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, maio 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)

ÁREA DE APLICAÇÃO

Linguística (14:110)

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

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

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

PRIBERAN. 2008-2021. *"pseudopalavra", in Dicionário Priberam Da língua Portuguesa*. <https://dicionario.priberam.org/pseudopalavra>. [Acesso em: 01-05-2021].

Sumário

1 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 PRIBERAN (2008-2021), 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.

2 Objetivo(s)

O objetivo principal do trabalho é verificar se pseudopalavras criadas a partir de palavras reais, aqui denominadas palavra-alvo, podem sofrer um processo análogo e ter o mesmo padrão acentual da referência real. Além disso, buscamos entender quais são os conhecimentos linguísticos do falante utilizados nesse processo de acentuação tônica das pseudopalavras e determinar um modelo que explique a classificação das pseudopalavras.

Algumas perguntas a serem respondidas pela análise estatística são:

1. A classificação tônica das pseudopalavras pode recuperar o acento das palavras-alvo semelhantes a elas? Em outras palavras, entender se a tonicidade da palavra-alvo tem papel na predição do acento da pseudopalavra.

2. As características fonológicas e estruturais das palavras-alvo têm influência na acentuação tônica das pseudopalavras?

3. A taxa de similaridade influencia na associação acentual? 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.

3 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.

3.1 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 4.4.

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.

3.2 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 palavras similares e valores maiores ou iguais a 5 determinam palavras dissimilares, e para vogais, valores acima de 1 determinam palavras dissimilares.

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 real a qual eles associavam a palavra criada. 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.

Da mesma forma, 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. Ressalta-se que, nesse caso, a associação não precisava ser exclusivamente com a palavra-alvo.

4 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.

4.1 Variável Dependente

A variável resposta de interesse é **Tonicidade da pseudopalavra**, ou seja, a classificação acentual tônica da pseudopalavra (oxítona, paroxítona e proparoxítona).

4.2 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
- **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 alvo: 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 um dos estímulos criados
- **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.

- **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)

4.3 Variáveis Extralinguísticas

- **Participante**: identifica os 34 participantes do experimento
- **Idade**: de 18 a 60 (anos)
- **Gênero**: feminino e masculino
- **Naturalidade**: indica a cidade em que o participante nasceu
- **Escolaridade**: ensino fundamental a mestrado
- **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

4.4 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).

5 Análise descritiva

5.1 Perfil dos participantes

5.1.1 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.

5.1.2 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.

5.1.3 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), 15 (55,9%) participantes são de outras áreas e 19 (44,1%) são da área de Letras; no entanto, cruzando 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.

5.2 Variáveis linguísticas

A tonicidade de produção (classificação da pseudopalavra) do tipo proparoxítona é a menos expressiva na base,enquanto a categoria paroxítona aparece com maior frequência nas respostas dos participantes, como visto na Tabela 5.1 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 5.1: Distribuição das respostas nos níveis da variável Tonicidade de produção.

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

Observa-se na Tabela A.6 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.

Comparando a tonicidade das pseudopalavras e a tonicidade das palavras-alvo (vide Tabela 5.2), 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.7). Esse fluxo pode ser analisado graficamente na Figura B.1.

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 5.2: Frequência da tonicidade das pseudopalavras 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%)

Como observação subsequente, podemos ver que a divisão de estrutura da palavra-alvo não é bem distribuída entre as classificações de tonicidades das palavras-alvo. A quantidade de palavras com estrutura CV-CV-CVC em proparoxítonas é pequena uma vez que a frequência de palavras com essas características no *Corpus Brasileiro* é baixa (vide Tabela A.8).

Tabela 5.3: Frequência da Tonicidade da pseudopalavra 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%)

Na Tabela 5.3, 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 influência da estrutura na atribuição da tonicidade em palavras do português brasileiro. A partir disso, vemos a possibilidade de testar a significância dessa discrepância através de testes univariados entre a tonicidade da produção e estrutura da palavra.

Tabela 5.4: Tonicidade de pseudopalavras por Grupo, Estrutura e Tonicidade da palavra-alvo

			Tonicidade Alvo		
Grupo de Classificação	**Estrutura da Palavra**	**Tonicidade Produção**	**Oxítona**	**Paroxítona**	**Proparoxítona**
1	CV-CV-CV	oxítona	80 (1.7%)	30 (0.6%)	17 (0.6%)
1	CV-CV-CV	paroxítona	520 (10.7%)	562 (11.6%)	464 (16.4%)
1	CV-CV-CV	proparoxítona	4 (0.1%)	17 (0.4%)	125 (4.4%)
1	CV-CV-CVC	oxítona	558 (11.5%)	392 (8.1%)	NA (-)
1	CV-CV-CVC	paroxítona	50 (1.0%)	209 (4.3%)	NA (-)
1	CV-CV-CVC	proparoxítona	1 (0.0%)	6 (0.1%)	NA (-)
2	CV-CV-CV	oxítona	41 (0.8%)	44 (0.9%)	41 (1.4%)
2	CV-CV-CV	paroxítona	553 (11.4%)	546 (11.3%)	544 (19.2%)
2	CV-CV-CV	proparoxítona	7 (0.1%)	13 (0.3%)	21 (0.7%)
2	CV-CV-CVC	oxítona	539 (11.1%)	487 (10.1%)	NA (-)
2	CV-CV-CVC	paroxítona	66 (1.4%)	109 (2.3%)	NA (-)
2	CV-CV-CVC	proparoxítona	1 (0.0%)	6 (0.1%)	NA (-)
3	CV-CV-CV	oxítona	191 (3.9%)	86 (1.8%)	33 (1.2%)
3	CV-CV-CV	paroxítona	411 (8.5%)	510 (10.5%)	510 (18.0%)
3	CV-CV-CV	proparoxítona	7 (0.1%)	12 (0.2%)	65 (2.3%)
3	CV-CV-CVC	oxítona	563 (11.6%)	360 (7.4%)	75 (2.6%)
3	CV-CV-CVC	paroxítona	43 (0.9%)	232 (4.8%)	54 (1.9%)
3	CV-CV-CVC	proparoxítona	1 (0.0%)	9 (0.2%)	73 (2.6%)
4	CV-CV-CV	oxítona	128 (2.6%)	39 (0.8%)	18 (0.6%)
4	CV-CV-CV	paroxítona	457 (9.4%)	562 (11.6%)	561 (19.8%)
4	CV-CV-CV	proparoxítona	18 (0.4%)	6 (0.1%)	28 (1.0%)
4	CV-CV-CVC	oxítona	542 (11.2%)	501 (10.4%)	139 (4.9%)
4	CV-CV-CVC	paroxítona	54 (1.1%)	94 (1.9%)	52 (1.8%)
4	CV-CV-CVC	proparoxítona	10 (0.2%)	3 (0.1%)	11 (0.4%)
Total	•	•	4845 (100.0%)	4835 (100.0%)	2831 (100.0%)

Analisando a divisão por grupo de classificação, na Tabela **??**, é possível notar que há um lapso faltante de dados pois não existe, no português brasileiro, palavras com tonicidade proparoxítona e de alta frequência no Corpus com estrutura CV-CV-CVC. Isso pode ser um problema a ser contornado na análise, uma vez que existem combinações de categorias das variáveis grupo e estrutura por nível da tonicidade da palavra-alvo sem informação.

Por fim, explorando os dados no âmbito de validação, vemos, na Tabela 5.5, que aproximadamente 73% das pseudopalavras foram validadas de acordo com o processo descrito anteriormente no capítulo 3 (Descrição).

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 (5.6. 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.

Tabela 5.5: Frequência da tonicidade das pseudopalavras, por nível da variável Validação.

	Status 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%)

Tabela 5.6: Frequência das pseudopalavras nos Grupos de classificação, por status de Validação.

	Status Validação		
Grupo de classificação	**Não Validada**	**Quase Validada**	**Validada**
1	772 (36.4%)	202 (15.0%)	2061 (22.8%)
2	135 (6.4%)	372 (27.6%)	2511 (27.8%)
3	1044 (49.3%)	203 (15.1%)	1988 (22.0%)
4	167 (7.9%)	571 (42.4%)	2485 (27.5%)
Total	2118 (100.0%)	1348 (100.0%)	9045 (100.0%)

5.3 Conclusão da análise descritiva e próximos passos (ideia 1)

A partir dos resultados presentes nas tabelas da seção anterior, percebe-se um comportamento não homogêneo entre níveis de algumas variáveis.Portanto, o próximo passo do estudo consite em realizar uma análise univariada, com testes qui-quadrado, que permite analisar a relação de independência entre variáveis qualitativas. Algumas hipóteses iniciais são:

1. Entender se os grupos de classificação afetam a acentuação das pseudopalavras.

H_0: O grupo de classificação não determina a atribuição acentual da pseudopalavra.
H_1: O grupo de classificação determina a atribuição acentual da pseudopalavra.

2. Observar se a tonicidade de produção é coincidente com a tonicidade alvo.

H_0: Não há associação entre os padrões acentuais da palavra-alvo e da pseudopalavra. H_1: Há associação entre os padrões acentuais da palavra-alvo e da pseudopalavra.

3. Entender se quanto mais similar a pseudopalavra for da palavra-alvo, maiores são as chances de atribuição do mesmo padrão acentual da palavra-alvo.

H_0: A taxa de similaridade não influencia na atribuição acentual da pseudopalavra.
H_1: A taxa de similaridade influencia na atribuição acentual da pseudopalavra.

4. A validação da similaridade e dissimilaridade em relação à palavra alvo afeta a atribuição acentual das pseudopalavras?

H_0: A atribuição acentual da pseudopalavra independe da validação da similaridade/dissimilaridade da mesma com relação à palavra-alvo. H_1: A atribuição acentual da pseudopalavra depende da validação da similaridade/dissimilaridade da mesma com relação à palavra-alvo.

5. A estrutura da palavra-alvo têm influência na acentuação tônica das pseudopalavras?

H_0: Não há associação entre a estrutura da palavra-alvo e a atribuição acentual da pseudopalavra. H_1: Há associação entre a estrutura da palavra-alvo e a atribuição acentual da pseudopalavra.

6. Entender se o conhecimento em Música influencia na acentuação tônica.

H_0: Não há associação entre conhecimento em Música e a atribuição acentual. H_1: Há associação entre conhecimento em Música e a atribuição acentual.

7. Entender se a estudantes no início da graduação em Letras tem um comportamento diferente no processo de acentuação tônica

H_0: Não há associação entre graduação em letras e a atribuição acentual da pseudopalavra. H_1: Há associação entre graduação em letras e a atribuição acentual da pseudopalavra.

Após os testes, decidimos investigar mais minunciosamente a relação de dependência entre as variáveis cujo teste qui-quadrado foi significante. Essa etapa foi realizada através de análises de correspondência, que visa medir o grau de associação de variáveis categorizadas comparadas. O objetivo é construir uma representação gráfica da associação entre as variáveis.

5.4 Conclusão da análise descritiva e próximos passos (ideia 2)

Dado que temos dados 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, surgiu a necessidade de aplicar testes de concordância do tipo Kappa para entender a relevância de variáveis inicialmente consideradas importantes para o modelo.

Como os níveis da variável resposta (tonicidade da pseudopalavra) tem 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, comumente utilizada em amostras desbalanceadas.

Observando o resultado geral (vide Tabela **??**), é possível notar que aparentemente há concordância na variável resposta entre os falantes. Já na Tabela **??**, vemos que há semelhança entre os grupos 1 e 3 (similares) e 2 e 4 (dissimilares), sendo nas dissimilares os maiores níveis de concordância.

Estatística	Valor da Estatística
Gwet's AC1	0.6775543

6 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.

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 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.7: Frequência da tonicidade das palavras-alvo por nível de tonicidade das pseudopalavras.

	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.8: Frequência da Tonicidade alvo por Estrutura das palavras.

	Estrutura da palavra		
Tonicidade palavra-alvo	**CV-CV-CV**	**CV-CV-CVC**	**Total**
oxítona	8 (40%)	12 (60%)	20 (100%)
paroxítona	12 (67%)	6 (33%)	18 (100%)
proparoxítona	12 (92%)	1 (8%)	13 (100%)

Tabela A.9: Tonicidade de pseudopalavras por Grupo 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%)

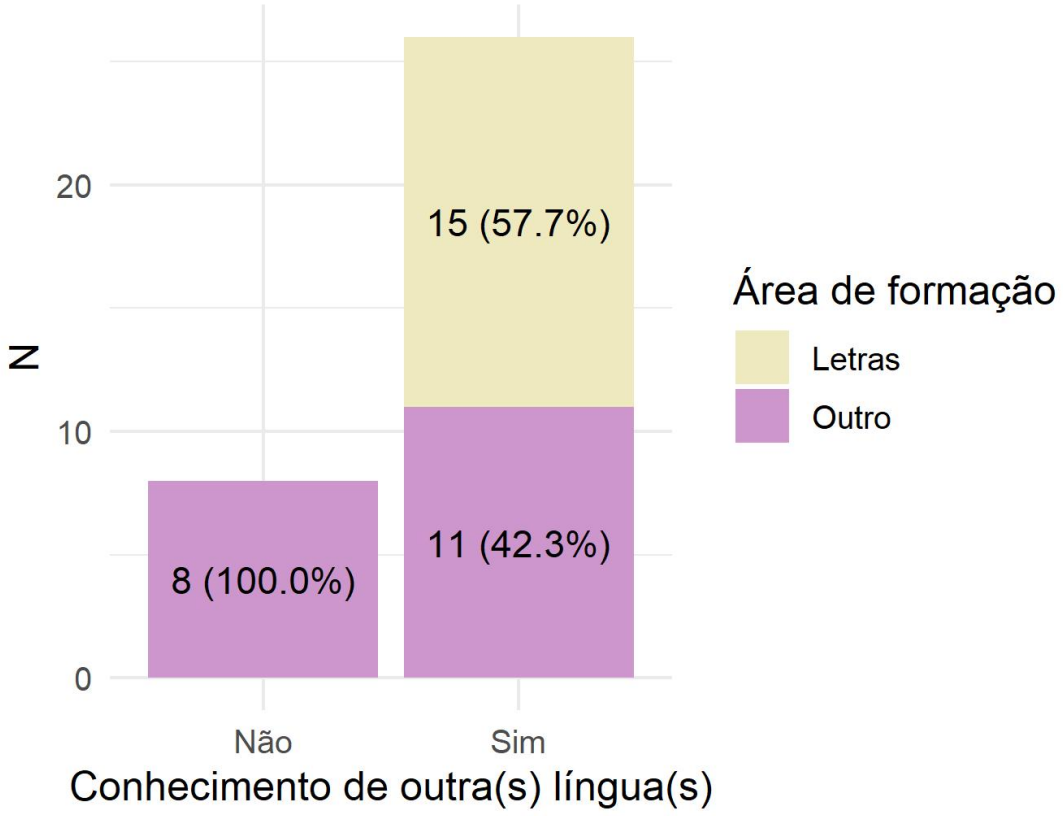
Grupo de Classificação	**Valor da Estatística**
Grupo 1	0.6641838
Grupo 2	0.7549719
Grupo 3	0.5874817
Grupo 4	0.7063545

Estrutura da Palavra	**Valor da Estatística**

CV-CV-CVC	0.6582021
CV-CV-CV	0.7626996

Segmento Modificado	**Valor da Estatística**
vogal	0.6725362
consoante	0.6826297

B Gráficos



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

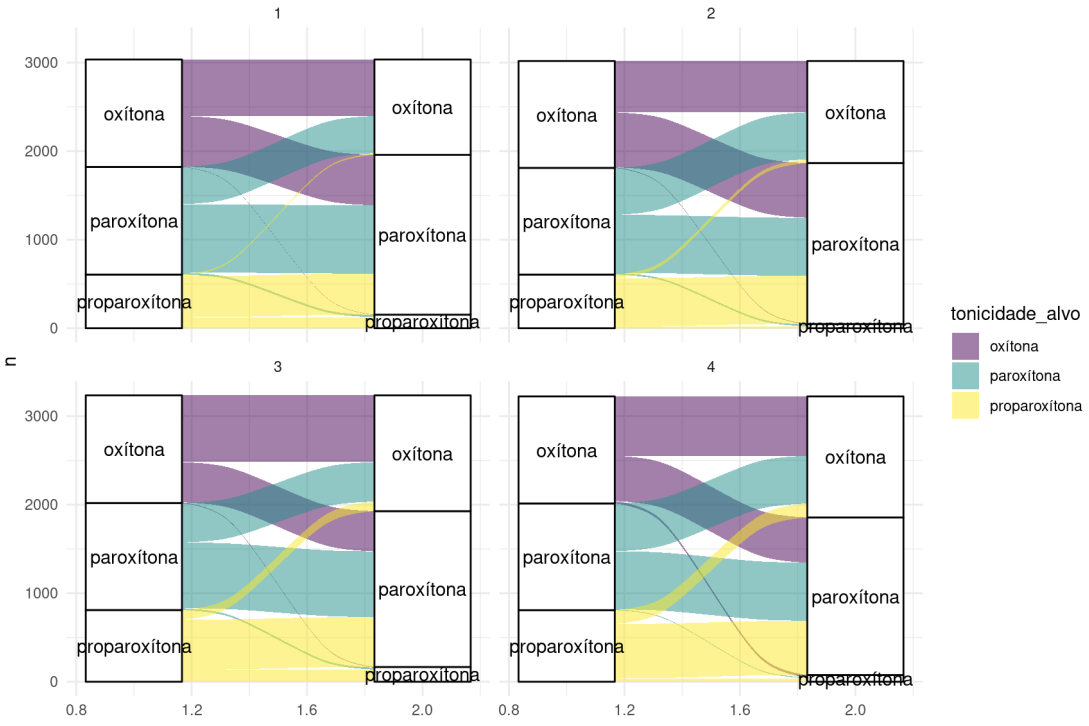


Figura B.1: Tonicidade de palavra-alvo e pseudopalavra

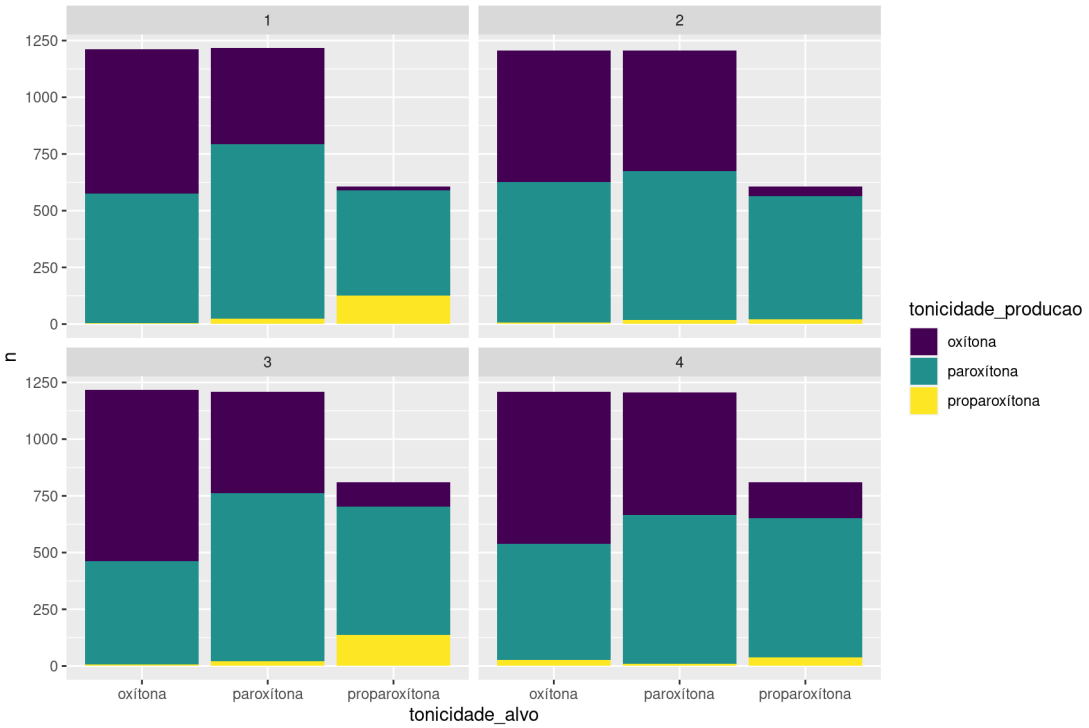


Tabela B.1: Distribuição das respostas nos níveis da variável Tonicidade de produção.

Grupo	Tonicidade de produção	Oxítona	Paroxítona	Proparoxítona
1 (Similar de alta freq.)	oxítona	638	422	17
1 (Similar de alta freq.)	paroxítona	570	771	464
1 (Similar de alta freq.)	proparoxítona	5	23	125
2 (Dissimilar de alta freq.)	oxítona	580	531	41

2 (Dissimilar de alta freq.)	paroxítona	619	655	544
2 (Dissimilar de alta freq.)	proparoxítona	8	19	21
3 (Similar de baixa freq.)	oxítona	754	446	108
3 (Similar de baixa freq.)	paroxítona	454	742	564
3 (Similar de baixa freq.)	proparoxítona	8	21	138
4 (Dissimilar de baixa freq.)	oxítona	670	540	157
4 (Dissimilar de baixa freq.)	paroxítona	511	656	613
4 (Dissimilar de baixa freq.)	proparoxítona	28	9	39