Do Varbrul ao R: uma reanálise das vogais médias pretônicas em Fortaleza-CE

Ronaldo Mangueira Lima Júnior, Pablo Arantes, Guilherme Garcia, Luciana Lucente, Renata Passetti

https://doi.org/10.1590/SciELOPreprints.15222

Submetido em: 2026-02-25
Postado em: 2026-05-28 (versão 2)
(AAAA-MM-DD)

Justificativa da versão: Detalhes metodológicos e analíticos foram acrescentados a pedido de pareceristas externos a fim de garantir replicabilidade.

Categoria da Contribuição: Protocolo de Relato Registrado - Envio inicial

Do Varbrul ao R: uma reanálise das vogais médias pretônicas em Fortaleza-CE

From Varbrul to R: a reanalysis of pretonic mid vowels in Fortaleza-CE

Ronaldo Lima Jr.
Doutor em Linguística; Universidade de Brasília; Departamento de Linguística, Português e Línguas Clássicas; Brasília-DF; ronaldo.junior@unb.br; https://orcid.org/0000-0002-8610-0306
Pablo Arantes
Doutor em Linguística; Universidade Federal de São Carlos; Departamento de Letras; São Carlos-SP; pabloarantes@ufscar.br; https://orcid.org/0000-0001-9707-8493
Guilherme Garcia
Doutor em Linguística; Université Laval; Département de Langes, Linguistique et Traduction; Québec-QC, Canadá; guilherme.garcia@lli.ulaval.ca; https://orcid.org/0000-0003-1412-3856
Luciana Lucente
Doutora em Linguística; Universidade Federal de Minas Gerais; Faculdade de Letras; Belo Horizonte-MG; lucente@ufmg.br; https://orcid.org/0000-0001-6325-0531
Renata Passetti
Doutora em Linguística; Pesquisadora independente; Sumaré-SP; re.passetti@gmail.com; https://orcid.org/0000-0002-1547-2831

Resumo

Propomos reanalisar os dados da tese *As vogais médias pretônicas no falar popular de Fortaleza: uma abordagem variacionista* (Araújo, 2007), que investigou a realização das vogais pretônicas sob a perspectiva da Sociolinguística Variacionista. O estudo original utilizou o VARBRUL para modelar o efeito de variáveis linguísticas e sociais sobre os processos de alteamento, abaixamento e manutenção das vogais. Nossa proposta é utilizar modelos de regressão logística de efeitos mistos no R, incorporando avanços metodológicos da análise estatística, como a inclusão de efeitos aleatórios para informante e item lexical. Pretendemos verificar a robustez dos resultados originais frente a técnicas mais recentes, bem como discutir as implicações da modelagem mista para a interpretação de fenômenos variáveis. A reanálise busca contribuir para a promoção de práticas de reprodutibilidade na Linguística, oferecendo um exemplo de replicação com dados reais e disponibilização transparente do código e dos resultados.

Palavras-chave: fonética, R, sociolinguística, vogais

Abstract

We propose to reanalyze the data from the thesis "Pretonic Mid Vowels in the Popular Speech of Fortaleza: A Variationist Approach" (Araújo, 2007), which investigated the

realization of pretonic vowels from the perspective of Variationist Sociolinguistics. The original study used VARBRUL to model the effect of linguistic and social variables on the processes of raising, lowering, and maintaining vowels. Our proposal is to use mixed effects logistic regression models in R, incorporating methodological advances in statistical analysis, such as the inclusion of random effects for informants and lexical items. We intend to verify the robustness of the original results against more recent techniques, as well as discuss the implications of mixed modeling for the interpretation of variable phenomena. The reanalysis seeks to contribute to the promotion of reproducibility practices in Linguistics, offering an example of replication with real data and transparent availability of the code and results.

Keywords: phonetics, R, sociolinguistics, vowels

Introdução

Este trabalho propõe reanalisar os dados originais da tese de Araújo (2007), que investigou a realização das vogais médias pretônicas no falar popular de Fortaleza, Ceará, à luz da Sociolinguística Variacionista. O estudo original utilizou o programa VARBRUL (*Variable Rule Program*), modelo clássico da Sociolinguística Quantitativa desenvolvido por David Sankoff no início dos anos 1970, destinado à modelagem estatística de fenômenos linguísticos variáveis (Cedergren; Sankoff, 1974).

A escolha de Araújo (2007) pelo modelamento estatístico no software VARBRUL reflete a prática majoritária dos estudos sociolinguísticos variacionistas da época, cuja literatura evidencia uma tradição de análises quantitativas conduzidas com essa ferramenta até o final da primeira década dos anos 2000 (Oushiro, 2022; Tagliamonte, 2011). A combinação entre recursos que permitiam a condução de análises inferenciais de dados variáveis e uma interface amigável e de uso intuitivo contribuiu para a ampla adoção do VARBRUL nesses estudos (Oushiro, 2022).

No entanto, a partir do final dos anos 2000, consolida-se uma mudança de paradigma nos métodos quantitativos da pesquisa variacionista, motivada por críticas às limitações das análises realizadas pelo VARBRUL e pela adoção de novas ferramentas estatísticas consideradas mais robustas para a modelagem de fenômenos linguísticos variáveis (Torres Vieira, 2022). O modelo mais recente do VARBRUL permite a condução de um único tipo de análise, a regressão logística, adequado para variáveis de resposta binárias (Lima Jr; Garcia, 2021). Além disso, as variáveis preditoras devem ser necessariamente nominais. Por isso, nos estudos sociolinguísticos que utilizam essa ferramenta para modelagem estatística dos dados, variáveis originalmente contínuas, como idade, são tratadas de maneira discreta, isto é, através de sua redução a faixas etárias (Oushiro, 2022). Ademais, o VARBRUL trabalha com um processo de seleção de variáveis, mantendo no modelo apenas aquelas que atingem significância estatística. Essa prática já não é mais aconselhada, uma vez que o modelo estatístico deve ser definido antes da análise, com base no modelo científico da área, e mesmo as variáveis sem resultados significativos devem ser mantidas no modelo a ser relatado, pois essa informação também é relevante para a área de conhecimento (McElreath, 2020).

Outra crítica às análises conduzidas pelo VARBRUL, e que está diretamente relacionada aos objetivos da presente pesquisa, está relacionada à condução de um modelo estatístico de efeitos fixos. Modelos de efeitos fixos trabalham com o pressuposto de independência das observações, ou seja, pressupõem que não há correlação entre os dados provenientes de diferentes indivíduos ou unidades de análise. Esse, porém, raramente é o caso de dados linguísticos (Lima Jr; Garcia, 2021). Por essa razão, fatores relacionados aos indivíduos que compõem a amostra e às suas escolhas lexicais, por exemplo, são denominados efeitos aleatórios.

Modelos estatísticos que consideram a existência de efeitos aleatórios, comumente referidos pela expressão "modelos de efeitos mistos", são considerados mais robustos em comparação aos modelos de efeitos fixos, pois levam em conta a não independência na coleta de dados, o que, por sua vez, reduz a chance de erro do tipo I (Lima Jr; Garcia, 2021). Além disso, por calcularem tanto estimativas globais para o grupo como estimativas individuais por efeito aleatório, neste caso para cada indivíduo e para cada item lexical, os modelos de efeitos mistos permitem observar, reportar e prever comportamentos linguísticos intra e interindividual, aspecto tão caro para estudos sociolinguísticos.

As vantagens de modelos estatísticos de efeitos mistos sobre o modelo de efeito fixo conduzido pelo programa VARBRUL motivam o questionamento da replicação proposta: os resultados sobre a variação das vogais médias pretônicas no falar popular de Fortaleza, Ceará, obtidos por Araújo (2007), serão mantidos quando reanalisados com modelos de regressão logística de efeitos mistos? Nossa hipótese é que a inclusão de efeitos aleatórios para informante e para item lexical não alterará substancialmente as conclusões principais do estudo original, mas fornecerá estimativas mais robustas e generalizáveis, como aquelas advindas dos participantes e dos itens lexicais como efeitos aleatórios. Consequentemente, esta reanálise de dados não busca trazer contribuições teóricas para o campo da sociolinguística, mas contribuir metodologicamente ao demonstrar como modelos mais recentes são capazes de incluir mais informações empíricas em suas estimativas.

Métodos

Trata-se da reanálise de um corpus oral de fala espontânea, constituído por entrevistas sociolinguísticas, utilizados na tese *As vogais médias pretônicas no falar popular de Fortaleza* (Araújo, 2007) e provenientes do Projeto Norma Oral do Português Popular de Fortaleza (NORPORFOR). O corpus é composto por gravações de 72 informantes, conforme a amostra original, totalizando 5.848 dados analisados (3.337 de /e/ e 2.511 de /o/), estratificados por sexo, faixa etária e escolaridade, conforme descrito na tese original. Não serão incluídos novos participantes nem haverá exclusão adicional de informantes. As entrevistas sociolinguísticas foram realizadas em Fortaleza (CE), representando o falar popular da região.

A escolha da tese de Araújo (2007) como objeto desta reanálise se justifica por três razões principais: (i) o estudo investiga não apenas o alçamento, mas também o abaixamento das vogais médias pretônicas; (ii) trata-se de um dialeto de região historicamente com menor cobertura nas investigações linguísticas de larga escala no Brasil; e (iii) a autora concordou em ceder os dados originais para fins desta reanálise.

Serão seguidas as seguinte etapas metodológicas:

1. Organização e formatação dos dados originais em planilha compatível com o R;
2. Codificação das variáveis preditoras e de resposta conforme a tese original;
3. Ajuste de modelos de regressão logística de efeitos aleatórios, incluindo informante e item lexical como efeitos aleatórios;
4. Comparação dos resultados com os obtidos pelo VARBRUL na tese original;
5. Divulgação do código e dos resultados em repositório aberto.

Os dados serão analisados no ambiente R, utilizando modelos de regressão logística de efeitos aleatórios. As variáveis de resposta serão a realização das vogais pretônicas, e as variáveis preditoras serão as mesmas testadas na tese original (9 fatores linguísticos: tipo de vogal acentuada, tipo de vogal átona seguinte, nasalidade, contexto fonológico precedente, contexto fonológico seguinte, tipo de sílaba, distância em relação à tônica, tipo de atonicidade da pretônica, estrutura morfológica da palavra; e 3 fatores sociais: sexo, faixa etária e escolaridade). Assim como no trabalho original, será primeiramente analisado o alçamento das vogais e depois o abaixamento, em modelos separados. Serão incluídos efeitos aleatórios para informante e item lexical, visando incorporar a variação intrafalantes e intrapalavras nas estimativas do modelo. É possível que o modelo não convirja ou apresente aviso de singularidade (*singular fit*) com a inclusão de item lexical como efeito aleatório por aumentar muito a quantidade de parâmetros do modelo. Caso isso ocorra, serão adotados dois caminhos: i) será ajustado um modelo apenas com participante como efeito aleatório, por ser o efeito aleatório mais importante para se observar a variação intra e interindividual e por demandar menos o modelo; e ii) será ajustado um modelo bayesiano com participante e item lexical como efeitos aleatórios, visto que esse tipo de modelo converge independentemente da quantidade de parâmetros contanto que seus critérios de ajuste sejam estipulados adequadamente (Garcia; Lima Jr, 2021). Nesse caso, os dois modelos serão reportados e comparados aos resultados originais, discutindo as limitações e vantagens de cada.

Sendo assim, os modelos a serem ajustados serão:

- Para vogal /e/:
    - p(alçamento) ~ vogal acentuada + vogal átona seguinte + nasalidade + contexto fonológico precedente + contexto fonológico seguinte + tipo de sílaba + distância da tônica + tipo de atonicidade da pretônica + estrutura morfológica + sexo + faixa etária + escolaridade + (1|participante) + (1|item lexical)
    - p(abaixamento) ~ vogal acentuada + vogal átona seguinte + nasalidade + contexto fonológico precedente + contexto fonológico seguinte + tipo de sílaba + distância da tônica + tipo de atonicidade da pretônica + estrutura morfológica + sexo + faixa etária + escolaridade + (1|participante) + (1|item lexical)

- Para vogal /o/:
    - p(alçamento) ~ vogal acentuada + vogal átona seguinte + nasalidade + contexto fonológico precedente + contexto fonológico seguinte + tipo de sílaba + distância da tônica + tipo de atonicidade da pretônica + estrutura morfológica + sexo + faixa etária + escolaridade + (1|participante) + (1|item lexical)
    - p(abaixamento) ~ vogal acentuada + vogal átona seguinte + nasalidade + contexto fonológico precedente + contexto fonológico seguinte + tipo de sílaba + distância da tônica + tipo de atonicidade da pretônica + estrutura morfológica + sexo + faixa etária + escolaridade + (1|participante) + (1|item lexical)

Apesar da vantagem de modelos de regressão ajustados com efeitos mistos permitirem a inclusão de variáveis preditoras contínuas, as variáveis preditoras nesta reanálise continuarão sendo tratadas como categóricas por não haver as informações dos valores contínuos nos dados e para não adicionar mudanças demasiadas à análise original, possibilitando uma comparação mais direta. Os resultados serão comparados aos da análise original (VARBRUL), verificando se os efeitos principais (altura da vogal tônica, contexto fonológico, fatores sociais) se mantêm significativos e com direção semelhante. Será considerada reprodução bem-sucedida a obtenção dos mesmos padrões qualitativos e tendências gerais reportadas na tese, ainda que pequenas diferenças nos valores de probabilidade sejam esperadas pela inclusão de efeitos aleatórios. Optaremos por métodos mais recentes (modelos mistos) por serem estatisticamente mais robustos e amplamente aceitos atualmente, permitindo inferências mais generalizáveis que o VARBRUL, que não incorpora efeitos aleatórios.

Cronograma

A conclusão do estudo está prevista para 6 meses após a publicação do relato registrado. As atividades serão realizadas de acordo com o cronograma a seguir.

Tempo	**Etapa**
1º mês após publicação do relato registrado	Obtenção dos dados, adequação da planilha para ser utilizada em ambiente R.
2º mês após publicação do relato registrado	Análise descritiva dos dados. Confecção de tabelas de proporção e de gráficos e conferência com os achados no estudo original.
3º ao 5º mês após publicação do relato registrado	Ajuste e avaliação dos modelos de regressão logística com efeitos aleatórios; conferência com os resultados do estudo original.
6º mês após publicação	Redação e submissão do relato de pesquisa.

do relato registrado	

Link para Preprint

https://doi.org/10.1590/SciELOPreprints.15222

Declaração de disponibilidade de dados

Todos os arquivos e materiais necessários para replicar esta reanálise (planilha de dados, scripts de análise e documentação da análise) serão disponibilizados publicamente no Open Science Framework, sob DOI 10.17605/OSF.IO/HYXNK. Os participantes serão anonimizados na planilha de dados.

Declaração de uso de inteligência artificial

Os autores declaram que nenhuma ferramenta de inteligência artificial foi utilizada na criação deste manuscrito.

Declaração ética

Os dados utilizados pertencem ao corpus Norma do Português Oral Popular de Fortaleza (NORPORFOR), cuja utilização não carece de nova aprovação por comitê de ética, conforme parecer 4.025.173 do CAAE 31405620.3.0000.5054, uma vez que não há forma de identificação dos falantes, cuja contribuição com o corpus cumpriu com os requisitos éticos. Além disso, esta reanálise não utilizará os áudios dos participantes, apenas os dados já extraídos e tabulados de maneira anonimizada.

Conflito de interesses

Os autores declaram ausência de conflitos de interesse.

Contribuição de autoria

Funções do CRediT: Ronaldo Lima Jr.: Administração do projeto, Conceitualização, Curadoria de dados, Análise formal, Metodologia, Software, Redação – rascunho original; Pablo Arantes: Análise formal, Investigação, Metodologia, Software, Validação, Visualização, Redação – rascunho original, Redação – revisão e edição; Guilherme Garcia: Conceitualização, Curadoria de dados, Análise formal, Investigação, Metodologia, Software, Visualização, Redação – rascunho original, Redação – revisão e edição; Luciana Lucente: Análise formal, Investigação, Validação, Visualização, Redação – rascunho original, Redação – revisão e edição; Renata Passetti: Análise formal,

Investigação, Metodologia, Supervisão, Validação, Visualização, Redação rascunho original, Redação – revisão e edição.

Referências:

ARAÚJO, A. A. de. **As vogais médias pretônicas no falar popular de Fortaleza: uma abordagem variacionista**. [*s. l.*], 2007.

CEDERGREN, H. J.; SANKOFF, D. Variable rules: Performance as a statistical reflection of competence. **Language**, [*s. l.*], p. 333–355, 1974.

GARCIA, G. D.; LIMA JR., R. M. Introdução à estatística bayesiana aplicada à linguística. **Revista da ABRALIN**, [S. l.], v. 20, n. 2, p. 1–24, 2021. DOI: 10.25189/rabralin.v20i2.1914. Disponível em: https://revista.abralin.org/index.php/abralin/article/view/1914. Acesso em: 26 maio. 2026.

LIMA JR., R. M.; GARCIA, G. D. Diferentes análises estatísticas podem levar a conclusões categoricamente distintas. **Revista da ABRALIN**, [*s. l.*], p. 1, 2021.

MCELREATH, R. **Statistical rethinking**: A Bayesian course with examples in R and Stan. 2nd edition. Boca Raton: Chapman and Hall/CRC, 2020.

OUSHIRO, L. **Introdução à Estatística para Linguistas**. 1. ed. [*S. l.*]: Editora da Abralin, 2022. Disponível em: https://editora.abralin.org/. Acesso em: 27 out. 2025.

TAGLIAMONTE, S. A. **Variationist sociolinguistics: Change, observation, interpretation**. [*S. l.*]: John Wiley & Sons, 2011.

TORRES VIEIRA, N. M. **Monotongação de ditongos orais no português brasileiro: uma revisão sistemática da literatura**. 1. ed. [*S. l.*]: Editora da Abralin, 2022. Disponível em: https://editora.abralin.org/. Acesso em: 27 out. 2025.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.