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

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>.

PROGRAMAS COMPUTACIONAIS UTILIZADOS

Software R (versão 4.0.5)

ÁREA DE APLICAÇÃO

Linguística (14:110)

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 lexicais das palavras alvo têm influencia 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 pseudopalvra? Há associação entre conhecimento em música e a classificação tônica da pseudopalvra? 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 reunião individuais da 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 estímulos 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 valores menores de 4 determinam palavras similares e valores maiores ou iguais a 5 determinam palavras dissimilares. Logo, foi necessário definir um modo de validar essa classificação em similar e dissimilar.

Esse ponto está relacionado aos testes para validar se a pseudopalavra é similar à palavra a partir da qual ela foi criada (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 mais de dois falantes lembraram de uma mesma palavra ao ler a pseudopalavra, ela foi considerada dissimilar à sua palavra-alvo - pois entendeu-se que ela está “parecida” com uma palavra real. Pseudopalavras nas quais três falantes apresentaram a mesma associação foram classificadas como quase validadas. Ressalta-se que, nesse caso, a associação não necessariamente precisava ser 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 alta frequência e se possui menos de 2 mil ocorrências ela é classificada como 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 estímulos.

- **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 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

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**: 1, 2, 3 (grupos similares), 5, 6, 7, 8, 9, 10 (grupos dissimilares)

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**: 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 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). 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).