Caligrama, Belo Horizonte, v. 26, n. 2, p. 155-177, 2021

Monolíngues? Uma investigação sobre o reconhecimento de palavras cognatas português-inglês

Monolinguals? An Investigation on the Recognition of Cognate Words From Portuguese-English

Maylton Silva Fernandes
Universidade Federal da Paraíba (UFPB), João Pessoa, Paraíba / Brasil
maylton.fernandes@gmail.com
http://orcid.org/0000-0003-4245-6041

Gustavo Lopez Estivalet
Universidade Federal da Paraíba (UFPB), João Pessoa, Paraíba / Brasil
gustavoestivalet@hotmail.com
http://orcid.org/0000-0003-3462-4156

Márcio Martins Leitão
Universidade Federal da Paraíba (UFPB), João Pessoa, Paraíba / Brasil
profleitao@gmail.com
http://orcid.org/0000-0003-2385-1636

Resumo: Palavras cognatas são conhecidas por dividirem semelhanças formais e semânticas entre duas ou mais línguas, possivelmente dividindo representações no léxico mental. Nesse sentido, as palavras cognatas possuem diferentes graus de semelhança, como por exemplo pares do português-inglês: cognatos perfeitos "banana", cognatos de alto grau "momento-*moment*" e cognatas de baixo grau "noite-*night*". Focalizando a relação formal e independentemente do conhecimento bilíngue, como as palavras cognatas do português-inglês são reconhecidas por monolíngues? O presente artigo tem o objetivo de investigar o reconhecimento de palavras cognatas do português-inglês por monolíngues através do grau de semelhança ortográfica. Para tanto, aplicamos um experimento de julgamento de aceitabilidade entre pares de palavras cognatas.

eISSN: 2358-9787
DOI: 10.17851/2358-9787.26.1.155-177

Com o objetivo de se pesquisar o grau de similaridade, utilizou-se a Distância de Levenshtein Normalizada entre as palavras cognatas. Os resultados apontaram uma correlação significativa entre o julgamento de aceitabilidade e este coeficiente. Portanto, os resultados indicaram que mesmo participantes não-bilíngues são capazes de reconhecer a granularidade da semelhança ortográfica. Ainda, de forma exploratória, foi possível determinar o coeficiente a partir do qual as palavras podem ser consideradas pares cognatos. Enfim, espera-se que o presente estudo permita uma melhor compreensão das palavras cognatas assim como provoque uma reflexão do monolinguismo.

Palavras-chave: Cognatas; distância de Levenshtein; julgamento de aceitabilidade; bilinguismo.

Abstract: Cognate words are known to share formal and semantic similarities between two or more languages, possibly dividing representations in the mental lexicon. In this sense, cognate words have different degrees of similarity, as for example Portuguese-English pairs: perfect cognates "banana", high degree cognates "*momento*-moment" and low degree cognates "*noite*-night". Focusing on the formal relationship and regardless of bilingual knowledge, how are cognate words in Portuguese-English recognized by monolinguals? This article aims to investigate the recognition of cognate words in Portuguese-English by monolinguals through the degree of orthographic similarity. For that, we applied an acceptability judgment experiment between cognate word pairs. In order to investigate the degree of similarity, the Normalized Levenshtein Distance was used between cognate words. The results showed a significant correlation between the acceptability judgment and this coefficient. Therefore, the results indicated that even non-bilingual participants are able to recognize the granularity of orthographic similarity. Still, in an exploratory way, it was possible to determine the coefficient from which words can be considered cognate pairs. Therefore, it is hoped that the present study allows a better understanding of cognate words as well as provoking a reflection of monolinguals.

Keywords: cognate; Levenshtein distance; acceptability judgement task; bilingualism.

1 Introdução

As palavras, definidas aqui como itens lexicais dotados de traços ortográficos, fonológicos e semânticos (DIJKSTRA, 2005), têm papel fundamental na linguagem. Nesse sentido, como essas palavras são representadas na mente humana? Como conseguimos nos expressar usando palavras específicas, quando conhecemos milhares delas?

Como as palavras são organizadas? Na psicolinguística, há o interesse em se compreender como o ser humano usa e compreende a linguagem e quais os processos mentais envolvidos nisso (LEITÃO, 2008). Em especial, nos estudos em bilinguismo, há o interesse em se entender como palavras que compartilham um mesmo significado, como palavras cognatas, são reconhecidas e representadas na mente.

Para tal finalidade, utilizam-se métodos experimentais que podem ser classificados como métodos *offline* e métodos *online*. Conforme Oliveira e Sá (2013), os métodos *offline* se caracterizam por terem como objetivo a obtenção de dados experimentais do pós-processamento linguístico. Em outras palavras, isso significa que os dados obtidos através desse tipo de experimento não refletem o processamento linguístico no momento exato em que o participante foi exposto a um determinado estímulo linguístico, mas sim um dado obtido após o processamento já concluído. Testes de associação de palavras e julgamento de aceitabilidade são alguns exemplos de experimentos *offline*.

Já os experimentos denominados *online* podem ser definidos como métodos experimentais utilizados para se obter dados no momento em que o processamento está ocorrendo. Isso quer dizer que a atuação do participante durante o experimento proporciona dados que podem ser associados à forma como ele processa um determinado estímulo durante a realização do experimento. Como exemplos de experimentos *online*, podemos citar a decisão lexical, leitura automonitorada e nomeação de imagens.

O objetivo da presente pesquisa foi investigar a partir de um experimento *offline* de julgamento de aceitabilidade como falantes do português brasileiro como língua materna não-bilíngues associam palavras cognatas do inglês às suas respectivas traduções. A motivação para a sua realização surgiu do interesse de se verificar quanto a semelhança ortográfica entre pares de palavras cognatas influencia na apreensão de seu significado. Além disso, desejamos correlacionar os resultados do experimento com a Distância de Levenshtein Normalizada (DLN) (SCHEPENS; DIJKSTRA; GROOTJEN, 2012), com o intuito de se identificar um possível padrão para a qualificação de palavras cognatas com base no grau de semelhança ortográfica.

As perguntas norteadoras para esta investigação foram: i. É possível identificar o significado de uma palavra cognata sem instrução ou conhecimento prévio sobre ela? ii. Há correlação entre o julgamento dos participantes e a DLN? Nossas hipóteses são que é possível identificar o

significado de uma palavra cognata mesmo sem instrução explícita sobre o seu significado, considerando a semelhança ortográfica que um par cognato de palavras compartilha. Além disso, acreditamos que há uma correlação positiva significativa entre a DLN e os resultados obtidos no experimento de julgamento, confirmando que a semelhança ortográfica entre pares cognatos influencia na identificação de seu significado, diferente dos falsos cognatos, como veremos adiante..

Esta pesquisa se justifica por apresentar uma forma prática e eficiente de validar palavras cognatas para outras investigações e experimentos, sejam elas no campo da psicolinguística, educação, ou áreas pertinentes às palavras cognatas.

O artigo está organizado da seguinte forma: na seção 2.1, apresentamos brevemente os conceitos de bilinguismo; na seção 2.2, conceituamos o termo cognato, assim como apresentamos exemplos e estudos psicolinguísticos sobre estas palavras; na seção 3.3, apresentamos a DLN; e, na seção 3.4, explicamos o funcionamento do experimental de julgamento de aceitabilidade, assim como sua aplicabilidade. Em seguida, descrevemos na seção 3 a metodologia aplicada para a realização do experimento: 3.1 participantes, 3.2 materiais e 3.3 procedimentos. Logo após, na seção 4, apresentamos e discutimos os resultados; e, na seção 5, discorremos as considerações finais do estudo.

2 Conceitos importantes

2.1 Bilíngues e monolíngues

O termo bilinguismo levanta diversas discussões entre os pesquisadores, em razão da complexidade de se definir quem é ou não bilíngue. Em sua forma mais simplória, podemos definir bilíngue alguém que conhece duas línguas (VALDÉS; FIGUEROA, 1994). Contudo, afirmar que alguém conhece duas línguas não nos permite concluir "o quão bilíngue" essa pessoa é, pois existem pessoas que apresentam alto nível de proficiência em duas línguas, enquanto outras apresentam um domínio maior sobre uma das duas línguas. Portanto, fica clara a importância de se levar em consideração a existência de diferentes níveis de bilinguismo.

Conforme Grant e Gottardo (2008), o que se deve considerar ao caracterizar bilíngues não é necessariamente o quanto ele conhece

das línguas, mas "quando" ele as adquiriu. Logo, bilíngues simultâneos são aqueles que foram expostos a duas línguas desde o nascimento (DE HOUWER, 2005), mas o termo também é aplicado àqueles que adquiriram a segunda língua (L2) entre os primeiros dois ou três anos de idade (COSTA; SEBASTIÁN-GALLÉS, 2014). Casos de bilinguismo simultâneo não são comuns, assim como não é comum que haja proficiência equivalente em ambas as línguas.

Ainda, os bilingues suscetíveis ou bilíngues sequenciais são aqueles que adquiriram a L2 após terem adquirido a sua primeira língua (L1). Dentro dessa perspectiva, é importante considerar quanto tempo depois da aquisição da L1 ocorreu a aquisição da L2. Bilíngues que adquirem a L2 até os cinco anos de idade são considerados bilíngues precoces, já aqueles que adquirem a L2 após esse período são chamados de bilíngues tardios (GRANT; GOTTARDO, 2008; COSTA; SEBASTIÁN-GALLÉS, 2014). Contudo, essa definição ainda gera discussões por linguistas e neurocientistas.

Sendo assim, pessoas que possuem conhecimentos triviais e gerais sobre uma L2 não seriam consideradas bilíngues conforme nenhuma dessas definições. Contudo, é fato que monolíngues que possuem um contato geral e sem instrução formal com uma L2, possuem alguns conhecimentos, ainda que limitados, sobre outras línguas, particularmente o inglês, língua franca presente nos dias atuais. Nesse sentido, monolíngues são capazes de perceber a semelhança entre as palavras de diferentes L2 e sua L1? Mais do que isso, monolíngues seriam capazes de reconhecer diferentes graus de semelhança formal, ou seja, ortográfica e/ou fonológica de palavras cognatas?

2.2 Palavras cognatas

Podemos definir palavras cognatas como sendo duas palavras que possuem um alto grau de semelhança ortográfica e/ou fonológica e que compartilham pelo menos um significado em comum entre duas línguas. Por exemplo, as palavras "cultura" e "*culture*" compartilham uma ortografia semelhante, bem como o mesmo significado. Semelhantemente, há palavras cognatas que apresentam a mesma representação ortográfica e sentido, as quais são definidas como cognatas perfeitas (e.g., "banana"). Além disso, conforme Schepens, Dijkstra e Grootjen (2012), palavras cognatas não precisam compartilhar todos os sentidos em ambas as

línguas, por exemplo, a palavra "*paper*" pode ser traduzida para "papel" ou "artigo" em português. Por sua vez, palavras que apresentam a mesma ortografia, mas diferentes significados, são definidas como falsas cognatas (e.g., "costume": "hábito" em português, "fantasia" em inglês).

Palavras cognatas têm sido exploradas em estudos psicolinguísticos com bilíngues, pois elas permitem que diversos fatores envolvidos na organização e processamento lexical sejam investigados. Nesse sentido, palavras cognatas normalmente apresentam tempos de reação mais curtos quando comparadas a outras palavras em estudos experimentais. Essa diferença, conhecida como efeito de facilitação de cognatas, pode ser observada em atividades de reconhecimento visual (DUÑABEITIA; PEREA; CARREIRAS, 2010), de compreensão auditiva (MARIAN; SPIVEY, 2003) e de produção oral (COSTA; CARAMAZZA; SEBASTIAN-GALLES, 2000; KROLL; STEWART, 1994).

Tendo em vista a sua relevância e pertinência em uma variedade de estudos sobre o processamento da L2, entende-se a importância em compreender o que são as palavras cognatas. Sendo assim, uma vez que tais palavras apresentam tamanha semelhança entre duas línguas, é possível que línguas que compartilhem um grande número de palavras cognatas sejam, consequentemente, mais fáceis de serem aprendidas, ainda que apenas em nos estágios iniciais (FIALHO, 2005).

Conforme Sánchez-Casas e García-Albea (2005), as palavras cognatas são representadas no léxico mental de forma distinta de outras palavras, assim como itens lexicais morfologicamente relacionados. Por exemplo, os substantivos "ajuda" e "ajudante" seriam representados juntos no léxico mental, uma vez que compartilham uma raiz comum e compartilharem semelhanças fonológicas, ortográficas e semânticas. Dessa forma, as palavras "música" e "*music*" também seriam representadas juntamente no léxico mental, uma vez que compartilham características formais e semânticas.

2.3 Distância de Levenshtein Normalizada

Uma série de estudos psicolinguísticos tem demonstrado que a distância de Levenshtein é a melhor medida para se quantificar e comparar a semelhança formal ortográfica e/ou fonológica entre palavras (DAVIS; PEREA; ACHA, 2009), assim como o número de vizinhos ortográficos e fonológicos das palavras de um léxico (ESTIVALET;

MEUNIER, 2017). A distância de Levenshtein é calculada a partir do número de modificações necessárias para transformar uma palavra em outra; estas modificações podem ser adições, apagamentos e substituições (e.g., adição: "acidental-*acc**i**dental*", apagamento: "momen**t**o-*momen**t***", substituição: "cultura-*cul**t**ure*") (LEVENSHTEIN, 1966).

Contudo, uma distância de Levenshtein pequena pode significar muita mudança em palavras pequenas (e.g., "oi-*on*" 50% da palavra) ou pouca mudança em palavras grandes (e.g., "independente-*independent*", 8,3% da palavra). Logo, apenas uma modificação em palavras pequenas pode influenciar consideravelmente sua representação formal, enquanto apenas uma modificação em palavras grande pode ser praticamente desconsiderada para sua representação formal. Sendo assim, com o objetivo de contornar esta limitação da distância de Levenshtein em relação ao tamanho das palavras comparadas, Schepens, Dijkstra e Grootjen (2012) propuseram a DLN, incluindo o número de caracteres ou fonemas para o cálculo do coeficiente, determinado pela fórmula:

$$-\frac{dist\hat{a}ncia}{tamanho}$$

onde distância = min. (adições, apagamentos, substituições) entre palavra 1 e palavra 2, e, tamanho = máx. (número de caracteres/fonemas). Logo, a DLN entre "oi-*on*" é

$$1-\frac{1}{2}=0{,}5$$

enquanto a DLN para "independente"-*independent* é

$$1-\frac{1}{12}=0{,}92$$

Assim, fica claro que a DLN do segundo par de palavras é muito maior que a DLN do primeiro par de palavras, evidenciando uma maior semelhança formal entre as palavras do segundo par de palavras do que do primeiro par de palavras.

Portanto, essa métrica corresponde à normalização da distância de Levenshtein relativa ao tamanho das palavras comparadas, considerando o número de modificações para se chegar da menor à maior palavra. Ou seja, calcula-se o menor número de adições, apagamentos e substituições

entre duas palavras para se transforma uma em outra e divide-se pelo número de caracteres/fonemas da maior palavra; enfim, estes resultados é subtraído de 1 para inverter o coeficiente. Sendo assim, a DLN apresenta valores entre 0 e 1, onde 0 representa nenhuma semelhança (e.g., "oi-*be*"

$$1-\frac{2}{2}=0$$

e 1 representa semelhança perfeita (e.g., "banana-*banana*"

$$1-\frac{0}{6}=1$$

2.4 Julgamento de aceitabilidade

O experimento de julgamento de aceitabilidade é comumente utilizado com o objetivo de validar palavras e sentenças para outros estudos (MORAES *et al*., 2016). Assim, ela funciona como uma forma de validar se os estímulos linguísticos selecionados correspondem às características esperadas e não foram selecionados de forma enviesada ou equivocada pelos pesquisadores.

Nesse sentido, o julgamento de aceitabilidade é um experimento *offline* que obtém dados do pós-processamento linguístico dos participantes. Ele busca verificar a compreensão dos participantes acerca de um determinado grupo de palavras, sentenças ou estruturas linguísticas. Além disso, Oliveira e Sá (2013, p. 7) destacam que essa metodologia não equivale ao teste de julgamento de gramaticalidade, pois esta última visa verificar se uma determinada estrutura linguística é gramaticalmente correta a partir de suas bases teóricas inerentes e independentes do indivíduo. Diferentemente, o julgamento de aceitabilidade "são relatos referentes às sensações dos participantes frente às construções em questão" e que considera não só a gramaticalidade de uma sentença, mas também seu custo no processamento, significado, o contexto em que se insere.

Comumente, utiliza-se a escala Likert como método de coleta de dados em julgamentos de aceitabilidade. Harpe (2015) menciona essa escala como uma boa forma de avaliar dados em grupos, como é o caso do teste em questão. Essa medida é "uma escala psicométrica que pode ser utilizada para se medir o nível de aceitabilidade de um indivíduo em relação a uma construção" (OLIVEIRA; SÁ, 2013, p. 8). Portanto, o participante pode atribuir uma pontuação para uma palavra

ou sentença que vai tipicamente de 1 a 5, contudo, escalas com diferentes gradações podem ser utilizadas. Essa pontuação corresponde ao nível de aceitabilidade do participante em relação a um estímulo linguístico apresentado. Assim, a pontuação 1 equivale a rejeição do estímulo apresentado, a pontuação 3 a uma aceitabilidade intermediária neutra a pontuação 5 a aceitação completa.

Moraes *et al*. (2016) realizaram um teste para verificar a aceitabilidade de frases no português brasileiro, considerando: i. nível de especificidade; ii. proeminência; e iii. perspectiva. Os autores observaram que tanto as frases em diferentes níveis de especificidade quanto de proeminência foram julgadas aceitáveis pelos participantes. No entanto, as frases com perspectiva espacial de localização foram julgadas como não-aceitáveis pelos participantes. Os autores avaliaram relevante a aplicação do experimento de julgamento de aceitabilidade na seleção dos estímulos, destacando sua importância para testagem de itens *a priori* de outros experimentos, além da exclusão do fator intuição para a seleção dos materiais e enviesamento dos mesmos.

Post e Leussen (2015) usaram um julgamento lexical com bilíngues altamente fluentes e a DLN com o objetivo de desenvolver um corpus de cognatos, especialmente baseando-se nas suas representações fonológicas. Os participantes tiveram que julgar a similaridade ortográfica entre as palavras, na pronúncia e sua familiaridade em uma escala Likert de 1 a 7. Em seus resultados, puderam identificar uma alta correlação entre a DLN e os julgamentos dos participantes.

Semelhantemente, em nosso estudo usaremos o julgamento de aceitabilidade para que participantes não-bilíngues avaliem o quão são cognatas as palavras a eles apresentadas.

3 Estudo experimental

3.1 Participantes

Os participantes do presente experimento foram selecionados através do compartilhamento de um *link* por meio das redes sociais. Antes de realizar o experimento, cada participante foi solicitado a consentir sua participar no experimento de forma voluntária e anônima por meio de um Termo de Consentimento Livre e Esclarecido (TCLE). Além disso, eles

foram orientados a apenas darem continuidade com o experimento se eles autodeclarassem conhecimento nenhum ou básico da língua inglesa. Assim, 117 participantes realizaram o experimento de julgamento de aceitabilidade, sendo a maioria do sexo feminino, entre 18 e 45 anos de idade, onde 93,3% dos participantes declarou não ter nenhum conhecimento de inglês e 6,7% dos participantes declarou ter nível básico de inglês.

Uma das nossas hipóteses foi que mesmo pessoas que não tiveram nenhum contato ou instrução formal do inglês seriam capazes de inferir o significado das palavras na língua inglesa, considerando apenas sua representação formal e semelhança com o português brasileiro.

3.2 Materiais

Utilizamos o Léxico do Português Brasileiro (ESTIVALET; MEUNIER, 2017) e o SUBTLEX-UK (VAN HEUVEN *et al.*, 2014) como corpora para seleção de palavras. Primeiramente, as palavras do português brasileiro foram selecionadas como estímulos de base, uma vez que é consideravelmente difícil manter a simetria entre as frequências de ambas as línguas, assim como devido às frequências das palavras do inglês corresponderem a frequências consideradas por nativos da língua, não refletindo o uso de um falante de inglês como L2.

Para a escolha das palavras do português, filtramos todos os substantivos singulares entre a frequência 1 e 1000 por milhão (f/M) de palavras. O objetivo dessa filtragem foi encontrar o maior número de palavras candidatas a cognatas com uma frequência semelhante, mas que não apresentassem nem alta e nem baixa frequência, minimizando possíveis efeitos desta métrica sobre os itens experimentais. Foram selecionadas 60 palavras com média de 260f/M e desvio padrão de 58, considerando-se o maior número de palavras desejadas no menor intervalo de frequência possível. Em seguida, controlamos o número de caracteres, sendo a média de caracteres de 6,67 e desvio padrão de 1,32, sendo a menor palavra composta por 5 caracteres e a maior palavra composta por 9 caracteres. Após essa seleção, verificamos as traduções das palavras do português brasileiro para o inglês, de maneira que a média de caracteres das palavras em inglês foi 6,52 e desvio padrão de 1,57, sendo a menor palavra composta por 5 caracteres e a maior palavra composta por 10 caracteres.

Realizamos um Teste-t para verificar possíveis diferenças entre o número de caracteres das palavras do português e do inglês $t(59) =$

1,21, p = 0,11, não evidenciando diferenças significativas. Além disso, recuperamos as frequências das palavras do inglês a partir do SUBTLEX-UK, calculamos a distância de Levenshtein (DAVIS; PEREA; ACHA, 2009) e a DLN (SCHEPENS; DIJKSTRA; GROOTJEN, 2012) entre todos os pares de palavras selecionados do português e do inglês para os nossos materiais.

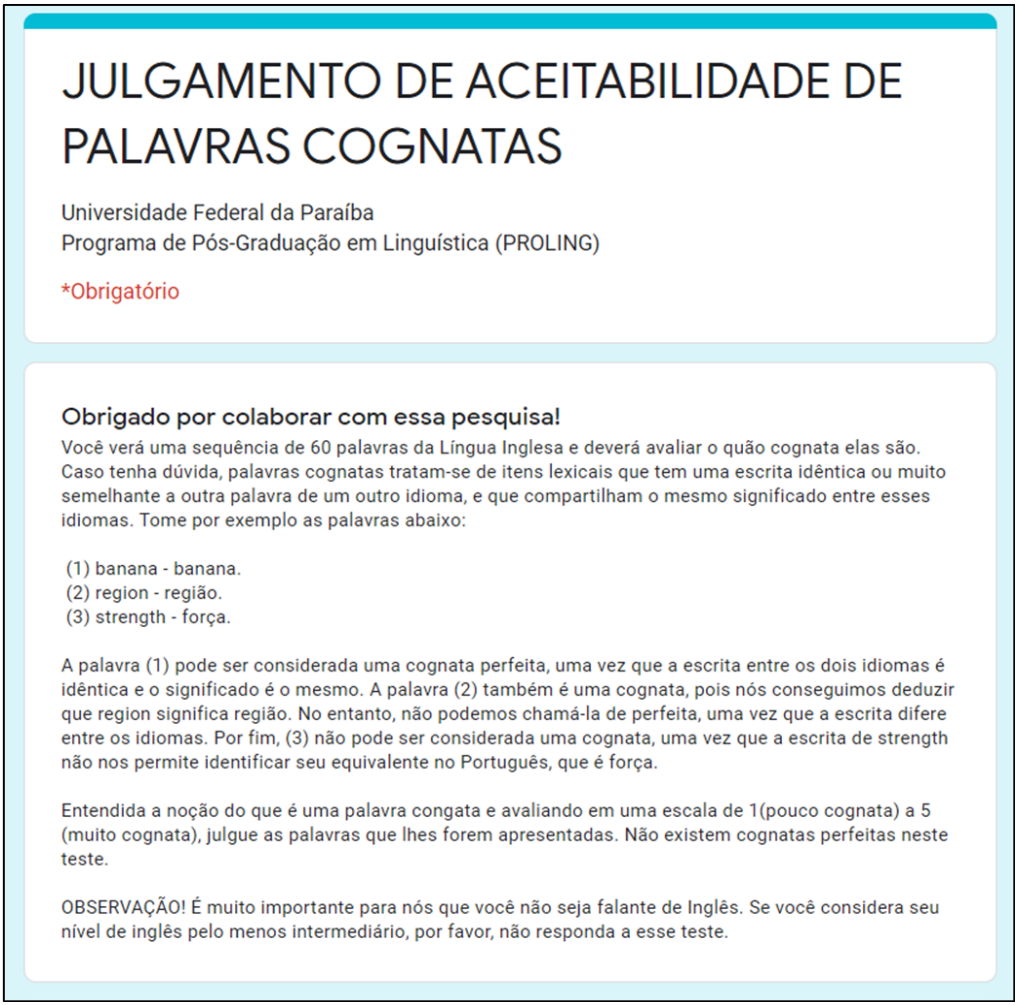
3.3 Procedimentos

O experimento foi realizado através da Internet e de forma remota por meio de um formulário elaborado no Google Forms. Nas instruções para o experimento, conforme a Figura 1, os participantes foram esclarecidos sobre a definição de palavra cognatas perfeitas "banana-*banana*" e palavras não-cognatas "casa-*house*". Ainda, as demais cognatas foram descritas como palavras que possuem ortografia semelhante entre as duas línguas e com o mesmo significado entre elas, como as palavras "computador-*computer*".

4 Resultados e discussão

Os dados foram analisados através do programa R (R CORE TEAM, 2014). Destaca-se que a análise realizada possui um caráter de mineração de dados com três objetivos específicos: i. análise descritiva da distribuição dos pares selecionados e dados coletados; ii. análise inferencial das diferenças das respostas em função da DLN; e iii. análise exploratória para determinação do status de palavra cognata. Inicialmente, foram contabilizados o número total de respostas, assim como a distribuição do número de respostas para cada um dos cinco pontos da escala Likert, conforme o Gráfico 1 e o Gráfico 2, respectivamente.

Figura 1 – Instruções do experimento de julgamento de aceitabilidade no Google Forms



JULGAMENTO DE ACEITABILIDADE DE PALAVRAS COGNATAS

Universidade Federal da Paraíba
Programa de Pós-Graduação em Linguística (PROLING)

*Obrigatório

Obrigado por colaborar com essa pesquisa!

Você verá uma sequência de 60 palavras da Língua Inglesa e deverá avaliar o quão cognata elas são. Caso tenha dúvida, palavras cognatas tratam-se de itens lexicais que tem uma escrita idêntica ou muito semelhante a outra palavra de um outro idioma, e que compartilham o mesmo significado entre esses idiomas. Tome por exemplo as palavras abaixo:

(1) banana - banana.
(2) region - região.
(3) strength - força.

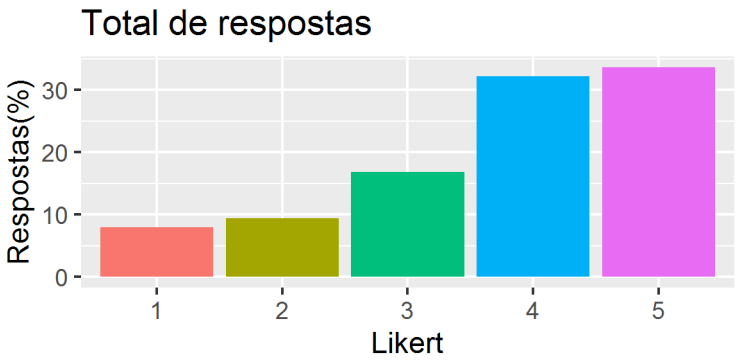
A palavra (1) pode ser considerada uma cognata perfeita, uma vez que a escrita entre os dois idiomas é idêntica e o significado é o mesmo. A palavra (2) também é uma cognata, pois nós conseguimos deduzir que region significa região. No entanto, não podemos chamá-la de perfeita, uma vez que a escrita difere entre os idiomas. Por fim, (3) não pode ser considerada uma cognata, uma vez que a escrita de strength não nos permite identificar seu equivalente no Português, que é força.

Entendida a noção do que é uma palavra congata e avaliando em uma escala de 1(pouco cognata) a 5 (muito cognata), julgue as palavras que lhes forem apresentadas. Não existem cognatas perfeitas neste teste.

OBSERVAÇÃO! É muito importante para nós que você não seja falante de Inglês. Se você considera seu nível de inglês pelo menos intermediário, por favor, não responda a esse teste.

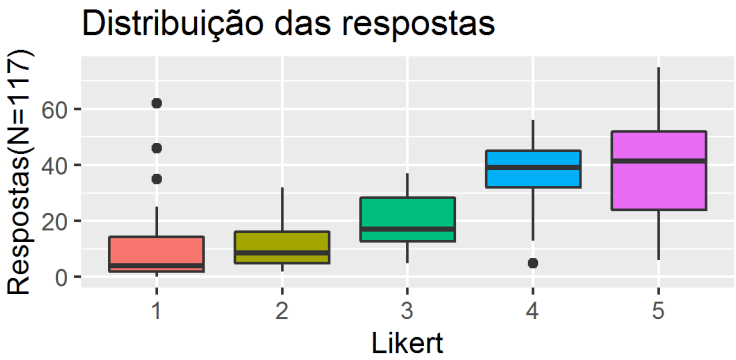
Fonte: Elaborado pelos autores.

Gráfico 1 – Número total de respostas



Fonte: Elaborado pelos autores.

Gráfico 2 – Distribuição do número de respostas por participante na escala Likert



Fonte: Elaborado pelos autores.

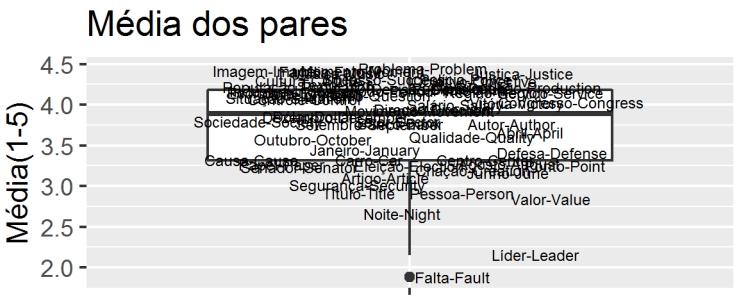
Em seguida, com objetivo de descrever as características lexicais dos pares de palavras utilizados no experimento de julgamento de aceitabilidade, foram calculadas as médias e desvios-padrão da média, mediana e modo da pontuação na escala Likert das palavras, assim como da distância de Levenshtein, da DLN, do número de caracteres e da frequência das palavras do português e do inglês, conforme a Tabela 1. As médias da escala Likert também podem ser observadas no Gráfico 3 e no Gráfico 5, e, as DLN também podem ser observadas no Gráfico 4.

Tabela 1 - Média e desvio padrão das respostas dos participantes e das características dos pares de palavras do experimento

Categoria	Média	Desvio-padrão
Média	3,74	0,58
Mediana	3,82	0,81
Modo	4,08	1,06
Distância de Levenshtein	2,30	1,06
Distância de Levenshtein Normalizada	0,67	0,14
Número de caracteres português	6,67	1,32
Número de caracteres inglês	6,52	1,53
Frequência português	5,40[1]	0,09
Frequência inglês	4,77	0,61

Fonte: Elaborado pelos autores.

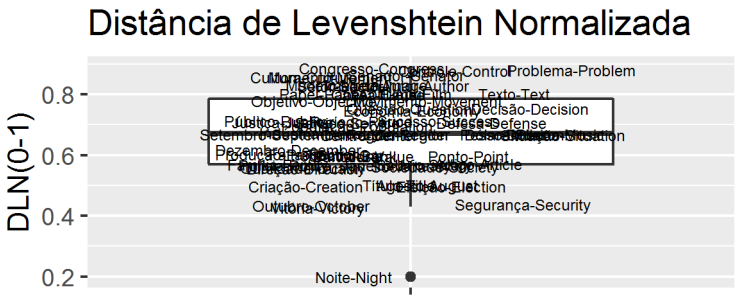
Gráfico 3 – Média dos pares de palavras



Fonte: Elaborado pelos autores.

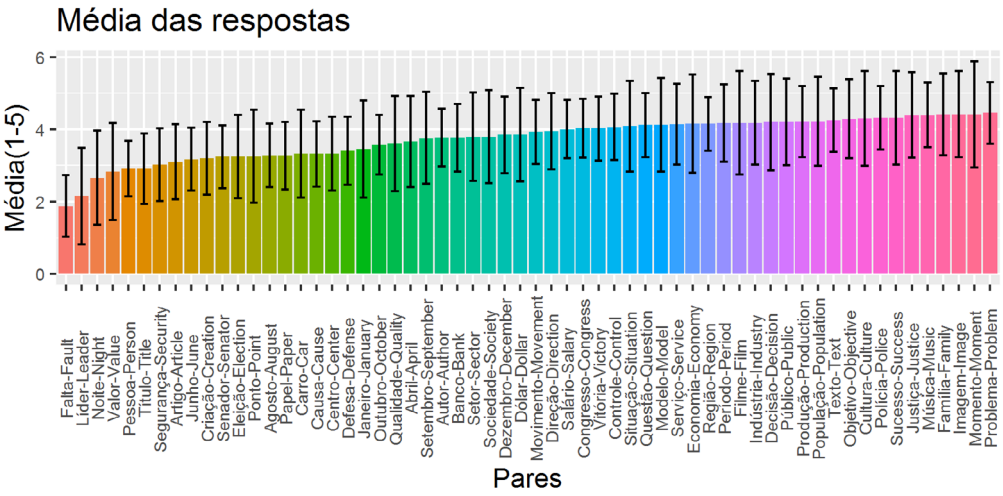
[1] As médias e os desvios padrão das frequências das palavras do português e do inglês nesta tabela são expressas através da escala Zipf (VAN HEUVEN et al., 2014).

Gráfico 4 – Distância de Levenshtein Normalizada para os pares de palavras



Fonte: Elaborado pelos autores.

Gráfico 5 – Médias e desvios-padrão das pontuações da escala Likert dos pares de palavras



Fonte: Elaborado pelos autores.

Através da Tabela 2, fica clara a média de 0,67 da DLN dos pares de palavras. Destaca-se a pequena diferença das médias e desvios padrão do número de caracteres das palavras do português e do inglês, assim como o pequeno desvio padrão das frequências do português, evidenciando

um bom controle dos materiais escolhidos para o experimento. A média e desvio padrão das frequências das palavras do inglês não puderam ser perfeitamente controladas justamente porque estes dados foram trazidos pelas traduções em inglês do SUBLEX-UK das palavras do português selecionadas primeiramente (VAN HEUVEN *et al.*, 2014).

Tabela 2 – Análise inferencial das simulações da divisão dos dados em não-cognatas e cognatas a partir da Distância de Levenshtein Normalizada

DLN	Pares de palavras	Teste-t	Kruskal-Wallis
0,4	1/59	-	$X^2(1) = 2,52$, $p = 0,11$
0,5	8/52	$t(11) = 3,47$, $p < 0,01$	$X^2(1) = 8,75$, $p < 0,01$
0,6	**23/37**	**$t(40) = 3,38$, $p < 0,01$**	**$X^2(1) = 11,55$, $p < 0,001$**
0,7	34/26	$t(55) = 3,87$, $p < 0,001$	$X^2(1) = 11,17$, $p < 0,001$
0,8	50/10	$t(19) = 2,61$, $p < 0,05$	$X^2(1) = 3,74$, $p = 0,05$

Fonte: Elaborado pelos autores.

Tendo em vista que a escala Likert de cinco pontos utilizadas no presente experimento sugere um contínuo na escala de aceitabilidade e semelhança das palavras, utilizou-se como coeficiente da variável dependente a média das respostas dos participantes, ainda, a média foi o coeficiente que apresentou menor desvio padrão de resposta. Este coeficiente foi calculado a partir da soma de pontuações das respostas da escala Likert (1-5) dividida pelo número total de participantes do experimento (117) (HARPE, 2015).

Em seguida, observou-se a distribuição da média das respostas, conforme o Gráfico 3 e o Gráfico 5. O Gráfico 3 evidencia que os pares "falta-*fault*" e "líder-*leader*" foram *outliers* em relação às demais médias e são analisados qualitativamente posteriormente. Portanto, através do Gráfico 5, fica evidente que a maior parte (90%) dos pares de palavras obtiveram médias de pontuação entre 3 e 4,5, ou seja, acima da metade superior da escala Likert utilizada. Estes resultados corroboram nossa hipótese de que mesmo participantes não-bilíngues podem reconhecer e classificar as palavras cognatas com boa assertividade baseados na semelhança formal ortográfica.

Em seguida, o teste de Kolmogorov-Smirnov ($D = 0,14$, $p = 0,18$) evidenciou que os dados apresentam uma distribuição normal, logo, as análises subsequentes foram realizadas a partir de testes paramétricos. Entre todas as variáveis apresentadas acima (média, mediana, modo, distância de Levenshtein, DLN, número de caracteres e frequência do

português e do inglês), o par de variáveis que apresentou maior correlação e resultado mais significativo foi justamente a correlação entre a média das pontuações da escala Likert e a DLN $t(58) = 3,84$, $p < 0,001$, $r = 0,45$, IC(95%) = 0,22-0,63, conforme o Gráfico 6.

Em seguida, foi aplicado um modelo generalizado de regressão linear multifatorial com a média das pontuações da escala Likert como variável dependente, e, distância de Levenshtein, DLN, número de caracteres e frequência do português e do inglês como variáveis independentes. A ANOVA do modelo linear apresentou resultados significativos apenas para as variáveis DLN $F(1,53) = 18,67$, $p < 0,001$ e distância de Levenshtein $F(1,53) = 12,17$, $p < 0,001$ (DAVIS; PEREA; ACHA, 2009). Este resultado já era esperado justamente porque ambas as variáveis se correlacionam fortemente, tendo em vista que a primeira é a normalização da segunda em função do número de caracteres das palavras comparadas. Enfim, o modelo linear evidenciou como único resultado significativo a DLN $t(53) = 2,09$, $p < 0,05$, $r^2 = 0,36$.

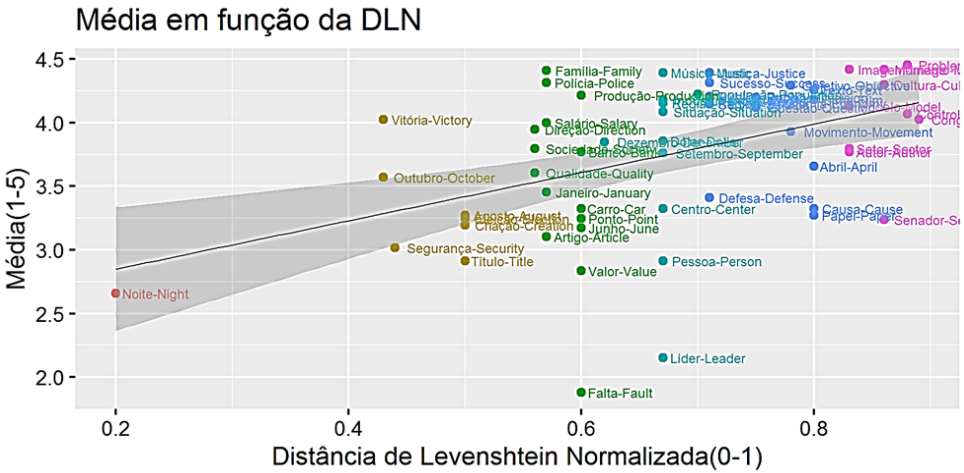
Estes resultados corroboram nossa hipótese de trabalho da correlação positiva significativa entre a média o julgamento de aceitabilidade das palavras da escala Likert e a DLN dos pares de palavras $t(58) = 12,97$, $p < 0,001$, $r = 0,86$, IC(95%) = 0,77-0,91, como no trabalho de Post e Leussen (2015). Ou seja, quanto menor o número de modificações entre um par de palavras, mais elas são parecidas e, consequentemente, mais facilmente elas são identificadas como palavras cognatas, explicando as médias no julgamento de aceitabilidade.

Logo após, foi realizada uma análise da variável numérica independente DLN com o objetivo de se definir o coeficiente que permite a interpretação do status de palavra cognata (SCHEPENS; DIJKSTRA; GROOTJEN, 2012). Para tanto, realizaram-se análises inferenciais com o teste paramétrico Teste-t e o teste não-parâmetrico Kruskal-Wallis de cinco simulações de divisão do conjunto dos 60 pares de palavras em dois (i.e., não-cognatas e cognatas), a partir dos diferentes coeficientes de DLN possíveis, conforme a Tabela 2.

A análise da Tabela 2 aponta que pares de palavras com DLN de 0,5 ou mais já podem ser consideradas cognatas, ou pelo menos, já há algum grau de semelhança formal de semelhança entre as palavras. Contudo, destaca-se que o coeficiente de 0,6 parece ser mais seguro na determinação do status de palavras cognatas, especialmente observando-se os resultados do teste não-paramétricos. Portanto, além da análise formal indicando que DLN maiores que 0,5 deveriam ser consideradas palavras cognatas (SCHEPENS; DIJKSTRA; GROOTJEN, 2012), os resultados

da Tabela 2 corroboram esta hipótese através de um teste de percepção e julgamento de aceitabilidade através da semelhança formal ortográfica.

Gráfico 6 – Correlação e regressão linear das variáveis média de pontuação da escala Likert em função da Distância de Levenshtein Normalizada.



Fonte: Elaborado pelos autores.
Legenda: As cores das palavras representam as divisões dos coeficientes da Tabela 3

Finalmente, aplicou-se o teste Tukey a partir as divisões 0-0,4-0,5-0,6-0,7-0,8-0,9 da DLN (conforme as cores do Gráfico 6) para se definirem as diferenças significativas dos diferentes graus de pares de palavras cognatas. Conforme a Tabela 3, observa-se que as palavras cognatas com DLN entre 0,6 e 0,7 são as que possuem diferença mais significativa, seguidas pelas cognatas com DLN entre 0,5 e 0,6. Isso indica mais uma vez que existe uma diferença significativa no quão cognatas as palavras são entre as escalas de 0,5 a 0,7 em uma escala crescente (quanto maior, mais cognata). Contudo, essa diferença não existe nas escalas entre 0,7 e 0,8, nem entre 0,8 e 0,9, sugerindo que a DLN de 0,6 seria a melhor escolha para se determinar o status cognato das palavras.

Em relação à DLN, o Gráfico 4 e o Gráfico 6 destacam que a DLN de 0,2 de "noite-*night*" foi um *outlier* e que todos os demais pares de palavras possuem uma distância entre 0,4-0,9. Mesmo sendo uma palavra frequentemente encontrada no Brasil, "*night*" recebeu uma pontuação baixa no teste, sobretudo, em função da baixa DLN. Logo, é possível

que os participantes reconheçam esta palavra, mas não a consideram cognata da palavra "noite", tendo em vista sua baixa semelhança formal ortográfica, sugerindo o bom funcionamento e confiabilidade do experimento aplicado, assim como das hipóteses propostas.

Tabela 3 – Teste Tukey (IC 95%) dos diferentes níveis de Distância de Levenshtein Normalizada

NLD	Diferença	Limite inferior	Limite superior	Valor p
0,4-0,5	0,66	-0,99	2,31	0,81
0,5-0,6	0,68	-0,04	1,38	0,05
0,6-0,7	**0,74**	**-0,02**	**1,51**	**0,03**
0,7-0,8	0,28	-0,32	0,89	0,73
0,8-0,9	0,06	-0,56	0,68	0,91

Fonte: Elaborado pelos autores.

Por outro lado, o par de palavras "congresso-*congress*" apresentou a maior DLN de 0,89 em função de ter apenas uma modificação e serem palavras grandes, recebendo média de 4,03. De forma semelhante, o par de palavras "problema-*problem*" também apresenta alta DLN de 0,88 e apresentou a maior média de 4,45 em função de apresentar a modificação apenas na última letra da palavra. Ainda, destacamos que o par de palavras "senado-*senator*" também apresenta DLN alta de 0,86, mas recebeu média de 3,24 em função de apresentar a modificação no meio da palavra (DAVIS; PEREA; ACHA, 2009).

Já em relação às médias da escala Likert, o par de palavras fonologicamente semelhantes "líder-*leader*" com DLN de 0,67 e média baixa de 2,15 indica que o julgamento de aceitabilidade foi feito exclusivamente na forma ortográfica. Diferentemente do inglês, o português apresenta transparência entre ortografia e fonologia, logo, este resultado sugere que não houve influência da forma fonológica semelhante desse par de palavras, uma vez que os participantes não-bilíngues não apresentam nível de inglês suficiente para tal assimilação. Assim, é provável que os participantes leram a palavra "*leader*" usando o padrão fonológico do português, dificultando a identificação da semelhança formal com a palavra "líder" (COSTA; CARAMAZZA; SEBASTIAN-GALLES, 2000).

Ainda, o par de palavras "falta-*fault*" apresentou tanto uma baixa DLN de 0,60 quanto uma baixa média de 1,88. A diferença formal entre essas palavras parece pequena, mas ao se considerar o número de modificações, assim como o seu tamanho pequeno, justifica-se a dificuldade dos participantes em identificar a semelhança ortográfica. Portanto, corrobora-se que o tamanho das palavras é um fator determinante na medida da semelhança entre pares, onde palavras maiores são menos afetadas do que palavras menores por poucas modificações.

Enfim, destacamos o par de palavras "família-*family*", recebendo DLN baixa de 0,57 e média alta de 0,41. Interessantemente, sua média foi maior que palavras como "modelo-*model*" que apresenta apenas uma modificação. Esse resultado pode estar relacionado a fatores como a familiaridade dos participantes com o termo. Outra possibilidade é que, apesar da letra "y" ser incomum em palavras do português, a palavra "*family*" já se tornou conhecida entre falantes do português não-bilíngues e, portanto, foi julgada como altamente semelhante, além do fato da letra "y" apresentar a mesma pronúncia da vogal "i" do português.

5 Considerações finais

De forma geral, os resultados apresentados corroboram nossas hipóteses acerca do julgamento de aceitabilidade de palavras cognatas do português-inglês como uma forma eficaz de se validar tais palavras para estudos linguísticos. Uma de nossas perguntas norteadoras foi: seria possível identificar o significado de uma palavra cognata sem instrução ou conhecimento prévio sobre a mesma? Sim, nossos dados evidenciaram que mesmo não-bilíngues são capazes de reconhecer e julgar palavras como cognatas em diferentes graus. Outro questionamento que levantamos foi se haveria correlação entre o julgamento dos participantes do experimento e a DLN? Mais do que isso, os resultados significativos dos testes paramétricos e não-paramétricos sugeriram que existe uma correlação significativa entre a percepção dos pares de cognatos e a DLN. Enfim, ainda foi possível investigar o coeficiente de 0,5-0,6 que define o status de palavra cognata a partir das médias do julgamento de aceitabilidade em função da DLN.

Como limitações, nosso experimento não apresentou muitos pares de palavras com baixa DLN. Para investigações futuras, sugere-se utilizar estímulos em toda gama de DLN para verificar o espectro de aceitabilidade ao longo deste coeficiente. Além disso, aumentar e controlar o número de

palavras e suas características lexicais em cada coeficiente da DLN pode ajudar a qualificar de forma mais aprofundada a distinção entre palavras cognatas e não-cognatas. Outro aspecto interessante para futuras pesquisas é investigar como a localização das modificações (i.e., início, meio ou fim das palavras) influencia no julgamento dos participantes. Também, é possível se manipular além da forma ortográfica a forma fonológica das palavras, assim como introduzirem-se ao experimento falsos-cognatos.

Agradecimentos

Agradecemos à CAPES pela bolsa de mestrado do primeiro autor (processo 88887.352423/2019-00), ao CNPq pela bolsa de Produtividade em Pesquisa do terceiro autor. Agradecemos à comissão organizadora do II Encontro Mineiro de Psicolinguística. Enfim, agradecemos enormemente aos dois revisores da Revista Caligrama pela leitura atenta e sugestões.

Referências

COSTA, A.; CARAMAZZA, A.; SEBASTIÁN-GALLÉS, N. The cognate facilitation effect: implications for models of lexical access. *Journal of experimental psychology:* learning memory and cognition, [*S. l.*], v. 26, n. 5, p. 1283–1296, 2000. DOI: https://doi.org/10.1037/0278-7393.26.5.1283

COSTA, A.; SEBASTIÁN-GALLÉS, N. How does the bilingual experience sculpt the brain? *Nature reviews neuroscience,* [*S. l.*], v. 15, n. 5, p. 336–345, May 2014. DOI: https://doi.org/10.1038/nrn3709

DAVIS, C. J.; PEREA, M.; ACHA, J. Re(de)fining the orthographic neighborhood: the role of addition and deletion neighbors in lexical decision and reading. *Journal of experimental psychology*: human perception and performance, [*S. l.*], v. 35, n. 5, p. 1550–1570, 2009. DOI: https://doi.org/10.1037/a0014253

DE HOUWER, A. Early bilingual acquisition: focus on morphosyntax and the separate development hypothesis. *In:* KROLL, Judith F.; DE GROOT, Annette M. B. *Handbook of bilingualism:* psycholinguistic approaches. New York: Oxford University Press USA, 2005. p. 30–48.

DUÑABEITIA, J. A.; PEREA, M.; CARREIRAS, M. Masked translation priming effects with highly proficient simultaneous bilinguals. *Experimental psychology*, [*S. l.*], v. 57, n. 2, p. 98–107, 2010. DOI: https://doi.org/10.1027/1618-3169/a000013.

ESTIVALET, G. L.; MEUNIER, F. Corpus psicolinguístico Léxico do Português Brasileiro. *Revista SOLETRAS*, Rio de Janeiro, n. 33, p. 212-229, 2017. DOI: https://doi.org/10.12957/soletras.2017.29702

FIALHO, V. Proximidade entre línguas: algumas considerações sobre a aquisição do espanhol por falantes nativos de português brasileiro. *Espéculo*: revista de estudios literarios, Madri, v. 1, n. 31, p. 1-15, 2005. Disponível em: https://webs.ucm.es/info/especulo/numero31/falantes.html. Acesso em: 22 jul. 2021.

GRANT, A.; GOTTARDO, A. *Defining bilingualism*. London, ON: Encyclopedia of Language and Literacy Development, 2008.

HARPE, S. E. How to analyze Likert and other rating scale data. *Currents in pharmacy teaching and learning*, [*S. l.*], v. 7, n. 6, p. 836-850, 2015. DOI: https://doi.org/10.1016/j.cptl.2015.08.001

KROLL, J. F.; STEWART, E. Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. *Journal of memory and language*, [*S. l.*], v. 33, n. 2, p. 149-174, 1994. DOI: https://doi.org/10.1006/jmla.1994.1008

LEITÃO, M. M. Psicolinguística experimental: focalizando o processamento da linguagem. *In:* Martelotta, Mario Eduardo. (org.) *Manual de linguística.* São Paulo: Contexto, 2008.

LEVENSHTEIN, V. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady,* Moscow, v. 10, n. 8, p. 707-710, 1966.

MARIAN, V.; SPIVEY, M. Bilingual and monolingual processing of competing lexical items. *Applied psycholinguistics*, [*S. l.*], v. 24, n. 2, p. 173–193, 2003. DOI: https://doi.org/10.1017/S0142716403000092

MORAES *et al.* A importância do teste de plausibilidade na validação de frases em experimentos psicolinguísticos. *Prolingua*, [*S. l.*] v. 11, n. 1, p. 17-26, 2016. Disponível em: https://periodicos.ufpb.br/index.php/prolingua/article/view/30627. Acesso em: 22 jul. 2021.

OLIVEIRA, C. S. F.; SÁ, T. M. M. Métodos off-line em psicolinguística: julgamento de aceitabilidade. *Revele:* revista virtual dos estudantes de Letras, [*S. l.*], v. 5, p. 77-96, 2013. DOI: https://doi.org/10.17851/2317-4242.5.0.77-96.

POST, A.; LEUSSEN, J. Generating a bilingual lexical corpus using interlanguage Normalized Levenshtein Distances. *In:* INTERNATIONAL CONGRESS OF PHONETIC SCIENCES, 18., 2015, Glasgow. *Proceedings* […], 2015. [*S. l.: s. n.*], 2015.

R CORE TEAM. *R*: A language and environment for statistical computing. Vienna: [*s. n.*], 2014. *E-book*.

SÁNCHEZ-CASAS, R.; GARCÍA-ALBEA, J. The representation of cognate and noncognate words in bilingual memory: can cognate status be characterized as a special kind of morphological relation? *In:* KROLL, Judith F.; DE GROOT, Annette M. B. *Handbook of bilingualism:* psycholinguistic approaches. New York: Oxford University Press USA, 2005. p. 226–250.

VALDÉS, G.; FIGUEROA, R. A. *Bilingualism and testing:* a special case of bias. Westport: Ablex Publishing, 1994. *E-book.*

VAN HEUVEN, W. J. B. *et al.* SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly journal of experimental psychology*, Canterbury, v. 67, n. 6, p. 1176-1190, 2014. DOI: https://doi.org/10.1080/17470218.2013.850521.

Recebido em: 10 de abril de 2021.

Aprovado em: 19 de junho de 2021.

VARIA