significativas: **sexo** x **escolaridade** e **faixa etária** x escolaridade.

Desta feita, realizamos, ainda, uma verificação de infracionamento dos parâmetros de previsão do modelo, a partir de um "sobreajuste" (= *overfitting*). Realizamos esse teste por meio da função validate() (do pacote *rms*), função que se utiliza apenas de modelos sem interação. Ao final, o teste informa quantas variáveis

[25] Princípio da Navalha de Occam é também chamado de princípio da economia, é um princípio de investigação que preza pela maior parcimônia em termos de complexidade: o simples é melhor.

foram selecionadas, quantas vezes e qual o melhor modelo sem inflação dos parâmetros, que, em nosso modelo sem interação, a variável sexo foi excluída.

Assim, nosso melhor modelo conteve as variáveis **faixa etária, escolaridade e região.** E como modelo com interação, um modelo com as variáveis sexo, faixa etária, escolaridade e região. Esclarecemos ao leitor que a insistência de uso da variável *sexo* se dá pela possível interação que viesse a ter e ser destaque nesta análise.

Assim, seguindo a recomendação de checagem de pressupostos para as regressões logísticas (Oushiro, 2022), verificamos o seguinte:

a) Quanto à relação entre o *logit*[26] e as variáveis previsoras numéricas, é uma relação linear?

b) Há multicolinearidade entre as variáveis previsoras?

c) Há independência entre as observações/dados?

Atendendo o pressuposto (*a*), não temos variáveis previsoras numéricas, então, não foi preciso realizar qualquer teste, pois o pressuposto não foi violado. Quanto ao pressuposto (*b*), encontramos, sim, presença de multicolinearidade entre as variáveis ***região*** e ***localidade*** (0,44), como já vimos no Gráfico 2, e, para atender ao pressuposto, retiramos a variável **localidade** do modelo de regressão.

Por último, e muito importante, verificamos o pressuposto (*c*) que avalia a independência entre os dados, sendo que, em se tratando de dados linguísticos, dificilmente encontramos independência entre eles, pois são, em sua grande parte, inúmeros dados de um mesmo indivíduo. Para resolvermos esse problema e atender ao pressuposto, acrescentamos uma variável aleatória: a variável **informante**, que nos fez chegar a um modelo de regressão logística com efeitos mistos (Oushiro, 2022).

Dessa maneira, chegamos aos modelos finais de regressão logística: um modelo sem interação (mod4)[27], um modelo com interação (modint). Para a modelagem dos

[26] O logit pode ser compreendido, aqui, como a variável dependente de ordem categórica, que assume apenas dois valores como sucesso ou fracasso, sim ou não, 1 ou 0.

[27] A nomenclatura "mod4" e "modint" significam tão somente o "modelo número 4" e o "modelo com interação" desta análise.

efeitos mistos, utilizamos a função glmer()[28] que depende dos pacotes *lme4* e *lmerTest*.[29] Utilizamos os dois modelos finais, **com** e **sem interação** (modint e mod4), que já havíamos selecionado e acrescentamos, em ambos, a variável aleatória **informante**. Rodamos os modelos para podermos selecionar o que fosse estatisticamente significativo; assim, chegamos a outros dois modelos de regressão com efeitos mistos.

Para que pudéssemos selecionar, dentre esses, o melhor modelo, levamos em consideração os índices de AIC e BIC,[30] além do princípio da navalha de Occam, que nos revelaram como melhor modelo o mod.glmer, que foi o modelo com as variáveis **faixa etária, escolaridade** e **região**, juntamente à variável aleatória **informante**, sem interação. Vejamos o resumo do modelo selecionado a partir da Tabela 1.

Explicando a Tabela 1, visualizamos, na primeira coluna, os preditores, ou seja, as variáveis **faixa etária, escolaridad***e* e **região** e seus respectivos fatores do modelo; na segunda coluna, visualizamos as estimativas em *logodds*; na terceira, temos os intervalos de confiança para cada fator; e, na quarta e última coluna da tabela, os p-valores para cada estimativa, sendo os destacados em negrito, aqueles que possuem significância estatística, ou seja, p-valor < 0,05.

No final da tabela, temos os seguintes índices do modelo: a variância do modelo, representado pela letra grega sigma ao quadrado (σ^2); a variância gerada pelo informante, representada pela letra grega tau (τ_{00}) e o nome da variável aleatória (INF); o coeficiente de correlação intraclasse (*Intraclass Correlation Coefficient - ICC*), que mede a correlação entre amostras de avaliações entre dois ou mais avaliadores, no caso, dos

[28] Nome da função/comando estatístico que realiza o cálculo da regressão logística no programa *RStudio.*
[29] Chamamos de 'pacotes' um script de comandos que são utilizados por um programa estatístico para realizar um cálculo específico, como uma regressão logística com ou sem interação, ou ainda, uma regressão logística com variáveis aleatórias.
[30] De forma simples, o índice AIC (Akaike Information Criterion) procura estimar a qualidade relativa de modelos estatísticos de um determinado conjunto de dados e ajuda a selecionar aqueles modelos mais explicativos ou com menor perda de informação. Já o BIC (Bayesian Information Criterion), por outro lado, analisa mais rigorosamente a complexidade dos modelos de regressão, o que ajuda a selecionar modelos de regressão mais simples.

informantes; o número de "classes" da variável aleatória, ou seja, a quantidade de informantes (N INF) da amostra coletada.

Tabela 1 — Estimativas dos parâmetros do modelo (regressão logística binomial, modelo linear generalizado com efeitos mistos) da realização de botar e colocar (N = 3.040 – intercepto: -0.3721)[31].

	VD		
Predictors	*Log-odds*[32]	*CI*	*p*
(Intercept)	-0.3721	-0.16 – 0.90	0.170
FA IXA.ETARIA2	0.8710	-1.32 – -0.42	**<0.001**
ESCOLARIDADEsuperior	-0.9188	0.46 – 1.36	**<0.001**
REGIAONordeste	0.9392	-1.52 – -0.36	**0.001**
REGIAOCentro_Oeste	-1.0242	0.24 – 1.81	**0.011**
REGIAOSudeste	0.0228	-0.73 – 0.69	0.950
REGIAOSul	0.2602	-1.04 – 0.51	0.510
Random Effects			
σ^2	3.29		
$\tau_{00\ INF}$	1.85		
ICC	0.36		
N INF	200		
Observations	3040		
Marginal R^2 / Conditional R^2	0.131 / 0.444		
AIC	3259.375		
Modelo: VD ~ FAIXA.ETARIA + ESCOLARIDADE + REGIAO, data = dados, random = ~1 | INF			

Fonte: elaborada pelos autores.

[31] Na estatística, o intercepto ou, em inglês, intercept representa o ponto onde a linha de regressão (no eixo X) cruza o eixo Y em um gráfico. É um conceito fundamental na estatística e na análise de dados, especialmente em modelos de regressão linear.

[32] *Log-odds* é a medida em logaritmo das razões de chances de um evento acontecer ou não, número que vai do menos infinito ao mais infinito, passando pelo zero como ponto neutro.

Na linha logo abaixo, temos o número total de observações/dados (3.040); o R^2 marginal e R^2 condicional, que trazem o quão explicativo é o modelo, sendo o primeiro, somente com as variáveis fixas (13% de explicação da variação pelo modelo), e o segundo, com as variáveis fixas e aleatórias (44% de poder de explicação do modelo); e, por último, mas não menos importante, o AIC, já explicado anteriormente. Na última linha, há a sintaxe que utilizamos para calcular a regressão logística.

Para interpretar os dados, lembramos a célebre citação de Scherre e Naro (2004, p. 164), quando nos dizem que, para além dos dados da língua, de uma metodologia específica, através da estatística, o trabalho do sociolinguista não está concluído, pois "os resultados numéricos obtidos pelo programa só têm valor estatístico. O seu valor linguístico é atribuído e interpretado pelo linguista". Vejamos, assim, o que podemos compreender desse fenômeno variável.

Na regressão logística, o valor de *log-odds* é interpretado, conforme Gries (2019), dentre as variáveis com p-valor significativo, as estimativas positivas se correlacionam positivamente (favorecendo) com a variável de aplicação, enquanto as estimativas negativas se relacionam negativamente (desfavorecendo) com a variante de aplicação. Nos *log-odds*, o zero é o ponto neutro (Gries, 2019, p. 265).

Para chegar à estimativa de probabilidade de o fenômeno ocorrer dada uma determinada variável, é necessário somar o valor do *intercept* ao valor do coeficiente angular. A faixa etária 2, portanto, têm -0.3721 + 0.8710 = 0.4989 *log-odds* de emprego de **botar**. Aplicando a função ilogit()[33] a 0.4989 para transformar a medida de *logodds* em probabilidade, encontramos 0.6222008. O cálculo indica que a probabilidade de as pessoas da **faixa 2** empregarem o verbo **botar** em relação a **colocar** é de é 62,2%. Assim, esses resultados estatísticos corroboram a hipótese inicial de que os falantes mais velhos (*faixa etária – 45 a 60 anos*) tendem a favorecer o verbo **botar**, enquanto os mais jovens (*faixa etária* – 18 a 30 anos) beneficiam a variante **colocar**.

[33] Função criada por Gelman e Hill (2007, p. 80).

Dessa maneira, os indivíduos que estão na **faixa etária 2** têm probabilidade de 62,2% de uso de **botar**, assim como os residentes no **Nordeste** são mais propensos a usarem o verbo **botar** em lugar de **colocar** em sua fala com 63,8%[34] de probabilidade. Isso apresenta uma variação estável entre **botar** e **colocar**, pois, segundo a teoria variacionista, para podermos vislumbrar uma possível mudança em curso, a faixa etária dos mais jovens deve trazer a influência positiva sobre a variante de aplicação, o que não aconteceu, já que os mais velhos favorecem o verbo **botar**. Dessarte, esse resultado estatístico confirma a hipótese inicial de que a variação na comunidade estudada se trata de um caso de variação estável.

De influência negativa, ou seja, desfavorecedora, a probabilidade de indivíduos com **nível superior** usar **botar** é de 21,7%;[35] já os residentes no **Centro-Oeste** são menos propensos ao uso de **botar** com probabilidade de uso de 19,8%.[36] Logo, esses resultados estatísticos corroboram a hipótese de que os informantes menos escolarizados, nível fundamental, favorecem o uso do verbo **botar**, contrapondo-se aos informantes com maior escolaridade, nível universitário, que inibem o uso do verbo **botar** e, também, a hipótese de que as regiões Sudeste e Centro-Oeste inibem o uso do verbo **botar** favorecendo o uso do **colocar**.

Com esses resultados, pudemos traçar um paralelo entre o nosso texto e o de Lavor, Araújo e Viana (2018) que, com dados de informantes de Alagoas, Ceará e Piauí, trouxe variável **sexo** como relevante, sendo os homens aliados de **botar** (na pesquisa, PR de 0,624), o que não ocorreu com nossos dados; contudo, corrobora o resultado

[34] Cálculo que utilizamos para o fator **Região Nordeste**: intercept (-0.3721) + estimativa de nordeste (0.9392) = 0.5671. Calculando a probabilidade com a função ilogit(0.5671), chegamos a 0.6380937, ou seja, 63,8%.

[35] Cálculo que utilizamos para o fator **nível superior**: intercept (-0.3721) + estimativa de ensino superior (-0.9088) = -1.2809. Calculando a probabilidade com a função ilogit(-1.2809), chegamos a 0.2173971, ou seja, 21,7%.

[36] Cálculo que utilizamos para o fator **Região Centro-Oeste**: intercept (-0.3721) + estimativa de centro_oeste (-1.0242) = -1.3963. Calculando a probabilidade com a função ilogit(-1.3963) chegamos a 0.1984039, ou seja, 19,8%.

correlacionado à **faixa etária** 2, sendo, na pesquisa citada, favorecedora de **botar** (na pesquisa, PR de 0,650).

A não influência da variável **sexo** nesse processo nos faz refletir que sua influência histórica nos fenômenos variáveis linguísticos não tem sido mais verificada quando se trata de bancos de dados de fala mais recentes, mais precisamente, final do século XX e início do século XXI, com a ampliação da comunicação global e do empoderamento feminino. Como bem discorreu Labov (2008), as mulheres buscavam o uso das formas linguísticas mais prestigiadas, na perspectiva de se contrapor às condições de inferioridade nas quais foram historicamente colocadas.

Contudo, sabemos que tanto a teoria variacionista, quanto os próprios estudos trazem importante relevância para a variável **sexo**, haja vista ela representar importantes movimentos sociais nas comunidades de fala. Nesta pesquisa, essa variável mostrou-se irrelevante; fato que pode sugerir uma possível equiparação linguística. Quanto à **escolaridade** dos indivíduos também não se mostrou relevante, como pudemos verificar em Lavor, Araújo e Pereira (2020) que resenhamos neste texto. A seguir, vejamos nossas considerações finais para este estudo.

6 Considerações finais

O objetivo geral desta pesquisa foi verificar o processo variável em que se encontram os verbos **botar** e **colocar** no PB, tendo como *corpus* de análise os dados de fala constantes em 25 capitais brasileiras a partir do banco de dados do ALiB. Além desse, definiu-se como objetivos específicos, verificar qual variante em concorrência obteve mais resultados; em qual localidade regional a variante **botar** é mais frequente; em qual capital do Brasil a variante **botar** é mais frequente; qual ou quais grupos de fatores controlados favorece o uso do verbo botar; e ainda, se a concorrência entre os verbos **botar** e **colocar** representa um caso de variação estável ou um caso de mudança em progresso.

Respondendo a essas perguntas, temos que a variante **colocar** obteve mais dados (1.567 - 51,5%); **botar** foi mais frequente na região Norte com 571 de 1.165 dados para a região; assim, Manaus foi a capital do Brasil que apresentou mais dados da variante **botar** com 232 dados; e os grupos de fatores controlados que favoreceram o uso do verbo de aplicação foram **faixa etária** e **região**; e a variação entre os verbos em estudo representa um caso de variação estável.

Diante do exposto, podemos constatar que os resultados trazem confirmações e contradições. Nessas contradições está a não relevância estatística da variável **sexo**, variável muito cara à análise Sociolinguística e o verbo **colocar** como o mais frequente. Para isso, inferimos que, embora o banco de dados possua dados que perpassam muitos anos de coleta, o comportamento entre os sexos tem mudado nesses anos.

Romaine (2003) faz importantes reflexões sobre as relações de linguagem entre homens e mulheres. Segundo a autora, embora ainda nos tempos atuais a relação social entre homens e mulheres não seja equivalente, essa desigualdade permanece relevante para as mulheres, pois estas buscam cada vez mais a igualdade com os homens em vários aspectos, especialmente, os aspectos educacional e econômico, fato que é resultado do movimento das mulheres por reconhecimento. Assim, esses dados nos mostram que o sexo não é uma variável relevante.

Além disso, a pesquisa em que se baseia este artigo vem apresentar a preferência dos nordestinos pela variante **botar** em detrimento da comunidade da região Centro-Oeste que prefere **colocar**.

Por fim, essas conclusões servirão para a montagem da grande fotografia sociolinguística sobre os verbos **botar** e **colocar** no PB. Sabemos que ainda há muito o que ser feito, pois, mesmo analisando 200 informantes, sendo oito por capital, pesquisas que tragam mais informantes por capital ou mesmo por estado ou região, certamente trarão mais conclusões. Ou, ainda, novas pesquisas que abarquem a testagem de mais variáveis sociais e linguísticas.

Referências

BATORÉO, H. J.; CASADINHO, M. Botar as mãos na massa? Estudo Cognitivo da produtividade lexical do verbo 'botar' no PE e PB". *In:* SIMPÓSIO MUNDIAL DE ESTUDOS DE LÍNGUA PORTUGUESA, 2., 2009, Évora. **Anais Eletrônicos [...]**. Évora: Universidade de Évora, 2009. p. 37-55. Disponível em: http://www.simelp2009.uevora.pt/pdf/slg4/04.pdf

BECHARA, E. **Moderna gramática portuguesa**. 38. ed. rev. ampl. Rio de Janeiro: Nova Fronteira, 2015.

CAMACHO, R. G. Sociolinguística parte II. *In:* MUSSALIM, F; BENTES, A. C. (org.). **Introdução à linguística**: domínios e fronteiras. 6. ed. São Paulo: Cortez, 2006. p. 49-75.

CARDOSO, S. A. M. **Geolinguística**: tradição e modernidade. São Paulo: Parábola, 2010.

CARDOSO, S. A. M. O Atlas Linguístico do Brasil: uma visão crítica dos caminhos seguidos e perseguidos. Comunicação apresentada no **II Congresso Internacional de Linguística Histórica**, Universidade de São Paulo, São Paulo, fev. 2012.

COELHO, I. L.; GÖRSKI, E. M.; SOUZA, C. M. N. de; MAY, G. H. **Para conhecer Sociolinguística.** São Paulo: Editora Contexto, 2015.

FERREIRA, M. **Aprender e praticar gramática**. São Paulo: FTD, 2003.

FISHER, J. L. Social influences on the choice of linguistic variant. **Word**, New York, n.14, p. 47-56, 1958. DOI https://doi.org/10.1080/00437956.1958.11659655

GELMAN, A.; HILL, J. **Data Analysis Using Regression and Multilevel/Hierarchical Models**. Cambridge: Cambridge University Press, 2007. DOI https://doi.org/10.32614/CRAN.package.arm

GIL, A. C. **Métodos e Técnicas de Pesquisa Social**. São Paulo: Atlas, 2008.

GRIES, S. T. **Estatística com R para a Linguística**: uma introdução prática. Tradução de Heliana R. Mello et al. Belo Horizonte: FALE/UFMG, 2019.

GUY, G. R.; ZILLES, A. **Sociolinguística Quantitativa:** instrumental de análise. São Paulo: Editora Parábola, 2007.

HOSMER JR., David W.; LEMESHOW, Stanley. **Applied logistic regression**. 2. ed. New York: Wiley-Interscience Publication, 2000. DOI https://doi.org/10.1002/0471722146

LABOV, W. **Padrões Sociolinguísticos.** Tradução de Marcos Bagno, Maria Marta Pereira Scherre, Caroline R. Cardoso. São Paulo: Parábola Editorial, 2008.

LABOV, W. The intersection of sex and social class in the course of linguistic change. **Language Variation and Change**, [*s. l.*], v. 2, n. 2, p. 205–254, 1990. DOI https://doi.org/10.1017/S0954394500000338

LAVOR, C. M. A. de; ARAÚJO, A. A. de; VIANA, R. B. de M. Uma fotografia sociolinguística dos verbos *botar*, *colocar* e *pôr* em Alagoas, Ceará e Piauí a partir de dados do ALiB. **Polifonia**, Cuiabá, v. 25, n. 37, p. 171-310, jan./abr. 2018.

LAVOR, C. M. A. de; VIANA, R. B. de M.; ARAÚJO, A. A. de. A variação dos verbos botar e colocar no Ceará em amostra do Atlas Linguístico do Brasil. **Polifonia**, Cuiabá, v. 26, n. 43, p. 01 – 357, jul./set., 2019. Disponível em: https://periodicoscientificos.ufmt.br/ojs/index.php/polifonia/article/view/7999/pdf

LAVOR, C. M. A. de; VIEIRA; V. da S.; ARAÚJO, A. A. de. Os verbos botar e colocar em Salvador e Porto Alegre: um estudo variacionista nos dados do Atlas Linguístico do Brasil. **Miguilim**, Crato-CE, v. 8, n. 3, p. 493-511, set./dez., 2019. DOI https://doi.org/10.47295/mgren.v8i3.1996

LAVOR, C. M. A. de; ARAÚJO, A. A. de; PEREIRA, M. L. de S. Os verbos botar e colocar no estado do Maranhão em dados do ALiB: uma pesquisa variacionista. **Leitura**, Maceió, v. 1, n. 66, p. 146 – 164, set – dez, 2020. DOI https://doi.org/10.28998/2317-9945.202066.146-164

MOTA, J. A.; CARDOSO, S. A. M. A construção de um Atlas Linguístico do Brasil: o percurso do ALiB. **Signum**: Estudos da Linguagem, Londrina, v. 12, n. 1, p. 237-256, jul. 2009. DOI https://doi.org/10.5433/2237-4876.2009v12n1p237

OUSHIRO, L. **Introdução à Estatística para Linguistas**. Campinas: Editora da Abralin, 2022. *E-book*. DOI https://doi.org/10.25189/9788568990209

OUSHIRO, L. Tratamento de Dados com o R para Análises Sociolinguísticas. *In:* FREITAG, R. M. K. (ed.). **Metodologia de Coleta e Manipulação de Dados em Sociolinguística**. São Paulo: Blücher, 2014. p. 133–176. DOI https://doi.org/10.5151/BlucherOA-MCMDS-10cap

PRODANOV, C. C.; FREITAS, E. C. **Metodologia do trabalho científico [recurso eletrônico]**: métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: Freevale, 2013.

R CORE TEAM. **R**: A language and environment for statistical computing. Viena R Foundation for Statistical Computing, 2021. Disponível em: https://www.r-project.org/

RSTUDIO TEAM. **RStudio**: Integrated Development for RBostonRStudio, PBC, 2020. Disponível em: http://www.rstudio.com. Acesso em: 7 jan. 2024

ROCHA LIMA, C. H. da. **Gramática normativa da língua portuguesa.** 31. ed. Rio de Janeiro: José Olympio, 1992.

ROMAINE, S. Variation in Language and Gender. *In*: HOLMES, J.; MEYERHOFF, M. (org.). **The Handbook of Language and Gender**. Oxford: Blackwell, 2003. p. 69–97.

SANKOFF, D.; TAGLIAMONTE, S. A.; SMITH, E. **Goldvarb X**: A variable rule application for Macintosh and Windows. Department of Linguistics, University of Toronto, 2005. Disponível em: http://individual.utoronto.ca/tagliamonte/goldvarb.html. Acesso em: 09 fev. 2023.

SCHERRE, M. M. P. **Introdução ao Pacote VARBRUL para microcomputadores**. Brasília: UNB, 1993.

SCHERRE, M. M. P.; NARO, A. J. Análise quantitativa e tópicos de interpretação do Varbrul. *In*: MOLLICA, M. de M.; BRAGA, M. L. (org.). **Introdução à sociolinguística**: o tratamento da variação. 2. ed. São Paulo: Contexto, 2004. p. 147–178.

SILVEIRA, D. T.; CÓRDOVA, F. P. A pesquisa científica. *In*: GERHARD, T. E.; SILVEIRA, D. T. (org.). **Métodos de Pesquisa**. Porto Alegre: Editora da UFRGS, 2009. p. 31-42.

WEINREICH, U.; LABOV, W.; HERZOG, M. I. **Fundamentos empíricos para uma teoria da mudança linguística**. Tradução Marcos Bagno. São Paulo: Parábola Editorial, 2006.

Anexos

ANEXO 1 - Características dos informantes do ALiB

Todos os informantes do ALiB são nascidos na localidade examinada e apresentam as seguintes características:

Idade:
Faixa 1 – 18 a 30 anos
Faixa 2 – 45 a 60 anos
Escolaridade:
Fundamental
Graduado

Informantes das capitais (8 por capital)
Exemplo: Fortaleza (041)
041.1 – Homem, faixa 1, fundamental;
041.2 – Mulher, faixa 1, fundamental;
041.3 – Homem, faixa 2, fundamental;
041.4 – Mulher, faixa 2, fundamental;
041.5 – Homem, faixa 1, graduado;
041.6 – Mulher, faixa 1, graduado;
041.7 – Homem, faixa 2, graduado;
041.8 – Mulher, faixa 2, graduado

Informantes do interior do Brasil (4 por município)
Exemplo: Camocim (039) (interior do estado do Ceará)
039.1 – Homem, faixa 1, fundamental;
039.2 – Mulher, faixa 1, fundamental;
039.3 – Homem, faixa 2, fundamental;
039.4 – Mulher, faixa 2, fundamental

ANEXO 2 - Códigos das capitais do ALiB

Código	Capital
002	Macapá-AP
003	Boa Vista-RR
006	Manaus-AM
012	Belém-PA
020	Rio Branco-AC
021	Porto Velho-RO
026	São Luis-MA
034	Teresina-PI
041	Fortaleza-CE
053	Natal-RN
061	João Pessoa-PB
070	Recife-PE
077	Maceió-AL
079	Aracajú-SE

093	Salvador-BA
108	Cuiabá-MT
115	Campo Grande-MS
123	Goiânia-GO
138	Belo Horizonte-MG
179	São Paulo-SP
190	Vitória-ES
202	Rio de Janeiro-RJ
220	Curitiba-PR
230	Florianópolis-SC
243	Porto Alegre-RS