Avaliação do conhecimento lexical: dados experimentais e hipóteses de análise

Alina Villalva*
http://orcid.org/0000-0002-7798-5034

Carina Pinto**
http://orcid.org/0000-0003-1904-5922

Como citar este artigo: VILLALVA, A.; PINTO, C. Avaliação do conhecimento lexical: dados experimentais e hipóteses de análise. *Todas as Letras – Revista de Língua e Literatura*, São Paulo, v. 22, n. 1, p. 1-14, jan./abr. 2020. DOI 10.5935/1980-6914/eLETDO2012748

Submissão: setembro de 2019. **Aceite:** fevereiro de 2020.

Resumo: Embora existam hipóteses sobre a estrutura interna do léxico e o seu estatuto em modelos de gramática, nenhuma abordagem teórica se pronuncia sobre quais são as palavras que os falantes realmente conhecem ou utilizam. Neste texto, revelamos os resultados de um teste *off-line* realizado com o objetivo de medir o conhecimento individual das palavras. Esses resultados mostram que o conhecimento das palavras é independente do seu comprimento ou dos valores de frequência. Essa é uma conclusão muito interessante, pois permite definir uma nova variável para testes *on-line*, adaptada para cada teste específico.

Palavras-chave: Acesso lexical. Conhecimento lexical. Frequência. Comprimento da palavra. Morfologia.

* Universidade de Lisboa, Lisboa, Portugal. *E-mail*: alinavillalva@campus.ul.pt

** Escola Superior de Saúde do Politécnico de Leiria, Leiria, Portugal. *E-mail*: pintocarinaal@gmail.com

Introdução

Diversos quadros de análise linguística consideram o léxico um lugar de idiossincrasias, no qual a matéria-prima de cada língua pode ser encontrada, seja na forma de palavras, de radicais e afixos, ou mesmo de expressões lexicalizadas. Embora existam algumas hipóteses sobre a estrutura interna do léxico e o seu estatuto em modelos de gramática, e embora a representação das especificações lexicais seja ocasionalmente ensaiada, nenhuma abordagem teórica parece estar disposta a assumir o risco de especular sobre quais são as palavras que os falantes realmente conhecem ou utilizam.

Diversos estudos experimentais que envolvem o processamento de palavras afirmam que a frequência de ocorrência de uma palavra numa determinada língua ou a sua dimensão são critérios relevantes no acesso lexical (por exemplo, CLAHSEN; NEUBAUER, 2010; FORD; DAVIS; MARSLEN-WILSON, 2010), mas essas afirmações carecem de evidências convincentes, particularmente no que diz respeito ao português europeu.

A frequência de ocorrência de uma palavra é medida no âmbito de um determinado *corpus* ou de diversos *corpora*. Os *corpora* textuais permitem emular o léxico de uma língua, mas a distância entre o conjunto das palavras que integram qualquer *corpus* e as palavras do léxico individual de cada sujeito pode ser enorme. Os *corpora* textuais são, de fato, influenciados pelas suas próprias características definidoras. Na constituição de um *corpus* várias escolhas são feitas, como, por exemplo, entre registro escrito, oral ou misto; escolha de tipo ou tipos de texto (por exemplo, literário, administrativo, jornalístico etc.); entre discurso infantil, adolescente ou adulto; entre discurso contemporâneo ou recolhas em intervalos de tempo passados; entre recolhas em variedades prestigiadas do uso da língua ou em variedades, dialetais ou socioletais menos prestigiadas. Os *corpora* textuais são ferramentas muito poderosas, mas eles permitem acesso apenas a um subconjunto do uso real de determinada língua. Eles abrem uma janela sobre o léxico da língua na sua dimensão coletiva e patrimonial, mas nunca na sua manifestação protagonizada por cada falante. Por esse conjunto de razões, é importante ter em mente que as conclusões baseadas na frequência lexical extraídas do *Corpus* X podem diferir das conclusões extraídas a partir do *Corpus* Y. Ainda mais importante é que elas certamente diferirão de conclusões que possam ser tiradas a partir da análise do léxico dos falantes, se ou quando essa avaliação do léxico mental estiver ao nosso alcance.

Podemos, pois, suspeitar da relevância dos valores de frequência tão sistematicamente utilizados no trabalho experimental, e também podemos admitir a hipótese de que a centralidade da frequência nesse tipo de trabalho pode estar na origem de algumas das distorções que são frequentemente visíveis nos resultados absurdos de muitos experimentos – a investigação sobre processamento lexical é baseada em respostas individuais, enquanto os valores de frequência dependem de conjuntos enviesados de dados linguísticos. A avaliação da frequência de *corpora* textuais, que estão em constante crescimento e sofisticação, tornar-se-á progressivamente mais confiável, mas ainda não chegamos a esse estágio.

Quanto à dimensão das palavras, é necessário considerar duas unidades de medida: o número de grafemas, tomando a grafia como referência central, quer

DOI 10.5935/1980-6914/eLETDO2012748

a observação diga respeito a processos de leitura e escrita, quer não[1], e o número de sílabas. Sabe-se que nos processos de leitura, dependendo da proficiência do leitor, os sujeitos têm janelas de percepção que abrangem de três a quatro grafemas à esquerda do ponto de fixação a 14-15 grafemas à direita desse mesmo ponto de fixação (cf. RAYNER, 1998, 2009). Como a dimensão das palavras, pelo menos em línguas como o português, é geralmente menor do que essa janela de leitura (entre 17 e 19 grafemas), a relevância do tamanho da palavra para o processamento morfológico, sendo o tamanho medido em número de grafemas, precisa ser demonstrada, tendo em conta que os testes realizados se centram habitualmente na leitura de palavras isoladas.

O mesmo se pode dizer relativamente à medida do número de sílabas, mas essas duas medidas não produzem os mesmos resultados. Dadas as características da ortografia do português, com a existência de sequências consonânticas como <lh> ou <rr>, ou o registro das vogais nasais (cf. <am>, <en>), verifica-se que existem distorções entre essas duas medidas que não são negligenciáveis. Os exemplos seguintes mostram que a medida duas sílabas comporta quatro medidas em número de grafemas (quatro a sete) e, concomitantemente, que o mesmo número de grafemas se pode repartir por diferentes classes de número de sílabas:

(1) *doar* 2 sílabas: 4 grafemas
dotar 2 sílabas: 5 grafemas
formar 2 sílabas: 6 grafemas
plantar 2 sílabas: 7 grafemas

comparar 3 sílabas; 8 grafemas
atribuir 4 sílabas; 8 grafemas

A medida relativa ao número de sílabas é frequentemente usada no domínio dos estudos sobre o processamento fonológico (cf. HUDSON; BERGMAN, 1985; BERTRAM; HYONA, 2002; NEW *et al.*, 2006), mas é difícil encontrar, na literatura, evidências de que essa medida possa ser uma variável útil para a observação do processamento morfológico. Torna-se, assim, necessário observar se existem contrastes produtivos na avaliação do conhecimento das palavras em função do seu tamanho, medido em número de sílabas.

Em face dessas considerações, será possível conceber algum outro critério que possa ocupar o lugar central até agora atribuído à frequência ou à dimensão da palavra, no que diz respeito a estudos sobre processamento morfológico? Em pesquisas anteriores (cf. PINTO, 2017; VILLALVA; PINTO, 2018), sugerimos que o conhecimento individual de palavras pode ser relevante para o processamento visual de palavras. Neste artigo, relatamos uma primeira tentativa de definir esse conhecimento individual como uma variável que permita refinar a seleção de um *corpus* linguístico para pesquisa em estudos sobre processamento de palavras.

1 A consideração do número de segmentos sonoros (fonológicos ou fonéticos) introduz um maior grau de complexidade, dado que é necessário estipular um tipo de débito (fala pausada, normal, rápida), a escolha de uma dada realização fonética, de natureza dialetal, ou escolhas teóricas sobre a natureza das representações fonológicas. Por essa razão, damos preferência à consideração do número de grafemas.

DOI 10.5935/1980-6914/eLETDO2012748

Metodologia

A hipótese que vamos apresentar baseia-se na análise de resultados de um trabalho experimental. Para essa experiência, projetamos um teste *off-line* relacionado com a interpretação dos nomes deverbais em português europeu, que assenta em dois pressupostos: o processamento de palavras derivadas não está relacionado com a frequência da sua ocorrência na língua; o conhecimento do significado da palavra base e a composicionalidade da palavra derivada são critérios predominantes para o processamento desse tipo de palavras.

Para a construção dessa experiência, começamos por realizar e catalogar um *corpus* que compreende 152 substantivos de ação deverbais composicionais (cf. *utilização*) que contêm um tema verbal (cf. *utiliza*) e o sufixo *-ção*. Esse conjunto de dados foi codificado de acordo com o número de sílabas da base (duas a seis sílabas, correspondente a três a sete sílabas nos derivados)[2], com a frequência de ocorrência do lexema e a estrutura morfológica do verbo. Considerando que os valores de frequência não são tão facilmente codificáveis, selecionamos um subconjunto de 96 palavras que se distribuem de forma sistemática por três faixas: baixa, média e alta frequência[3]. Em seguida, excluímos as palavras que se comportam de maneira semelhante. Assim, cada grupo final inclui 20 itens[4]:

Quadro 1 – Número de palavras por número de sílabas.

Número de sílabas da base	Seleção inicial	Seleção final
2-síl	25 itens	20 itens
3-síl	28 itens	20 itens
4-síl	45 itens	20 itens
5-síl	38 itens	20 itens
6-síl	16 itens	16 itens
Total	152 itens	96 itens

Fonte: Elaborado pelas autoras.

O teste *off-line* teve como objetivo medir a familiaridade dos sujeitos com o conjunto de derivados deverbais e dos seus verbos-base. Foi solicitado aos sujeitos que respondessem a uma pergunta (Xção significa o ato de X?). Em seguida, era-lhes pedido que executassem uma tarefa de produção:

2 Decidimos controlar a dimensão das palavras a partir do número de sílabas porque estabelece um menor número de categorias (uma sílaba a seis sílabas).

3 Tomando o CRPC como base de trabalho, consideramos as seguintes faixas de valores de ocorrência dos lexemas:
palavras de baixa frequência 0 - 100
palavras de média frequência 100 - 1.000
palavras de alta frequência > 1.000

4 Verbos com seis sílabas são escassos. Pelo que apenas encontramos 16 pares adequados de derivados com sete sílabas.

DOI 10.5935/1980-6914/eLETDO2012748

a. Os informantes que respondessem "sim" teriam de usar o verbo-base (e.g. *utilizar*) numa frase, o que nos permitiria avaliar tanto o seu conhecimento do verbo-base como o do derivado composicional.
b. Os informantes que respondessem "não" eram convidados a escrever uma frase com o derivado (e.g. *utilização*). Essa frase permitiria que avaliássemos o seu conhecimento dos derivados lexicalizados.
c. Os informantes que respondessem "não sei" eram então convidados a explicar o significado do verbo, o que nos permitiria estabelecer:
   i. se os sujeitos conheciam o significado do verbo, embora não conhecessem o significado do derivado;
   ii. se eles não conheciam o significado do verbo e, portanto, não poderiam conhecer o significado do derivado.

Para realizar o teste *off-line*, usamos os formulários *on-line* da plataforma Google. Todos os informantes estavam fisicamente presentes na mesma sala, em simultâneo, e cada processo de teste individual foi concluído em cerca de 90 minutos.

Amostra

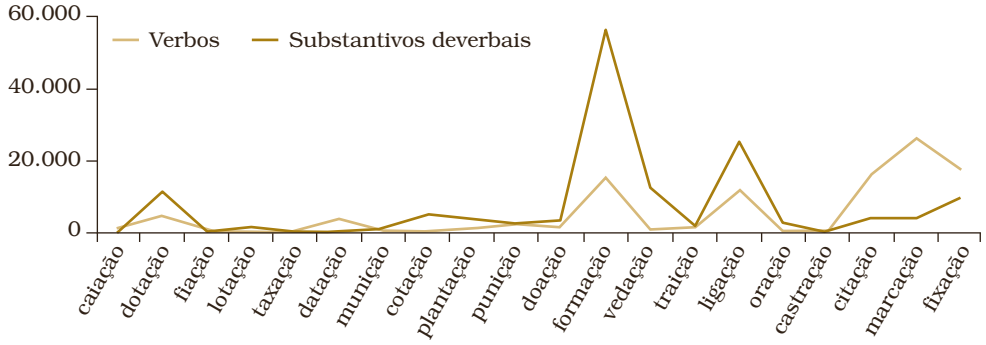
A amostra foi composta por 51 falantes nativos de português europeu, estudantes universitários, cuja média de idade era igual a 21,74 ± 5,1 anos. Todos os informantes tinham visão normal ou corrigida e não apresentavam qualquer patologia de linguagem. Essas informações foram obtidas previamente à realização do teste, através de perguntas presentes no questionário realizado na plataforma Google. Antes do início do teste, foi pedido a todos os sujeitos o consentimento esclarecido e informado, também por via web, por meio da aceitação na colaboração no estudo.

No total, obtivemos entre 37 e 51 respostas para cada item testado. Esse número varia consoante o item, uma vez que alguns sujeitos desistiram da tarefa, o que era permitido, tal como anunciado no consentimento.

Interação entre o comprimento e a frequência das palavras

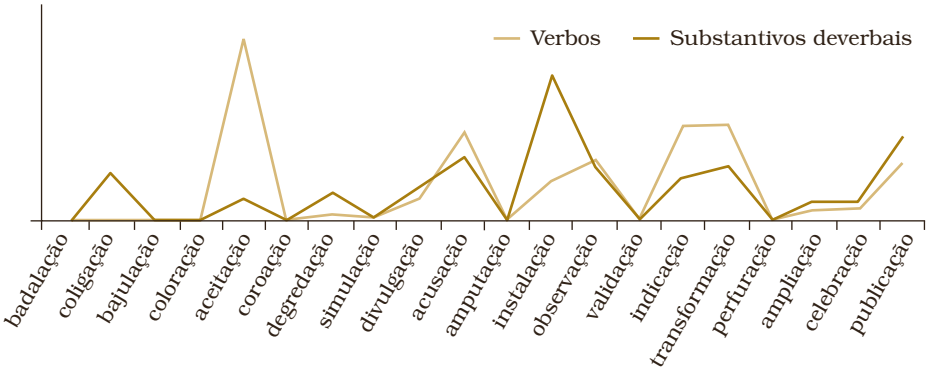
Antes de olhar para os resultados do teste *off-line*, analisamos a interação entre o comprimento da palavra (número de sílabas) e a sua frequência. Os gráficos que se seguem (1-5) mostram que, embora as palavras mais longas apresentem valores de frequência mais baixos, todas as categorias de número de sílabas incluem palavras de alta frequência e palavras de baixa frequência:

Gráfico 1 – Frequência dos verbos de duas sílabas e dos seus derivados



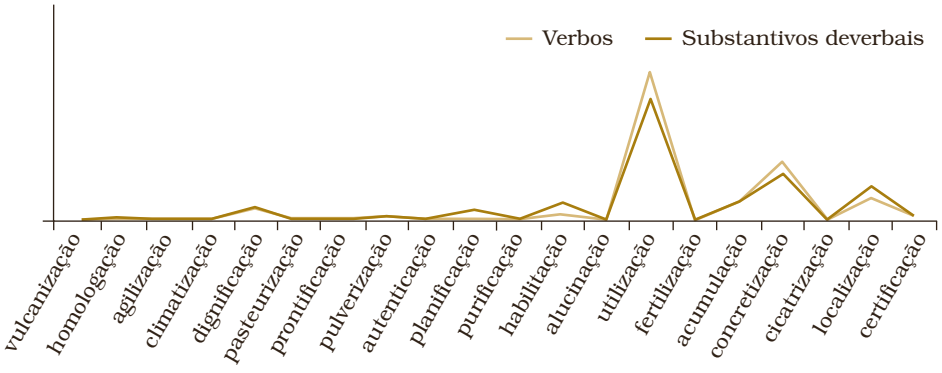
Fonte: Elaborado pelas autoras.

Gráfico 2 – Frequência dos verbos de três sílabas e dos seus derivados



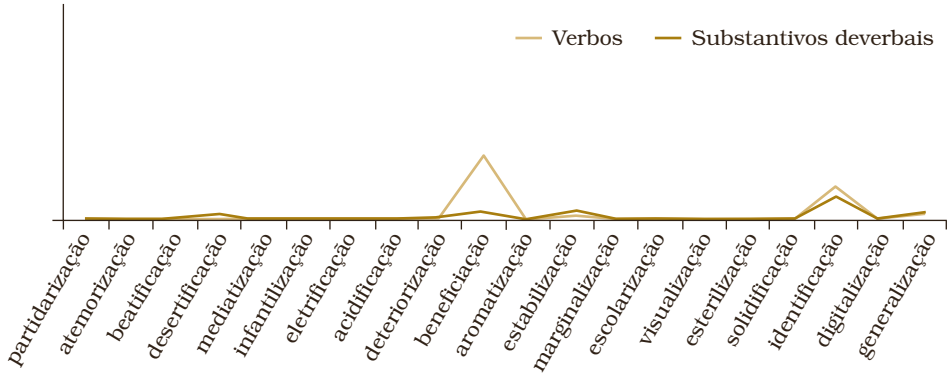
Fonte: Elaborado pelas autoras.

Gráfico 3 – Frequência dos verbos de quatro sílabas e dos seus derivados



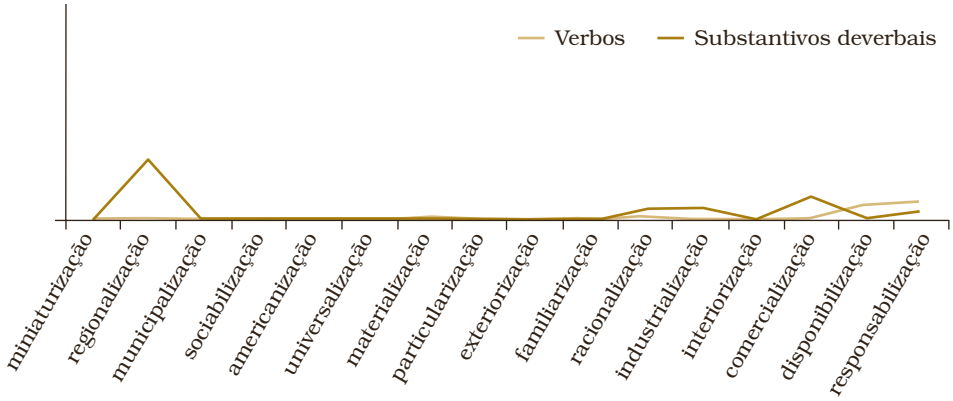
Fonte: Elaborado pelas autoras.

Gráfico 4 – Frequência dos verbos de cinco sílabas e dos seus derivados



Fonte: Elaborado pelas autoras.

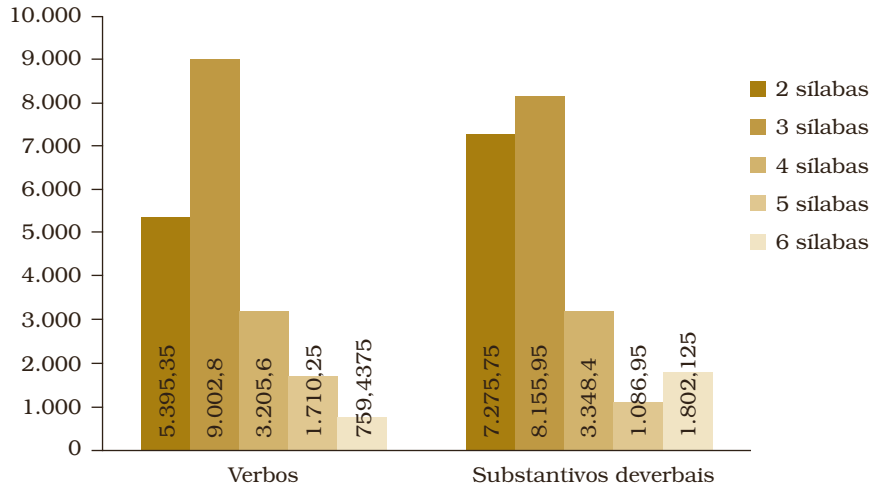
Gráfico 5 – Frequência dos verbos de seis sílabas e dos seus derivados



Fonte: Elaborado pelas autoras.

O próximo gráfico mostra a relação global entre frequência média por categoria de número de sílabas, nos verbos e nos substantivos derivados deverbais. É interessante verificar que os verbos de três sílabas exibem a frequência mais alta, e o mesmo se aplica aos derivados de verbos de três sílabas. Nota-se ainda que a frequência média das palavras tende a diminuir na inversa proporção do aumento do seu comprimento – os resultados são muito consistentes para os verbos e ligeiramente menos consistentes para os derivados, no que diz respeito aos substantivos derivados de verbos com seis sílabas.

Gráfico 6 – Média da frequência dos verbos e dos substantivos deverbais



Fonte: Elaborado pelas autoras.

Por último, o Quadro 2 lista um subconjunto de verbos-base e de substantivos derivados deverbais com a frequência mais alta e a mais baixa, por número de sílabas. Esses exemplos mostram que há pares de verbo-base e derivado deverbal que têm um comportamento coincidente (cf. *utilizar* versus *utilização*; *validar* versus *validação*) e pares que têm um comportamento distinto (cf. *citar* versus *ligação*; *orar* versus *datação*). Com efeito, os dados analisados permitem concluir que a frequência do verbo-base e a frequência do substantivo derivado deverbal são independentemente estabelecidas.

Quadro 2 – Verbos e derivados com alta e baixa frequência

Número de sílabas da base	Verbos com frequência alta	Derivados com frequência alta	Verbos com frequência baixa	Derivados com frequência baixa
2-síl	*citar*	*ligação*	*orar*	*datação*
3-síl	*aceitar*	*publicação*	*validar*	*validação*
4-síl	*utilizar*	*utilização*	*alucinar*	*agilização*
5-síl	*identificar*	*identificação*	*digitalizar*	*eletrificação*
6-síl	*disponibilizar*	*regionalização*	*americanizar*	*miniaturização*

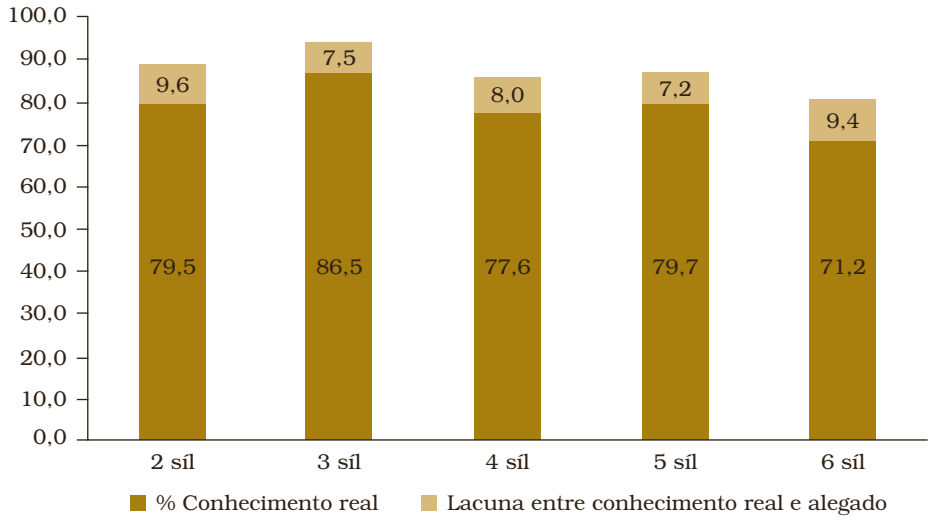
Fonte: Elaborado pelas autoras.

DOI 10.5935/1980-6914/eLETDO2012748

Resultados e discussão

O teste *off-line* foi objeto de uma análise qualitativa para podermos determinar se a resposta dos sujeitos à pergunta inicial que diz respeito ao conhecimento dos derivados deverbais, classificada como conhecimento "alegado", é consistente com o conhecimento "real" das palavras, demonstrado pelo seu uso em frases, que os informantes eram convidados a construir. O Gráfico 7 mostra a existência de alguma inconsistência entre o conhecimento "alegado" e o conhecimento "real", que oscilará entre os 7% e os 10%. Por outras palavras, as respostas dos falantes apresentam um grau de fiabilidade próximo dos 90%. Os dados permitem concluir ainda que a margem de erro é menor em palavras de tamanho médio, ou seja, há menor discrepância nas palavras derivadas de verbos com três, quatro ou cinco sílabas do que nas palavras derivadas de verbos com duas ou seis sílabas.

Gráfico 7 – Alegado conhecimento e real conhecimento dos derivados por número de sílabas

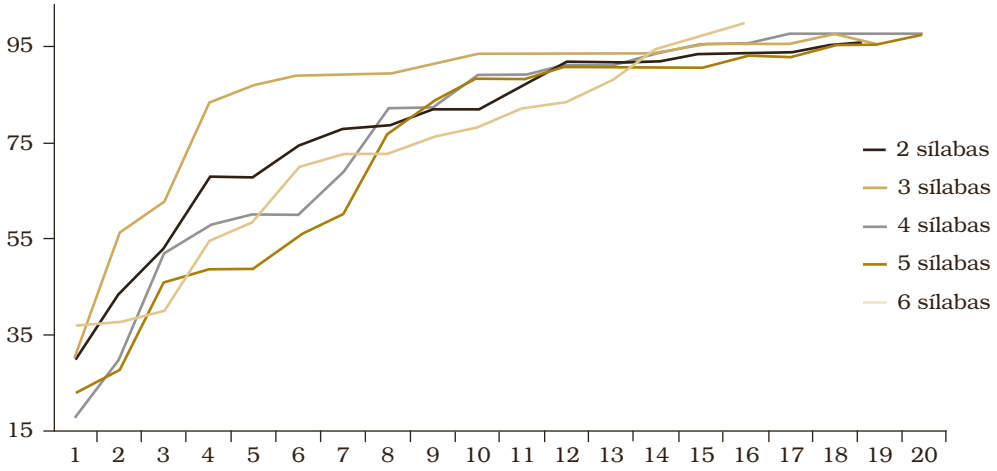


Fonte: Elaborado pelas autoras.

Tendo em conta esses resultados preliminares, decidimos que a análise posterior deveria considerar apenas os resultados do conhecimento "real".

O Gráfico 8 mostra que os subgrupos definidos pelo número de sílabas se comportam consistentemente em relação ao conhecimento de palavras. De fato, todos os subgrupos incluem palavras que se distribuem por todo o espectro de resultados (bons, médios e maus). Podemos, então, concluir que o conhecimento das palavras não está relacionado com o seu comprimento:

Gráfico 8 – Conhecimento das palavras por número de sílabas da base



Fonte: Elaborado pelas autoras.

O Quadro 3 mostra-nos exemplos dos resultados apresentados no Gráfico 8: *marcação* é uma palavra curta que os falantes conhecem bem; *responsabilização* é uma palavra longa igualmente bem conhecida. Pelo contrário, *coligação* é uma palavra que os falantes conhecem menos, tal como *atemorização*, independentemente de a primeira ser uma palavra de menor dimensão do que a segunda.

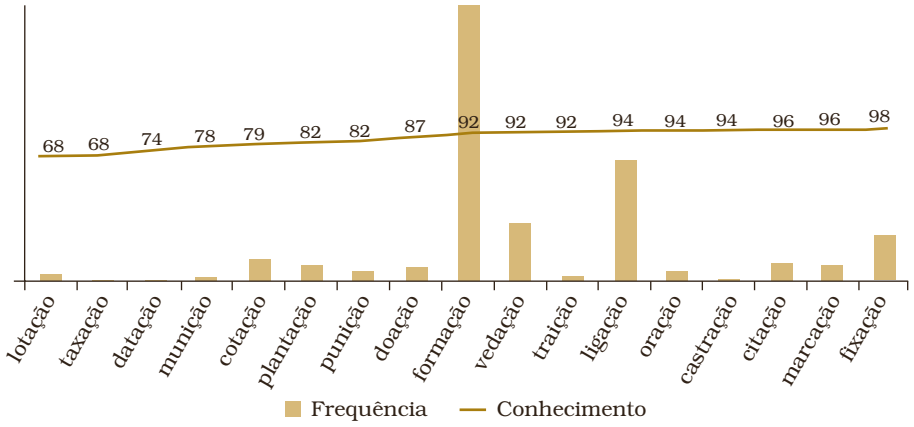
Quadro 3 – Nível de conhecimento das palavras por número de sílabas.

Número de sílabas da base	Conhecimento bom	Conhecimento médio	Conhecimento ruim
2-síl	*marcação*	*datação*	*fiação*
3-síl	*aceitação*	*bajulação*	*coligação*
4-síl	*concretização*	*prontificação*	*agilização*
5-síl	*identificação*	*acidificação*	*atemorização*
6-síl	*responsabilização*	*universalização*	*regionalização*

Fonte: Elaborado pelas autoras.

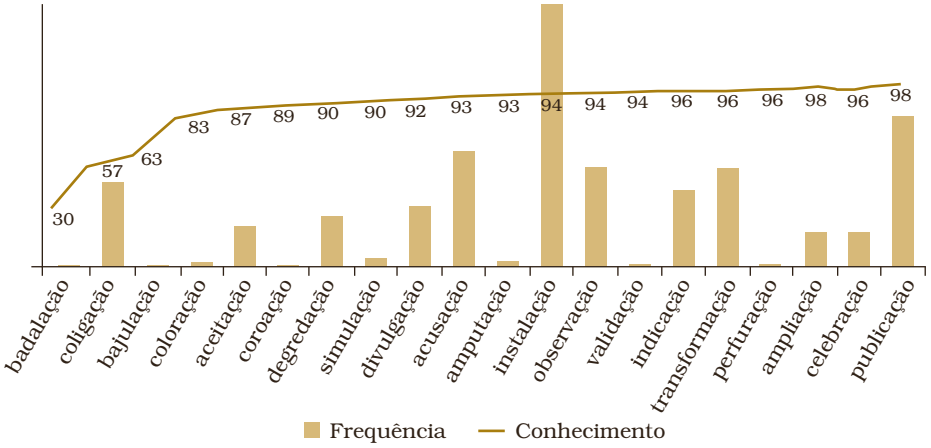
Os gráficos de 9 a 13 apresentam os resultados que correlacionam os valores de frequência das palavras (a partir do CRPC) com os resultados do teste *off-line* sobre o conhecimento das palavras (em percentagem). Cada um desses gráficos mostra que palavras que os sujeitos conhecem mal tanto podem ser muito frequentes (cf. *regionalização*) como raras (cf. *caiação*). Inversamente, as palavras que os sujeitos conhecem bem também podem ser muito frequentes (cf. *identificação*) ou raras (cf. *perfuração*). Consequentemente, também podemos afirmar que o conhecimento das palavras não está relacionado com a sua frequência.

Gráfico 9 – Frequência *versus* conhecimento real (base duas sílabas, derivado três sílabas)



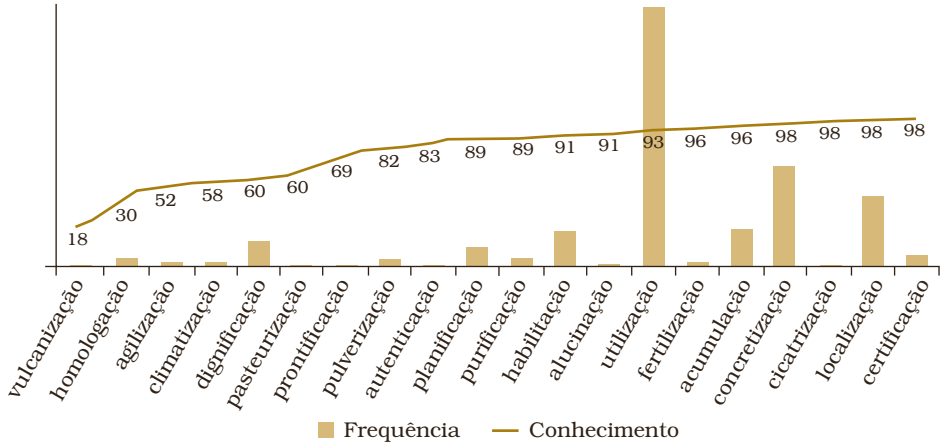
Fonte: Elaborado pelas autoras.

Gráfico 10 – Frequência *versus* conhecimento real (base três sílabas, derivado quatro sílabas)



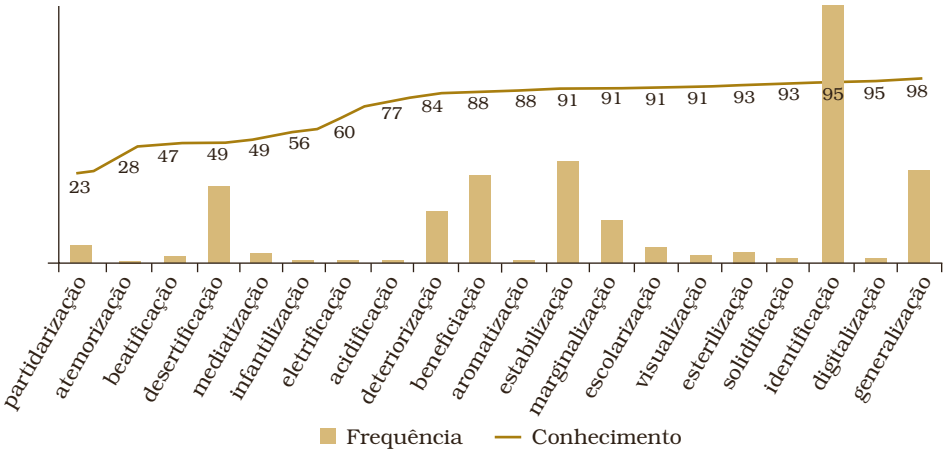
Fonte: Elaborado pelas autoras.

Gráfico 11 – Frequência *versus* conhecimento real (base quatro sílabas, derivado cinco sílabas)



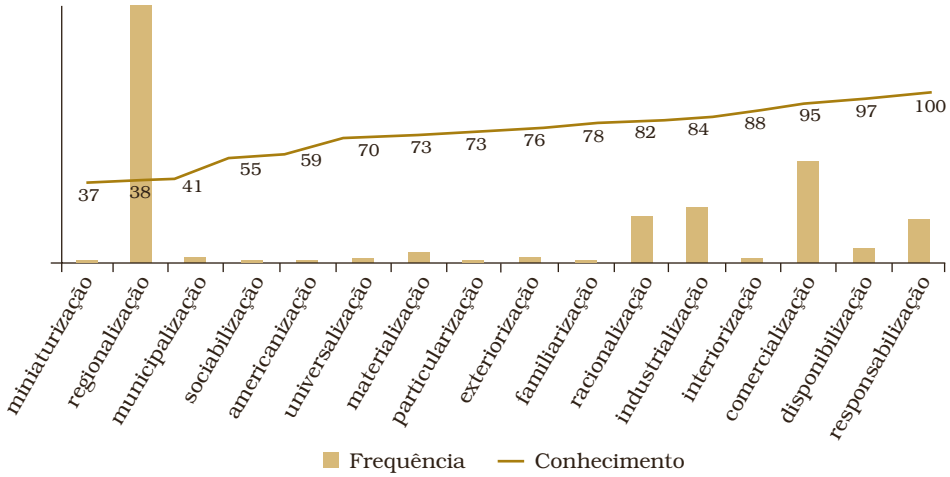
Fonte: Elaborado pelas autoras.

Gráfico 12 – Frequência *versus* conhecimento real (base cinco sílabas, derivado seis sílabas)



Fonte: Elaborado pelas autoras.

Gráfico 13 – Frequência *versus* conhecimento real (base seis sílabas, derivado sete sílabas)



Fonte: Elaborado pelas autoras.

Conclusões

Essa nossa pesquisa foi inicialmente motivada pela dúvida em relação à confiabilidade dos critérios normalmente aceitos nos testes morfológicos relacionados com o processamento visual das palavras, a saber, os valores de frequência encontrados num *corpus* estável e representativo, como o CRPC, e a dimensão das palavras, aqui medida em número de sílabas. Partimos do pressuposto de que esses fatores não refletem necessariamente o conhecimento lexical dos falantes, e da necessidade de criar melhores instrumentos de avaliação desse conhecimento.

Nesse sentido, decidimos aplicar um teste *off-line* dirigido à avaliação do conhecimento individual de palavras derivadas e das suas bases derivantes. A seleção das palavras apresentadas aos sujeitos obedeceu aos critérios geralmente considerados, ou seja, a frequência e o número de sílabas. Analisamos previamente a correlação entre comprimento e frequência de palavras, e concluímos que as palavras com maior número de sílabas, quer os verbos quer os derivados deverbais, são sistematicamente menos frequentes.

Os resultados do teste *off-line* foram bastante consistentes, mostrando que o conhecimento das palavras é independente tanto do seu comprimento quanto da frequência registrada no CRPC. Essa é uma conclusão muito interessante, pois permite definir uma nova variável para testes *on-line*, adaptada para cada teste específico.

Lexical knowledge assessment: experimental data and hypotheses

Abstract: Although there are some hypotheses about the internal structure of the lexicon and its status in grammar models, no theoretical approach about what words speakers really know or use is available. In this text, we display the

results of an offline test, conducted to measure individual word knowledge. These results show that word knowledge is independent of word length or word frequency values. This is a very interesting conclusion because it allows us to define a new variable for online testing, tailored for each specific test.

Keywords: Lexical access. Lexical knowledge. Frequency. Word length. Morphology.

Referências

BERTRAM, R.; HYONA, J. The length of a complex word modifies the role of morphological structure: evidence from eye movements when reading short and long Finnish compounds. *Journal of Memory and Language*, v. 48, n. 3, p. 615-634, 2002. DOI 10.1016/S0749-596X(02)00539-9

CLAHSEN, H.; NEUBAUER, K. Morphology, frequency, and the processing of derived words in native and non-native speakers. *Lingua*, v. 120, n. 11, p. 2627-2637, 2010. DOI 10.1016/j.lingua.2010.06.007

CRPC. *Reference Corpus of Contemporary Portuguese*. Disponível em: http://www.clul.ul.pt/pt/recursos/183-reference-corpus-of-contemporary-portuguese-crpc. Acesso em: 13 nov. 2018.

FORD, M. A.; DAVIS, M. H.; MARSLEN-WILSON, W. D. Derivational morphology and base morpheme frequency. *Journal of Memory and Language*, v. 63, n. 1, p. 117-130, 2010. DOI 10.1016/j.jml.2009.01.003

HUDSON, P. T.; BERGMAN, M. W. Lexical knowledge in word recognition: word length and word frequency in naming and lexical decision tasks. *Journal of Memory and Language*, v. 24, n. 1, p. 46-58, 1985. DOI 10.1016/0749-596X(85)90015-4

NEW, B. *et al.* Reexamining the word length effect in visual word recognition: new evidence from English Lexicon Project. *Psychonomic Bulletin & Review*, v. 13, n. 1, p. 45-52, 2006. DOI 10.3758/BF03193811

PINTO, C. *O papel da estrutura morfológica nos processos de leitura*. Lisboa: Faculdade de Letras da Universidade de Lisboa, 2017.

RAYNER, K. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, v. 124, n. 3, p. 372-422, 1998. DOI 10.1037//0033-2909.124.3.372

RAYNER, K. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, v. 62, n. 8, p. 1457-1506, 2009. DOI 10.1080/17470210902816461

VILLALVA, A.; PINTO, C. Morphological complexity and lexical processing costs. *Alfa*, v. 62, n. 1, p. 149-168, 2018. DOI: 10.1590/1981-5794-1804-7