https://doi.org/10.31513/linguistica.2020.v16nEsp.a20134
Edição Especial Comemorativa

EXPANSÃO LEXICAL: O ACESSO A EMPRÉSTIMOS DO INGLÊS COM MORFOLOGIA DO PORTUGUÊS

LEXICAL EXPANSION: ACCESS TO ENGLISH LOAN-WORDS WITH PORTUGUESE MORPHOLOGY

Fernando Luis Sabatini[1]

Thiago Oliveira da Motta Sampaio[2]

RESUMO

A Expansão Lexical é um tema recorrente na linguística histórica e na sociolinguística. Até onde sabemos, esse estudo é a primeira tentativa de compreender o tema por meio de uma abordagem experimental. Neste artigo, reportamos os resultados de um experimento de decisão lexical baseado em palavras emprestadas do Inglês que aceitam afixos do Português do Brasil (ex. clipagem, cropar). Uma árvore de regressão CRT indica que os participantes reconhecem essas palavras com mais facilidade do que pseudopalavras. Porém, essas respostas são mais lentas e não podem ser comparadas às respostas a palavras do Inglês ou do Português do Brasil.

PALAVRAS CHAVE: Expansão Lexical, Decisão Lexical, processamento de palavras, empréstimos lexicais

ABSTRACT

Lexical Expansion is a recurring theme in historical linguistics and sociolinguistics. As far as we know, the present study is the first attempt to pursuit the subject through an experimental approach. In the present paper, we report the results of a lexical decision test based on loan words from English that accept affixes from Brazilian Portuguese (eg. clipagem, cropar). A CRT regression tree indicates that participants recognize these words better than pseudowords. However, their responses are slower and cannot be compared to responses for English or for Brazilian Portuguese words.

KEYWORDS: Lexical Expansion, Lexical Decision, Word Processing, Loan Words

[1] Universidade Estadual de Campinas (UNICAMP). Graduando em Letras – Língua Portuguesa. Contato: sabatini.nando@gmail.com.

[2] Universidade Estadual de Campinas (UNICAMP). Professor do Programa de Pós-Graduação em Linguística e do Departamento de Linguística. Contato: thiagomotta@iel.unicamp.br.

1. Introdução

Como funciona a expansão do léxico em uma língua? Esse tema é bastante recorrente e importante para compreender, em estudos históricos, a variação e a mudança nas diferentes línguas do mundo (Algeo, 1980; Arnaud, 2013; Kriaučiūnienė & Sangailaitė 2016). Porém, até onde sabemos, nunca foi alvo de investigação psicolinguística — o que é compreensível.

A investigação sobre a expansão do léxico parece se tornar mais simples e concreta ao analisarmos, em um corpus, a ocorrência de determinados termos em diferentes momentos históricos. Através desse método, é possível apontar com razoável precisão quando uma palavra surgiu na língua, de onde veio e as mudanças em seu uso ao longo do tempo. Dessa forma, os estudos históricos sobre o tema adotam uma perspectiva diacrônica e focada nas línguas. Por outro lado, a psicolinguística estuda os processos mentais que nos permitem usar de forma natural as línguas e suas possibilidades. Isso configura a abordagem da psicolinguística como sincrônica e focada no falante. Essas características parecem afastar a possibilidade de testes psicolinguísticos focados na expansão lexical — ou ao menos dificultá-los.

Ainda assim, acreditamos que a psicolinguística seja uma abordagem bastante produtiva e que pode nos oferecer *insights* sobre o processo da expansão lexical. A variação e a mudança linguística ocorrem pois os falantes de uma língua a utilizam de formas variadas sem que isso prejudique a intercompreensão. Junto com os diferentes contextos de uso, a variação conduz a língua a uma mudança ao longo do tempo. Portanto, uma abordagem que analise a ativação lexical de novos termos pode vir a trazer dados que nos permitam compreender melhor o processo de expansão lexical na própria mente do falante.

Considerando que uma das formas de expansão do léxico de uma língua é o uso de palavras emprestadas de outras línguas, o presente artigo se propõe a investigar o acesso lexical de uma categoria curiosa de empréstimo, aquelas palavras da língua inglesa que aceitam material morfológico do português (ex. cropar, clipar, twittar). Até onde sabemos, essa é a primeira tentativa de investigação experimental desse tipo de empréstimos no português, o que nos traz um grande desafio metodológico, especialmente no que diz respeito às frequências de acesso das palavras.

A próxima seção inicia a discussão sobre a formação de novas palavras; as duas seções seguintes abordam os estudos de acesso lexical na psicolinguística e os testes de decisão lexical, respectivamente. A seção 5 apresenta nossa proposta de abordagem psicolinguística dos empréstimos com camadas morfológicas. A seção 6 apresenta nosso experimento e finalizamos o artigo com as discussões gerais e considerações finais na seção 7.

2. Sobre a formação de novas palavras

Você já criou neologismos ou compreendeu algum assim que foi proferido sem nunca tê-lo escutado antes? É muito comum a criação, espontânea ou não, de novas palavras no léxico de uma língua. Compreender como acontecem esses acréscimos no nosso léxico mental é o objetivo dos estudos em Expansão Lexical (*Lexical Expansion*[3]). De uma forma geral, sabemos que os neologismos são criados a partir de alguma manipulação que toma como base alguma palavra já existente. Mas que tipos de manipulações seriam essas?

Segundo Kriaučiūnienė & Sangailaité (2016), a formação de uma nova palavra pode ocorrer de três maneiras. O primeiro caso é a ocorrência de uma nova composição morfológica com peças já existentes — seja por meio da adição de afixos (descafeinar), seja por composição (porta-temperos), seja por blending (Reiberto Carlos) —, processo esse que é responsável pela maioria dos casos de neologismos e, por consequência, alvo de grande parte das pesquisas linguísticas sobre o tema.

A segunda forma de neologismo consiste na mudança semântica de palavras já existentes, geralmente resultado de mudanças na sociedade que causam o enfraquecimento do conteúdo semântico de uma palavra e, por vezes, a utilização dessas palavras em contextos diferentes dos tradicionais, como nos casos de semantic bleaching (Sweetzer, 1988) (ex. “esse rapaz é foda!”, que pode ser usado tanto de forma elogiosa quanto pejorativa).

A terceira forma de expansão lexical consiste em tomar emprestada uma palavra de outra língua. Os empréstimos, por vezes, serão adotados em sua forma original, como em ‘download’ ou ‘shopping’, mantendo ortografia e fonologia independente das regras da língua alvo. Em outras situações, a língua alvo adaptará a palavra original à sua ortografia/fonologia, como em ‘blecaute’ (de blackout) ou ‘surfe’ (de surf). Um terceiro caso acontece quando palavras emprestadas ganham material morfológico da língua alvo, como em ‘linkado’ e ‘crackeado’. É este terceiro caso o tema principal de nosso experimento.

Esse campo de pesquisa linguística trata de um tema recorrente na Linguística Teórica (Algeo, 1980; Arnaud, 2013; *apud* Kriaučiūnienė & Sangailaité, 2016) mas que, até onde sabemos, não foi foco de alguma abordagem na Linguística Experimental. Este estudo consiste

[3] Não confundir com a área de expansão lexical na Linguística Computacional, (ex. Miller et al., 2012), que tratam da expansão lexical da máquina e possuem objetivos mais aplicados como, por exemplo, melhorar a desambiguação de palavras e sentenças no processamento de linguagem natural.

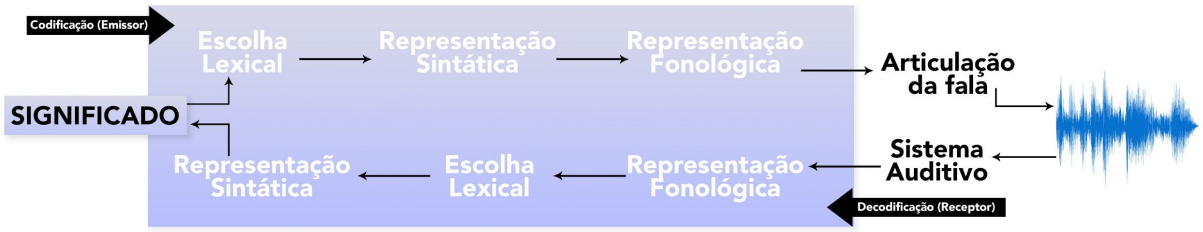
na apresentação de uma pesquisa psicolinguística da expansão lexical por composição morfológica e por empréstimos.

3. Acesso Lexical: Sobre a produção e a compreensão das palavras

Antes de compreender como o léxico é expandido, é importante ter noção dos mecanismos relacionados ao acesso do léxico na mente, tema que possui uma literatura já bem consolidada na psicolinguística.

Quando produzimos uma palavra, é necessário ter a intenção de produzi-la, converter a sua forma semântica em uma forma fonológica e enviar um comando para que os músculos do aparelho fonador se movimentem de maneira adequada e coordenada. Ao compreender essa mesma palavra, o ouvinte deve escutá-la através de seus ouvidos, converter as ondas acústicas em formas fonológicas que façam sentido em sua língua e combiná-las para formar a palavra que nos fará ativar em seu léxico mental o(s) sentido(s) correspondente(s) (para uma revisão sobre percepção da fala, ver Phillips, 2001). O mesmo ocorre ao ler e escrever, bastando alterar a modalidade auditiva para a visual.

Figura 1. Processamento de palavras (inspirada em Fernández, 2010).



Embora essa explicação possa parecer simples, existem diversos níveis e diversas dúvidas sobre o funcionamento dos mecanismos mentais envolvidos na percepção, na produção e no armazenamento de uma palavra, o que fomenta a produtividade desse campo de pesquisa.

Uma das maiores discussões da área diz respeito à organização das palavras na mente. Observe que muitas das palavras que usamos no dia a dia são computacionais, a ponto de que um falante proficiente da língua consegue acessar seu sentido mesmo que nunca tenha sido exposto a ela. Considere o caso da palavra inconstitucionalissimamente. Um falante de português identifica imediatamente que sua raiz é a palavra 'constituição' em (1a), que pode ser adjetivada (1b), que em seguida recebe o prefixo que indica negação do conceito em (1c), até alcançarmos a forma final em (1e).

1. Computação morfológica da palavra ‘inconstitucionalissimamente’
    a. constituição
    b. [[constitucion]-al]]
    c. [in-[constitucion[-al]]]
    d. [in-[constitucion[-al[-issimo]]]]
    e. [in-[constitucion[-al[-issima[-mente]]]]]

Devido a essa característica computacional das línguas, foi proposta a hipótese de que nossa mente funciona também de forma computacional e que, sempre que ouvimos uma palavra semelhante a outra já conhecida, a compreendemos identificando todas as suas partes e juntando-as em uma só. Essa hipótese foi batizada de *Decomposição Plena* (*Full Decomposition* ou *Full Parsing Hypothesis*; Taft, 1979, 2004; Halle & Marantz, 1993; Stockall & Marantz, 2006) e considera que, ao usar linguagem, nossa mente precisa de mais processamento do que de memória (armazenamento), bastando guardar as peças da língua e juntá-las quando preciso. A hipótese imediatamente concorrente foi batizada de *Listagem Plena* (*Full Listing Hypothesis*; Butterworth, 1983; Rumelhart & McCleland, 1986; Bybee, 1995), na qual a mente usaria mais memória para guardar todas as formas das palavras, mas precisaria de menos processamento, uma vez que bastaria escolher, sem realizar sua decomposição.

Ambas as hipóteses são fundadas em evidências experimentais, apesar de terem naturezas praticamente opostas. Isso levou a uma proposta intermediária, a Hipótese da Dupla Rota ou os Modelos Duais (*Double-Route Models*, Caramazza, Laudanna & Romani, 1988; Pinker & Prince, 1992; Schereuder & Baayen, 1995). Estes modelos propõem, por exemplo, que palavras mais frequentes e/ou irregulares geralmente são acessadas de forma direta independente de sua composição.

Outro tipo de questionamento diz respeito à relação entre as palavras. Diversos tipos de experimento buscam compreender os efeitos de facilitação de uma palavra em relação a outras, efeito que chamamos de *priming* (Meyer & Schvaneveldt, 1971) e que também pode ser observado em imagens e outros estímulos sensoriais. Muitos trabalhos experimentais vêm sendo desenvolvidos nos últimos anos para explorar diversos tipos de relações e de medidas. Isel & Bacri (1999) buscam entender relações semânticas e fonológicas, Kroll & Sunderman (2003) buscam pelas relações nas palavras homógrafas entre duas línguas em participantes bilíngues, Maia, Lemle & França (2007) exploram o efeito stroop e monitoram o movimento ocular da leitura dos participantes enquanto França et al., (2009) partem de uma abordagem

com base em medidas eletrofisiológicas (ERP) para analisar eventuais variações no tempo de processamento relacionadas ao tamanho e à complexidade morfológica das palavras.

4. Testes de Decisão Lexical

Uma forma de buscar evidências sobre o processamento de palavras são os testes de Decisão Lexical. Esses experimentos geralmente se baseiam em monitorar o tempo de resposta a determinados tipos de palavras, a depender de suas características.

O teste de Decisão Lexical apresenta, na tela de um computador, uma sequência de letras ao participante. Esses estímulos podem ser palavras reais da língua ou não. As palavras reais podem ser categorizadas a depender do objetivo do teste, sendo a categoria mais comum a faixa de frequência de uso. Já as palavras não reais se dividem entre as não-palavras, que se caracterizam por uma sequência de letras impossível na língua (ex. dtumaech), e as pseudopalavras, ou logatomas, que são sequências não usadas mas que respeitam todas as regras da língua e poderiam, em algum momento, ganhar o status de palavra (ex. notelo). A tarefa dos participantes consiste em responder se a sequência de letras lida é ou não uma palavra da língua, enquanto o computador registra o seu tempo de resposta (RT). Alternativamente, as medidas coletadas podem ser os movimentos oculares dos participantes enquanto leem as palavras ou padrões de ativação cerebral (*f*MRI, PET, M/EEG).

A estrutura morfológica de uma palavra, assim como a sua frequência de uso pelo falante, interfere diretamente na velocidade de ativação. Isso é evidenciado por testes psicolinguísticos baseados em tempo de resposta (Garcia, 2009), em movimentos oculares (Maia & Ribeiro, 2015) ou neurofisiológicos (França et al., 2008). Graças a esses estudos, observou-se que palavras utilizadas com mais frequência são ativadas mais rapidamente do que palavras menos frequentes (ex. laranja e cupuaçu, respectivamente); pseudopalavras, por sua vez, demoram mais tempo para serem reconhecidas como tal do que as palavras da língua. As não palavras costumam ser identificadas rapidamente como impossíveis. Petersen et al., (1990), usando técnicas de imageamento cerebral, demonstram que as mesmas áreas cerebrais são ativadas para as palavras reais e para as pseudopalavras, enquanto as não-palavras ativam uma área diferente do cérebro.

5. Empréstimos com camada morfológica: uma abordagem psicolinguística

Como vimos anteriormente, a expansão lexical pode ocorrer de três formas distintas: (i) composição morfológica, (ii) mudança semântica ou (iii) empréstimos. As palavras

emprestadas podem ser adaptadas para a língua, como em 'blecaute' (*blackout*), ou serem usadas na sua forma original, mantendo inclusive sua pronúncia, independente das regras da língua alvo, como em '*feedback*' ou '*delivery*'. Porém, existem aquelas palavras que chegaram na língua por empréstimo, mas que por alguma razão conseguem receber material morfológico (ex. sufixos) da língua alvo, como '*crop*ado' e '*twitt*aço'. Nosso objetivo neste experimento é iniciar um estudo psicolinguístico deste terceiro caso.

Os maiores desafios aos nossos objetivos foram, primeiramente, a identificação das formas ortográficas originais e adaptadas das palavras emprestadas que aceitam morfemas do português. O segundo foi o controle da frequência dos itens experimentais e de controle.

No que diz respeito às formas da escrita: (i) algumas das palavras levantadas para esse estudo possuem formas facilmente adequáveis ao português sem alteração da forma ortográfica original, como 'bugado' e 'veganismo'; (ii) outras palavras possuem formas adequáveis com pequenas alterações ortográficas, como 'tiquetagem' (de *ticket*), e 'craqueado' (de *to crack*); (iii) um terceiro grupo de palavras possui uma ortografia característica no inglês e precisam de maiores adequações ortográficas para se adaptarem às regras do português, como 'twittaço' ou 'tuitaço' (de *twitter*).

Esse terceiro grupo de palavras também nos traz um novo desafio, pois todas as suas palavras possuem duas formas de escrita, uma que considera a ortografia original (*twitt*aço, *link*ado, *crack*eado) e é bastante difundida em textos mais livres como blogs, posts de redes sociais e e-mails pessoais, enquanto a forma adequada ao português é mais difundida em mídias oficiais e textos jornalísticos (tuitaço, lincado, craqueado).

Neste estudo, fizemos a escolha de utilizar as palavras com suas formas originais concatenadas aos morfemas do português, da forma como são normalmente encontradas em textos espontâneos na internet. Apesar de não termos feito a escolha de controlar as duas formas ortográficas, consideramos que seja de extrema importância uma replicação deste estudo que aplique tal controle.

O segundo desafio diz respeito ao controle de frequência das palavras emprestadas com camada morfológica, uma vez que geralmente são palavras não muito difundidas e de nicho, como 'cropar'[4] (de *to crop*) e 'clipagem[5]' (de *to clip*). Nesse sentido, verificamos a frequência dessas palavras no Google buscando apenas por páginas em português com a ortografia não

[4] Da ferramenta *crop* em softwares de edição de imagens e de vídeos.

[5] Usado no meio jornalístico e significa, basicamente, buscar e republicar notícias, adaptando o texto o suficiente para não ser considerado plágio.

adaptada. Os demais detalhes do controle de frequência são descritos na seção de materiais e métodos.

A partir dessa seleção das palavras emprestadas que aceitam morfema do português brasileiro (PB), realizamos um teste de decisão lexical com 7 condições, a saber:

Quadro 1: Condições experimentais

a. **Palavras do PB** – Palavras simples do PB sem camada morfológica;
b. **Palavras do PB com sufixo** – Palavras do PB com camada morfológica;
c. **Pseudopalavras com sufixo** – Palavras inventadas que obedecem às regras do PB;
d. **Empréstimos** – Palavras emprestadas do Inglês, sem camadas morfológicas;
e. ***Empréstimos c/ sufixo** – Palavras emprestadas do inglês, com camada morfológica do PB;
f. **Palavras do Inglês** – Palavras do inglês que não foram emprestadas para o PB;
g. **Não-palavras** – Sequência de letras que não obedece às regras do PB nem do Inglês.

Como informado anteriormente, a condição principal desse experimento é a condição (*e*), palavras emprestadas que receberam morfologia do português. Essa condição deve ser comparada com a condição controle (*b*), palavras do português do Brasil com a mesma morfologia, para verificar se há diferença no comportamento dos participantes. Para efeitos de comparação, também inserimos a condição (*c*), pseudopalavras com a mesma morfologia.

Além das palavras com camadas morfológicas, também acreditamos ser importante comparar a condição (*e*) com a condição (*d*), palavras emprestadas do inglês que não aceitam morfologia do português, (*f*), composta de palavras do inglês não emprestadas, e (*a*), de palavras do PB sem morfemas, visando o controle do comportamento dos participantes enquanto leem palavras emprestadas e não emprestadas sem morfologia. A condição (*g*), composta de não-palavras, foi inserida apenas para balancear as respostas do experimento.

Nossas previsões consideram que as palavras do português (condições *a* e *b*) são processadas mais facilmente do que as palavras do inglês (condições *d*, *e* e *f*), em situação de empréstimo ou não, por ser a língua dominante de todos os participantes. As palavras emprestadas sem material morfológico (condição *d*) devem ter seu processamento facilitado em relação àquelas que com material morfológico (condição *e*), enquanto as palavras não emprestadas (condição *f*) devem ser mais custosas por não serem correntes na língua dominante. As pseudopalavras com morfema (condição *c*) podem ser consideradas palavras de baixa frequência, porém, a adição de morfemas pode facilitar a aceitação dessas palavras, tornando a resposta mais rápida.

6. Experimento: Teste de Decisão Lexical

6.1 Participantes:

A pesquisa foi aprovada pelo CEP [CAAE: 68163917.9.0000.5404]. Foram recrutados 32 participantes brasileiros, falantes nativos de português do Brasil, estudantes universitários, de ambos os sexos, com idades entre 18 e 28 anos, com nível intermediário da língua inglesa, destros, com visão normal ou corrigida. Foi vetada a participação de estudantes de Linguística e demais áreas dos Estudos da Linguagem, assim como quaisquer pessoas que tivessem conhecimento prévio prático ou teórico sobre o teste ou sobre os temas relacionados ao estudo.

6.2 Materiais:

Como indicado no quadro 1, sete categorias de palavras foram utilizadas nesse estudo: (i) palavras do PB sem morfema, (ii) palavras do PB com morfema, (iii) pseudo-palavras com morfema, (iv) empréstimos do inglês, (v) empréstimos do inglês com camada morfológica do PB, (vi) palavras do inglês não emprestadas e sem camadas morfológicas e (vii) não-palavras. Foram usadas 12 palavras para cada categoria. As não-palavras são usadas em maior número (28) para balancear o número de respostas no experimento. O quadro com as palavras utilizadas pode ser conferido no anexo 1.

6.3 Frequência das palavras:

Na ausência de um guia de frequência de palavras em português, buscamos controlá-la minimamente através de uma pesquisa no Google, realizada somente em páginas brasileiras e em português, no dia 28 de abril de 2017. Um desafio no controle das frequências é o fato de que as palavras de origem inglesa que usam morfologia do português costumam ser palavras de nicho, como gamificar (transformar em game), cropar (ex. cropar uma imagem num *software* de edição de imagem – *to crop*), crackear (ex.: quebrar a segurança eletrônica – *to crack*), setar (configurar uma máquina – *to set*). Uma vez que essas palavras são usadas em contextos muito específicos, a frequência de cada uma é mais baixa. Nesse sentido, nosso controle de frequência está descrito no quadro 2. Repare que (i) as palavras com camadas morfológicas foram buscadas a partir da soma da frequência de sua forma infinitiva (quando tinham) e a forma utilizada e (ii) a diferença de frequência entre elas ainda é bem alta, o que pode representar um viés para a pesquisa.

Quadro 2: Faixa de frequência por condição.

a. **Palavras do PB sem morfema** – 2 a 9 milhões;
b. **Palavras do PB com morfema** – 2 a 7 milhões;
c. **Pseudopalavras com morfema** – Palavras inventadas;
d. **Empréstimos** – 2 a 6 bilhões;
e. ***Empréstimos c/ morfema** – 300 mil a 2 milhões[6];
f. **Palavras do inglês [sem empréstimo]** – 1 a 5 bilhões;
g. **Não-palavras** – Sequência de letras inventadas.

As não-palavras foram inseridas para balancear o número de respostas esperadas SIM/NÃO ('palavra' e 'não-palavra', respectivamente), de forma a não viciar o participante em uma das duas respostas. Por essa razão, existem 16 não-palavras a mais do que nas demais condições. Todas as não-palavras foram elaboradas a partir de palavras de origem eslava adicionando algumas letras a mais, de modo que o número de letras estivesse balanceado em relação às categorias com morfemas.

6.4 Procedimentos:

O experimento foi elaborado no software PsychoPy 2 (v.1.82; Peirce, 2007), aplicado em ambiente Windows 10 e apresentado em um monitor de 15,6", posicionado a aproximadamente 60cm do participante. Uma versão de treinamento com 30 estímulos (palavras e pseudopalavras) era aplicada antes do teste principal para amenizar os efeitos da curva de aprendizagem.

Os participantes eram instruídos a responder se cada sequência de letras apresentada no monitor correspondia ou não a uma palavra do português ou do inglês. A resposta era registrada pelas setas do teclado identificadas com adesivos verde (esquerda; SIM) e vermelho (direita; NÃO), de forma que os participantes respondessem utilizando uma única mão. O limite para resposta era de 4 segundos.

Os estímulos eram apresentados de forma randomizada, na fonte Arial, no tamanho padrão do PsychoPy (0.1), em cor branca sobre um fundo cinza, centralizados na tela e separados por uma cruz de fixação com duração de 1 segundo, também centralizada. Cada participante levou cerca de 5 minutos para concluir o teste.

[6] Para completar o quadro, foi necessário inserir outras duas palavras. A dificuldade em achar essas palavras nessa faixa de frequência nos levou a usar (i) 'veganismo', que possui uma frequência de 3,5 milhões de ocorrências. Para buscar equilíbrio, usamos como contrapeso a palavra (ii) 'cropado', que, sendo ainda mais de nicho, apresenta frequência de 30 mil ocorrências.

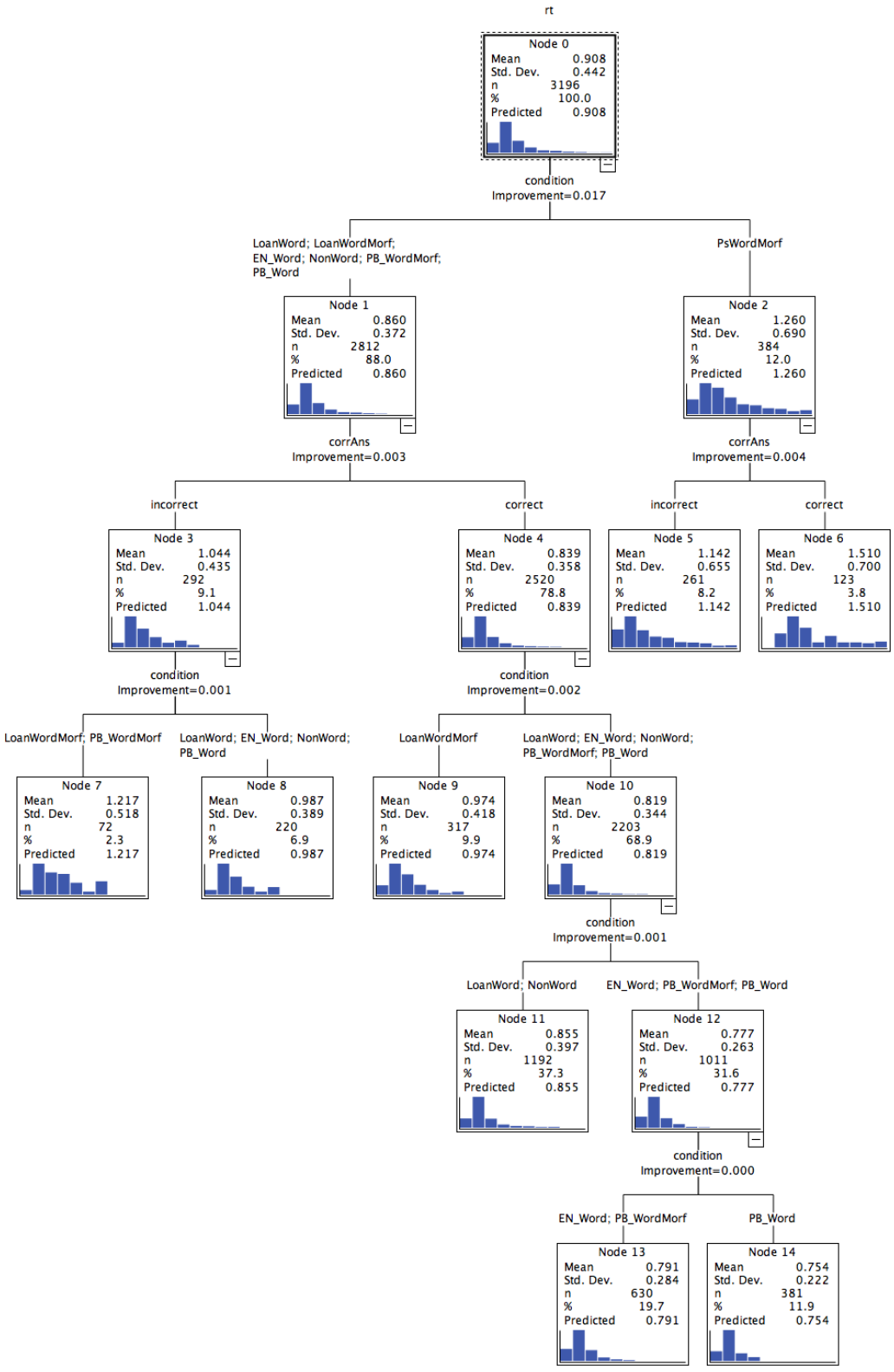
6.5 Resultados

Para a análise estatística dos dados, definimos os *outliers* pela *Outlier Labeling Rule*, utilizando g = 2,2 (Hoaglin, Iglewitz, Turkey, 1986). Os *outliers* foram tratados por winsorização[7] (Wilcox et al., 2003). Esse procedimento modificou aproximadamente 3% dos dados. O teste Shapiro-Wilk indicou que nenhum dos conjuntos de dados não relacionados apresenta distribuição normal nem antes nem após o tratamento com winsorização ($W < .95$ e $p < .000$ em todos os grupos), nos levando a utilizar o *Mann-Whitney* como teste de hipótese.

Para traçar a proximidade entre as condições, aplicamos uma árvore de regressão CRT (Classification and Regression Tree; Breiman et al., 1984), que realiza agrupamentos automáticos binários dos dados conforme a proximidade entre suas variáveis. Foi realizado um agrupamento considerando a curva dos dados e as respostas corretas/incorretas. O resultado pode ser observado na árvore ilustrada na figura 2.

7 No meio termo da discussão sobre cortar os dados que desviam do padrão ou considerá-los como dados relevantes para a análise, a winsorização iguala os outliers aos dados limítrofes (os menores e maiores dados não outliers). Dessa forma, damos peso aos outliers ao mesmo tempo em que mantemos uma base de dados sem dados desviantes.

Figura 2: Árvore de regressão CRT dos RT x Resposta. O gráfico agrupa as condições por proximidade do comportamento de resposta dos participantes.



O algoritmo de classificação automática isolou inicialmente os dados das pseudopalavras dos demais dados iniciais (node 0, 1 e 2), pois as pseudopalavras possuem um maior índice de erros na tarefa de decisão lexical (8.2% do node 0, contra 3.8% de acertos), o que é previsível uma vez que foram criadas para colocar os participantes em dúvida. As demais condições foram, então, divididas entre as respostas corretas e incorretas (node 3 e 4), assim como as pseudopalavras (node 5 e 6).

Das condições com respostas corretas, os empréstimos com morfema foram isolados das demais condições (node 9 e 10), demonstrando que as respostas a essa condição, de fato, possuem um comportamento distinto. As demais condições foram, então, agrupadas em dois conjuntos que indicam que o comportamento dos participantes nos trials em que acertaram ao responder às palavras do inglês foi semelhante ao dos trials em que acertaram as não-palavras (node 11) e um conjunto de palavras do inglês, palavras do PB e palavras do PB com morfema (node 12). Esses agrupamentos indicam que estas foram as condições em que os participantes responderam com maior rapidez e precisão. A árvore de regressão ainda criou uma quinta e última camada na qual isola as palavras do PB (node 14) das palavras do inglês e palavras do PB com morfema (node 13), indicando que a distribuição das respostas corretas de palavras do PB com morfema se assemelha mais às respostas para palavras do inglês do que para as palavras simples do PB.

A tabela 1, abaixo, indica o índice de respostas para cada condição. A figura 3 apresenta a média dos tempos de resposta dos participantes para as palavras de cada condição experimental, independente da resposta. Observe que todos os nossos gráficos possuem limite inferior de 0,5 segundos para melhorar a visualização e contam com barras de erro padrão que, entre outras vantagens, auxiliam a manter uma base de comparação entre os dados independente da escala no eixo y. Observe também que as respostas incorretas não foram descartadas para manter a comparabilidade com as pseudopalavras. Ao comparar a condição de empréstimos com morfema com as demais condições, obtemos sempre uma diferença significativa que pode ser observada na tabela 2.

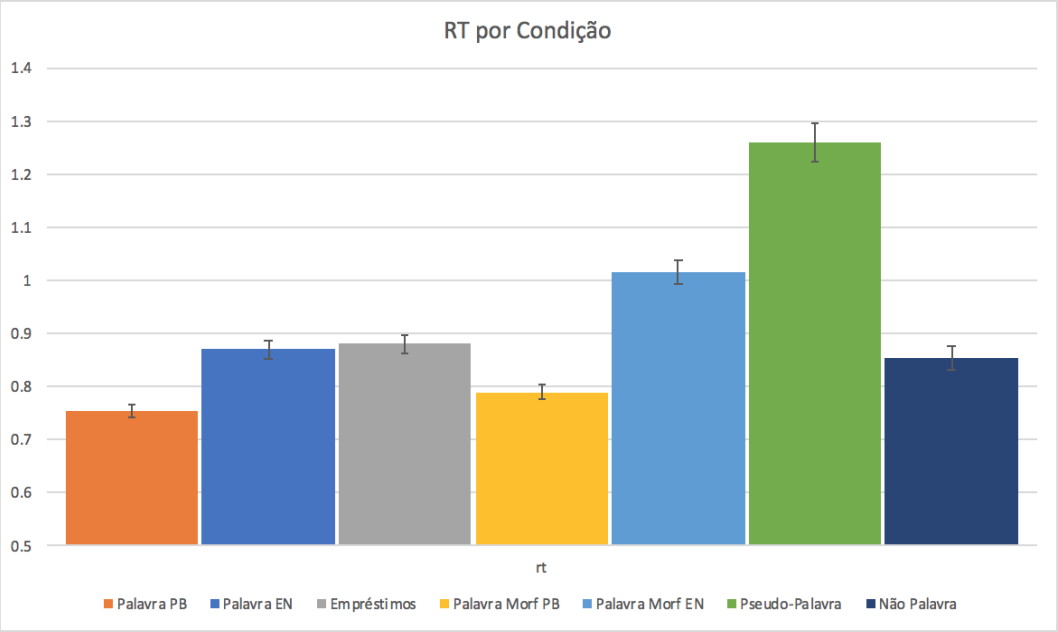
Tabela 1: Número de respostas por condição

Condição / Resposta	**Palavra**	**Não Palavra**
Palavra PB c/ morfema	379	5
Pseudopalavra c/ morfema	261	123
Empréstimo c/ morfema	317	67
Não-Palavra	14	882
Empréstimo	314	70
Palavra inglês	251	133
Palavra PB	381	3

Tabela 2: Teste Mann Whitney para as seguintes condições comparadas com palavras com empréstimo que aceitam morfologia do Português.

Condição	**U / Z**	**p**
Palavra PB	$U = 44.659$; $Z = -5.930$	$p < .000$
Palavra PB c/ morfema	$U = 44.769$; $Z = -5.793$	$p < .000$
Empréstimos	$U = 42.213$; $Z = -3.300$	$p < .001$
Pseudopalavra c/ morfema	$U = 50.632$; $Z = -3.834$	$p < .000$
Palavra Inglês	$U = 31.959$; $Z = -4.028$	$p < .000$

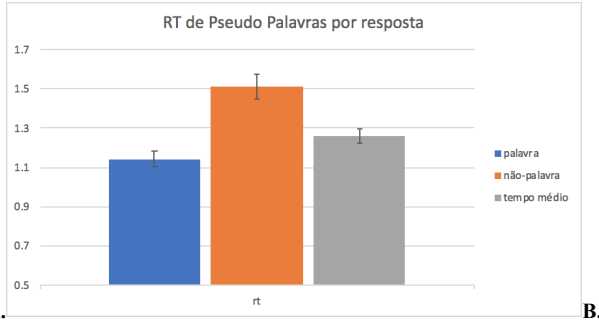
Figura 3: Tempo de resposta médio de todas as condições. Todas apresentam diferenças significativas se comparadas à condição de empréstimos com sufixos do português do Brasil (Palavra Morf EN).



Repare que no caso das pseudopalavras é esperado que os participantes fiquem em dúvida quanto à resposta, o que é refletido no número de respostas [palavra] (261) e [não-palavra] (123). Nesse sentido, acreditamos ser importante separar as respostas corretas das incorretas nessa condição. A figura 4A representa os tempos médios para cada resposta nas pseudopalavras com camadas morfológicas. A tabela 3 indica que há diferença significativa entre os tempos médios comparado com os tempos de cada resposta e, também, entre os tempos de cada uma das respostas nessa condição.

Figura 4: Tempos médios de resposta. **Painel A**: RT [palavra] e [não-palavra] na condição com pseudopalavras com camada morfológica. **Painel B:** RT das pseudopalavras com camada morfológica indicadas como palavras pelos participantes contra os tempos médios de resposta para a condição palavras do PB com camada morfológica.

A.



B.

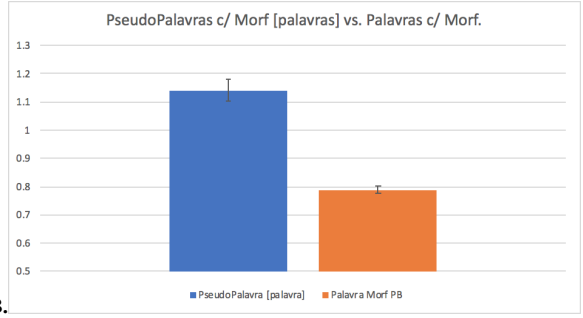


Tabela 3: Teste Mann Whitney entre o tempo médio e os tempos das respostas [palavra] e [não palavra] na condição de pseudopalavras com camada morfológica

Resposta	**U / Z**	**p**
Média vs. Palavra	$U = 18599$; $Z = -3.548$	$p < .000$
Média vs. Não-Palavra	$U = 45095$; $Z = -2.216$	$p = .031$
Palavra vs. Não-palavra	$U = 11035$; $Z = -4.943$	$p < .000$

Nesse ponto nos parece razoável uma comparação entre as palavras com morfema e as pseudopalavras com morfema que foram indicadas como palavras da língua. O teste de hipótese Mann Whitney indica que as palavras ainda apresentam tempos de resposta significativamente menores ($U = 38.108$; $Z = -5.168$; $p < .000$) do que as pseudopalavras, como podemos observar na figura 4A.

Repare também que a condição com pseudopalavras não foi a que apresentou respostas mais divididas, mas sim a condição com palavras do inglês que não sofreram empréstimo (251-133), apesar de os participantes serem instruídos a indicar que esses casos se tratam de palavras. Por este motivo, a mesma análise interna foi realizada também nesta condição.

Tabela 4: Teste Mann Whitney entre o tempo médio e os tempos das respostas [palavra] e [não-palavra] na condição de palavras do inglês não emprestadas ao PB.

Resposta	**U / Z**	**p**
Média vs. Palavra	$U = 21.194$; $Z = -2.934$	$p = .03$
Média vs. Não-Palavra	$U = 44.223$; $Z = -1.756$	$p = .08$
Palavra vs. Não-palavra	$U = 12.475$; $Z = -4.074$	$p < .000$

Figura 5: Tempos médios das respostas. **Painel A:** RT de [palavra] e [não-palavra] na condição com palavras do inglês não emprestadas ao PB. **Painel B:** RT das respostas para as condições com palavras em inglês com e sem empréstimo ao PB.

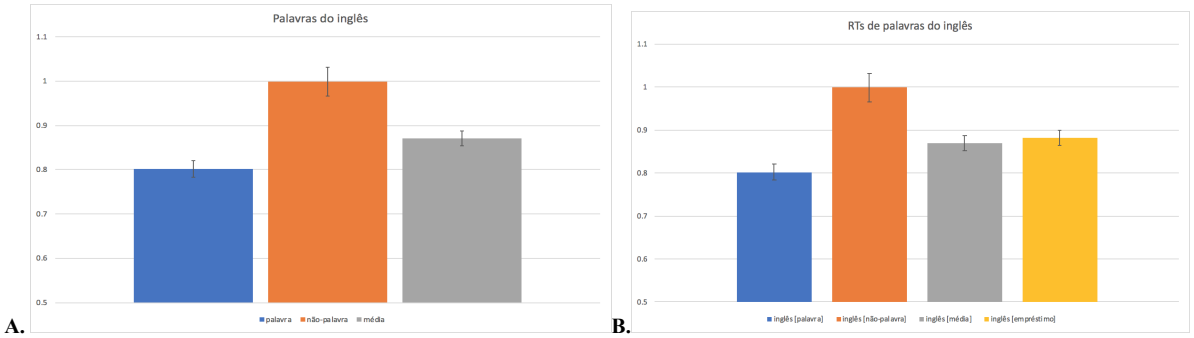


Tabela 5: Teste Mann Whitney entre os tempos de resposta para as condições com palavras sem empréstimo comparadas com a condição com empréstimo ao PB.

Resposta	**U / Z**	**p**
Inglês [palavra]	$U = 43.767$; $Z = -1.958$	$p = .05$
Inglês [não-palavra]	$U = 21.212$; $Z = -2.912$	$p = .004$
Inglês [média]	$U = 73.627$; $Z = -.033$	$p = .974$

Visto que essas palavras emprestadas também são do inglês e também não possuem camadas morfológicas, analisamos as condições com e sem empréstimo, como pode ser observado na figura 5B e na tabela 5. Os resultados indicam que, considerando a média de respostas, não há diferença significativa entre as condições. Porém, quando os participantes respondem corretamente às instruções, ou seja, indicam que se trata de uma palavra, os tempos de resposta são mais rápidos do que nas situações de empréstimo. O contrário acontece quando os participantes indicam que a sequência de letras não é uma palavra, quando os tempos são maiores do que na situação de empréstimo.

6.6 Discussão

Copiamos o quadro 1, abaixo, de forma a facilitar a discussão dos resultados.

Quadro 1: Condições experimentais

a. **Palavras do PB sem morfema** – Palavras simples do PB sem camada morfológica;
Palavras do PB com morfema – Palavras do PB com camada morfológica;
Pseudopalavras com morfema – Palavras inventadas que obedecem às regras do PB;
Empréstimos – Palavras emprestadas do Inglês, sem camadas morfológicas;
***Empréstimos c/ morfema do PB** – Palavras emprestadas do inglês, com camada morfológica do PB;
Palavras não emprestadas – Palavras do inglês que não foram emprestadas para o PB;
Não-palavras – Sequência de letras que não obedece às regras do PB nem do inglês.

Ao realizar um agrupamento automático através da árvore de regressão CRT, considerando a resposta e a distribuição dos tempos de resposta em cada condição, o algoritmo indica que as pseudopalavras têm um comportamento diferenciado devido à distribuição das respostas corretas (32%; 3.8% dos dados totais) e incorretas (68%; 8.2% dos dados totais). No mesmo nível hierárquico, o agrupamento das demais condições também é dividido entre corretas (90%; 78.8% dos dados totais) e incorretas (10%; 9.1% dos dados totais).

As demais condições apresentam agrupamentos por distribuição dos tempos de resposta que isolam, primeiramente, os empréstimos com morfemas do português e, num segundo nível, os empréstimos e as não-palavras, indicando que os participantes tiveram dificuldade semelhante para responder aos empréstimos e às não-palavras. Esses dados estão de acordo com o ilustrado na figura 3. Em um nível abaixo, indicando as condições com maior facilidade de resposta, encontramos dois agrupamentos que dividem as palavras do inglês e as palavras do português com camada morfológica das palavras simples do português. Ao contrário do nível acima, esse agrupamento desmente o ilustrado na figura 3 em que, considerando apenas a média de respostas e sem a eliminação das respostas incorretas, as palavras do inglês (≅.9s) seriam acessadas mais lentamente do que as palavras do português com camadas morfológicas (≅.8s).

Independente da análise, conforme nossa previsão, as palavras simples do português, foram as condições mais rápidas, indicando uma maior facilidade dos participantes em acessar palavras de sua língua nativa e dominante. Contrariando nossa hipótese, a categoria das palavras do português com camada morfológica parece ter uma diferença na distribuição dos dados que a aproxima das palavras simples do inglês, segundo a árvore de regressão.

Como informado anteriormente, nossa condição experimental principal é a condição (e), contendo palavras emprestadas do inglês que se tornaram palavras de nicho e receberam morfologia do português, como cropar, clipar e twittar. Como a observação visual da figura 3

mais os testes de hipótese descritos na tabela 2 apontam, os tempos de resposta para essa condição é significativamente mais lenta do que nas demais condições, indicando que o processo de ativação deste tipo de palavra na memória seja diferente das outras condições. A árvore de regressão também aponta para uma diferença significativa entre esta categoria de palavras e as demais condições. Esses dados poderiam indicar, por exemplo, uma intersecção mais custosa entre dois sistemas de composição de palavras nos participantes falantes nativos de Português do Brasil.

Vale ressaltar a dificuldade de controlar a frequência desses itens, como descrito na seção 4.3. Nesse sentido, não descartamos a possibilidade de que o maior tempo de resposta nessa condição seja relacionado ao fato de essas palavras possuírem maior variação de frequência, serem utilizadas em contextos muito específicos. Outro fator relevante pode ser a ortografia, uma vez que mantivemos essas palavras com sua raiz original, com encontros de letras que não correspondem às regras do Português (ex.: twittaço, linkado, trollagem). Seria importante a aplicação de um novo experimento que controle a variável ortográfica e verifique se a ortografia aportuguesada dessas palavras apresentaria tempos de resposta menores. De todo modo, é importante notar que essas palavras obtiveram um baixo índice de erros (67/384). A palavra 'cropado' foi a menos aceita pelos participantes, correspondendo a 15 (1.33s) dos 67 erros nessa condição, contando com 17 acertos (.98s). Essas respostas são semelhantes às apresentadas pela condição de empréstimos sem morfologia, com 314 acertos e 70 erros.

As palavras do inglês, com ou sem empréstimos não apresentam diferença em seus tempos de resposta. Porém, é importante observar que as palavras que não sofreram empréstimos apresentam respostas bastante divididas. Apesar de os participantes serem instruídos a informarem que as palavras do inglês também são palavras, 133 dos 384 trials obtiveram a resposta contrária [não-palavra]. Dessa forma, achamos importante separar essas condições entre as duas respostas e a média. Essa comparação é apresentada na figura 5 e na tabela 4, indicando uma diferença significativa nos tempos das respostas [palavra] comparadas as respostas [não-palavra] e a média. Entre a resposta [não-palavra] e a média existe uma diferença apenas marginal, o que parece indicar uma tendência dos participantes a considerar palavras, apenas aquelas do Português. De todo modo, é importante observar que os participantes que respondem que as palavras do inglês são palavras o fazem com maior convicção do que aqueles que respondem que não são.

Considerando o comportamento interno da condição com palavras do inglês que não sofreram empréstimo, é importante compará-lo agora com a condição de empréstimo. Vimos anteriormente que, ao comparar as médias das condições com palavras inglesas emprestadas e

não emprestadas, não há diferença significativa. Na figura 6 e na tabela 5, porém, verificamos uma diferença marginal entre as respostas [palavra] e as palavras emprestadas. Isso indica que, apesar de ambas serem consideradas palavras, o fato de algumas já terem entrado no léxico do português facilita ligeiramente a sua ativação mnemônica. Uma diferença significativa é apresentada ao comparar as respostas [não-palavra].

As pseudopalavras com camada morfológica também ficaram bastante divididas entre as respostas [palavra] e [não-palavra]. Dos 384 trials, 261 indicaram que se tratam de palavras. Repetindo o que foi feito com as palavras do inglês, decidimos realizar uma comparação interna à condição, verificando a diferença entre as respostas. Essas diferenças são apresentadas na figura 3 e na tabela 3. Na média, as pseudopalavras apresentam o maior tempo de resposta do experimento, o que indica a dificuldade em processar formas que respeitam as regras fonotáticas e a morfológicas do português, mas são semanticamente opacas. Esse tempo aumenta ainda mais quando os participantes as marcam como [não-palavras], que apresenta uma diferença marginal comparada à média. Isso indica que os participantes que rejeitaram as pseudopalavras o fizeram com muito custo. Já os participantes que aceitaram as pseudopalavras como palavras do português o fazem com maior convicção, apresentando tempos de resposta menores e diferenças significativas comparadas com as rejeições e com a média da condição. Por outro lado, mesmo as pseudopalavras aceitas apresentam tempos de resposta maiores do que as demais condições experimentais.

7. Considerações Finais

Os tempos de resposta não configuram um comportamento estático de cada palavra ou participante. Cada novo input, vale destacar, leva a uma nova ativação e, consequentemente, altera a própria frequência de uso e outros fatores que interferem na facilidade e no tempo de acesso lexical. Isso pode fazer com que palavras como as que originaram nosso estudo, apesar de menos frequentes e de nicho, podem vir a se tornar mais frequentes no futuro com a expansão de seu uso para outros domínios e a consequente expansão e/ou reconfiguração do léxico mental de cada falante. Por essa razão, acreditamos ser importante abordar, também experimentalmente, o fenômeno da expansão lexical.

Fazemos aqui uma primeira tentativa desse tipo de estudo, mas destacamos que o tema precisa ser expandido, precisamos encontrar melhores controles, bem como é necessária a replicação desses e de futuros resultados. Ressaltamos também que os desafios relacionados ao controle do experimento apontam para a necessidade de estudos mais detalhados com

participantes que dominam o Inglês como L2 ou que tenham duas L1, além do maior controle sobre a influência da forma ortográfica (ex. twittar vs. tuitar) e frequência das palavras emprestadas ou mesmo que comparem o tempo de resposta para *input* visual (forma escrita) e fonológico (forma oral) dos empréstimos, evitando a influência de forma entre essas palavras.

Num caminho semelhante, o alto índice de aceitação das pseudopalavras (261 - 123) que se adequam fonotática e morfologicamente à língua parece se configurar em uma pré-condição para que estejamos aptos a adicionar novas palavras ao nosso léxico. Esse fenômeno levanta questões sobre como ocorre a atribuição de sentido a novas palavras e a sua acomodação no dicionário mental dos falantes, o que pretendemos investigar nos próximos trabalhos.

Contribuição

O experimento foi desenhado pelos dois autores, aplicado pelo primeiro autor, analisado e escrito pelos dois autores.

Agradecimentos

Agradecemos aos participantes por cederem seu tempo para esta pesquisa. Esse projeto teve apoio da bolsa FAPESP 18/15718-8 do primeiro autor e do auxílio FAPESP 16/13920-9 do segundo autor.

Disponibilidade dos Dados:

Os dados deste experimento estão disponíveis via Open Science Framework: [https://osf.io/vqwux/?view_only=2e015dfee1cf438fb6bf89337db38b7f]

REFERÊNCIAS

Algeo, J. (1980). Where Do All the New Words Come from?, *American Speech*, v.55, n.4, p. 264-277.

Arnaud, P. J. L. (2013). Word-Formation and Word-Creation: A Data-driven Exploration of Inventiveness in Neologisms. *Theoretical and Empirical Advances in Word Formation*, 18, p. 97–113.

Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R. A. (1984). Classification and Regression Trees. CRC Press, Boca Raton.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and cognitive processes*, v. 10, p. 425-455.

Butterworth, B. (1983). Lexical Representation. In: Butterworth, B. (Ed.). *Language production*. London: Academic Press, v. 2, p. 257-294.

Caramazza, A.; Laudanna, A.; Romani, C. (1988). Lexical access and inflectional morphology. *Cognition*, 28, p. 297–332.

França, A. I.; Lemle, M; Gesualdi, A.; Cagy, M.; Infantosi, A. F. C. (2008). A neurofisiologia do acesso lexical: palavras em português. *Veredas, Revista do Programa de Pós-Graduação em Lingüística da Universidade Federal Juiz de Fora*, Rio de Janeiro, v. 12, n. 2, julho-dezembro.

Garcia, D. C. (2009). Elementos estruturais no acesso lexical: o reconhecimento de palavras multimorfêmicas no português brasileiro. 108 f. Dissertação (Mestrado no Programa de Pós-Graduação em Linguística) – Faculdade de Letras, Universidade Federal do Rio de Janeiro.

Halle, M.; Marantz, A. (1993). Distributed morphology and the pieces of inflection. In: Hale, K. L.; Keyser, S. J.; Bromberger, S. (Eds.). *The view from Building 20: Essays in linguistics in honor of Sylvain Bromberger*. Massachusetts: MIT Press.

Hoaglin, D. C.; Iglewicz, B.; Turkey, J. W. (1986). Performance of some resistant rules for outlier labeling, *Journal of American Statistical Association*, n. 81, p. 991-999.

Kriauciuniene, R.; Sangailaité, V. (2016). An inquiry into the processes of lexical expansion in current English, *Verbum*, v. 7.

Kroll, J. F.; Sunderman, G. (2003). Cognitive Processes in Second Language Learners and Bilinguals: The development of Lexical and Conceptual Representations, In: Doughty C.J. & Long, M.H. *Handbook of Second Language Acquisition*, 1st Edition, Wiley-Blackwell.

Maia, M.; Ribeiro, J. (2015). Jabuticaba Liboramina lê mais fácil do que Jornaleiro Norbalense: um estudo de rastreamento ocular de palavras e pseudopalavras mono e polimorfêmicas. In: Buchweitz, Augusto; Mota, Mailce Borges (Org.). *Linguagem e Cognição: processamento, aquisição e cérebro.* 1ed. Porto Alegre: EDIPUC-RS.

Maia, M.; Lemle, M.; França, A. I. (2007). Efeito Stroop e rastreamento ocular no processamento de palavras. *Ciências & cognição*, v. 4, p. 2-17.

Meyer, D. E.; Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, v. 90, n.2, p.227-234.

Miller, T.; Biemann, C.; Zesch, T.; Gurevych, I. (2012). Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. Proceedings of COLING 2012, *Technical Papers*, p. 1781-1796. Mumbai.

Phillips, C. (2001). Levels of Representation in the electrophysiology of Speech Perception. *Cognitive Science*, 25, p.711-731.

Peirce, J. W. (2007) PsychoPy - Psychophysics software in Python. J Neurosci Methods, 162(1-2):8-13

Petersen, S. E.; Fox, P. T.; Snyder, A. Z.; Raichie, M.E. (1990). Activation of extrastriate and frontal cortical areas by visual words and word-like stimuli. *Science*, v. 249, p.1041–1044.

Pinker, S. (1999). Words and rules: the Ingredients of Language. New York: Basic Books.

Pinker, S. (2004). Morphological decomposition and the reverse base frequency effect. *Quarterly journal of experimental Psychology*, v. 57, p. 745-765.

Pinker, S.; Prince, A. (1992). Regular and irregular morphology and the psychological status of rules of grammar, in: Sutton, L.A.; Johnson, C.; Shields, R. (Eds.). *ANNUAL MEETING OF THE BERKELEY LINGUISTICS SOCIETY*. Proc. 17th.. Berkeley, CA.: Berkeley Linguistics Society.

Rumelhart, D. E.; McClelland, J. L. (1986). On learning the past tenses of English verbs: implicit rules or parallel distributed processing? In: Rumelhart, D. E.; McClelland J. L.; PDP Research Group (Eds.). *Parallel distributed processing: explorations in the microstructure of cognition*. Cambridge, MA: MIT Press, v. 2, p. 216–271.

Schreuder, R.; Baayen, R. H. (1995). Modeling morphological processing. In: Feldman, L. B. (Ed.). *Morphological aspects of language processing*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.. p. 131-154.

Stockall, L.; Marantz, A. (2006). A single route, full decomposition model of morphological complexity: MEG evidence. *The Mental Lexicon*, v. 1, n. 1, p. 85-123.

Sweetzer, E. (1988). Grammaticalization and semantic bleaching. *Berkeley Linguistics Society*, Berkeley, n. 14, p. 389-405.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory and cognition*, v. 7, p. 263- 272.

Wilcox, R. R.; Keselman, H. J. (2003). Modern Robust Data Analysis Methods: Measures of Central Tendency. *Psychological Methods*, v. 8, n. 3, p.254-274.

ANEXO 1

Palavras usadas no experimento. As não palavras foram elaboradas em maior número para balancear o número de respostas corretas e incorretas. Tabela disponível via Open Science Framework: [https://osf.io/vqwux/?view_only=2e015dfee1cf438fb6bf89337db38b7f]

N	Palavra Morf	Pseudopalavras Morf	Empréstimos Morf	Não Palavras	Empréstimos	Palavras Inglês	Palavras PB
1	Bailado	Ampado	Linkado	Ogrodan	Marketing	Bird	Açougue
2	Canoagem	Litonagem	Ticketagem	Rynekien	Network	Movie	Colírio
3	Exorcismo	Arelismo	Fashionismo	Pieknakn	Tablet	Food	Vassoura
4	Ostracismo	Aplacismo	Veganismo	Soboeki	Background	Sugar	Beterraba
5	Ferragem	Cairagem	Clipagem	Vrlieceb	Feedback	Screen	Flanela
6	Marcado	Briscado	Cropado	Fjaekome	Business	Table	Seringa
7	Golaço	Reslaço	Twittaço	Vlcekom	Fitness	Paper	Colóquio
8	Rasgado	Taspado	Bugado	Dzienan	Software	Mind	Fusível
9	Penteado	Amorsado	Crackeado	Pdomnet	Password	Head	Maçaneta
10	Bagagem	Percagem	Trollagem	Zdrowiem	Shopping	Train	Alfazema
11	Cartilagem	Poteragem	Pixelagem	Dtumaech	Wireless	Bottle	Mandioca
12	Ancoragem	Caromagem	Remixagem	Wroclemil	Delivery	Square	Alfinete
13				Stypendiach			
14				Zamieszczane			
15				Ogloszenia			
16				Znajdziesz			
17				Wydarzen			
18				Wygodzie			
19				Dobrze			
20				Wszedzie			
21				Czasem			
22				Wszelkie			
23				Znalek			
24				Kwestie			
25				Ktorzy			
26				Myslimy			
27				Rozwijania			
28				Zawodowe			