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

Sumário

List of Tables

List of Figures

1 1. Introdução

De acordo com Margarida Petter (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 Florin (2019), 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 fonético, morfológico, sintático, semântico, social e psicológico no português brasileiro. Dentro disso, existe o conceito de *pseudopalavra* que, de acordo com o dicionário de Priberan, 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 acento tônico: 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 trabalho investiga a maneira como os falantes nativos do português atribuem a acentuação tônica em pseudopalavras correspondentes à palavras existentes na língua.Em outras palavras, queremos compreender como o indivíduo define a sílaba tônica de uma palavra inventada e desconhecida. Além disso, busca-se entender quais são os conhecimentos linguísticos do falante que influenciam esse processo de classificação das pseudopalavras.

Assim, é possível inferir se um falante do português ao se deparar com uma palavra nova, compara-a com palavras conhecidas para acentuá-la.

2 2. Objetivo(s)

O objetivo do trabalho é verificar se uma pseudopalavra que se assemelha a uma palavra real pode sofrer um processo analógico e ter o mesmo padrão acentual da palavra real, além de entender quais são os conhecimentos linguísticos do falante utilizados nesse processo de acentuação tônica das pseudopalavras. Algumas perguntas a a serem respondidas nessa etapa são:

- Quanto mais similar a palavra inventada for da palavra real, maior as chances do falante atribuir o mesmo padrão acentual da palavra real?
- Há diferença entre os grupos similares (i e iii) e os grupos dissimilares (ii e iv)?
- Há diferença entre os grupos frequentes (i e ii) e os infrequentes (iii e iv)?
- A tonicidade da palavra alvo tem papel na predição do acento?
- As variáveis selecionadas pelo modelo estão em concordância com a literatura da área?
- Identificar se há algum falante ou pseudopalavra tiveram comportamento destoante dos demais e qual o seu papel na atribuição acentual

3 3. Descrição do experimento

O estudo foi realizado durante o 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 fonologia, 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.

O experimento consistia em apresentar aos participantes, através do software Psychopy, uma série de palavras inventadas e registrar, via Google Meets, 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

Ao todo, foram criadas 372 pseudopalavras aleatorizadas em 4 blocos de 93 pseudopalavras, um total de 12.648 dados registrados na base. Dentro de cada bloco a ordem em que as pseudopalavras eram 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 aleatorização prévia. Em outras palavras, todos os indivíduos que em algum momento acompanharam o experimento pelos slides seguiram com blocos com a mesma aleatorização. (explicar a perda de respostas por causa de barulho etc, n temos todas as respostas de cada participante)

As palavras que deram origem às pseudopalavras foram classificadas em frequência de acordo com a sua ocorrência no *Corpus brasileiro*, corpus linguístico coordenado pelo pesquisador Antonio Paulo Berber Sardinha. Se possue mais de 100 mil ocorrências no corpus ela é classificada como alta frequência e se possue 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

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 nais qual sete falantes apresentaram a associação correta foram classificadas como quase validadas.

Da mesma forma, para validar se a pseudopalavra é dissimilar a sua palavra alvo, considerou-se como validadas as pseudopalavras cuja associação foi a mesma na resposta (não necessariamente a palavra alvo) de, no máximo, dois indivíduos.Pseudopalavras com um número de associações maior que dois foram consideradas não validadas,porém,pseudopalavras nas qual três falantes apresentaram a mesma associação foram classificadas como quase validadas.

Por fim, destacamos que os falantes e as pseudopalavras são variáveis aleatórias, pois podem indicar se algum participante ou pseudopalavra apresentam um comportamento muito específico.

4 4. Descrição das variáveis

Foram selecionadas variáveis linguísticas, extralinguística e experimentais, que podem, segundo a literatura da área, influenciar o comportamento acentual no português.

Variáveis Linguísticas

- **Tonicidade**: oxítona, paroxítona e proparoxítona (variável dependente)
- **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**: indica qual é a estrutura da pseudopalavra (CV-CV-CV ou CV-CV-CVC) - **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**: indica se a pseudopalavra foi validada ou não em testes prévios (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

Variáveis Extralinguísticas São as variáveis que caracterizam a amostra

- **Participante**: indica os 34 participantes do experimento.
- **Idade**: de 18 a 60
- **Gênero**: feminino e masculino
- **Naturalidade**: indica a cidade em que o participante nasceu
- **Escolaridade**: ensino eundamental a mestrado
- **Área de Formação**: indica a área em que o participante é formado (0 = outros e 1 = letras)
- **Línguas**: indica quais línguas o falante declara que fala ou já estudou. (categorizar ?)
- **Música**: indica se o participante já fez aula ou tem algum conhecimento em música

Variáveis Experimentais São as variáveis relacionadas ao método experimental

- **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**: 1, 2, 3 e 4
- **Ordem de apresentação**: indica em qual ordem a pseudopalavra foi apresentada dentro do bloco de apresentação (1 a 93)

5 5. Análise descritiva

6 6. Plano de análise

Pode-se considerar pseudopalavra uma variável aleatória/mista, pois pode indicar se alguma pseudopalavra apresenta um comportamento muito específico.

6.1 Organização do trabalho

O relatório foi organizado em XX capítulos, além desta introdução. No Capítulo 7, apresentamos as decisões metodológicas, procedimento de análise e organização dos dados. No Capítulo 8, apresentamos os principais resultados da análise, organizados de acordo com as questões norteadoras.

7 Metodologia

7.1 Coleta dos dados

7.2 Variáveis

As variáveis

7.3 Análise

8 Resultados