

## **Centro de Estatística Aplicada**

### **Relatório de Análise Estatística**

**RELATÓRIO DE ANÁLISE ESTATÍSTICA SOBRE O PROJETO:**

***“O processamento de pseudopalavras no Português Brasileiro”***

**Giovanna Vendeiro Vilar**

**Mariana Almeida**

**Renata Massami Hirota**

**Viviana Giampaoli**

**São Paulo, julho de 2021**

# CENTRO DE ESTATÍSTICA APLICADA - CEA - USP

**TÍTULO:** Relatório de Análise Estatística sobre o Projeto: “O processamento de pseudopalavras no Português Brasileiro”

**PESQUISADORA:** Aline Benevides

**ORIENTADORA:** Profa. Dra. Raquel Santana Santos

**INSTITUIÇÃO:** Faculdade de Filosofia e Ciências Humanas da Universidade de São Paulo

**FINALIDADE DO PROJETO:** Doutorado

**RESPONSÁVEIS PELA ANÁLISE:**

Giovanna Vendeiro Vilar

Mariana Almeida

Renata Massami Hirota

Viviana Giampaoli

**REFERÊNCIA DESTE TRABALHO:** ALMEIDA, M.C.; GIAMPAOLI, V.; HIROTA, R.M.; VILAR, G.V. **Relatório de análise estatística sobre o projeto: “O processamento de pseudopalavras no Português Brasileiro”**. São Paulo, IME-USP, ano. (RAE–CEA-21P02)

## FICHA TÉCNICA

### REFERÊNCIAS BIBLIOGRÁFICAS

### PROGRAMAS COMPUTACIONAIS UTILIZADOS

Software R (versão 4.0.5)

### 8.1 TÉCNICAS ESTATÍSTICAS UTILIZADAS

Análise Descritiva Unidimensional (03:010) Análise de Dados Categorizados  
(06:030) Outros (07:990)

### ÁREA DE APLICAÇÃO

Linguística (14:110)

COLLISCHONN, G. 1999. *Acento Em Português. In BISOL, I. (ed.) Introdução a Estudos de Fonologia Do Português Brasileiro*. 1st ed. Porto Alegre: EDIPUCRS.

FIORIN, J. L. 2019. *Linguística? Que é Isso?* 1st ed. São Paulo: Contexto.

PETTER, M. 2007. *Introdução à Linguística i: Objetos Teóricos*. 5th ed. São Paulo: Contexto.

PRIBERAM. 2021. "pseudopalavra", in *Dicionário Priberam Da língua Portuguesa*. <<https://dicionario.priberam.org/pseudopalavra>>. [Acesso em: 01-05-2021].

## Sumário

1	Resumo	1
2	Introdução	1
3	Objetivo(s)	2
4	Descrição do estudo	3
4.1	Limitações do estudo	4
4.2	Conceitos de Similaridade e Validação	4
5	Descrição das variáveis	5
5.1	Variável Dependente	6
5.2	Variáveis Linguísticas	6
5.3	Variáveis Extralinguísticas	7
5.4	Variáveis Experimentais	7
6	Análise descritiva	8
6.1	Perfil dos participantes	8
6.1.1	Sexo e idade	8
6.1.2	Naturalidade	9
6.1.3	Área de formação	9
6.2	Variáveis linguísticas	10
6.3	Análise de Concordância	13
6.4	Resumo da Análise Descritiva	16
6.5	Análise dos efeitos aleatórios preditos	16
7	Metodologia	16
7.1	Modelos Ajustados	18
7.1.1	Modelo com Grupo de classificação (Modelo 1)	19
7.1.2	Modelo com Similaridade e Frequência (Modelo 2)	24
7.1.3	Problema de Aleatorização	29
8	Conclusão	30
<b>Apêndices</b>		
A	Tabelas	31
B	Gráficos	37

## 1 Resumo

A literatura na área de linguística afirma, em geral, que o acento tônico na penúltima sílaba (paroxítona) é o padrão, enquanto que acentos na última (oxítona) e na antepenúltima (proparoxítona) sílabas são desvios.

O estudo analisado neste relatório estatístico busca entender como se dá o processo da acentuação tônica (Oxítona, Paroxítona e Proparoxítona) em falantes nativos do português brasileiro, através de um experimento utilizando pseudopalavras e se, de fato, o acento paroxítona é o padrão da língua. A partir desse objetivo, foram construídos modelos de regressão multinomial com efeitos aleatórios levando em conta a estrutura de medidas repetidas dos dados. Concluiu-se que, de modo geral, existem variáveis além da estrutura da palavra que influenciam no processo de acentuação tônico no português brasileiro.

## 2 Introdução

De acordo com PETTER (2007), o interesse pela linguagem é antigo e vem sendo expresso por meio de mitos, lendas, cantos, rituais e trabalhos eruditos que buscam conhecer essa capacidade humana como sistema de comunicação. A autora aponta que, a partir do século XX, os estudos linguísticos passaram a ter um caráter científico, ou seja, centrados na observação dos fatos a partir de pressupostos teóricos da linguagem, no estabelecimento de hipóteses e na examinação mediante experimentos.

Conforme descrito por FIORIN (2019), a linguística é uma ciência da linguagem porque, ao contrário da gramática, ela tem como objetivo estabelecer o que uma língua é e por que é de uma determinada maneira. Logo, a área estuda os aspectos fonéticos, morfológicos, sintáticos, semânticos, sociais e psicológicos de uma língua, e neste caso, o português brasileiro. Dentro deste contexto, existe o conceito de *pseudopalavra*, que, de acordo com o dicionário PRIBERAM (2021), é uma

Sequência regular e pronunciável de caracteres que não tem um significado numa língua, apesar de obedecer às regras ortográficas,

morfológicas ou de pronúncia

No português brasileiro existem três classes de palavras segundo sua tonicidade: oxítona, paroxítona e proparoxítona. Essas denominações estão relacionadas à intensidade dada a determinadas sílabas na pronúncia das palavras. Aquela que é pronunciada de forma mais acentuada é a sílaba tônica. Assim, oxítonas são as palavras cuja sílaba tônica é a última; paroxítonas são as palavras cuja sílaba tônica é a penúltima; e proparoxítonas são as palavras cuja sílaba tônica é a antepenúltima.

O intuito do trabalho é investigar a maneira como os falantes nativos do português atribuem a tonicidade em pseudopalavras parcialmente relacionadas a vocábulos existentes no idioma. Em outras palavras, busca-se compreender como o indivíduo, ao se deparar com uma palavra nova, nesse caso, uma pseudopalavra, define a sílaba tônica. Além disso, busca-se entender quais são os outros fatores, tais como os conhecimentos linguísticos do falante e as estruturas linguísticas das palavras, que podem influenciar nesse processo de classificação e portanto da determinação da entonação.

### **3 Objetivo(s)**

O objetivo principal do trabalho é entender o processo de acentuação no Português Brasileiro, com base em um experimento utilizando pseudopalavras. Ou seja, buscamos compreender quais as características da palavra e quais as características do falante que influenciam nesse processo de acentuação tônica quando o indivíduo se depara com uma palavra nova.

Algumas perguntas a serem respondidas pela análise estatística são:

1. A classificação tônica das pseudopalavras pode recuperar o acento das palavras-alvo de referência? Em outras palavras, queremos entender se a tonicidade da palavra-alvo tem papel na predição do acento da pseudopalavra.
2. As características estruturais da pseudopalavra têm influência no processo de acentuação tônica? Essa é a única característica que realmente tem impacto

nesse processo?

3. A similaridade entre palavra-alvo e pseudopalavra influencia na associação acentual? Buscamos entender se quanto mais similar a pseudopalavra for da palavra-alvo, maiores são as chances de atribuição do mesmo padrão acentual da palavra-alvo.
4. As variáveis selecionadas pelo modelo estão em concordância com a literatura da área? Existem variáveis linguísticas, extralinguísticas -relacionadas aos participantes- e experimentais -relacionadas ao estudo-, que podem, segundo a literatura da área, influenciar o comportamento acentual no português.
5. Há associação entre graduação em letras e a classificação tônica da pseudopalavra? Há associação entre conhecimento em música e a classificação tônica da pseudopalavra? Espera-se que indivíduos com conhecimento em música ou que entraram recentemente em letras tenham um comportamento de classificação das pseudopalavras distinto dos demais.

## **4 Descrição do estudo**

O estudo foi realizado de maneira remota com reuniões individuais entre a pesquisadora e cada um dos participantes via Google Meet. Consistiu em apresentar aos participantes, através do software Psychopy, 372 pseudopalavras agrupadas nos denominados grupos de classificação e registrar a forma como eles reproduziam verbalmente tais palavras criadas. A seguir, as respostas dos participantes foram classificadas de acordo com as três classes de acentuação tônica: oxítona, paroxítona e proparoxítona.

A coleta dos dados foi realizada no início do primeiro semestre de 2020 com 34 indivíduos que, por meio de divulgações em redes sociais e de colegas, se voluntariaram a participar do experimento. Os voluntários tiveram como pré-requisitos, ser maior de 18 anos, ser falante nativo do português brasileiro e não ter estudado linguística.



Entre os participantes da pesquisa encontram-se estudantes do primeiro semestre da faculdade de Letras da Universidade de São Paulo, músicos, alguns residentes de fora do estado de São Paulo, entre outros. Supõe-se que os alunos do primeiro semestre do curso de Letras ainda não têm conhecimento na área.

#### **4.1 Limitações do estudo**

Identificamos dois eventuais problemas -um de caráter técnico e outro de aleatorização- na coleta de dados que tentaremos contornar nas análises. O primeiro é descrito a seguir, enquanto o outro será mencionado na seção 5.4.

##### **Problemas técnicos e interferência externa**

Destaca-se a perda de algumas respostas durante o processo de coleta de dados, visto que ruídos externos impediram que algumas entonações fossem captadas e registradas na gravação. Logo, na base de dados não temos 372 registros de pseudopalavras para todos os participantes. Portanto, um total de 12.511 dados serão utilizados na análise, em vez dos 12.648 esperados, o qual não representa uma perda substancial.

#### **4.2 Conceitos de Similaridade e Validação**

O conceito de similaridade entre palavra-alvo e pseudopalavra foi construído com base nas mudanças feitas na palavra-alvo até a obtenção da pseudopalavra. Essas alterações estão relacionadas à mudanças de ponto, modo e/ou vozeamento.

De acordo com o tipo e a quantidade de alterações foi estabelecido um valor de 1 a 10 -chamado de taxa de similaridade- onde, para mudanças de consoantes, valores menores do que 4 determinam pseudopalavras similares a sua palavra-alvo e valores maiores ou iguais a 5 determinam pseudopalavras dissimilares a sua palavra-alvo. Já para mudanças de vogais, valores acima de 1 determinam pseudopalavras dissimilares a sua palavra-alvo.

Diante disso, foi necessário definir um modo de validar essa classificação em similar e dissimilar, ou seja, verificar se a pseudopalavra classificada como similar -ou

dissimilar- é, de fato, similar -ou dissimilar- à palavra da qual ela se originou (palavra-alvo). Nessa etapa -chamada de validação- pediu-se para 10 falantes do português, que não fazem parte do estudo final, listarem a palavra do português a qual eles associavam cada uma das pseudopalavras.

Dentre as pseudopalavras consideradas similares a palavras alvo, considerou-se validadas como “similar à palavra-alvo” as pseudopalavras cuja associação foi a palavra-alvo na resposta de, no mínimo, oito indivíduos. Porém, pseudopalavras nas quais sete falantes apresentaram a associação correta foram classificadas como quase validadas e quando menos de sete falantes apresentaram a associação “correta” entendeu-se que a classificação em similar não foi validada (ela não está, de fato, parecida com sua palavra-alvo).

Da mesma forma, dentre as pseudopalavras consideradas dissimilares a palavra-alvo, foram consideradas validadas como “dissimilar à palavra-alvo” pseudopalavras não associadas a uma mesma resposta por mais de dois indivíduos. Em outras palavras, se até dois falantes associaram uma mesma palavra à pseudopalavra, ela foi considerada dissimilar à sua palavra-alvo. Pseudopalavras nas quais três falantes apresentaram a mesma associação foram classificadas como quase validadas e quando mais de três falantes lembraram de uma mesma palavra do português ao ler a pseudopalavra entendeu-se que a classificação em dissimilar não foi validada (ela parece estar parecida com uma palavra real).

Ressalta-se que, nesse caso, a associação não precisava ser exclusivamente com a palavra-alvo. Por exemplo, suponha que pesquisadora criou a pseudopalavra *Fanama* através de modificações na palavra *Parána* (palavra-alvo) e definiu que essa pseudopalavra era dissimilar a sua palavra-alvo. Se, na etapa de validação, 5 dos 10 falantes lembraram da palavra *Canáda* ao ler *Fanama* então considerou-se que a dissimilaridade não foi validada, pois muitos falantes associaram a pseudopalavra a uma mesma palavra do português (ela parece ser similar a uma palavra real.)

## 5 Descrição das variáveis

Foram coletadas variáveis linguísticas -relacionadas às pseudopalavras-, extralinguísticas -relacionadas aos participantes- e experimentais -relacionadas ao estudo-, que podem, segundo a literatura da área, influenciar o comportamento acentual no português. A seguir, listamos as variáveis pré-selecionadas para o estudo.

### 5.1 Variável Dependente

A variável resposta de interesse é **Tonicidade de produção**, ou seja, a classificação acentual tônica da pseudopalavra (oxítônica, paroxítônica e proparoxítônica).

### 5.2 Variáveis Linguísticas

As pseudopalavras foram construídas com três sílabas de extensão para que os três padrões acentuais do português brasileiro pudessem ser produzidos. As palavras que deram origem às pseudopalavras, definidas como palavras-alvo, foram classificadas em dois níveis de acordo com a sua ocorrência no Corpus brasileiro, corpus linguístico coordenado pelo pesquisador Antonio Paulo Berber Sardinha. Se a palavra possui mais de 100 mil ocorrências no corpus ela é classificada como de alta frequência e se possui menos de 2 mil ocorrências ela é classificada como de baixa frequência. A junção da ideia de frequência e similaridade entre a palavra alvo e a pseudopalavra resultou na criação de uma variável com 4 categorias chamada grupo de classificação.

- **Validação:** s = sim, n = não validada e q = quase validada
- **Palavra alvo:** palavra real que deu origem à pseudopalavra
- **Tonicidade da palavra-alvo:** oxítônica, paroxítônica e proparoxítônica
- **Estrutura da palavra:** indica qual é a estrutura da pseudopalavra (CV-CV-CV ou CV-CV-CVC), sendo que C indica Consoante e V indica Vogal
- **Pseudopalavra:** refere-se a cada uma das palavras criadas

**Segmento modificado:** indica qual letra foi modificada na criação da pseudopalavra a partir da palavra real (consoante ou vogal)

- **Grupo de classificação:** indica o efeito da similaridade (entre a pseudopalavra e a palavra real) e da frequência (alta e baixa) na produção acentual
  - 1 = pseudopalavras similares de alta frequência
  - 2 = pseudopalavras dissimilares de alta frequência
  - 3 = pseudopalavras similares de baixa frequência
  - 4 = pseudopalavras dissimilares de baixa frequência

Ressaltamos que essa variável não foi controlada durante a coleta de dados, ou seja, não foi pré-definido uma quantidade de palavras de cada categoria em cada conjunto apresentado aos participantes.

### 5.3 Variáveis Extralinguísticas

- **Participante:** identifica os 34 participantes do experimento
- **Idade:** de 18 a 60 (anos)
- **Gênero:** feminino e masculino
- **Escolaridade:** categorizada em Fundamental Completo, Superior Completo, Superior Incompleto e Pós-Graduação (Completo ou Incompleta)
- **Área de formação:** 0 = outros e 1 = letras
- **Línguas:** 0 = não tem conhecimento em línguas e 1 = tem conhecimento em línguas
- **Música:** 0 = não tem conhecimento em música e 1 = tem conhecimento em música

### 5.4 Variáveis Experimentais

As palavras foram aleatorizadas no Excel e divididas em 4 conjuntos - variável Bloco de apresentação- a serem apresentados aos participantes com um intervalo de tempo entre cada conjunto. Por limitações do software, a ordem de apresentação

desses conjuntos não pôde ser aleatorizada, apenas a ordem das palavras dentro de cada conjunto.

- **Bloco de apresentação:** indica em qual bloco (ou conjunto) a pseudopalavra foi inserida (1, 2, 3 ou 4)

### **Problemas de aleatorização**

Para alguns participantes o software Psychopy apresentou problemas e eles tiveram que continuar o experimento a partir de slides com uma ordem aleatória pré-estabelecida. Em outras palavras, todos os indivíduos que em algum momento acompanharam o experimento pelos slides seguiram com palavras apresentadas na mesma ordem (a primeira aleatorização retirada do Excel).

- **Aleatorização:** codifica se o bloco de apresentação foi aleatorizado para o indivíduo ou não, e portanto foi considerada a aleatorização prévia ( $s = o$  estímulo foi aleatorizado e  $n = o$  estímulo não foi aleatorizado).

## **6 Análise descritiva**

A análise descritiva do projeto foi dividida em duas partes. A seguir, apresentamos alguns gráficos e tabelas juntamente com interpretações retiradas das volumetrias vistas. Ao final, discutimos uma idéia de análise de concordância envolvendo os falantes e a classificação tônica das pseudopalavras.

### **6.1 Perfil dos participantes**

#### *6.1.1 Sexo e idade*

Os 34 participantes do estudo estão divididos entre 21 mulheres e 13 homens, com idades que variam entre 18 e 60 anos. A Figura @ref(fig:idade\_genero) mostra a distribuição dos respondentes segundo a faixa etária e gênero. A faixa etária foi dividida a partir dos quartis da variável idade, para resumir as informações da amostra. Com exceção da faixa etária de 31 a 38 anos, há mais mulheres do que homens no experimento.

### 6.1.2 Naturalidade

A Tabela [A.1](#) exibe o perfil dos informantes segundo sua naturalidade (variável agrupada em razão da baixa volumetria de indivíduos por UF). É possível perceber que a maior proporção (82%) dos participantes reside no estado de São Paulo, portanto, essa variável não será considerada na análise.

### 6.1.3 Área de formação

Parte considerável dos participantes (11) são ingressantes do curso de Letras, logo, essa variável foi categorizada em dois níveis. A Tabela [A.2](#) mostra esse agrupamento da formação dos voluntários. A influência do curso de Letras na categorização das pseudopalavras é um dos fatores de possível interesse na análise dos dados.

Analisando a distribuição por formação (Tabela [A.3](#)), 19 (55,9%) participantes são de outras áreas e 15 (44,1%) são da área de Letras; no entanto, cruzando essa variável com as informações de escolaridade, notamos que o grupo majoritário são estudantes de Letras com Ensino Superior Incompleto (11 pessoas, o que representa 32,4% do total). Também destacamos que a amostra a nível de indivíduos é pequena e possivelmente não representativa da população brasileira, pois apenas um indivíduo tem nível de escolaridade abaixo do universitário.

Outras variáveis relacionadas à linguagem (línguas e música) também foram transformadas em variáveis binárias, pois há interesse em entender se o conhecimento nessas áreas afeta a percepção da tonicidade de palavras. A distribuição das variáveis já agrupadas também pode ser observada nas Tabelas [A.4](#) e [A.5](#), respectivamente.

Observamos no Gráfico [@ref\(fig:area\\_linguas\)](#) que, dos participantes que têm conhecimento de outras línguas, mais de metade (57,7%) são do curso de Letras, e todas as pessoas sem nenhum conhecimento de outro idioma são de outras áreas. Isso pode levar a um confundimento do efeito dessas duas variáveis, pois não há nenhum indivíduo que curse Letras sem conhecimento de outras línguas no estudo.

## 6.2 Variáveis linguísticas

A tonicidade de produção (classificação da pseudopalavra) do tipo proparoxítona é a menos expressiva na base (4%), enquanto a categoria paroxítona aparece com maior frequência nas respostas dos participantes (57%), como visto na Tabela 6.1 a seguir. Portanto, o processo de acentuação das pseudopalavras nesse estudo parece ir de acordo com a afirmação de COLLISCHONN (1999) de que

Podemos considerar que o acento proparoxítono é marcado, no sentido de que é menos usual. É um acento especial, contrário à tendência geral de acentuar a penúltima sílaba.

Tabela 6.1: Distribuição das respostas nos níveis da variável Tonicidade de produção.

Tonicidade de produção	Total
oxítona	4904 (39%)
paroxítona	7163 (57%)
proparoxítona	444 (4%)

De forma geral, comparando a tonicidade da produção e a tonicidade das palavras-alvo (vide Tabela 6.2), nota-se que aproximadamente 73% das pseudopalavras classificadas como proparoxítonas tinham uma palavra-alvo de tonicidade proparoxítona. Entretanto, cerca de 77% das pseudopalavras que têm como alvo uma palavra proparoxítona foram categorizadas pelos participantes como paroxítonas (Tabela A.6).

Além disso, nota-se também que pseudopalavras originadas de palavras-alvo oxítonas e paroxítonas foram acentuadas de forma razoavelmente uniforme entre oxítonas e paroxítonas, o que não ocorre com as pseudopalavras criadas a partir de palavras-alvo proparoxítonas. Logo, pode-se pensar que pseudopalavras derivadas de proparoxítonas não tendem a ser acentuadas com a mesma tonicidade da palavra-alvo (11,4%), enquanto pseudopalavras lidas como proparoxítonas tendem a ser derivadas de proparoxítonas (73%).

Tabela 6.2: Frequência da Tonicidade de produção por nível de Tonicidade das palavras-alvo.

Tonicidade produção	Tonicidade da palavra-alvo			Total
	oxítona	paroxítona	proparoxítona	
oxítona	2642 (53.9%)	1939 (39.5%)	323 (6.6%)	4904 (100.0%)

paroxítona	2154 (30.1%)	2824 (39.4%)	2185 (30.5%)	7163 (100.0%)
proparoxítona	49 (11.0%)	72 (16.2%)	323 (72.7%)	444 (100.0%)

Observa-se na Tabela A.7 que os grupos de classificação 2 e 4 (ou seja, os grupos em que as palavras sofreram mais alterações em relação à referência original) apresentam uma proporção menor de proparoxítonas produzidas quando comparados aos grupos 1 e 3.

O fluxo entre tonicidade da palavra-alvo e tonicidade da produção pode ser analisado graficamente por grupo de classificação na Figura 6.1.

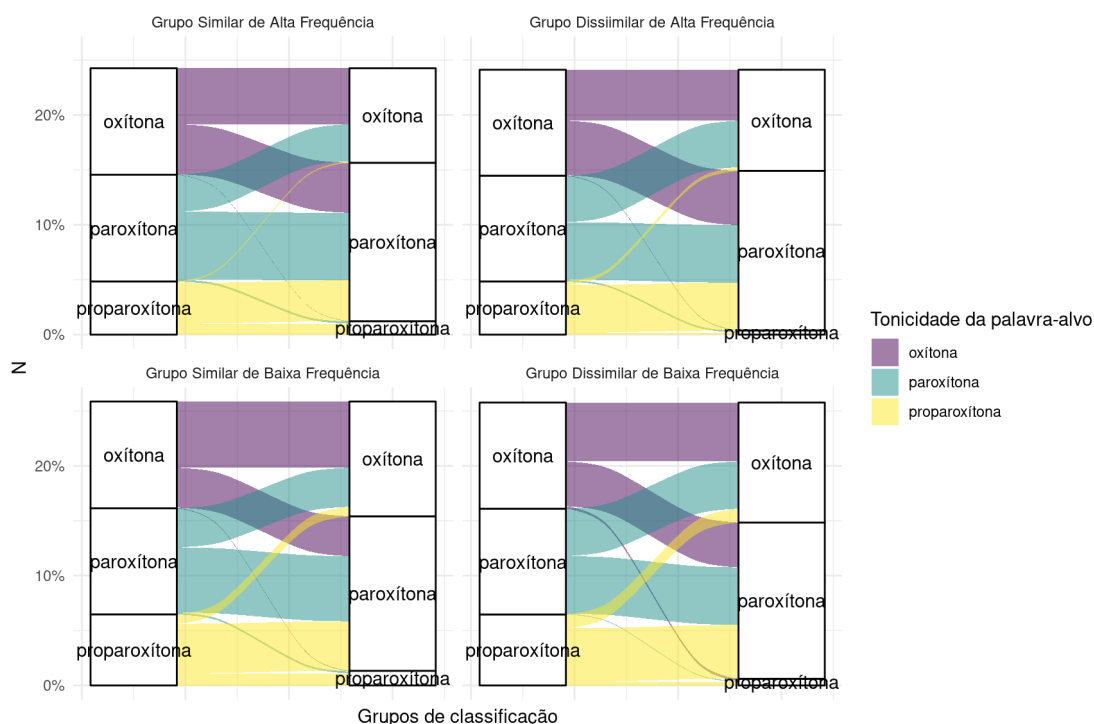


Figura 6.1: Tonicidade de palavra-alvo e pseudopalavra

No geral, podemos ver que há um “confundimento” entre oxítonas e paroxítonas em todos os grupos de classificação, ou seja, muitas pseudopalavras cuja palavra de referência é paroxítona foram classificadas pelos falantes como oxítonas, e vice-versa.

Consegue-se perceber que nos grupos de baixa frequência, a quantidade de pseudopalavras que foram classificadas como oxítona e cuja tonicidade da palavra-alvo era proparoxítona é consideravelmente maior do que nos grupos de alta frequência. Supomos que o fato de a palavra ser de baixa frequência faz com que



seja mais difícil ter uma referência na mesma estrutura. Além disso, para os grupos dissimilares a quantidade de pseudopalavras acentuadas como proparoxítonas é ainda menor do que nos grupos similares, o que pode concordar com a hipótese de que quando o falante perde a referência de uma palavra real, ele acaba acentuando-a de acordo com um padrão, que seria o paroxítono. Cada um dos gráficos de fluxo pode ser visto detalhadamente no Apêndice B-Gráficos.

Na Tabela 6.3, vemos que 85% das pseudopalavras cuja estrutura era CV-CV-CV foram classificadas como paroxítonas, e 79% das pseudopalavras cuja estrutura era CV-CV-CVC foram classificadas como oxítonas, o que corrobora com a afirmação da literatura de que uma palavra típica da língua portuguesa é formada por sílabas CV e com a tonicidade recaindo na penúltima sílaba (paroxítona), uma vez que o padrão silábico canônico do português é CV e o padrão tônico é o paroxítono. Logo, parece existir forte influência da estrutura na atribuição da tonicidade em palavras do português brasileiro.

Tabela 6.3: Frequência da Tonicidade de produção por Estrutura das palavras.

Tonicidade produção	Estrutura da palavra		Total
	CV-CV-CV	CV-CV-CVC	
oxítona	748 (10.3%)	4156 (79%)	4904 (39%)
paroxítona	6200 (85.3%)	963 (18%)	7163 (57%)
proparoxítona	323 (4.4%)	121 (2%)	444 (4%)
Total	7271 (100.0%)	5240 (100%)	12511 (100%)

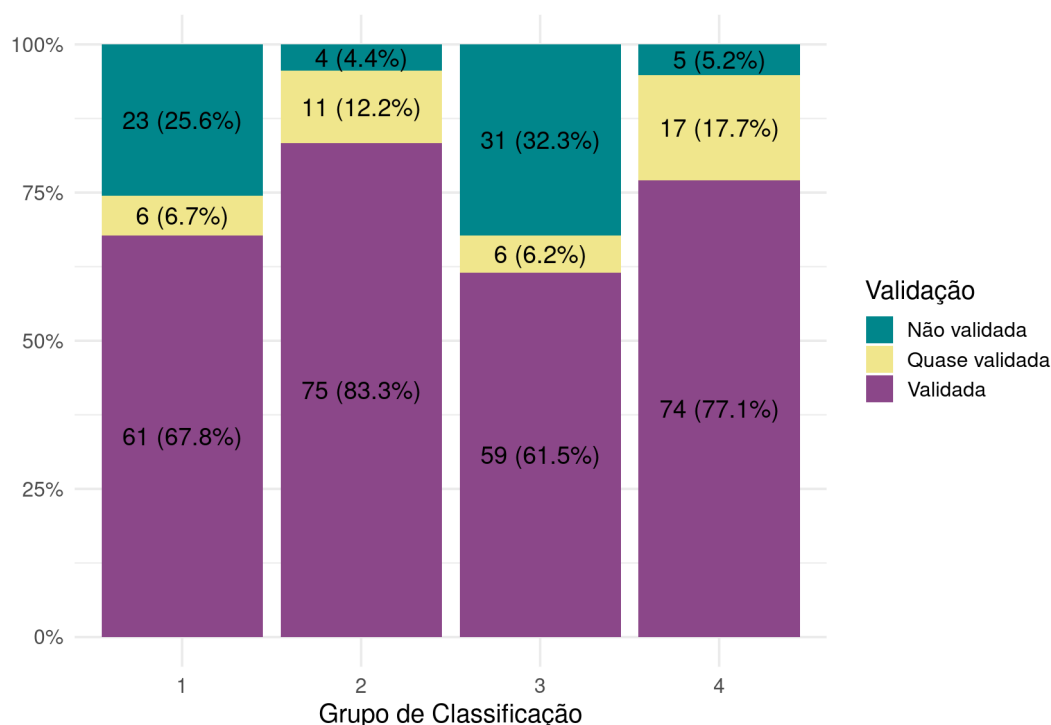
Por fim, explorando os dados no âmbito de validação, vemos, na Tabela 6.4, que aproximadamente 73% das pseudopalavras foram validadas de acordo com o processo descrito anteriormente no capítulo 4 (Descrição).

Tabela 6.4: Frequência da Tonicidade de produção, por nível da variável Validação.

Tonicidade produção	Validação		
	Não Validada	Quase Validada	Validada
oxítona	335 (15.8%)	529 (39.2%)	4040 (44.7%)
paroxítona	1727 (81.5%)	771 (57.2%)	4665 (51.6%)
proparoxítona	56 (2.6%)	48 (3.6%)	340 (3.8%)
Total	2118 (100.0%)	1348 (100.0%)	9045 (100.0%)

Cerca de 81% das pseudopalavras cuja similaridade/dissimilaridade não foi validada foram classificadas como paroxítonas. Dentre as não validadas, aproximadamente 85% das pseudopalavras pertencem ao grupo das similares (??). Ou seja, percebe-

se que nos grupos 1 (pseudopalavras similares de alta frequência) e 3 (pseudopalavras similares de baixa frequência) a porcentagem de palavras não validadas é superior aos demais grupos, o que parece indicar que a validação de palavras similares é mais complexa e difícil, uma vez que pseudopalavras classificadas como similares a palavra-alvo podem ser associadas a mais de uma palavra real pelos falantes.



### 6.3 Análise de Concordância

Como nossa base de dados está no formato de medidas repetidas (falantes acentuando diversas pseudopalavras) e uma mesma palavra-alvo gerou diversas pseudopalavras, o pressuposto de independência entre as observações não é verdadeiro. Logo, optamos por medir a concordância entre a acentuação dos falantes através de uma Estatística do tipo Kappa, separando as análises em cada nível das seguintes variáveis: Grupo de Classificação, Estrutura da palavra e Segmento Modificado.

Busca-se investigar o grau de concordância entre os falantes quanto a classificação das 372 pseudopalavras em oxítona, paroxítona e proparoxítona. Como os níveis da

variável resposta (tonicidade da produção) têm uma distribuição desbalanceada, houve a necessidade de utilizar um fator corretor no cálculo da estatística Kappa. Assim, para esse conjunto de dados, foi adotado a estatística de Gwet, medida que, de acordo com a literatura, pode ser utilizada em amostras desbalanceadas (Gwet 2002,2008).

**AC1 de Gwet:** Coeficiente para n experimentos com r avaliadores que utilizam um sistema de classificação em Q categorias

$$AC1 = \frac{p_a - p_{e\gamma}}{1 - p_{e\gamma}}$$

com

$$p_a = \frac{1}{n} \sum_{i=1}^n \left( \sum_{q=1}^Q \frac{r_{iq}(r_{iq} - 1)}{r - 1} \right)$$

$$p_{e\gamma} = \frac{1}{Q - 1} \sum_{q=1}^Q \pi_q(1 - \pi_q)$$

$$\pi_q = \frac{1}{n} \sum_{i=1}^n \frac{r_{iq}}{r}$$

Temos que

$p_{e\gamma}$  = probabilidade de concordância ao acaso  $p_a$  = probabilidade de concordância geral  $\pi_k$  = probabilidade de classificação na categoria k  $r_{ik}$  = probabilidade de classificação na categoria k

Quanto mais próximo de 1 menor a probabilidade de a concordância acontecer devido ao acaso, ou seja, melhor a concordância entre os falantes na classificação das pseudopalavras. (Altman,1991) define os seguintes pontos de corte para a interpretação da estatística Kappa (e, conseqüentemente, do AC1)

Tabela 6.5: Pontos de corte para medidas do tipo Kappa

AC1 < 0,20	pobre
0,2 <= AC1 < 0,4	razoável
0,4 <= AC1 < 0,6	moderada
0,6 <= AC1 < 0,8	boa
AC1 >= 0,8	muito boa

Além da estimação da medida geral (sem segmentar a base de acordo com alguma característica), separou-se as pseudopalavras em cada uma das categorias da variável escolhida e calculou-se a medida AC1 de concordância em cada um dos níveis.

Observando o resultado da medida sem abertura dos níveis (vide Tabela 6.6), os falantes apresentaram uma concordância boa na atribuição do acento tônico pois a estatística de Gwet produziu um coeficiente de, aproximadamente, 0.68.

Tabela 6.6: Estatística de Gwet para a base total (sem abertura por nível de uma variável).

Estatística	Valor da Estatística
Gwet's AC1	0.6775543

Já na Tabela 6.7, vemos que há semelhança de concordância entre os grupos similares e dissimilares, sendo nas dissimilares os maiores níveis de concordância (0.75 e 0.71). Porém em todos os casos pode se considerar que a concordância foi boa.

Uma hipótese para o fato de os falantes concordarem mais nos grupos dissimilares seria de que quanto menos referência entre a pseudopalavra e uma palavra real, maior a chance de o falante utilizar outra característica para definir a tonicidade (como por exemplo a estrutura da palavra).

Tabela 6.7: Estatística de Gwet para cada nível da variável Grupo de classificação).

Grupo de classificação	Valor da Estatística
Grupo 1	0.6641838
Grupo 2	0.7549719
Grupo 3	0.5874817
Grupo 4	0.7063545

Em relação a estrutura da palavra, há uma maior concordância nas respostas dos falantes cujas pseudopalavras estão no nível CV-CV-CV, como visto na ???. Dado que dentro da estrutura CV-CV-CV 85% das pseudopalavras nesse nível foram classificadas como paroxítonas, parece existir uma concordância alta entre os falantes de seguirem essa tonicidade.

Estrutura da palavra	Valor da Estatística
CV-CV-CVC	0.6582021
CV-CV-CV	0.7626996

Por fim, em relação ao segmento modificado, vemos que há semelhança na medida de concordância entre os dois níveis de mudança (0.67 e 0.68). Ou seja, os falantes concordam na tonicidade da produção de forma muito parecida entre os dois níveis como visto na ??.

#### 6.4 Resumo da Análise Descritiva

No geral, podemos ver que há um grande “confundimento” entre oxítonas e paroxítonas, além dos participantes acentuarem as pseudopalavras de forma paroxítona com bastante frequência. Logo, isso reforça a teoria de que o acento na penúltima sílaba é o padrão seguido pelos falantes do português brasileiro.

Além disso, dificilmente um falante acentua uma palavra desconhecida de forma proparoxítona e as volumetrias parecem ir ao encontro das teorias fonológicas de que se a palavra terminar em consoante o acento será oxítono e se terminar em vogal, o acento é paroxítono.

A análise de concordância parece indicar que as variáveis Grupo de classificação e Estrutura da palavra são importantes para a atribuição do acento tônico, visto que seus níveis apresentaram estatísticas de Gwet levemente diferentes entre si.

A partir desses resultados, serão desenvolvidos **Modelos de Regressão Mistos Multinomiais** com o falante como efeito aleatório, pois há interesse em analisar variáveis relacionadas aos indivíduos.

#### 6.5 Análise dos efeitos aleatórios preditos

Os efeitos aleatórios seguem o pressuposto de distribuição  $N(0, \Sigma_k)$ , logo, foram construídos gráficos de quantis-quantis para verificar a normalidade dos efeitos aleatórios preditos. No entanto, a literatura aponta que modelos mistos são robustos

Efeitos aleatórios preditos: gráfico de quantis-quantis para verificar a normalidade (no entanto, o modelo é robusto para esse pressuposto)

## 7 Metodologia

Nessa seção é feita a análise das variáveis linguísticas, extralinguísticas e experimentais com o intuito de caracterizar o processo de acentuação tônica. Para entender quais variáveis de fato têm influência construímos modelos multinomiais mistos com os falantes definidos como efeito aleatório. Assim, além de considerar a correlação das respostas do mesmo falante e o desbalanceamento na quantidade de respostas de cada um, essa estrutura possibilita a inclusão de variáveis relacionadas às características dos falantes.

As seguintes variáveis foram consideradas no modelo completo: - **Variável resposta:** Tonicidade de produção - **Efeito aleatório:** Efeito de cada falante \_  
**Variáveis com efeitos fixos:** Tonicidade alvo, Estrutura da palavra, Grupo de classificação, Segmento modificado, Bloco de apresentação, Aleatorização, Música, Línguas, Idade, Gênero, Escolaridade, Área de formação

Com isso, foram construídos modelos considerando o nível Paroxítona como categoria de referência da regressão multinomial. Ou seja, estaremos comparando através de uma função de ligação logito os níveis Oxítona/Paroxítona e Proparoxítona/Paroxítona. A escolha da categoria Paroxítona como referência aconteceu por ela ser considerada a acentuação tônica mais comum na língua portuguesa. A seguir, apresentamos as formulações matemáticas do modelo no contexto do trabalho

$$\ln\left[\frac{\pi_{ijOxítona}}{\pi_{ijParoxítona}}\right] = X'\beta_{(O-Pa)} + b_{i(O-Pa)}$$

com

$i = 1, \dots, 34$  = Falantes

$j = 1, \dots, n_i$  =  $j$ -ésima pseudopalavra acentuada pelo  $i$ -ésimo falante

$X'$  = Matriz de variáveis independentes

$\beta_{(O-Pa)}$  = Vetor de parâmetros

$b_{i(O-Pa)}$  = Efeito aleatório relacionado ao  $i$ -ésimo falante

$$\ln\left[\frac{\pi_{ijProparoxítóna}}{\pi_{ijParoxítóna}}\right] = X'\beta_{(Pro-Pa)} + b_{i(Pro-Pa)}$$

com

$i = 1, \dots, 34$  = Falantes

$j = 1, \dots, n_i$  =  $j$ -ésima pseudopalavra acentuada pelo  $i$ -ésimo falante

$X'$  = Matris de variáveis independentes

$\beta_{(Pro-Pa)}$  = Vetor de parâmetros

$b_{i(Pro-Pa)}$  = Efeito aleatório relacionado ao  $i$ -ésimo falante

Além disso, devido ao experimento relacionado à validação da similaridade/dissimilaridade entre palavra-alvo e pseudopalavra no qual algumas dessas relações não foram de fato validadas (Seção 4), testamos se a exclusão de pseudopalavras não validadas afetaria o resultado do modelo ajustado à base total. Dessa forma, foi possível entender se essa classificação afetava os valores inferenciais ou se poderíamos excluir as observações cuja similaridade/dissimilaridade entre palavra-alvo e pseudopalavra não foi validada.

Retirando amostras aleatórias contendo 25% de pseudopalavras não validadas, observou-se que o resultado do modelo é alterado de formas diferentes, a depender da amostra. Logo, entendemos que não seria possível remover apenas uma porcentagem aleatória das pseudopalavras cuja similaridade/dissimilaridade não for validada.

Como explicado anteriormente, essa variável foi utilizada para validar outra variável da pesquisa (Grupo de classificação). Então, para reduzir a incerteza em relação à definição do Grupo de classificação, decidiu-se por ajustar um modelo apenas com as pseudopalavras validadas e quase validadas, excluindo as não validadas.

## 7.1 Modelos Ajustados

Consideramos dois modelos possíveis: um com a variável Grupo de classificação, que foi a variável construída pela pesquisadora, e outro em que essa variável é separada em duas: Similaridade e Frequência.

Com esse segundo modelo, podemos identificar o efeito marginal da similaridade ou da frequência de forma separada, o que não é possível fazer com a variável Grupo de classificação, que já considera a interação entre essas duas variáveis. Portanto, o processo de ajuste do modelo e seleção de variáveis foi feito para esses dois modelos, descritos a seguir.

#### *7.1.1 Modelo com Grupo de classificação (Modelo 1)*

Como o modelo multinomial é constituído de duas regressões neste caso, em que a variável resposta contém três níveis, foi feito um processo de seleção de variáveis *backwards* manual, considerando o valor-p das estimativas nos dois modelos de forma simultânea.

Adotando 5% como nível de significância, verificou-se que a maior parte das variáveis relacionadas aos falantes não apresentaram significância nos modelos ajustados. Depois de retirar uma a uma as variáveis não significantes nos dois modelos, testamos retirar a variável extralinguística que apresentou apenas um nível significativo (Escolaridade), já que essa variável possuía apenas uma observação em um dos níveis.

Comparamos o critério de informação de Akaike (AIC) dos modelos com e sem Escolaridade, o modelo sem Escolaridade apresentou menor AIC, logo, concluímos que a exclusão apresentou melhora no modelo.

Portanto, as variáveis que apresentaram significância em pelo menos um nível foram: *tonicidade\_alvo*, *estrutura\_palavra*, *grupo*, *segmento\_modificado*, *bloco\_apresentação* e *aleatorização*.

As Tabelas A.8 e A.9 contêm as estimativas, erros padrões e valores-p dos efeitos fixos obtidos pelo modelo final ajustado. A Tabela A.8 apresenta os valores para o modelo dos níveis Oxítone/Paroxítone e a Tabela A.9 apresenta os valores para o modelo dos níveis Proparoxítone/Paroxítone. É importante observar que, como as variáveis foram selecionadas de forma simultânea para os dois modelos, nem sempre os coeficientes das variáveis serão representadas em ambas as equações,



pois podem ter níveis de significância diferentes entre os modelos

Oxítona/Paroxítona e Proparoxítona/Paroxítona.

term	estimate	std.error	statistic	p.value
oxítona~(Intercept)	-1.9497069	0.2708188	-7.199305	0.0000000
oxítona~tonicidade_alvoparoxítona	-1.3922195	0.0754364	-18.455547	0.0000000
oxítona~tonicidade_alvoproparoxítona	-2.0137152	0.1056674	-19.057103	0.0000000
oxítona~estrutura_palavraCV-CV-CVC	4.2931359	0.0797580	53.827055	0.0000000
oxítona~grupo2	0.1052218	0.0996410	1.056009	0.2909640
oxítona~grupo3	0.5556489	0.1031689	5.385818	0.0000001
oxítona~grupo4	0.4185702	0.0983536	4.255768	0.0000208
oxítona~segmento_modificadovogal	0.1874888	0.0644308	2.909926	0.0036151
oxítona~bloco_apresentacao2	0.4266194	0.0923129	4.621449	0.0000038
oxítona~bloco_apresentacao3	0.1412840	0.0938383	1.505611	0.1321670
oxítona~bloco_apresentacao4	0.3999452	0.0964728	4.145680	0.0000339
oxítona~aleatorizacaon	-0.6377245	0.1662861	-3.835104	0.0001255

term	estimate	std.error	statistic	p.value
proparoxítona~(Intercept)	-3.6852650	0.3509559	-10.5006500	0.0000000
proparoxítona~tonicidade_alvoparoxítona	-0.6209199	0.2111748	-2.9403118	0.0032788
proparoxítona~tonicidade_alvoproparoxítona	1.6686983	0.1865004	8.9474227	0.0000000
proparoxítona~estrutura_palavraCV-CV-CVC	1.8140990	0.1478927	12.2663211	0.0000000
proparoxítona~grupo2	-1.3384948	0.1845553	-7.2525410	0.0000000
proparoxítona~grupo3	0.1631648	0.1518148	1.0747622	0.2824812
proparoxítona~grupo4	-1.1391548	0.1618900	-7.0365964	0.0000000
proparoxítona~segmento_modificadovogal	0.2943657	0.1178233	2.4983655	0.0124767
proparoxítona~bloco_apresentacao2	-0.7940043	0.1957112	-4.0570191	0.0000497
proparoxítona~bloco_apresentacao3	-0.2160771	0.1583277	-1.3647457	0.1723330
proparoxítona~bloco_apresentacao4	-0.1794640	0.1594542	-1.1254896	0.2603816
proparoxítona~aleatorizacaon	0.2171223	0.2734835	0.7939138	0.4272456

### 7.1.1.1 Interpretações dos parâmetros (Modelo 1)

A partir da saída percebe-se que, mantidas as demais variáveis em seus respectivos níveis de referência:

1. A probabilidade de produção oxítona (ao invés de produção paroxítona) quando a estrutura da palavra é CV-CV-CVC é, aproximadamente, 73 vezes a probabilidade de produção oxítona (ao invés de produção paroxítona) quando a estrutura da palavra é CV-CV-CV. Já a probabilidade de produção proparoxítona (ao invés de produção paroxítona) quando a estrutura da palavra é CV-CV-CVC é, aproximadamente, 6 vezes a probabilidade de produção proparoxítona (ao invés de produção paroxítona) quando a estrutura da palavra é CV-CV-CV.
2. A probabilidade de produção oxítona (ao invés de produção paroxítona) quando a palavra-alvo é paroxítona é, aproximadamente, 0,25 vezes a

probabilidade de produção oxítônica (ao invés de produção paroxítônica) quando a palavra-alvo é oxítônica. Já a probabilidade de produção proparoxítônica (ao invés de produção paroxítônica) quando a palavra-alvo é paroxítônica é, aproximadamente, 0,54 vezes a probabilidade de produção proparoxítônica (ao invés de produção paroxítônica) quando a palavra-alvo é oxítônica.

3. A probabilidade de produção oxítônica (ao invés de produção paroxítônica) quando a palavra-alvo é proparoxítônica é, aproximadamente, 0,13 vezes a probabilidade de produção oxítônica (ao invés de produção paroxítônica) quando a palavra-alvo é proparoxítônica. Já a probabilidade de produção proparoxítônica (ao invés de produção paroxítônica) quando a palavra-alvo é proparoxítônica é, aproximadamente, 5 vezes a probabilidade de produção proparoxítônica (ao invés de produção paroxítônica) quando a palavra-alvo é proparoxítônica.

4. A probabilidade de produção oxítônica (ao invés de produção paroxítônica) quando o segmento modificado foi uma vogal é, aproximadamente, 1,2 vezes a probabilidade de produção oxítônica (ao invés de produção paroxítônica) quando o segmento modificado foi uma consoante. Já a probabilidade de produção proparoxítônica (ao invés de produção paroxítônica) quando o segmento modificado foi uma vogal é, aproximadamente, 1,3 vezes a probabilidade de produção proparoxítônica (ao invés de produção paroxítônica) quando o segmento modificado foi uma consoante.

5. A probabilidade de produção oxítônica (ao invés de produção paroxítônica) quando a pseudopalavra é do grupo 2 é 0 vezes a probabilidade de produção oxítônica (ao invés de produção paroxítônica) quando a pseudopalavra é do grupo 1. Já a probabilidade de produção proparoxítônica (ao invés de produção paroxítônica) quando a pseudopalavra é do grupo 2 é 0,26 vezes a probabilidade de produção proparoxítônica (ao invés de produção paroxítônica) quando a pseudopalavra é do grupo 1.

A probabilidade de produção oxítônica (ao invés de produção paroxítônica) quando a pseudopalavra é do grupo 3 é 1,74 vezes a probabilidade de produção oxítônica (ao invés de produção paroxítônica) quando a pseudopalavra é do grupo 1. Já a

probabilidade de produção proparoxítona (ao invés de produção paroxítona) quando a pseudopalavra é do grupo 3 é 0 vezes a probabilidade de produção proparoxítona (ao invés de produção paroxítona) quando a pseudopalavra é do grupo 1.

A probabilidade de produção oxítona (ao invés de produção paroxítona) quando a pseudopalavra é do grupo 4 é 1,52 vezes a probabilidade de produção oxítona (ao invés de produção paroxítona) quando a pseudopalavra é do grupo 1. Já a probabilidade de produção proparoxítona (ao invés de produção paroxítona) quando a pseudopalavra é do grupo 4 é 0,32 vezes a probabilidade de produção proparoxítona (ao invés de produção paroxítona) quando a pseudopalavra é do grupo 1.

Podemos resumir que, na regressão oxítona/paroxítona, a estrutura da pseudopalavra é o que mais influencia na produção oxítona, além da tonicidade da palavra-alvo (quando a mesma difere de oxítona, a probabilidade de produção oxítona diminui entre 75% e 90%, se comparada à da tonicidade alvo oxítona) e o segmento modificado (se a mudança ocorre em uma vogal, a probabilidade de produção oxítona aumenta em 20% se comparada à probabilidade de produção oxítona com mudança em consoante). Outro ponto de destaque vai para a influência dos grupos de classificação: para os grupos 3 (similar de baixa frequência) e 4 (dissimilar de baixa frequência), a probabilidade de produção oxítona aumenta em 70% e 52%, respectivamente, quando comparada ao grupo 1 (similar de alta frequência).

#### **7.1.1.2 Avaliação do Modelo (Modelo 1)**

Não há, na literatura, explicações de métodos para analisar resíduos e encontrar pontos aberrantes em modelos multinomiais, apenas para modelos logísticos (conforme - citar artigo da prof viviana).

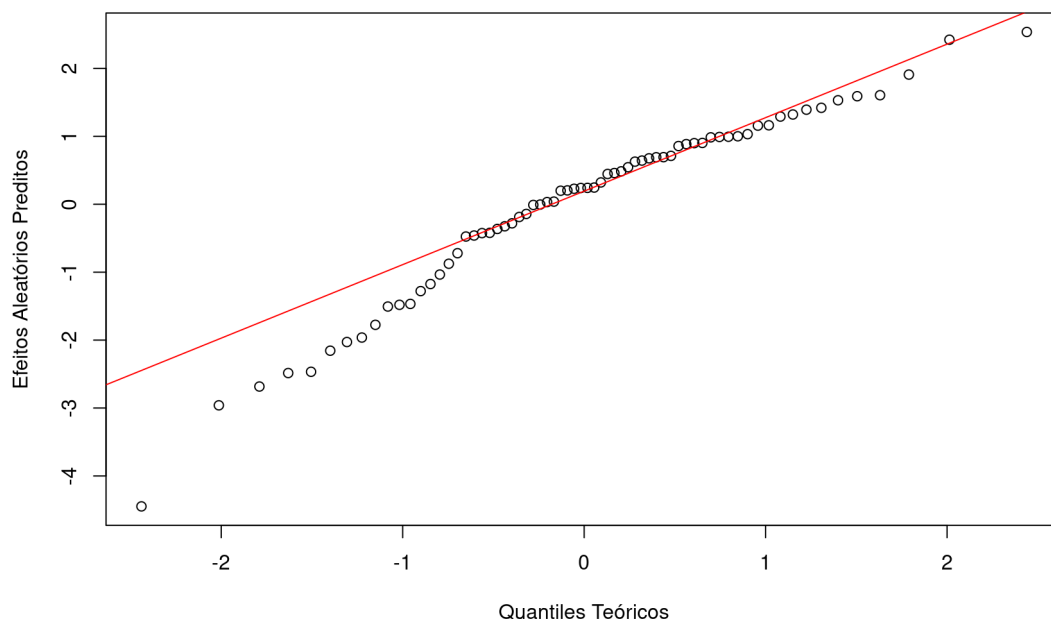
Como alternativa, conseguimos prever, para cada observação, a probabilidade a probabilidade da mesma ser classificada em cada nível da variável dependente. Assim, é possível avaliar a qualidade do modelo através do cruzamento da variável dependente com a maior probabilidade predita de cada observação.

Calculamos a sensibilidade (probabilidade de classificação no nível  $i$  dado que a tonicidade de produção foi avaliada no nível  $i$ ) e especificidade (probabilidade de não ser classificada no nível  $i$  dado que a tonicidade de produção não foi avaliada no nível  $i$ ) de cada nível da variável resposta. As medidas são feitas agrupando dois níveis da variável dependente e os comparando com o terceiro. Os valores podem ser analisados nas Tabelas A.10, A.11 e A.12, para os níveis Oxítone, Paroxítone e Proparoxítone, respectivamente.

Vemos que para Oxítone e Paroxítone, os valores de Sensibilidade e Especificidade são altos, o que indica um bom ajuste do modelo aos dados. Olhando para Proparoxítone, vemos que o valor da Sensibilidade é baixo, o que pode ser explicado pela baixa frequência dessa classificação na base. Concluímos então que o ajuste do modelo está satisfatório.

#### **7.1.1.3 Análise dos Componentes Aleatórios**

Em modelos mistos, os efeitos aleatórios seguem o pressuposto de distribuição  $N(0, \Sigma)$ . Embora a figura aponte uma distribuição com caudas pesadas, a literatura indica que essa é uma premissa robusta, logo, podemos considerar que houve uma boa aproximação para a distribuição normal.



### 7.1.2 Modelo com Similaridade e Frequência (Modelo 2)

Como o modelo multinomial é constituído de duas regressões neste caso, em que a variável resposta contém três níveis, foi feito um processo de seleção de variáveis *backwards* manual, considerando o valor-p das estimativas nos dois modelos de forma simultânea.

Adotando 5% como nível de significância, verificou-se que a maior parte das variáveis relacionadas aos falantes não apresentaram significância nos modelos ajustados. Depois de retirar uma a uma as variáveis não significantes nos dois modelos, testamos retirar a variável extralinguística que apresentou apenas um nível significativo (Escolaridade), já que essa variável possuía apenas uma observação em um dos níveis.

Comparamos o critério de informação de Akaike (AIC) dos modelos com e sem Escolaridade, o modelo sem Escolaridade apresentou menor AIC, logo, concluímos que a exclusão apresentou melhora no modelo.

Portanto, as variáveis que apresentaram significância em pelo menos um nível foram: *tonicidade\_alvo*, *estrutura\_palavra*, *similaridade*, *frequência*, *segmento\_modificado*, *bloco\_apresentação* e *aleatorização*.

As Tabelas A.8 e A.9 contêm as estimativas, erros padrões e valores-p dos efeitos fixos obtidos pelo modelo final ajustado. A Tabela A.8 apresenta os valores para o modelo dos níveis Oxítona/Paroxítona e a Tabela A.9 apresenta os valores para o modelo dos níveis Proparoxítona/Paroxítona. É importante observar que, como as variáveis foram selecionadas de forma simultânea para os dois modelos, nem sempre os coeficientes das variáveis serão representadas em ambos as equações, pois podem ter níveis de significância diferentes entre os modelos Oxítona/Paroxítona e Proparoxítona/Paroxítona.

term	estimate	std.error	statistic	p.value
oxítona~(Intercept)	-1.4738815	0.2589150	-5.6925302	0.0000000
oxítona~tonicidade_alvoparoxítona	-1.3974574	0.0754794	-18.5144243	0.0000000
oxítona~tonicidade_alvoproparoxítona	-2.0159961	0.1054633	-19.1156129	0.0000000
oxítona~estrutura_palavraCV-CV-CVC	4.2845534	0.0795045	53.8907137	0.0000000
oxítona~similaridade1	0.0288691	0.0680648	0.4241417	0.6714625
oxítona~frequencia1	-0.4121105	0.0663835	-6.2080245	0.0000000
oxítona~segmento_modificadovogal	0.1898682	0.0644222	2.9472490	0.0032061
oxítona~bloco_apresentacao2	0.4180387	0.0922466	4.5317501	0.0000058
oxítona~bloco_apresentacao3	0.1377549	0.0937723	1.4690355	0.1418231
oxítona~bloco_apresentacao4	0.3848272	0.0960415	4.0068854	0.0000615
oxítona~aleatorizacaon	-0.6357210	0.1662005	-3.8250242	0.0001308

term	estimate	std.error	statistic	p.value
proparoxítona~(Intercept)	-4.8315109	0.3531843	-13.6798577	0.0000000
proparoxítona~tonicidade_alvoparoxítona	-0.6212719	0.2109469	-2.9451579	0.0032279
proparoxítona~tonicidade_alvoproparoxítona	1.6623087	0.1857496	8.9491894	0.0000000
proparoxítona~estrutura_palavraCV-CV-CVC	1.8092469	0.1480389	12.2214326	0.0000000
proparoxítona~similaridade1	1.3074342	0.1232610	10.6070387	0.0000000
proparoxítona~frequencia1	-0.1553983	0.1215618	-1.2783473	0.2011270
proparoxítona~segmento_modificadovogal	0.2956344	0.1177929	2.5097808	0.0120806
proparoxítona~bloco_apresentacao2	-0.7940670	0.1942528	-4.0878016	0.0000435
proparoxítona~bloco_apresentacao3	-0.2176893	0.1583112	-1.3750718	0.1691092
proparoxítona~bloco_apresentacao4	-0.1810994	0.1589141	-1.1396055	0.2544507
proparoxítona~aleatorizacaon	0.2172405	0.2732678	0.7949728	0.4266293

### 7.1.2.1 Interpretações dos parâmetros (Modelo 2)

A partir da saída percebe-se que, mantidas as demais variáveis em seus respectivos níveis de referência:

- 1.

A probabilidade de uma pseudopalavra ser acentuada como oxítone (e não como paroxítone) quando sua palavra-alvo é paroxítone é aproximadamente 0.25 da probabilidade dela ser acentuada como como oxítone (e não como paroxítone) quando sua palavra-alvo é oxítone. Já a probabilidade de uma pseudopalavra ser acentuada como proparoxítone (e não como paroxítone) quando sua palavra alvo é paroxítone é aproximadamente 0.54 da probabilidade dela ser acentuada como como proparoxítone (e não como paroxítone) quando sua palavra alvo é oxítone.

2. A probabilidade de uma pseudopalavra ser acentuada como oxítone (e não como paroxítone) quando sua palavra alvo é proparoxítone é 0.13 vezes a probabilidade dela ser acentuada como como oxítone (e não como paroxítone) quando sua palavra alvo é oxítone. Já a probabilidade de uma pseudopalavra ser acentuada como proparoxítone (e não como paroxítone) quando sua palavra alvo é proparoxítone é 5.27 vezes a probabilidade dela ser acentuada como proparoxítone (e não como paroxítone) quando sua palavra alvo é oxítone.
3. A probabilidade de uma pseudopalavra ser acentuada como oxítone (e não como paroxítone) quando sua estrutura é no formato CV-CV-CVC é aproximadamente 73 vezes a probabilidade de uma pseudopalavra ser acentuada como oxítone (e não como paroxítone) quando sua estrutura é no formato CV-CV-CV. Já a probabilidade de uma pseudopalavra ser acentuada como proparoxítone (e não como paroxítone) quando sua estrutura é no formato CV-CV-CVC é 6.10 vezes a probabilidade de uma pseudopalavra ser acentuada como proparoxítone (e não como paroxítone) quando sua estrutura é no formato CV-CV-CV.
4. A probabilidade de uma pseudopalavra ser acentuada como oxítone (e não como paroxítone) quando sua palavra alvo é de alta frequência é 0.66 vezes a probabilidade de uma pseudopalavra ser acentuada como oxítone (e não como paroxítone) quando sua sua palavra alvo é de baixa frequência. Já a probabilidade de uma pseudopalavra ser acentuada como proparoxítone (e

não como paroxítona) quando sua palavra alvo é de alta frequência é 0.85 vezes a probabilidade de uma pseudopalavra ser acentuada como proparoxítona (e não como paroxítona) quando sua palavra alvo é de baixa frequência. (não significativa)

5) A probabilidade de uma pseudopalavra ser acentuada como oxítona (e não como paroxítona) quando sua palavra alvo é de dissimilar é 1.02 vezes a probabilidade de uma pseudopalavra ser acentuada como oxítona (e não como paroxítona) quando sua palavra alvo é dissimilar (não significativa). Já a probabilidade de uma pseudopalavra ser acentuada como proparoxítona (e não como paroxítona) quando ela foi classificada como similar à sua palavra alvo é 3.7 vezes a probabilidade de uma pseudopalavra ser acentuada como proparoxítona (e não como paroxítona) quando ela foi classificada como dissimilar à sua palavra alvo.

6. A probabilidade de uma pseudopalavra ser acentuada como oxítona (e não como paroxítona) quando a mudança ocorreu em uma vogal é 1.21 vezes a probabilidade de uma pseudopalavra ser acentuada como oxítona (e não como paroxítona) quando a mudança ocorreu em uma consoante. Já a probabilidade de uma pseudopalavra ser acentuada como proparoxítona (e não como paroxítona) quando a mudança ocorreu em uma vogal é 1.34 vezes a probabilidade de uma pseudopalavra ser acentuada como proparoxítona (e não como paroxítona) quando a mudança ocorreu em uma consoante.

Podemos resumir que, na regressão oxítona/paroxítona, a estrutura da pseudopalavra é o que mais influencia na produção oxítona, além da tonicidade da palavra-alvo (quando a mesma difere de oxítona, a probabilidade de produção oxítona diminui entre 75% e 90%, se comparada à da tonicidade alvo oxítona) e o segmento modificado (se a mudança ocorre em uma vogal, a probabilidade de produção oxítona aumenta em 21% se comparada à probabilidade de produção oxítona com mudança em consoante). Outro ponto de destaque vai para a variável frequência: pseudopalavras com palavras alvo classificadas como de alta frequência tem uma redução de  $\frac{1}{3}$  na probabilidade de serem acentuadas como oxítonas (e



não paroxítonas) - ou seja, pseudopalavras com palavras alvo de baixa frequência tem maior probabilidade de serem acentuadas como oxítonas-.

Para a regressão proparoxítona/paroxítona, é possível notar que grande parte das variáveis se mostrou de grande impacto na tonicidade atribuída proparoxítona. Podemos destacar a tonicidade alvo proparoxítona, que aumenta a probabilidade de produção proparoxítona -e não paroxítona- em mais de 400% (em relação a probabilidade de produção proparoxítona com tonicidade alvo oxítona), a estrutura da palavra CV-CV-CVC, que aumenta a probabilidade de produção proparoxítona -e não paroxítona- em 500% (em relação à probabilidade de produção proparoxítona com estrutura CV-CV-CV) e a similaridade que, para pseudopalavras classificadas como similares à sua palavra alvo, sua probabilidade de produção proparoxítona -e não paroxítona- aumenta em aproximadamente 270% (quando comparada à probabilidade de produção proparoxítona de pseudopalavras dissimilares).

#### **7.1.2.2 Avaliação do Modelo (Modelo 2)**

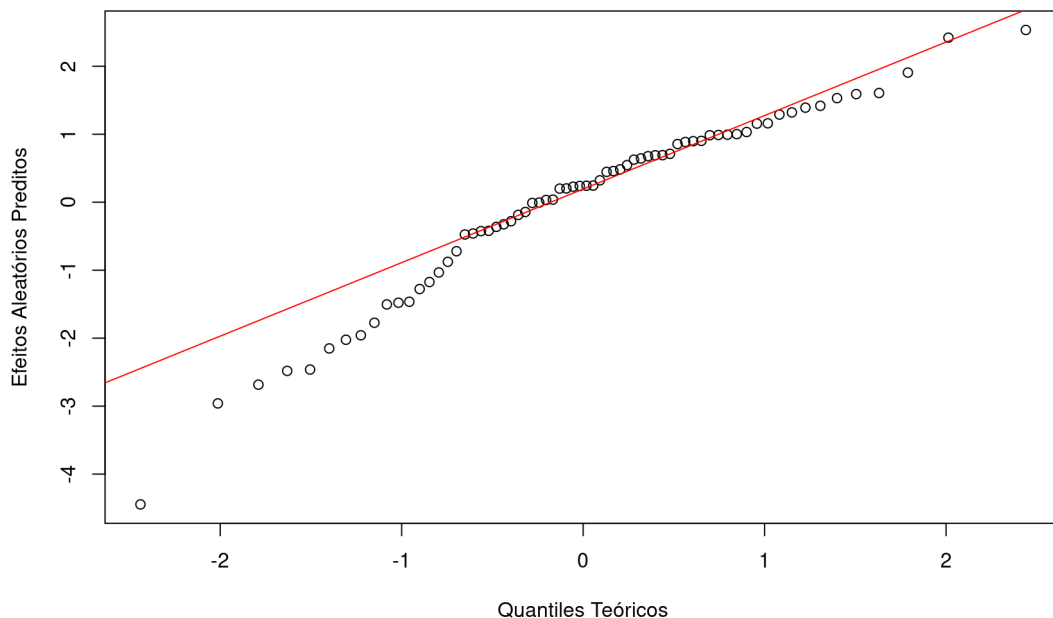
Não há, na literatura, explanações de métodos para analisar resíduos e encontrar pontos aberrantes em modelos multinomiais, apenas para modelos logísticos (conforme - citar artigo da prof viviana). Como alternativa, conseguimos prever, para cada observação, a probabilidade de resposta em cada nível da variável dependente. Assim, é possível avaliar a qualidade do modelo através do cruzamento da variável dependente com a maior probabilidade predita de cada observação. Com base nisso, calculamos a sensibilidade (probabilidade de classificação no nível  $i$  dado que a tonicidade de produção foi avaliada no nível  $i$ ) e especificidade (probabilidade de não ser classificada no nível  $i$  dado que a tonicidade de produção não foi avaliada no nível  $i$ ) de cada nível da variável resposta. As medidas são feitas agrupando dois níveis da variável dependente e os comparando com o terceiro. Os valores podem ser analisados nas Tabelas A.10, A.11 e A.12, para os níveis Oxítona, Paroxítona e Proparoxítona, respectivamente.

Vemos que para Oxítona e Paroxítona, os valores de Sensibilidade e Especificidade são altos, o que indica um bom ajuste do modelo aos dados. Olhando para

Proparoxítona, vemos que a o valor da Sensibilidade é baixo, o que pode ser explicado pela baixa frequência dessa classificação na base. Concluimos então que o ajuste do modelo está satisfatório.

### 7.1.2.3 Análise dos Componentes Aleatórios

Em modelos mistos, os efeitos aleatórios seguem o pressuposto de distribuição  $N(0, \Sigma)$ . Embora a figura aponte uma distribuição com caldas pesadas, a literatura indica que essa é uma premissa robusta, logo, podemos considerar que houve uma boa aproximação para a distribuição normal.



### 7.1.3 Problema de Aleatorização

Um ponto de atenção é a variável Aleatorização que apresentou significância estatística na regressão Oxítona/Paroxítona, assim como alguns níveis da variável Bloco de Apresentação em ambas as regressões -Oxítona/Paroxítona e Proparoxítona/Paroxítona-. Ambas variáveis estão relacionadas ao experimento e a Aleatorização tem impacto direto no Bloco de Apresentação. Ou seja, há evidências de que os problemas ocorridos com os participantes que prejudicaram a apresentação inicial das pseudopalavras impactaram na acentuação atribuída pelos

falantes. Não foi possível contornar esse problema durante a modelagem, o que talvez indique a necessidade de replicar o experimento garantindo a aleatoriedade de todos os blocos para todos os falantes.

## **8 Conclusão**

A partir da análise inferencial concluímos que há indícios de que existem outros fatores, além da estrutura da palavra, que interferem na acentuação tônica. Por meio dos modelos foi evidenciado o papel da similaridade e da tonicidade da palavra-alvo na determinação da tonicidade de produção. No entanto, como a interferência na aleatorização durante o processo de coleta dos dados mostrou-se significativa para o modelo final, acreditamos que seria importante uma replicação do estudo, no qual se garantisse a aleatorização de todos os blocos para todos os falantes.

## Apêndices

### A Tabelas

Tabela A.1: Perfil dos participantes de acordo com a Naturalidade.

Naturalidade	Total
São Paulo, SP	16
Outros municípios de SP	12
Outras UF	6

Tabela A.2: Frequência da variável dicotômica Área de formação.

Área de formação	Total
Outro	19 (56%)
Letras	15 (44%)

Tabela A.3: Escolaridade dos participantes pela Área de formação.

Escolaridade	Área de formação		
	Outro	Letras	Total
1. Fundamental Completo	1 (2.9%)	0 (0.0%)	1 (2.9%)
2. Superior Incompleto	6 (17.6%)	11 (32.4%)	17 (50.0%)
3. Superior Completo	5 (14.7%)	1 (2.9%)	6 (17.6%)
4. Pós-Graduação (Completo ou Incompleto)	7 (20.6%)	3 (8.8%)	10 (29.4%)
Total	19 (55.9%)	15 (44.1%)	34 (100.0%)

Tabela A.4: Frequência da variável dicotômica Línguas.

Conhecimento em línguas	Total
Sim	26
Não	8

Tabela A.5: Frequência da variável dicotômica Música.

Conhecimento em música	Total
Sim	21
Não	13

Tabela A.6: Frequência da tonicidade das palavras-alvo por nível de tonicidade de produção

Tonicidade produção	Tonicidade da palavra-alvo			Total
	oxítona	paroxítona	proparoxítona	
oxítona	2642 (54.5%)	1939 (40.1%)	323 (11.4%)	4904 (39.2%)
paroxítona	2154 (44.5%)	2824 (58.4%)	2185 (77.2%)	7163 (57.3%)

proparoxítona	49 (1.0%)	72 (1.5%)	323 (11.4%)	444 (3.5%)
Total	4845 (100.0%)	4835 (100.0%)	2831 (100.0%)	12511 (100.0%)

Tabela A.7: Frequência do Grupo de classificação por Tonicidade de produção (da pseudopalavra).

Tonicidade produção	Grupo de classificação				Total
	1 (Similar de alta freq.)	2 (Dissimilar de alta freq.)	3 (Similar de baixa freq.)	4 (Dissimilar de baixa freq.)	
oxítona	1077 (22.0%)	1152 (23.5%)	1308 (26.7%)	1367 (27.9%)	4904 (100%)
paroxítona	1805 (25.2%)	1818 (25.4%)	1760 (24.6%)	1780 (24.8%)	7163 (100%)
proparoxítona	153 (34.5%)	48 (10.8%)	167 (37.6%)	76 (17.1%)	444 (100%)

Tabela A.8: Tonicidade de produção por Grupo de classificação e Validação.

Grupo de classificação	Tonicidade produção	Status Validação		
		Não Validada	Quase Validada	Validada
1	oxítona	82 (3.9%)	44 (3.3%)	951 (10.5%)
1	paroxítona	673 (31.8%)	141 (10.5%)	991 (11.0%)
1	proparoxítona	17 (0.8%)	17 (1.3%)	119 (1.3%)
2	oxítona	38 (1.8%)	154 (11.4%)	960 (10.6%)
2	paroxítona	96 (4.5%)	210 (15.6%)	1512 (16.7%)
2	proparoxítona	1 (0.0%)	8 (0.6%)	39 (0.4%)
3	oxítona	183 (8.6%)	109 (8.1%)	1016 (11.2%)
3	paroxítona	824 (38.9%)	87 (6.5%)	849 (9.4%)
3	proparoxítona	37 (1.7%)	7 (0.5%)	123 (1.4%)
4	oxítona	32 (1.5%)	222 (16.5%)	1113 (12.3%)
4	paroxítona	134 (6.3%)	333 (24.7%)	1313 (14.5%)
4	proparoxítona	1 (0.0%)	16 (1.2%)	59 (0.7%)
Total	•	2118 (100.0%)	1348 (100.0%)	9045 (100.0%)

Segmento Modificado	Valor da Estatística
vogal	0.6725362
consoante	0.6826297

FALSE	Outcome +	Outcome -	Total
FALSE Test +	3958	692	4650
FALSE Test -	611	5132	5743
FALSE Total	4569	5824	10393
FALSE			

FALSE Point estimates and 95 % CIs:

FALSE	-----
FALSE Apparent prevalence	0.45 (0.44, 0.46)
FALSE True prevalence	0.44 (0.43, 0.45)
FALSE Sensitivity	0.87 (0.86, 0.88)
FALSE Specificity	0.88 (0.87, 0.89)
FALSE Positive predictive value	0.85 (0.84, 0.86)
FALSE Negative predictive value	0.89 (0.89, 0.90)
FALSE Positive likelihood ratio	7.29 (6.79, 7.83)
FALSE Negative likelihood ratio	0.15 (0.14, 0.16)
FALSE	-----

FALSE	Outcome +	Outcome -	Total
FALSE Test +	4836	843	5679
FALSE Test -	600	4114	4714
FALSE Total	5436	4957	10393

FALSE

FALSE Point estimates and 95 % CIs:

FALSE	-----
FALSE Apparent prevalence	0.55 (0.54, 0.56)
FALSE True prevalence	0.52 (0.51, 0.53)
FALSE Sensitivity	0.89 (0.88, 0.90)
FALSE Specificity	0.83 (0.82, 0.84)
FALSE Positive predictive value	0.85 (0.84, 0.86)
FALSE Negative predictive value	0.87 (0.86, 0.88)
FALSE Positive likelihood ratio	5.23 (4.92, 5.57)
FALSE Negative likelihood ratio	0.13 (0.12, 0.14)
FALSE	-----

FALSE	Outcome +	Outcome -	Total
FALSE Test +	42	22	64
FALSE Test -	346	9983	10329
FALSE Total	388	10005	10393

FALSE

FALSE Point estimates and 95 % CIs:

FALSE	-----
FALSE Apparent prevalence	0.01 (0.00, 0.01)
FALSE True prevalence	0.04 (0.03, 0.04)
FALSE Sensitivity	0.11 (0.08, 0.14)
FALSE Specificity	1.00 (1.00, 1.00)
FALSE Positive predictive value	0.66 (0.53, 0.77)
FALSE Negative predictive value	0.97 (0.96, 0.97)
FALSE Positive likelihood ratio	49.23 (29.69, 81.63)
FALSE Negative likelihood ratio	0.89 (0.86, 0.93)
FALSE	-----

FALSE	Outcome +	Outcome -	Total
FALSE Test +	3955	697	4652
FALSE Test -	614	5127	5741
FALSE Total	4569	5824	10393

FALSE

FALSE Point estimates and 95 % CIs:

FALSE	-----
FALSE Apparent prevalence	0.45 (0.44, 0.46)
FALSE True prevalence	0.44 (0.43, 0.45)
FALSE Sensitivity	0.87 (0.86, 0.88)
FALSE Specificity	0.88 (0.87, 0.89)
FALSE Positive predictive value	0.85 (0.84, 0.86)

FALSE Negative predictive value	0.89 (0.88, 0.90)
FALSE Positive likelihood ratio	7.23 (6.74, 7.76)
FALSE Negative likelihood ratio	0.15 (0.14, 0.16)
FALSE -----	

FALSE	Outcome +	Outcome -	Total
FALSE Test +	4832	847	5679
FALSE Test -	604	4110	4714
FALSE Total	5436	4957	10393

FALSE

FALSE Point estimates and 95 % CIs:

FALSE	-----
FALSE Apparent prevalence	0.55 (0.54, 0.56)
FALSE True prevalence	0.52 (0.51, 0.53)
FALSE Sensitivity	0.89 (0.88, 0.90)
FALSE Specificity	0.83 (0.82, 0.84)
FALSE Positive predictive value	0.85 (0.84, 0.86)
FALSE Negative predictive value	0.87 (0.86, 0.88)
FALSE Positive likelihood ratio	5.20 (4.89, 5.54)
FALSE Negative likelihood ratio	0.13 (0.12, 0.14)
FALSE	-----

FALSE	Outcome +	Outcome -	Total
FALSE Test +	42	20	62
FALSE Test -	346	9985	10331
FALSE Total	388	10005	10393

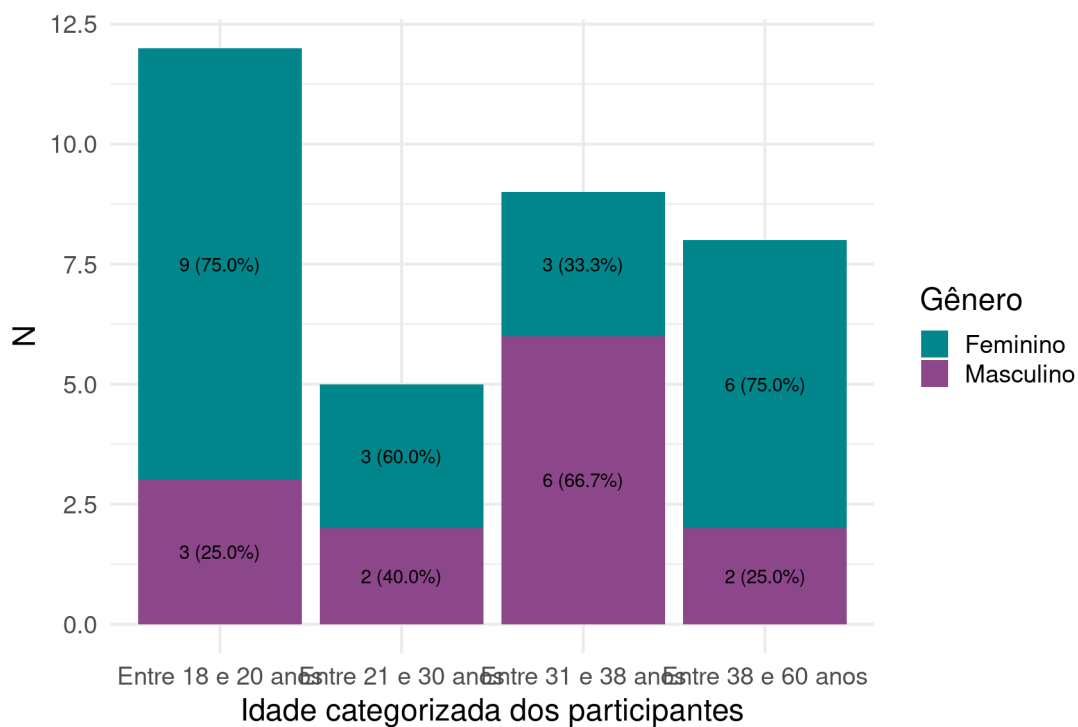
FALSE

FALSE Point estimates and 95 % CIs:

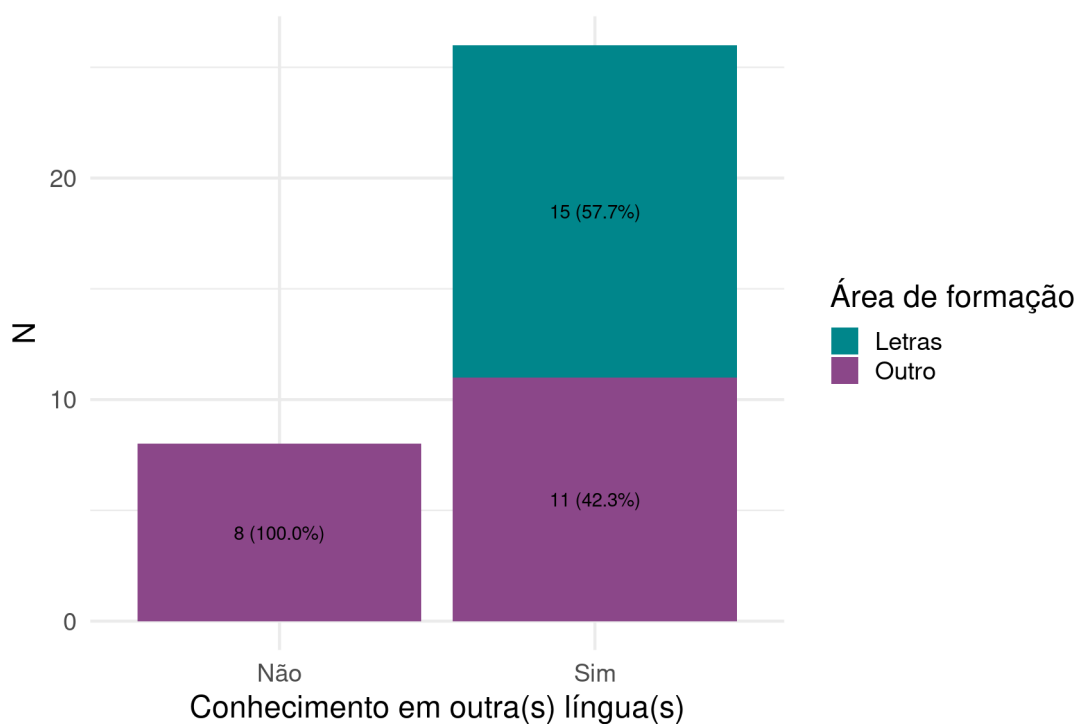


FALSE	-----
FALSE Apparent prevalence	0.01 (0.00, 0.01)
FALSE True prevalence	0.04 (0.03, 0.04)
FALSE Sensitivity	0.11 (0.08, 0.14)
FALSE Specificity	1.00 (1.00, 1.00)
FALSE Positive predictive value	0.68 (0.55, 0.79)
FALSE Negative predictive value	0.97 (0.96, 0.97)
FALSE Positive likelihood ratio	54.15 (32.11, 91.33)
FALSE Negative likelihood ratio	0.89 (0.86, 0.93)
FALSE	-----

## B Gráficos



(#fig:idade\_genero) Idade dos participantes distribuída pelo Gênero



(#fig:area\_linguas) Conhecimento em outras línguas por área de formação

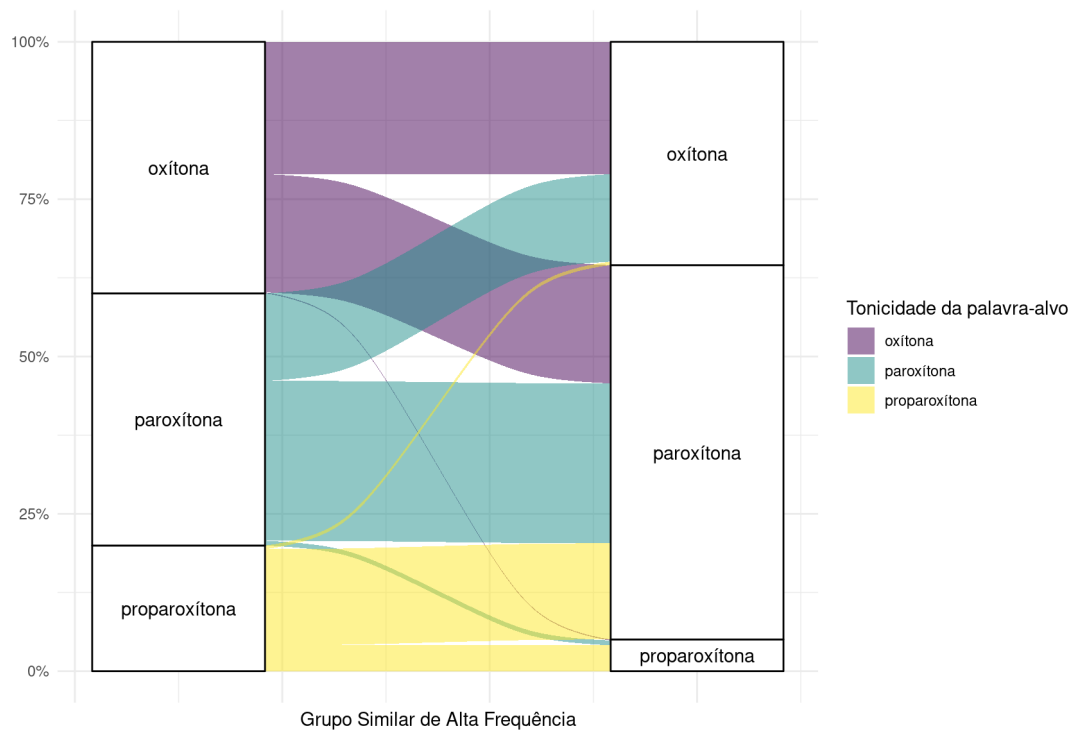


Figura B.1: Fluxo de tonicidade no Grupo Similar de Alta Frequência

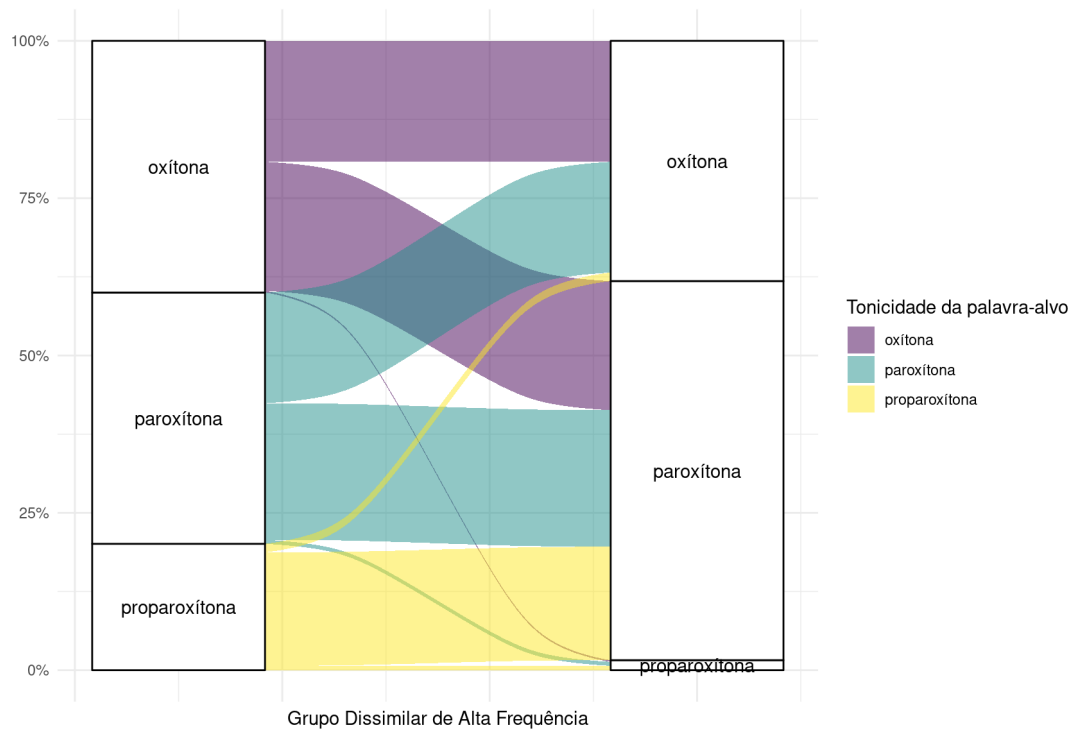


Figura B.2: Fluxo de tonicidade no Grupo Dissimilar de Alta Frequência

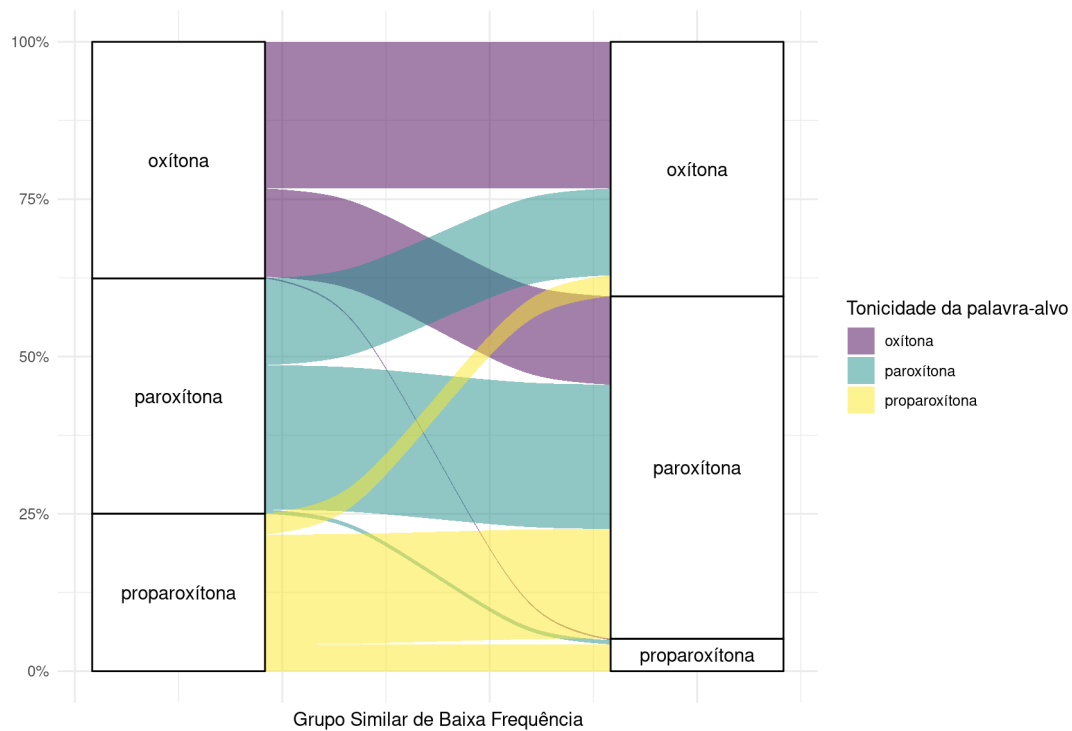


Figura B.3: Fluxo de tonicidade no Grupo Similar de Baixa Frequência

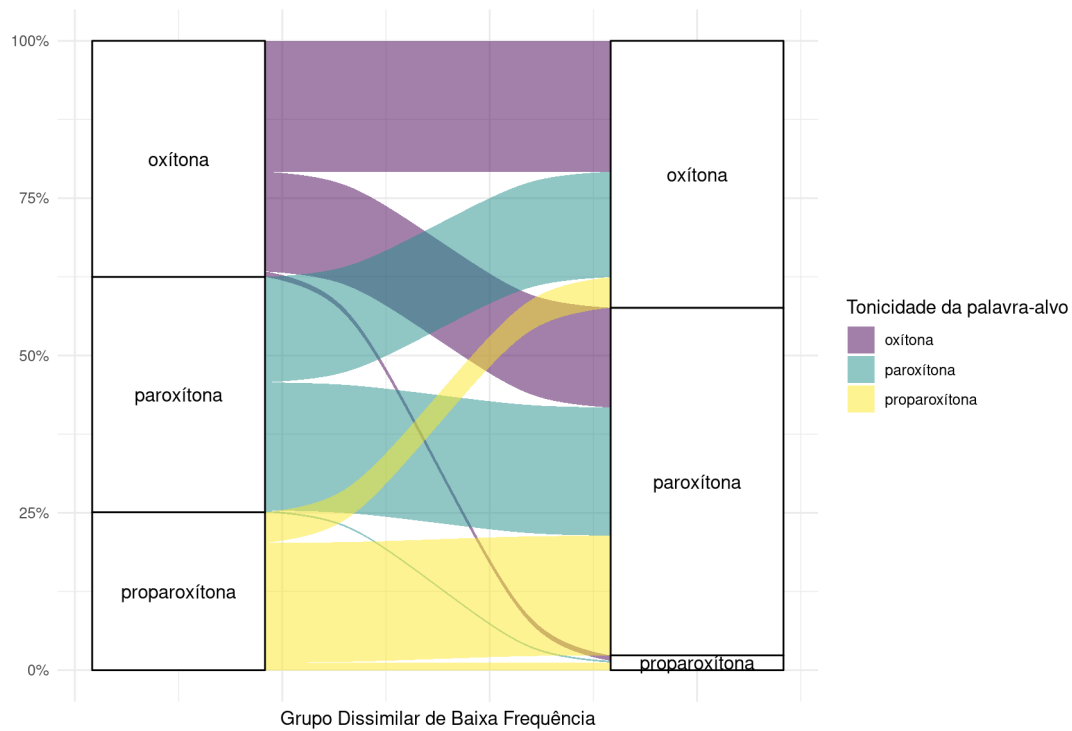


Figura B.4: Fluxo de tonicidade no Grupo Dissimilar de Baixa Frequência