Universidade de São Paulo
Instituto de Matemática e Estatística

# Centro de Estatística Aplicada

# Relatório de Análise Estatística

RAE-CEA–21P02

# 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, junho 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)

## 7.1 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, l. (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

A literatura na área de linguistíca 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 (CÂMARA JR 1970, LEITE 1974, ANDRADE 1994, BISOL 1992, 1994, MATEUS 1996, MASSINI-CAGLIARI 1999, CAGLIARI 1999, SÂNDALO 1999, LEE 1995, 2004, AMARAL 2002,inter alios)

O estudo analisado neste relatório estatístico busca entender como se dá o processo da acetuação tônica (Oxítona, Paroxítona e Proparoxíton) em falantes nativos do português brasileiro, através de um experimento utilizando pseudopalavras e se, de fato, o acento paroxítona é o padrão d alíngua. 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,

# 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 linguistícos 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.

As características estruturais da pseudopalavra têm influência no processo de acentuação tônico? 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 pseudpalavras 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 similiar -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 pseudolavras.

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 *Paráná* (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 (oxítona, paroxítona e proparoxítona).

### 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
- **Vizinhança tonicidade**: indica qual foi o padrão acentual das palavras que os participantes julgaram similares a pseudopalavras criada; apenas para pseudopalavras não validadas
- **Taxa de validação**: indica quantas pessoas do estudo preliminar informaram que a palavra era similar ou dissimilar
- **Palavra alvo**: palavra real que deu origem à pseudopalavra

- **Tonicidade da palavra alvo**: oxítona, paroxítona e proparoxítona
- **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)
- **Sílaba modificada**: 1,2,3 ( 1º,2º,3º) ou 0 quando mais de uma sílaba foi modificada (caso das palavras dissimilares)
- **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.

- **Taxa de similaridade**: Para mudanças de consoante:1, 2, 3 (grupos similares), 5, 6, 7, 8, 9, 10 (grupos dissimilares); para mudanças de vogal :1 (grupos similares), 5, 6, 7, 8 (grupos dissimilares)

### 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)
- **Ordem de apresentação**: indica em qual ordem a pseudopalavra foi apresentada dentro do bloco de apresentação (1 a 93). Para os indivíduos que fizeram toda a dinâmica no software, a ordem das pseudopalavras era diferente dentro de cada bloco.

**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 da palavra-alvo** | | | |
|---|---|---|---|---|
| **Tonicidade produção** | **oxítona** | **paroxítona** | **proparoxítona** | **Total** |
| 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.

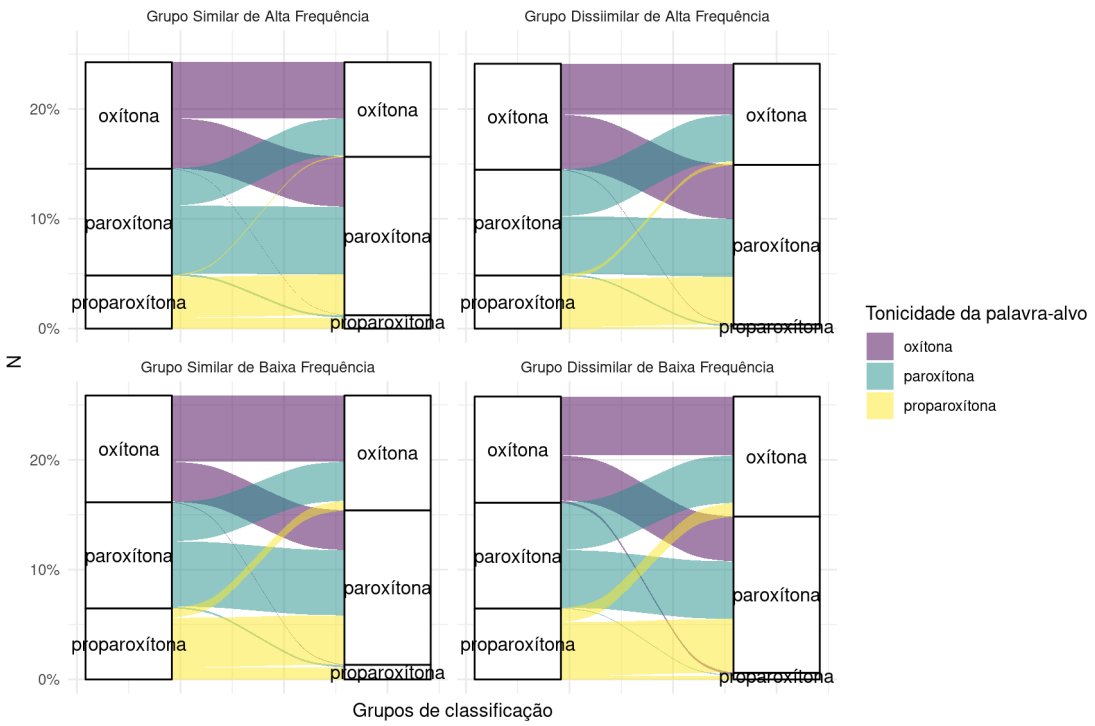



Figura 6.1: Tonicidade de palavra-alvo e pseudopalavra

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 dificil 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 concorda 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.

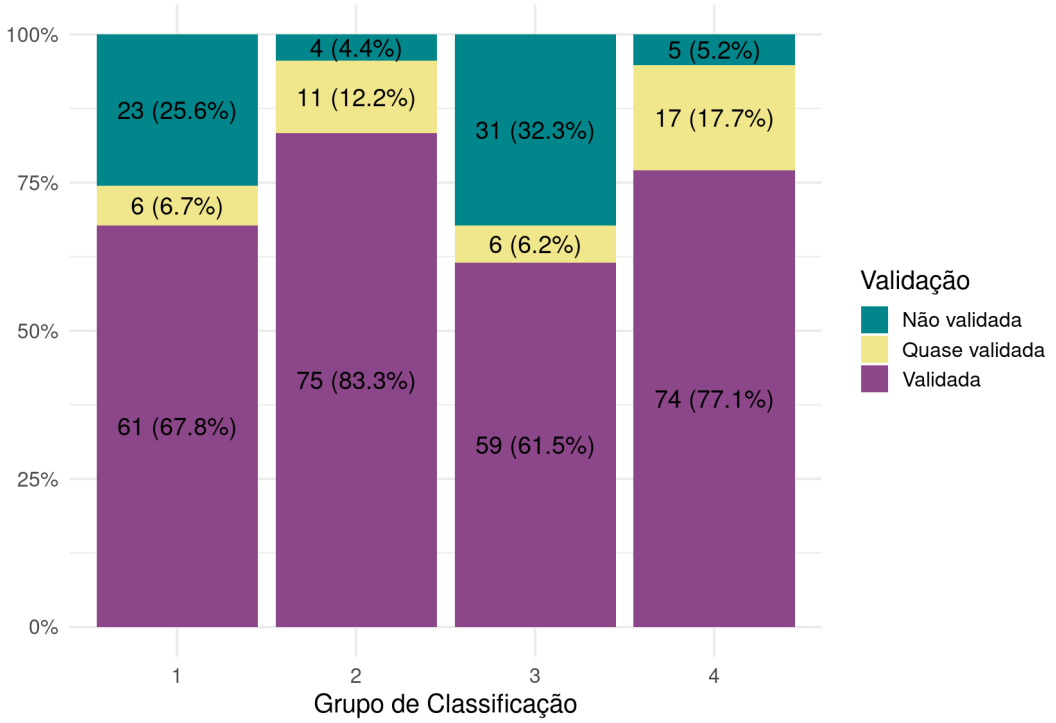
| | **Estrutura da palavra** | | |
|---|---|---|---|
| **Tonicidade produção** | **CV-CV-CV** | **CV-CV-CVC** | **Total** |
| oxítona | 748 (10.3%) | 4156 (79%) | 4904 (39%) |
| paroxítona | 6200 (85.3%) | 963 (18%) | 7163 (57%) |
| proparoxítona | 323 (4.4%) | 121 (2%) | 444 (4%) |
| Total | 7271 (100.0%)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

| | **Validação** | | |
|---|---|---|---|
| **Tonicidade produção** | **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 pseudolavras 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 pseupalavras, 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.

**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

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

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

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

Temos que

$p_{e\gamma}$ = probabilidade de concordância ao acaso $p_a$ = probabilidade de concordância geral $\pi_q$ = probabilidade de classificação na categoria q

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, DG (1991) define os seguintes pontos de corte para a interpretação da estatística Kappa (e, consequentemente, 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.

## 7 Análise univariada

# 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.

| | Área de formação | | |
|---|---|---|---|
| Escolaridade | 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 da palavra-alvo | | | |
|---|---|---|---|---|
| Tonicidade produção | oxítona | paroxítona | proparoxítona | Total |
| 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).

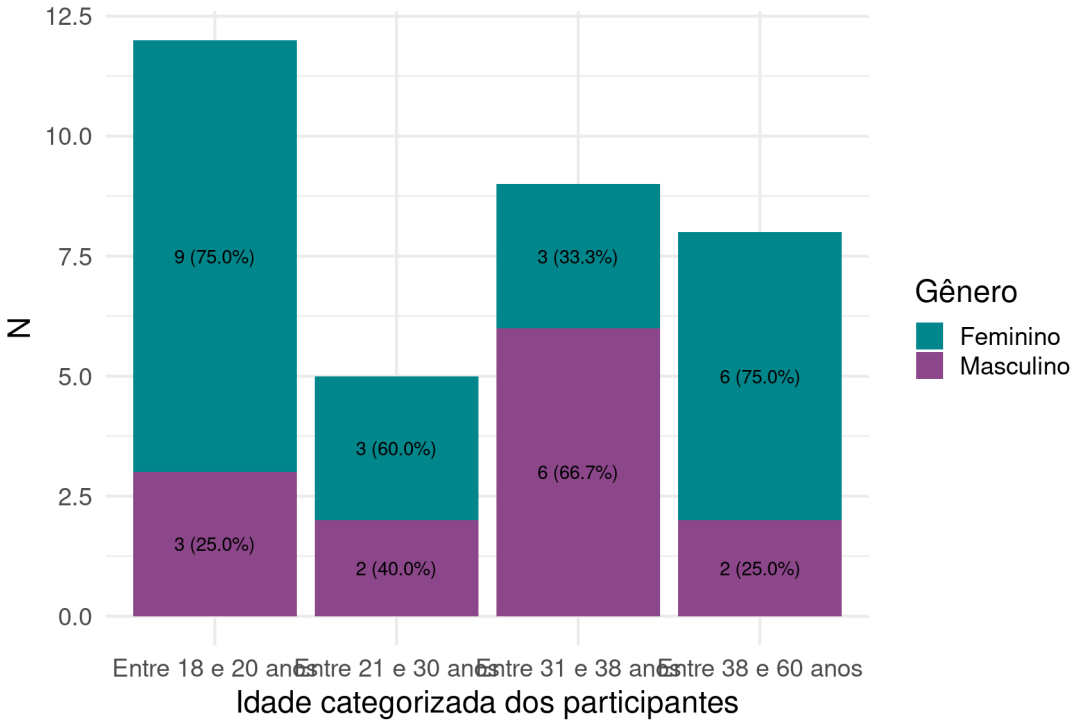
| | Grupo de classificação | | | | |
|---|---|---|---|---|---|
| **Tonicidade produção** | **1 (Similar de alta freq.)** | **2 (Dissimilar de alta freq.)** | **3 (Similar de baixa freq.)** | **4 (Dissimilar de baixa freq.)** | **Total** |
| oxítona | 1077 (22.0%) | 1152 (23.5%) | 1308 (26.7%) | 1367 (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.

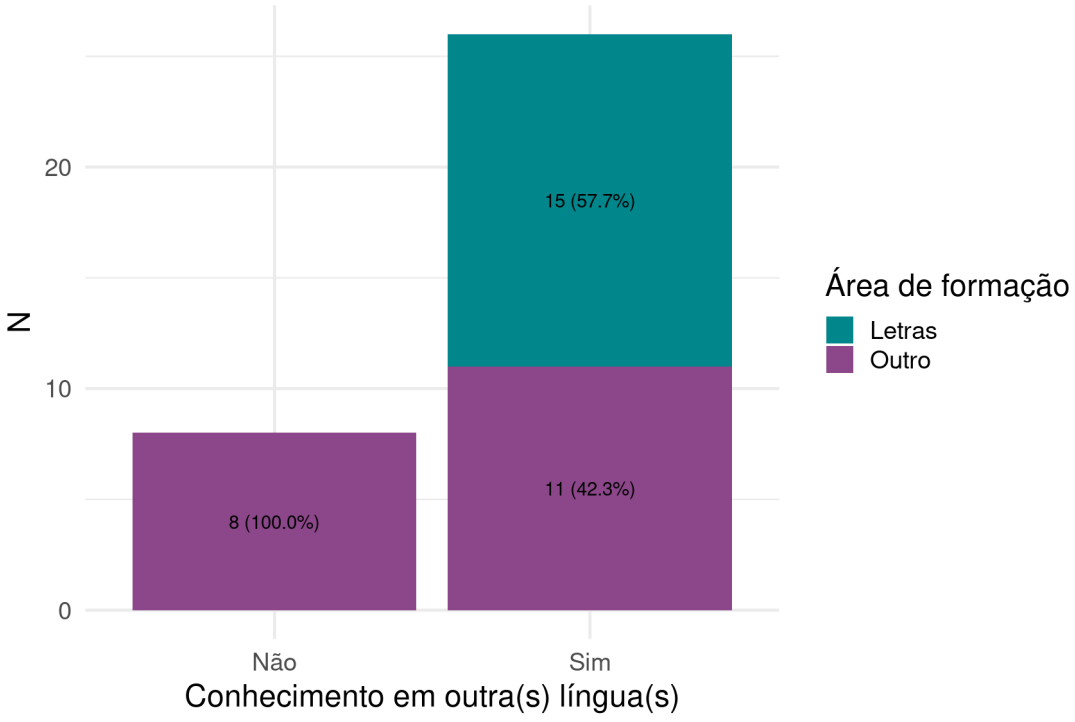
| | | Status Validação | | |
|---|---|---|---|---|
| **Grupo de classificação** | **Tonicidade produção** | **Não Validada** | **Quase Validada** | **Validada** |
| 1 | oxítona | 82 (3.9%) | 44 (3.3%) | 951 (10.5%) |
| 1 | paroxítona | 673 (31.8%) | 141 (10.5%) | 991 (11.0%) |
| 1 | proparoxítona | 17 (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 |

## B Gráficos




(#fig:idade_genero)Idade dos participantes distribuída pelo Genêro




(#fig:area_linguas)Conhecimento em outras línguas por área de formação

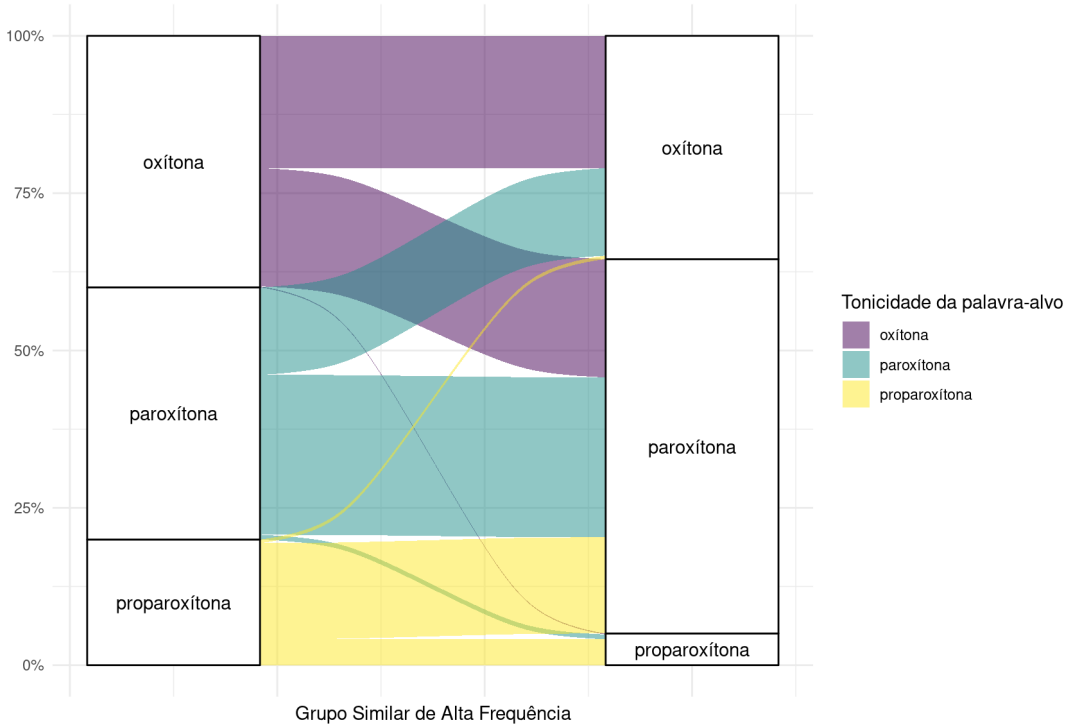



Figura B.1: Fluxo de tonicidade no Grupo Similar de Alta Frequência

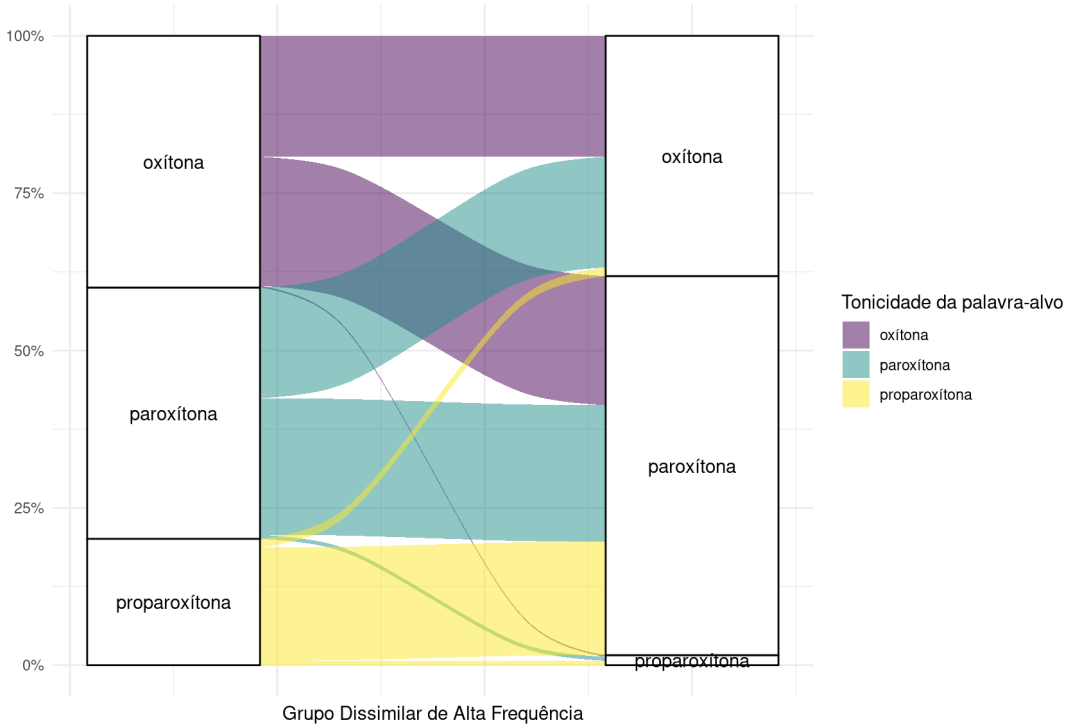



Figura B.2: Fluxo de tonicidade no Grupo Dissimilar de Alta Frequência

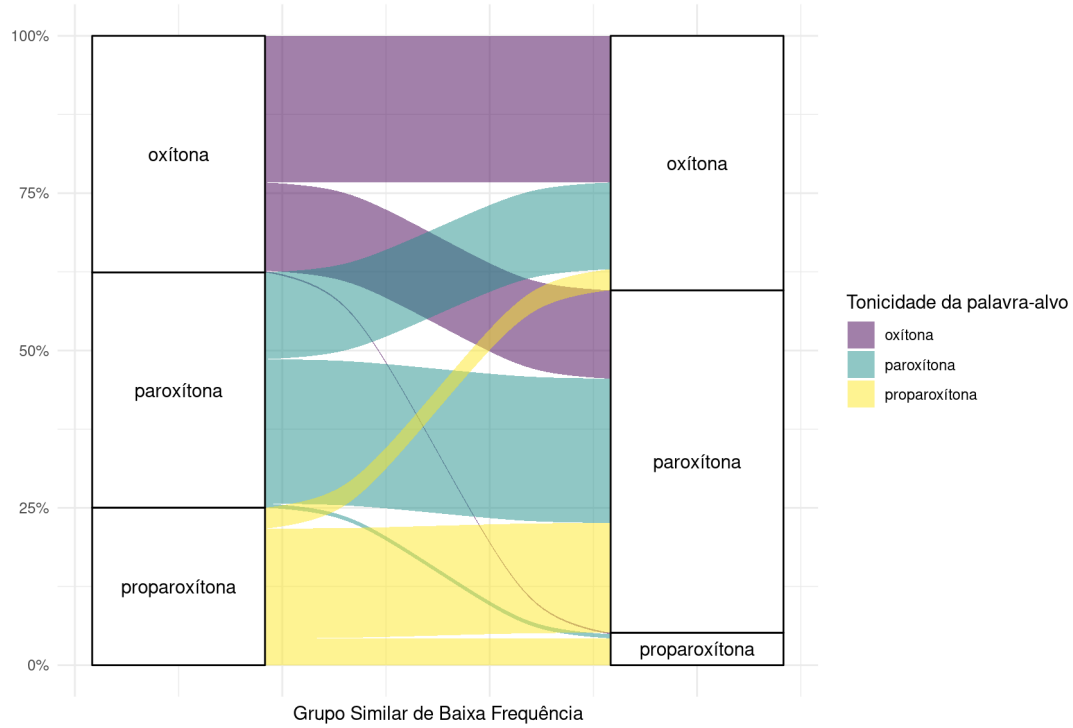



Figura B.3: Fluxo de tonicidade no Grupo Similar de Baixa Frequência

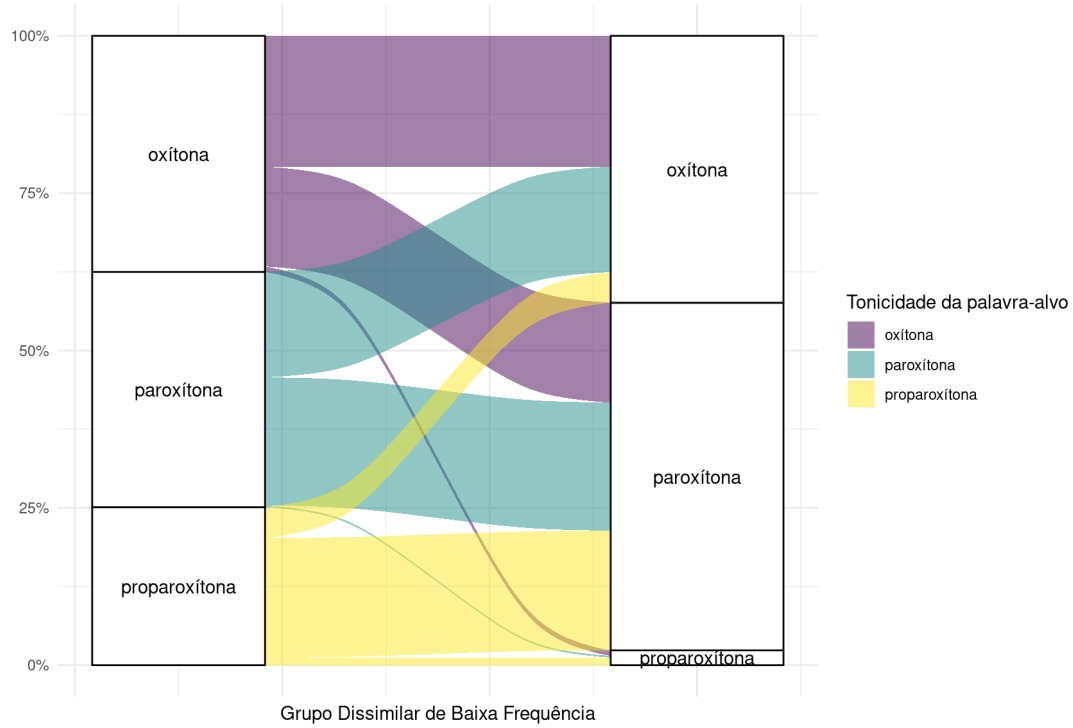



Figura B.4: Fluxo de tonicidade no Grupo Dissimilar de Baixa Frequência