Aperfeiçoando uma metodologia para extração semiautomática de parâmetros de um modelo dinâmico do ritmo

Pablo Arantes
Departamento de Letras, Laboratório de Fonética
Universidade Federal de São Carlos (UFSCar)
São Carlos, Brasil
pabloarantes@ufscar.br

Ronaldo Mangueira Lima Júnior
Departamento de Estudos da Língua Inglesa, suas Literaturas e Tradução
Universidade Federal do Ceará (UFC)
Fortaleza, Brasil
ronaldojr@letras.ufc.br

Resumo: Apresentamos aperfeiçoamentos a uma metodologia que permite a extração semiautomática de parâmetros do modelo dinâmico do ritmo de Barbosa (2006). A metodologia busca a combinação de três parâmetros do modelo - taxa de indução (α), taxa de decaimento (β) e força de acoplamento (w_0) - que minimiza a distância entre o contorno gerado pelo modelo e um contorno de duração de uma amostra de fala natural. Identificamos *dynamic time warping* como a medida de erro que mais minimiza a distância entre os contornos e comparamos as diferenças geradas por dois procedimentos de comparação entre o contorno natural e os gerados pelo modelo - comparação direta da duração posição a posição ao longo do enunciado ou comparação da mudança relativa na duração. Os resultados mostram que os dois métodos produzem estimativas diferentes para α e w_0, embora essa diferença tenha uma magnitude baixa, em torno de 4% da gama de valores que os parâmetros assumem na metodologia. Os valores de α estimados para amostras de duas variedades do português brasileiro (Ceará e São Paulo) apresentam uma diferença estatisticamente relevante em torno de 0,2, e a diferença observada para w_0 é de 0,01. Esse resultado se alinha à hipótese de Barbosa (2006), que prevê menor variabilidade em w_0 entre falantes de uma mesma língua e maior variabilidade em α em função de fatores como estilos de elocução. Outra novidade apresentada no presente estudo é a análise do efeito da taxa de elocução sobre a estimativa dos parâmetros para a variedade de São Paulo. A variação na taxa de elocução (normal/típica, lenta e rápida) gerou diferenças semelhantes e de baixa magnitude (entre 0,05 e 0,07) em α e w_0. A estimativa geral para w_0 gerada por modelos de regressão hierárquicos bayesianos é um intervalo entre 0,64 e 0,81, centrado em 0,72. Esse valor indicaria que o português brasileiro no estilo de elocução leitura de frases é uma língua de ritmo misto tendendo mais ao polo acentual do que ao silábico.

prosódia; ritmo linguístico; português brasileiro

I. INTRODUÇÃO

O ponto de partida do presente trabalho é o modelo dinâmico do ritmo desenvolvido por Barbosa [1], baseado na técnica matemática dos osciladores acoplados. Neste modelo, a silabicidade e acentuação são considerados os dois níveis básicos da organização temporal da fala e a duração sílaba a sílaba no enunciado é vista como o resultado da interação entre esses dois níveis, que são modelados como osciladores e a interação entre eles é entendida como um acoplamento no sentido matemático. O acoplamento dá conta da influência do oscilador acentual sobre o oscilador silábico, cujo resultado é o aumento no período do oscilador silábico, isto é, de unidades do tamanho da sílaba ao longo de um enunciado. O acoplamento é controlado por um certo número de parâmetros: α, chamado de taxa de indução, modula quão rapidamente se o período do oscilador silábico aumenta por influência do acentual; β, que modula a volta do oscilador silábico a seu estado não induzido após a batida do oscilador acentual e w_0, que indica o grau de acoplamento entre os dois osciladores.

ARANTES, Pablo; LIMA JÚNIOR, Ronaldo Mangueira. Aperfeiçoando uma metodologia para extração semiautomática de parâmetros de um modelo dinâmico do ritmo. In: CONGRESSO BRASILEIRO DE PROSÓDIA, 2., 2022, Florianópolis. **Anais [...]**. Florianópolis: Lbass, 2023. v. 1, p. 142-149. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/anais_coloquio/index.

Em trabalho anterior [2], apresentamos resultados iniciais de uma metodologia desenvolvida para extrair de forma semiautomática os parâmetros do modelo. Um dos usos relevantes possíveis que pode ser dado para estimativas confiáveis dos valores dos parâmetros é iluminar a discussão empírica e teórica a respeito da questão da tipologia rítmica, isto é, a localização de línguas ou variedades linguísticas dentro de um contínuo entre os tipos rítmicos acentual e silábico. Do ponto de vista do seu modelo, Barbosa (2006) propõe que a variável w_0 seja entendida como um índice tipológico. Por hipótese, a variação deste parâmetro entre os falantes de uma determinada língua é menor do que a variação entre línguas diferentes, especialmente aquelas que tradicionalmente são classificadas como tendo tipos rítmicos diferentes. Na formulação corrente do modelo [1], w_0 é um parâmetro que varia continuamente no intervalo entre 0 e 1 e valores de w_0 próximos ao piso do intervalo seriam indicativos do tipo silábico e valores próximos ao teto seriam encontrados em línguas de ritmo acentual. O parâmetro w_0 deve variar menos do que α e β em uma mesma comunidade linguística e a variação de α e β podem ser influenciada por fatores como estilo e taxa de elocução, assim como variação idiossincrática. Esse quadro interpretativo, ainda segundo a proposta de Barbosa (2006), precisa ser corroborado com dados experimentais. Com o aperfeiçoamento de nossa proposta de metodologia, esperamos facilitar a produção em maior escala do tipo de dado empírico que possa contribuir para a avaliação do quadro interpretativo proposto em [1].

No presente trabalho, nosso objetivo é introduzir aperfeiçoamentos à nossa formulação inicial e avaliar o efeito de dois fatores técnicos da metodologia, explicados em maior detalhe na seção II, que podem afetar o valor dos parâmetros estimados. Usamos os aperfeiçoamentos propostos para investigar o impacto de dois fatores linguísticos sobre a estimativa dos valores dos parâmetros α e β: as diferenças entre dois dialetos ou variedades linguísticas do português brasileiro (doravante PB) e a variação na taxa de elocução.

II. Materiais e métodos

A. Materiais de fala

Os materiais linguísticos analisados são dois conjuntos de dados de leitura de textos de duas variedades do português brasileiro, um produzido por falantes do estado do Ceará (CE) e outro por falantes do estado de São Paulo. Como os materiais foram coletados originalmente para projetos diferentes, os textos não são os mesmos, embora sejam comparáveis, o que os torna apropriados para os objetivos do trabalho.

O material dos falantes do CE consiste em gravações da leitura de um texto de 255 palavras, que é uma tradução para o PB de uma passagem usada para diagnóstico de leitura, retirada do livro de Celce-Murcia e colegas [3]. Os participantes foram instruídos a ler o texto em sua taxa de elocução normal/típica. A duração média das leituras é de 88 segundos. A idade dos participantes, quatro do sexo masculino e uma do sexo feminino, variava entre 18 e 20 no momento da gravação.

O material dos falantes de SP consiste da leitura de um trecho de 144 palavras do texto “A Menina do Narizinho Arrebitado”, do escritor Monteiro Lobato. Os participantes foram instruídos a ler o texto primeiramente em sua taxa de elocução normal/típica e posteriormente em uma taxa mais rápida e finalmente em uma taxa mais lenta em

comparação à sua taxa típica. A duração média das leituras é de 34 segundos. A idade dos participantes, cinco do sexo masculino e três do sexo feminino, variava entre 18 e 30 anos no momento da gravação.

B. Análise fonética

Cada amostra de áudio foi segmentada em unidades VV usando a convenção SAMPA-PB[1] em arquivos TextGrid do programa Praat. A segmentação foi usada posteriormente para identificar a posição dos acentos frasais ao longo do trecho de fala contido em cada áudio com o auxílio de um script para o programa Praat. Esse *script* implementa o procedimento em três passos descrito em [1].

C. Extração dos parâmetros do modelo

A metodologia que propomos parte de um conjunto de gravações segmentadas em unidades VV da maneira descrita na seção anterior e produz como resultado final uma tripla de estimativas para os parâmetros α, β e w_0 do modelo que geram o contorno simulado que mais se aproxima ao contorno de duração de uma amostra de fala natural. Uma descrição detalhada da metodologia é apresentada em trabalho anterior dos autores [2] e o código para o ambiente estatístico R que implementa o procedimento está disponível no repositório https://osf.io/w82ru/.

Para a geração de contornos simulados a serem comparados com o natural, são usadas triplas de valores para os parâmetros α, β e w_0 que varrem o espaço das combinações possíveis de uma gama de valores. Para α e β entre 0,05 e 1,5 em passos de 0,1 e para w_0 entre 0,05 e 1 em passos de 0,025. A cada tripla corresponde um contorno simulado que então é comparado ao contorno da fala natural.

Os aperfeiçoamentos introduzidos na versão descrita aqui foram: (1) a correção de um erro no código que calcula as medidas de erro entre os contornos simulados e os naturais e (2) a estimação dos valores da magnitude dos acentos frasais a partir do contorno natural.

Sobre o ponto (1), três medidas de erro foram empregadas: erro absoluto médio (*mean absolute error - MAE*), erro quadrático médio (*root mean squared error - RMSE*) e a técnica *dynamic time warping (DTW)*.

Sobre o ponto (2), na versão anterior, o valor da magnitude de todos os acentos frasais identificados nas amostras foi fixado em 1, seguindo [4]. Um exame posterior dos contornos simulados identificados como mais próximos aos naturais mostrou que essa decisão estava superestimando esse valor e causando discrepâncias de grande magnitude (ver figura 2 e 3 de [2]). Em testes preliminares, verificamos que a introdução da estimativa da magnitude a partir do próprio contorno natural mostrou-se eficiente para mitigar essa superestimativa.

Em nossa metodologia, o que chamamos de contorno da fala natural é o contorno de duração bruta que é posteriormente normalizado temporalmente e suavizado e, finalmente, tem sua duração restaurada para poder ser expressa em unidades de tempo real. No trabalho anterior, a comparação entre os contornos simulados e o natural foi feita segundo dois métodos. No primeiro, a que faremos referência pelo nome de *comparação direta*, as durações consecutivas de cada posição ao longo dos dois contornos são diretamente comparadas. No segundo, que chamamos

[1] Ver https://github.com/parantes/sampa-pb.

de *mudança relativa*, a mudança relativa da duração ao longo dos contornos é comparada. O segundo método foi testado por ter sido sugerido em [4].

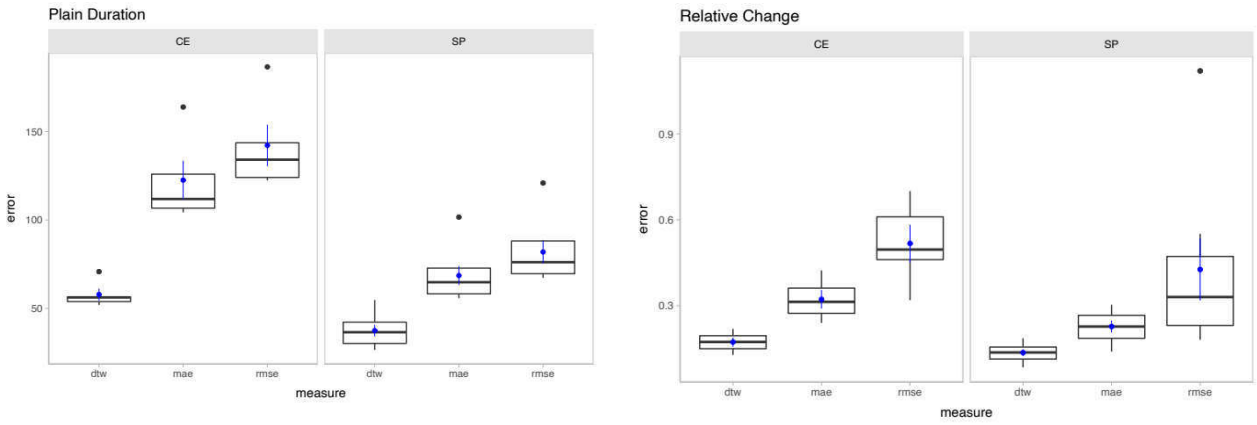
III. RESULTADOS

Para cada análise foram ajustados dois modelos de regressão linear hierárquicos bayesianos, um com os valores de α e o outro com valores de w_0 como variáveis resposta. Para analisar o efeito das medidas de erro, métodos de comparação de contornos e dialeto, foram utilizados os dados de leitura normal/típica dos falantes de SP, já que os falantes do CE foram gravados apenas com essa taxa de elocução. Posteriormente, foram analisados os dados de SP com as três taxas de elocução. Para os modelos de α foram utilizadas distribuições a priori não informativas – os valores padrão do pacote brms [5] para o R. Para os modelos de w_0 foram utilizados *priors* regularizadores, centrado em 0,5 e permitindo valores de w0 entre 0 e 1 – seguindo, portanto, o conhecimento prévio de que o português Brasil segue um ritmo misto, com um w_0 próximo a 0,5. Vale ressaltar que modelos bayesianos não apresentam estimativas pontuais (*point estimates*), mas distribuições de probabilidade dos seus coeficientes (vide [6] para explicações detalhadas).

A. Medidas de erro

Como pode ser observado na figura 1, DTW é a medida que produz o menor erro médio entre as três, considerando os dois métodos de comparação entre contornos.

Figura 1: Gráficos de caixas com valores das três medidas de erro por dialeto e por método. Os pontos e linhas azuis nas caixas indicam as médias e erros-padrão, respectivamente.



Fonte: elaborado pelos autores.

DTW gerou os menores erros e as menores dispersões. Para o método de comparação direta, o erro produzido por DTW é em torno 2 vezes menor do que o produzido por RMSE e MAE; para o método mudança relativa, o erro médio medido por DTW é 3 vezes menor do que o medido por RMSE e 1,8 vezes menor do que o gerado por MAE.

Foram ajustados dois modelos, um para α e outro para β, com medida de erro, método de comparação e dialeto como variáveis preditoras. O modelo para α não apresentou efeito de medida de erro (β = -0.01 [-0.04 – 0.02][2] para MAE e β = -0.00 [-0.03 – 0.03] para RMSE, ambos em comparação a DTW), e apresentou efeito plausível para DTW em relação às outras duas medidas, com erros claramente mais baixos para DTW (β = 0.03 [0.00 – 0.06] para MAE e β = 0.03 [0.00 – 0.06] para RMSE).

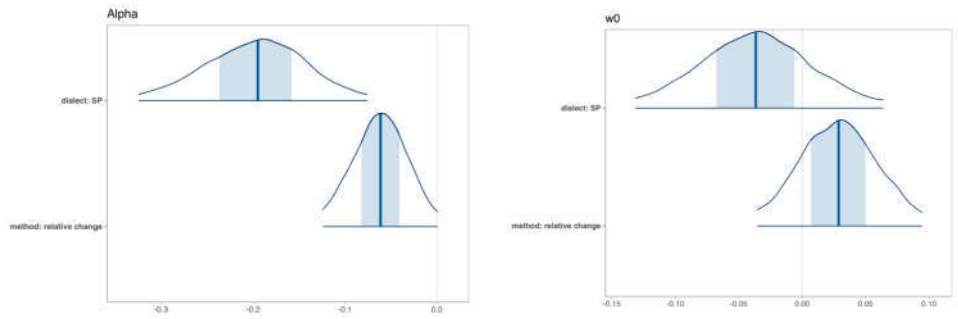
Baseado na observação de que DTW é a medida de erro que gera as estimativas com menores discrepâncias entre contornos simulados e naturais, os resultados sugerem a adoção de DTW como única medida de erro. Sendo assim, as demais análises reportadas foram realizadas apenas com os dados de DTW.

B. Método de comparação de contornos

O modelo com método de comparação de contornos e dialeto como variáveis preditoras estima um valor de α de 1,39 [1.29 – 1.49] para falantes do CE com o método de comparação direta. O modelo estima uma queda de aproximadamente 0,06 [-0.12 – 0.00] no valor de α para o método de mudança relativa. A diferença é crível do ponto de vista estatístico, embora a magnitude seja baixa (em torno de 4% da gama de valores testados para α) e possivelmente de relevância linguística muito limitada.

Semelhantemente, o modelo estima um valor de 0,74 [0.67 – 0.83] para o w_0 de um falante do CE com o método de comparação direta. O modelo estima um aumento de aproximadamente 0,03 [-0.04 – 0.09] no valor de w_0 para o método de mudança relativa, porém essa diferença não é crível do ponto de vista do modelo estatístico – vide a quantidade da distribuição *a posteriori* abaixo do zero na figura 2.

Figura 2: Distribuições a posteriori dos efeitos de dialeto e do método de comparação sobre os valores de alfa e de w0. O traço indica e mediana da distribuição, a área azul clara indica o intervalo de 50% de maior densidade (HDI - *High Density Interval*) da distribuição, e cada distribuição está cortada em seus limites de 95% de HDI.



Fonte: elaborado pelos autores.

[2] Os intervalos entre colchetes ao longo de todo o trabalho se referem à distribuição de 95% de credibilidade sobre o coeficiente.

Observa-se, como na versão anterior da metodologia [2], um efeito crível do ponto de vista estatístico motivado pelo método de comparação entre contornos para α, mas não para w_0 . O efeito tem a mesma direcionalidade nos dois estudos. A magnitude desse efeito na versão testada agora, no entanto, é menor: 0,06 agora contra 0,15 no caso de α; e 0,03 agora contra 0,19 no caso de w_0. Também se repete no presente estudo a relação de complementaridade na manifestação do efeito, isto é, α desce e w_0 sobe com o método de comparação duração relativa e o contrário acontece com o método que compara diretamente a duração. Consideramos positivo o fato de que na nova versão da metodologia o efeito tenha preservado sua estrutura, embora tenha sofrido uma redução grande, a ponto de tornar-se, do ponto de vista linguístico, possivelmente pouco relevante, de modo que a escolha por um ou outro método de comparação não tenha consequências importantes para a análise do fenômeno do ritmo.

Do ponto de vista da estimativa dos dois parâmetros, o valor dos dois na versão revisada da metodologia aumentou, de aproximadamente 0,97 para 1,4 em α e de 0,45 para 0,72 em w_0. Em termos da interpretação linguística dos resultados, a diferença indicaria que o PB tem um caráter mais acentual do que os resultados prévios indicaram.

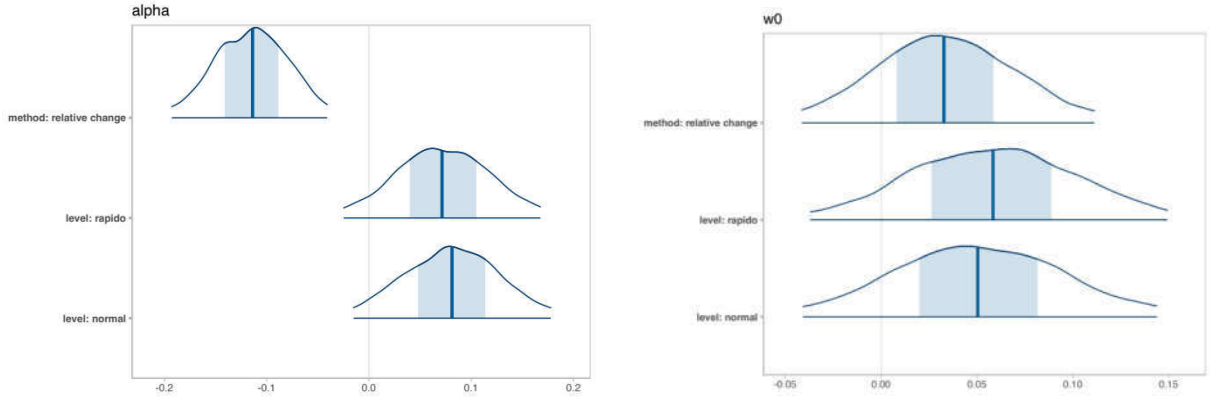
C. Efeito da variedade linguística

Para o parâmetro α, o modelo estima um valor em torno de 0,2 [-0.31 – -0.07] menor para falantes de SP do que para a variedade do CE, uma diferença crível nos termos da análise estatística (vide figura 2). Para w_0, a variedade de SP tem um valor estimado em torno de 0,04 [-0.13 – 0.06] menor do que o da variedade do CE, embora não seja uma diferença estatisticamente crível (vide figura 2). É interessante notar que a distribuição posterior gerada pela análise bayesiana para os dois parâmetros tem um grau de dispersão bastante grande em comparação com a do outro parâmetro (método), muito embora a diferença seja crível no caso de α, mas não para w_0.

D. Efeito da taxa de elocução

Para avaliar a taxa de elocução, foram ajustados modelos apenas com os dados de SP, desta vez incluindo os dados de leitura lenta e rápida. Para α, os resultados mostram que a taxa lenta diminui seu valor em torno de 0,07 em comparação com as outras duas, embora o modelo não apresente confiança sobre esse efeito, visto que parte da distribuição a posteriori, incluindo o intervalo de 95% de maior densidade (HDI), está do lado negativo para os níveis normal e rápido (veja figura 3). Para w_0, a taxa lenta provoca diminuição no valor em torno de 0,05 em comparação com as outras duas, porém essa diferença é ainda menos crível do ponto de vista do modelo estatístico.

Figura 3: Distribuições a posteriori dos efeitos de taxa de elocução e do método de comparação sobre os valores de alfa e de w0 para os falantes de SP. O traço indica e mediana da distribuição, a área azul clara indica o intervalo de 50% de maior densidade (HDI - *High Density Interval*) da distribuição, e cada distribuição está cortada em seus limites de 95% de HDI.



Fonte: elaborado pelos autores.

Note que nesses modelos os efeitos já discutidos sobre os métodos de comparação se mantêm: o método de mudança relativa diminui alfa (de maneira crível estatisticamente, mas com pequena magnitude do ponto de vista linguístico) e aumenta w0, porém sem credibilidade estatística.

IV. Discussão e conclusão

Do ponto de vista geral, consideramos que os resultados apresentados no presente trabalho mostram uma evolução positiva no desenvolvimento da nossa proposta de metodologia. Em primeiro lugar, foi possível definir em bases mais objetivas qual é a melhor medida de erro para a comparação entre os contornos natural e simulados, o que representa uma diminuição nos graus de liberdade em relação à proposta inicial. A diferença entre os métodos de comparação entre contornos persiste, embora os ajustes introduzidos tenham provocado o efeito de diminuir bastante a diferença, de modo que a relevância linguística dessa diferença pode ser considerada pouco relevante.

Do ponto de vista das hipóteses de Barbosa a respeito das possibilidades de uso de seu modelo para a discussão a respeito da tipologia linguística, os resultados relativos à mudança na variedade linguística parecem corroborar parte das assunções de Barbosa, em especial a que prevê uma variação intralinguística menor para o parâmetro w_0 em comparação a α. No caso dos nossos dados, a diferença de variedade linguística exerceria esse papel de aumentar a variabilidade de α, sem afetar significativamente o valor de w_0. A comparação das taxas de elocução, no entanto, mostra que essa variável provoca alterações semelhantes e de magnitude baixa tanto em α quanto em w_0, embora não completamente críveis do ponto de vista estatístico.

Um passo crucial a ser dado em etapas futuras é a aplicação da metodologia a materiais de fala de línguas tradicionalmente analisadas como representantes de tipos rítmicos diferentes do PB, em especial línguas consideradas

mais prototipicamente de tipo acentual (como o inglês) e silábico (como o espanhol), de modo que seja possível estabelecer se de fato o modelo dinâmico é capaz de gerar contornos de uma gama de tipos rítmicos como sugerido por Barbosa e se a metodologia que propomos é capaz de estimar os parâmetros capazes de mostrar essa flexibilidade.

AGRADECIMENTOS

Os autores agradecem Plínio Barbosa por ceder as gravações da variedade de São Paulo. O segundo autor agradece o CNPq por uma verba que financiou parcialmente a pesquisa sobre a variedade do Ceará (processo 438823/2018-4).

REFERÊNCIAS

[1] Barbosa PA. *Incursões em torno do ritmo da fala*. Campinas: Pontes, 2006.

[2] Arantes P, Lima Júnior RM. Using a coupled-oscillator model of speech rhythm to estimate rhythmic variability in two Brazilian Portuguese varieties (CE and SP). *Cadernos de Linguística* 2021; 2: 1–19.

[3] Celce-Murcia, M., Brinton, D. M., Goodwin, J. M., et al. *Teaching Pronunciation: a course book and reference guide*. Cambridge, UK: Cambridge University Press, 2010.

[4] Barbosa PA. Elementos para uma tipologia do ritmo (lingüístico) da fala à luz de um modelo de osciladores acoplados. *In Cognito - Cadernos Românicos em Ciências Cognitivas* 2004; 2: 31–58.

[5] Bürkner P. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 2017; 80: 1–28.

[6] Garcia GD, Lima Júnior RM. Introdução à estatística bayesiana aplicada à linguística. *Revista da ABRALIN* 2021; 20: 1–24.