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

Sumário

List of Tables

List of Figures

1 Introdução

De acordo com J. L. FIORIN (2007), o interesse pela linguagem é antigo e expresso através de mitos, lendas, cantos, rituais ou trabalhos eruditos que buscam conhecer essa capacidade humana. A autora aponta que, a partir do século XX, os estudos linguistícos passaram a ter um caráter cientifico, 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 dizer o que a língua é e por que é assim. Logo, a área estuda os aspectos fonéticos, morfológicos, sintáticos, semânticos, sociais e psicológicos no português brasileiro. Dentro deste contexto, existe o conceito de *pseudopalavra*, que, de acordo com o dicionário Priberan ("MS Windows NT "pseudopalavra", in Dicionário Priberam Da língua Portuguesa" (n.d.)), 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 a intensidade dada a determinadas sílabas na pronúncia das palavras. Aquela que é pronunciada de forma mais acentuada é a sílaba tônica. 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 relacionadas a vocábulos existentes no idioma. Em outras palavras, busca-se compreender como o indivíduo, ao se deparar com uma palavra nova e, nesse caso, inventada, define a sílaba tônica. Além disso, busca-se entender quais são os outros fatores, como os conhecimentos linguísticos

do falante e estruturas linguísticas das palavras que podem influenciar nesse processo de classificação da entonação.

2 Objetivo

O objetivo principal do trabalho é verificar se pseudopalavras que se assemelham a palavras reais podem sofrer um processo analógico e ter o mesmo padrão acentual da referência real, além de entender quais são os conhecimentos linguísticos do falante utilizados nesse processo de acentuação tônica das pseudopalavras.

Algumas perguntas a serem respondidas pela análise estatística são:

1. Quanto mais similar a palavra inventada for da palavra real, maiores são as chances de atribuição do mesmo padrão acentual da palavra real?
2. A tonicidade da palavra alvo, ou seja, a referência real em relação à pseudopalavra, tem papel na classificação da pseudopalavra?
3. As variáveis selecionadas pelo modelo estão em concordância com a literatura da área?
4. Há algum falante ou pseudopalavra com comportamento destoante dos demais? Qual o seu papel na atribuição acentual?

3 Descrição do estudo

O experimento consistia em apresentar aos participantes, através do software Psychopy, uma série de pseudopalavras e registrar, via Google Meet, a forma como eles reproduziam verbalmente tais palavras. A seguir, as respostas dos participantes foram classificadas de acordo com as três classes de acentuação tônica: oxítona, paroxítona e proparoxítona. Essa é a variável dependente do estudo, que será descrita junto às outras variáveis no capítulo **??**.

A coleta dos dados foi realizada no início do primeiro semestre de 2020 com 34 indivíduos que, através de divulgações em redes sociais e de colegas, se voluntariaram a participar do experimento. Os pré-requisitos eram de que tais voluntários não tivessem estudado linguística, fossem maior de 18 anos e falantes nativos do português brasileiro.

Na amostra 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. O perfil dos participantes é descrito nas tabelas e gráficos na seção 5.

3.1 Problemas na coleta dos dados

Identificamos duas potenciais falhas na coleta dos dados:

1. Problemas técnicos e interferência externa;
2. Problemas de aleatorização.

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 para todos os participantes. Portanto, um total de 12.511 dados serão utilizados na análise.

Problemas de aleatorização:

Inicialmente, as pseudopalavras foram aleatorizadas de forma geral no software *Excel* (ALEATORIo ENTRE) e separadas em 4 blocos com 93 pseudopalavras cada, resultando em um total de 372 palavras criadas. Dentro de cada bloco, a ordem em que as pseudopalavras foram apresentadas aos falantes também foi aleatorizada, porém, 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).

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 alta frequência e se possui menos de 2 mil ocorrências ela é classificada como baixa frequência. Ademais, as pseudopalavras foram construídas com três sílabas de extensão para que os três padrões acentuais do português brasileiro pudessem ser produzidos. 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 estímulos (alta frequência + similaridade, alta frequência + dissimilaridade, baixa frequência + similaridade, baixa frequência + dissimilaridade). Ressaltamos que essa variável não foi controlada durante a coleta de dados, ou seja, não foi definido uma quantidade de palavras de cada categoria em cada bloco

Outro ponto importante está relacionado aos testes para validar se a pseudopalavra é similar a palavra a partir da qual ela foi criada (palavra alvo). Nessa etapa, pediu-se para 10 falantes do português, que não participaram do estudo, listarem a palavra real a qual eles associavam a palavra inventada. Considerou-se como validadas as pseudopalavras cuja associação foi a palavra alvo na resposta de, no mínimo, oito indivíduos. Pseudopalavras com um número de associações corretas menor que oito foram consideradas não validadas, porém, pseudopalavras nas quais sete falantes apresentaram a associação correta foram classificadas como quase validadas.

Da mesma forma, foram consideradas validadas como dissimilares à palavra alvo pseudopalavras não associadas a uma mesma resposta por mais de dois indivíduos. Ressalta-se que a associação não necessariamente precisa ser com a palavra alvo. Pseudopalavras com um número de associações maior que dois foram consideradas não validadas,porém,pseudopalavras nas quais três falantes apresentaram a mesma associação foram classificadas como quase validadas

Por fim, destacamos que a variável relacionada aos falantes será considerada aleatória, pois são uma amostra da população dos falantes do português brasileiro e,

além disso, porque cada participante apresentou mais de uma resposta no estudo

4 Descrição das variáveis

Foram coletadas variáveis linguísticas (relacionadas as 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*, ou seja, aclassificação acentual tônica da pseudopalavra (oxítona, paroxítona e proparoxítona)

4.2 Variáveis Linguísticas

- **Grupo dos estímulos**: 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

- **Pseudopalavra**: refere-se a cada um dos estímulos criados
- **Palavra alvo**: palavra real que deu origem à pseudopalavra
- **Tonicidade da palavra alvo**: oxítona, paroxítona e proparoxítona
- **Estrutura da palavra**: CV-CV-CV e CV-CV-CVC
- **Tempo de resposta**: indica o tempo em milisegundos que os falantes levaram para apertar a tecla de espaço entre um estímulo e outro
- **Segmento modificado**: indica qual letra foi modificada na criação da pseudopalavra a partir da palavra real (consoante ou vogal)
- **Taxa de similaridade**: 1, 2, 3 (grupos similares), 5, 6, 7, 8, 9, 10 (grupos dissimilares)
- **Validação**: s = sim, n = não validada e q = quase validada

- **Taxa de validação**: indica quantas pessoas informaram que a palavra era similar ou dissimilar
- **Vizinhança Fonologica**: consiste na categorização por vizinhança fonológica apenas das pseudopalavras que não foram validadas
- **Vizinhança Tonicidade**: indica qual foi o padrão acentual das palavras que os participantes julgaram similares a pseudopalavras criada; apenas para as pseudopalavras que não foram validadas

4.3 Variáveis Extralinguísticas

- **Participante**: indica 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**: indica quais línguas o falante declara que fala ou já estudou. (categorizar ?)
- **Música**: 0 = não tem conhecimento em música e 1 = tem conhecimento em música)

4.4 Variáveis Experimentais

- **Aleatorizacao**: codifica se o bloco de apresentação foi aleatorizado para o indivíduo ou se foi a aleatorização prévia (s = o estímulo foi aleatorizado e n = não houve aleatorização)
- **Bloco de apresentação**: indica em qual bloco (ou rotina) 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)

Destacamos também algumas variável que foram excluídas inicialmente da análise e o motivo dessa escolha:

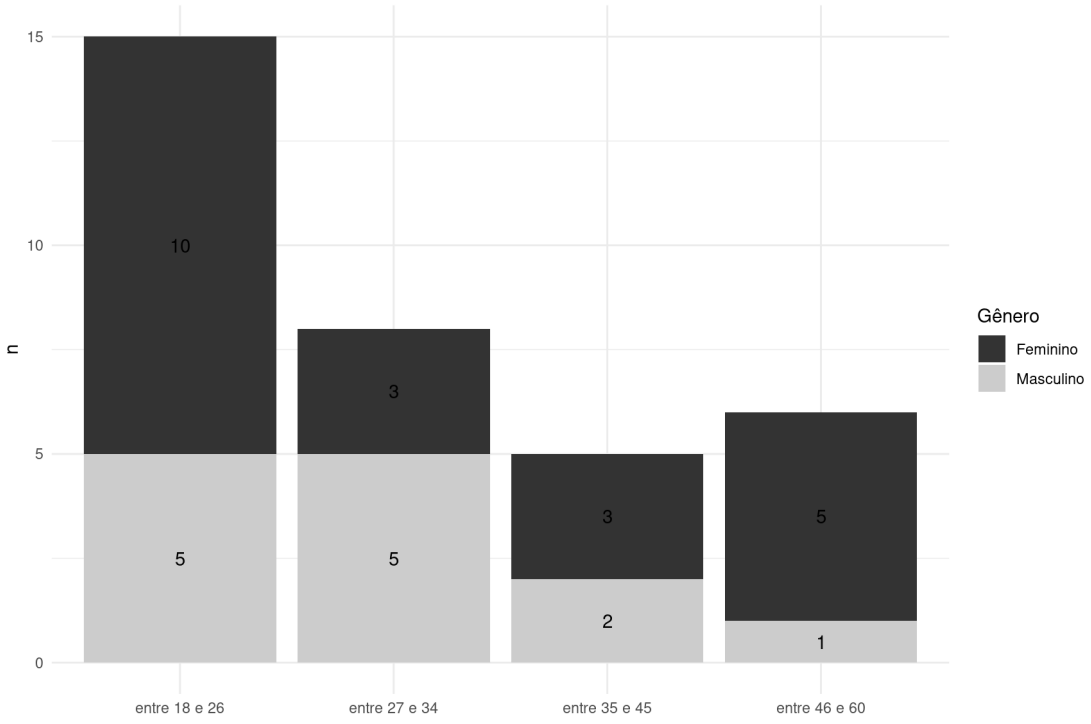
- **Tempo de resposta**: Só há tempo para a metade dos participantes devido ao problema com o software em que o experimento foi realizado. Portanto, essa variável pode ter sido influenciada por esse erros, além de possuir muitos dados faltantes
- **Taxa de similaridade**: 1, 2, 3 (grupos similares), 5, 6, 7, 8, 9, 10 (grupos dissimilares). Como o conceito de similaridade entre a palavra alvo e a pseudopalavra já está inserido na variável *Grupo de estímulos*, entendemos que ela não é necessária na análise

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. O gráfico **??** mostra a distribuição dos respondentes segundo a faixa etária e gênero.



5.1.2 Área de formação

Parte considerável dos participantes (10) são ingressantes do curso de Letras, logo, para a análise inferencial, essa variável foi categorizada em dois níveis. A tabela **??** mostra esse agrupamento da formação dos voluntários. Este é um dos fatores que a pesquisa busca entender se influencia ou não na categorização das pseudopalavras.

Outras variáveis relacionadas com a linguagem (conhecimento em música e de outras línguas) também foram transformadas em variáveis binárias, pois há interesse em saber se isso afeta na percepção da tonicidade de palavras. A distribuição inicial das variáveis por participante também pode ser observada nas tabelas **??** e **??**.

Área de formação	Número de participantes
Outro	19
Letras	15

musica	n
1	21
0	13

area_formacao	n
Outro	19
Letras	15

5.2 Variáveis linguísticas

Verifica-se que a proporção de proparoxítonas é inferior às outras categorias em todos os grupos, o que já era esperado devido à baixa frequência de palavras proparoxítonas de três sílabas na estrutura proposta (CV-CV-CV ou CV-CV-CVC). (PROCURAR REFERÊNCIA)

Também observamos na tabela **??** que os grupos 2 e 4 (ou seja, os grupos em que as palavras sofreram mais alterações em relação à referência original) apresentam uma proporção ainda menor de proparoxítonas quando comparados aos grupos 1 e 3.

	Grupo				
Tonicidade palavra-alvo	**1 (Similar de alta freq.)**	**2 (Dissimilar de alta freq.)**	**3 (Similar de baixa freq.)**	**4 (Dissimilar de baixa freq.)**	**Total**
oxítona	1213 (25.04%)	1207 (24.912%)	1216 (25.098%)	1209 (24.954%)	4845
paroxítona	1216 (25.15%)	1205 (24.922%)	1209 (25.005%)	1205 (24.922%)	4835

proparoxítona	606 (21.41%)	606 (21.406%)	810 (28.612%)	809 (28.576%)	2831

	Grupo				
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
paroxítona	1805 (25.2%)	1818 (25.4%)	1760 (24.6%)	1780 (24.8%)	7163
proparoxítona	153 (34.5%)	48 (10.8%)	167 (37.6%)	76 (17.1%)	444

Comparando a tonicidade das pseudopalavras e a tonicidade das palavras-alvo, nota-se que

	Tonicidade da palavra-alvo			
tonicidade_producao	**oxitona**	**paroxitona**	**proparoxitona**	**total**
oxítona	2642 (54%)	1939 (39.54%)	323 (7%)	4904
paroxítona	2154 (30%)	2824 (39.42%)	2185 (31%)	7163
proparoxítona	49 (11%)	72 (16.22%)	323 (73%)	444

Estrutura x tonicidade palavra-alvo

	Estrutura da palavra		
Tonicidade palavra-alvo	**CV-CV-CV**	**CV-CV-CVC**	**Total**
oxítona	2417 (49.89%)	2428 (50.11%)	4845
paroxítona	2427 (50.20%)	2408 (49.80%)	4835
proparoxítona	2427 (85.73%)	404 (14.27%)	2831

Estrutura x tonicidade pseudopalavra

A tabela 5.1 mostra o perfil dos informantes segundo sua naturalidade.

etc etc

Tabela 5.1: Perfil dos participantes segundo naturalidade

naturalidade	**n**
São Paulo, SP	16
Outros municípios de SP	12
Outras UF	6

6 Análise inferencial

6.1 Modelo proposto

6.2 Seleção de variáveis

6.3 Ajuste

6.4 Diagnóstico

7 Conclusões

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

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

“MS Windows NT "pseudopalavra", in Dicionário Priberam Da língua Portuguesa.” n.d. https://dicionario.priberam.org/pseudopalavra.