SOLETRAS
REVISTA
N. 33 ï 2017.1 ï GUSTAVO LOPEZ ESTIVALET
FANNY MEUNIER

Corpus psicolinguístico Léxico do Português Brasileiro

Gustavo Lopez Estivalet[1]
Fanny Meunier[2]

Resumo: O Léxico do Português Brasileiro foi desenvolvido com o objetivo de oferecer um *corpus* baseado em palavras para a pesquisa em psicolinguística no português brasileiro. Ele foi criado a partir de um corpus com mais de 32 milhões de palavras. Assim, o Léxico do Português Brasileiro contém mais de 215 mil entradas lexicais e apresenta 21 colunas com informações metalinguísticas e psicolinguísticas relevantes, como categoria gramatical, frequência ortográfica, número de letras, vizinhos ortográficos, entre outras. Ele é um corpus aberto e de livre acesso na internet, possuindo uma plataforma amigável e dinâmica para pesquisas simples e complexas. O Léxico do Português Brasileiro ainda disponibiliza uma série de dados já computados, oferece um motor de geração de pseudopalavras do português brasileiro e um conjunto de ferramentas de linguística e estatística. Sendo assim, o presente artigo tem como objetivo introduzir e apresentar o Léxico do Português Brasileiro, e servir como seu manual de utilização. Ainda, é realizada uma descrição do desenvolvimento e criação do *corpus*. Enfim, o Léxico do Português Brasileiro preenche uma enorme lacuna na pesquisa em psicolinguística e linguística computacional, oferecendo um *corpus* baseado em palavras com valiosas informações metalinguísticas e psicolinguísticas do português brasileiro.
Palavras-chave: Psicolinguística. Linguística computacional. *Corpus*. Lexicografia. Linguística. Português brasileiro.

Introdução

O principal objetivo do Léxico do Português Brasileiro (LexPorBR)[3] é oferecer um *corpus* baseado em palavras do português brasileiro (PB) que disponibilize o máximo de informações metalinguísticas e psicolinguísticas sobre as palavras do PB. O Léxico do Português Brasileiro é um *corpus* livre e aberto, consultado em uma plataforma simples e dinâmica através da internet. A partir de uma pesquisa, os resultados são apresentados de forma organizada e hierárquica, contendo dados metalinguísticos e psicolinguísticos das palavras ou grupos de palavras pesquisados.

Corpora psicolinguísticos são utilizados: (1) no controle, seleção e manipulação de palavras e critérios específicos para a criação de experiências psicolinguísticas; e (2) em

1 Doutor em Neurociências e Ciências Cognitivas na École Doctorale Neurosciences et Cognition (NSCo) da Université Claude Bernard Lyon 1 (UCBL) no Laboratoire sur le Langage, le Cerveau et la Cognition (L2C2), Lyon, França, com bolsa do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). E-mail: gustavoestivalet@hotmail.com.
2 Doutorado em Psicologia Cognitiva pela Universidade René Descartes (1997, Paris, França). Diretora de pesquisa do Centre National de la Recherche Scientifique (CNRS). E-mail: fanny.meunier@isc.cnrs.fr.
3 http://www.lexicodoportugues.com/.

análises em linguística computacional da distribuição e do comportamento lexical (BAAYEN, 2001).Alguns exemplos de *corpora* psicolinguísticos baseados em palavra são: francês - *Lexique*[4] (NEW et al., 2001, 2004), espanhol ï *Busca Palabras* (DAVIS; PEREA, 2005), inglês ï MRC [5] (COLTHEART, 1981), alemão, espanhol, francês, holandês e inglês - ClearPON[6](MARIAN et al., 2012), alemão, cirílico, holandês e inglês - CELEX[7](BAAYEN; PIEPENBROCK; VAN RIJN, 1995).

Esses *corpora* foram utilizados, por exemplo, em megaestudos que investigam o comportamento psicolinguístico no processamento de palavras e pseudopalavras: *English Lexicon Project* (BALOTA et al., 2007), *French Lexicon Project* (FERRAND et al., 2010), *Dutch Lexicon Project* (KEULEERS; DIEPENDAELE; BRYSBAERT, 2010), e *British Lexicon Project* (KEULEERS et al., 2012). Ainda, eles são utilizados na seleção, controle e manipulação de palavras para criação de experiências psicolinguísticas em inúmeros estudos e pesquisas específicas(GIMENES; NEW, 2015), assim como no desenvolvimento e simulação de modelizações linguísticas (SCHREUDER; BAAYEN, 1995).

NILC/São Carlos e Linguateca

O Léxico do Português Brasileiro foi desenvolvido a partir do *corpus* do Núcleo Interinstitucional de Linguística Computacional de São Carlos (NILC) [8] (PINHEIRO; ALUÍSIO, 2003) sediado no Instituto de Ciências Matemáticas e de Computação de São Carlos (ICMC/São Carlos) [9], da Universidade de São Paulo em São Carlos (USP/São Carlos)[10]. As listas de formas e lemas divididas em categorias gramaticais foram baixadas do site do Linguateca[11](SANTOS; BICK, 2000), onde se encontram informações do NILC, como

[4]http://www.lexique.org/.
[5]http://www.psych.rl.ac.uk/.
[6]http://clearpond.northwestern.edu/.
[7]http://celex.mpi.nl/.
[8]http://www.nilc.icmc.usp.br/nilc/index.php.
[9]http://www.icmc.usp.br/Portal/.
[10]http://www.saocarlos.usp.br/.
[11]http://www.linguateca.pt/.

dados quantitativos e estatísticos[12], descendência do corpus[13] e os arquivos de formas[14] e lemas[15] no formato .txt, separados por categorias gramaticais.

Lexique

A criação e o desenvolvimento do Léxico do Português Brasileiro foram inspirados no *corpus* psicolinguístico do francês *Lexique* (NEW et al., 2001, 2004). O *Lexique* tem oferecido dados sobre as palavras do francês a uma série de estudos e pesquisas, sendo um ótimo exemplo de *corpus* psicolinguístico simples e eficaz. Ele exemplifica as funcionalidades e utilidades que um *corpus* psicolinguístico deve e pode oferecer como recursos para a pesquisa em psicolinguística e linguística computacional. Oferece, ainda, uma série de informações indispensáveis para criação das experiências e análise dos resultados (categoria gramatical, frequência, número de letras, vizinhos ortográficos, entre outras), motores para criação de pseudopalavras, links e referências relacionadas ao corpus, assim como listas com dados já computados (FERRAND et al., 2010). Uma descrição detalhada desse *corpus* é encontrada no manual do *Lexique*[16].

Programa e pacotes R

O Léxico do Português Brasileiro foi desenvolvido com o programa R[17], com os dados originais importados a partir de arquivos .txt e cada coluna sendo criada e computada através de determinadas funções e algoritmos. O número de vizinhos ortográficos (Coltheartôs N) (COLTHEART et al., 1977) e a distância de Levenshtein ortográfica das 20 palavras mais próximas (OLD20) (YARKONI; BALOTA; YAP, 2008) foram calculados a partir das funções ñcoltheart.Nò e ñold20ò disponibilizadas no pacote ñvwrò[18] (KEULEERS, 2013).

[12]http://www.linguateca.pt/acesso/desc_corpus.php?corpus=SAOCARLOS.
[13]http://www.linguateca.pt/acesso/NILCsaocarlos.html.
[14]http://www.linguateca.pt/acesso/contabilizacao.php#listaPosSAOCARLOS.
[15]http://www.linguateca.pt/acesso/contabilizacao.php#listaLemasSAOCARLOS.
[16]http://www.lexique.org/docLexique.php.
[17]http://www.r-project.org/.
[18]http://cran.r-project.org/web/packages/vwr/index.html.

Uma série de funções do pacote ñlanguageRò[19](BAAYEN, 2013) também foram utilizadas no desenvolvimento do Léxico do Português Brasileiro.

Léxico do Português Brasileiro

O projeto de criar o Léxico do Português Brasileiro nasceu de uma necessidade, em 2013, quando começamos a investigar a representação e o processamento morfológico flexional verbal no PB, no francês e em bilíngues com PB como língua materna e francês como língua estrangeira. Para as experiências em francês, os estímulos foram selecionados a partir do *corpus Lexique* (NEW et al., 2004), quando começamos a preparar as experiências em PB, deparamo-nos com a completa falta de um *corpus* psicolinguístico do PB. Procurando suprir nossas necessidades, tivemos acesso ao site do Linguateca (SANTOS; BICK, 2000) que reúne vários *corpora* do português europeu e brasileiro. Entretanto, não encontramos nenhum corpus do PB com dados metalinguísticos e psicolinguísticos apropriados para a criação rigorosa de experiências psicolinguísticas em PB. Foi nesse momento que decidimos fazer o Léxico do Português Brasileiro, que apresenta a página principal conforme a **Figura 1**.

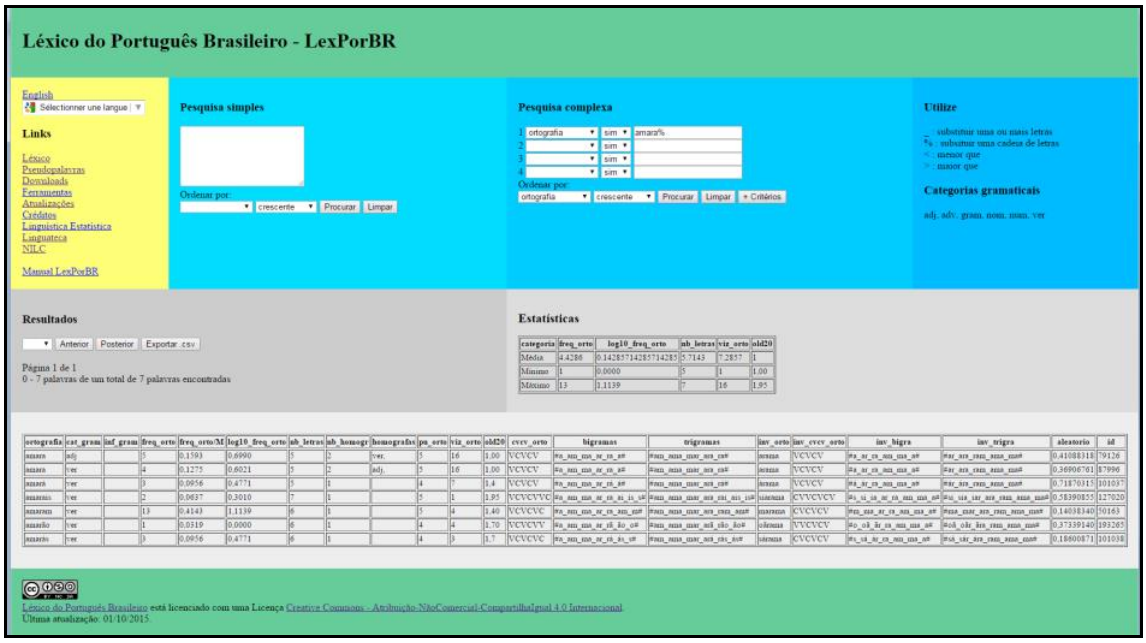


Figura 1. Página principal do Léxico do Português Brasileiro.

[19]https://cran.r-project.org/web/packages/languageR/index.html.

Categorias de informações

No início de 2014, o Léxico do Português Brasileiro começou a ser desenvolvido em quatro etapas: (1) construção do *corpus* com palavras e informações metalinguísticas e psicolinguísticas, (2) construção das páginas na internet em HTML, (3) importação do *corpus* para um banco de dados MySQL na internet e (4) programação em PHP do funcionamento do corpus. Além disso, foram criadas as demais páginas do site: atualizações, downloads, ferramentas, créditos. Em seguida, foi desenvolvido o motor de geração de pseudopalavras do PB e as ferramentas de linguística estatística.

Para tanto, foi desenvolvida uma série de conhecimentos de programação computacional em R, HTML[20], MySQL[21], PHP[22], Java[23] e CSS[24]. Selecionou-se no Linguateca o *corpus* do Núcleo Interdisciplinar de Linguística Computacional de São Carlos (NILC)[25]como o mais pertinente para a criação do Léxico do Português Brasileiro. Essa seleção foi baseada nos seguintes critérios: (1) número total de palavras (32 milhões) condizente com outros *corpora* psicolinguísticos (*Lexique*, CELEX, ClearPOND)(BRYSBAERT; NEW, 2009), (2) quantidade e tamanho dos arquivos (13 arquivos, tamanho total 49 MB), (3) organização do *corpus* em arquivos .txt separados por categorias gramaticais, (4) organização dos arquivos em duas colunas (ortografia e frequência) separadas por tabulação e (5) recursos e publicações já desenvolvidos pelo NILC (PINHEIRO; ALUÍSIO, 2003).

Para primeira etapa foi realizado o download dos 13 arquivos em formato .txt do corpus do NILC no site do Linguateca[26] separados por categorias gramaticais (6 arquivos de formas: adjetivos, advérbios, gramaticais, nomes, numerais e verbos; 7 arquivos de lemas: adjetivos, advérbios, gramaticais, nomes, nomes próprios, numerais e verbos). Em seguida, utilizou-se o programa R para a criação das categorias de informações de todas as palavras com menos de 30 letras. Criaram-se diferentes colunas com: (1) ortografia (orto) e

[20]http://pt.wikipedia.org/wiki/HTML.
[21]http://www.mysql.com/.
[22]http://www.php.net/.
[23]http://www.java.com/pt_BR/.
[24]http://pt.wikipedia.org/wiki/Cascading_Style_Sheets.
[25]http://www.linguateca.pt/acesso/corpus.php?corpus=SAOCARLOS.
[26]http://www.linguateca.pt/acesso/contabilizacao.php.

(2) categoria gramatical (cat_gram), além de uma coluna com o tipo de palavra (forma ou lema). Todas as palavras foram padronizadas em letras minúsculas e as formas repetidas foram somadas. Criou-se uma coluna com um (3) número de identificação (id) da palavra de acordo com a organização do corpus por frequência em ordem decrescente e ordem alfabéticaa-z.

Em seguida, as seguintes colunas com informações sobre as palavras foram contabilizadas: (4) frequência ortográfica (freq_orto), (5) frequência ortográfica por milhão de palavras (freq_orto/M, [1000000*freq_orto/freq_total]), (6) logaritmo natural da freq_orto/M (log10_freq_orto), (7) número de letras (nb_letras) (BRYSBAERT; NEW, 2009). Logo após, foram criadas colunas com: (8) número de formas homógrafas (nb_homogr) e (9) categorias gramaticais das formas homógrafas (homografas). Ainda, foram criadas colunas com: (10) informações gramaticais (inf_gram), (11) forma ortográfica invertida (inv_orto), (12) estrutura CVCV(CVCV_orto), (13) estrutura CVCV invertida (inv_CVCV_orto), (14) bigramas (bigramas), (15) bigramas invertidos (inv_bigra), (16) trigramas (trigramas), (17) trigramas invertidos (inv_trigra) e (18) número aleatório entre 0 e 1 com oito dígitos de precisão (aleatorio). Enfim, foi calculado: (19) ponto de unicidade ortográfico (pu_orto), (20) número de vizinhos ortográficos (viz_orto) (COLTHEART et al., 1977) e (21) distância de Levenshtein ortográfica (old20) (YARKONI; BALOTA; YAP, 2008) com a utilização do pacote ñvwrò (KEULEERS, 2013) e ñlanguageRò (BAAYEN, 2013) para o programa R.

Sendo assim, o Léxico do Português Brasileiro versão Alfa conta com 21 colunas de informações metalinguísticas e psicolinguísticas conforme o **Quadro 1**. Ele possui 215.175 linhas com diferentes palavras do PB. Portanto, cada linha do Léxico do Português Brasileiro contém uma palavra e cada coluna uma determinada informação sobre esta palavra. O corpus completo é disponibilizado em uma tabela em formato .csv com codificação UTF-8 com um tamanho de 45 MB.

1	**Ortografia (orto):** forma ortográfica da palavra em letras minúsculas (com exceção dos nomes próprios), respeitando os acentos específicos de cada palavra[27].
2	**Categoria gramatical (cat_gram):** categorial gramatical da palavra (adj, adv, gram, nom, num, prop, ver).
3	**Informação gramatical (inf_gram):** informações gramaticais sobre a palavra (e.g. singular/plural, masculino/feminino, passado/presente/futuro, 1/2/3 pessoas, etc.).
4	**Frequência ortográfica (freq_orto):** número de vezes que a palavra aparece no NILC.
5	**Frequência ortográfica por milhão (freq_orto/M):** número de vezes que a palavra aparece entre 1 milhão

[27] O corpus do NILC foi realizado em 1999, antes da reforma ortográfica do português.

	de palavras. Valor padrão para frequência de palavras(BRYSBAERT; NEW, 2009).
6	**Logaritmo natural da frequência ortográfica (log10_freq_orto):** logarítmico natural da frequência ortográfica. Utilizado para linearizar-se o comportamento da frequência das palavras no corpus(BAAYEN, 2001).
7	**Número de letras (nb_letras):** número de letras da palavra.
8	**Número de homógrafas (nb_homogr):** número de palavras homógrafas. Palavras que possuem a mesma ortografia ou diferenças de acentos, mas pertencem a categorias gramaticais diferentes.
9	**Homógrafas (homografas):** categorias gramaticais das palavras homógrafas.
10	**Ponto de unicidade ortográfico (pu_orto):** letra a partir da qual a palavra se dissocia das outras, ou seja, letra a partir da qual a palavra é única. Sentido da esquerda para direita.
11	**Vizinhos ortográficos (viz_orto):** número de vizinhos ortográficos a partir do N de Coltheart, ou seja, alterando-se apenas uma letra por vez (COLTHEART et al., 1977).
12	**Distância de Leveinshtein ortográfica (old20):** distância ortográfica de Leveinshtein das 20 palavras mais próximas calculadas a partir de regressões lineares (YARKONI; BALOTA; YAP, 2008).
13	**Estrutura CVCV (CVCV_orto):** estrutura CVCV da palavra, onde C para consoantes C e V para vogais. Ainda, A para acentos, P para pontuação, N para números e S para símbolos.
14	**Bigramas (bigramas):** bigramas que constituem a palavra separados por ñ_ò e limitados por ñ#ò. O número de bigramas é igual ao número de letras da palavra mais 1.
15	**Trigramas (trigramas):** trigramas que constituem a palavra separados por ñ_ò e limitados por ñ#ò. O número de trigramas é igual ao número de letras da palavra.
16	**Ortografia invertida (inv_orto):** forma invertida da ortografia (orto).
17	**Estrutura CVCV invertida (inv_CVCV_orto):** estrutura CVCV da palavra invertida a partir de (CVCV_orto).
18	**Bigramas invertidos (inv_bigra):** bigramas que constituem a palavra separados por ñ_ò e limitados por ñ#ò invertidos a partir de (bigramas).
19	**Trigramas invertidos (inv_trigra):** trigramas que constituem a palavra separados por ñ_ò e limitados por ñ#ò invertidos a partir de (trigramas).
20	**Número aleatório entre 0 e 1 (aleatório):** número aleatório entre 0 e 1 com oito algarismos de precisão.
21	**Número de identificação (id):** número de identificação da palavra designado a partir da organização do corpus por frequência decrescente e ordem alfabética a-z. O número de identificação é a posição da palavra no corpus.

Quadro 1. Nome e descrição das 21 colunas de informações metalinguísticas e psicolinguísticas do Léxico do Português Brasileiro versão Alfa.

Páginas e funcionamento

Para segunda etapa, utilizou-se o programa Notepad++[28]para o desenvolvimento de toda programação visual em HTML e CSS, e programação lógica em PHP e MySQL do site do Léxico do Português Brasileiro em um servidor local com o programa XAMPP[29], contendo os módulos Apache, MySQL, PHP e Perls pré-instalados.

A página principal do site do Léxico do Português Brasileiro possui dois motores de pesquisa: (1) pesquisa simples e (2) pesquisa complexa. A pesquisa simples contém uma área de texto onde se podem inserir uma palavra específica ou múltiplas palavras em forma de lista. Pode-se copiar e colar uma lista de palavras de uma planilha ou editor de texto. A

[28]http://notepad-plus-plus.org/.
[29]http://www.apachefriends.org/pt_br/index.html.

pesquisa complexa contém quatro campos de inserção de critérios específicos das palavras a serem pesquisadas. No primeiro campo, o usuário deve escolher a coluna de informação pela qual deseja realizar a pesquisa. No segundo campo, deve escolher se deseja considerar ñsimò ou desconsiderar ñnãoò o critério. E no terceiro campo, o usuário deve inserir os critérios específicos de sua pesquisa.

Os símbolos coringas ñ_ò para uma letra e ñ%ò para uma cadeia de letras são aceitos pelo MySQL e podem ser utilizados em ambos os motores de pesquisa. Ainda, os símbolos maior que ñ>ò e menor que ñ<ò podem ser utilizados para pesquisas numéricas de grupos de palavras na pesquisa complexa. Inicialmente, a pesquisa complexa apresenta quatro campos de critérios para pesquisa, clicando-se no botão ñ+Critériosò, o usuário é enviado a uma página que apresenta oito campos de critérios para a pesquisa. Cada motor de pesquisa possui um botão ñProcurarò para iniciar a pesquisa e apresentar os resultados e um botão ñLimparò para apagar os dados presentes nos campos. O usuário pode escolher a categoria utilizada para a organização e apresentação das palavras e o sentido de organização crescente ou decrescente, conforme a Figura 2:

Figura 2. Motores de pesquisa simples e pesquisa complexa do Léxico do Português Brasileiro.

Além da página principal do Léxico do Português Brasileiro: Léxico[30], as seguintes páginas ainda foram criadas para complementar o site: Pseudopalavras[31], Downloads[32], Ferramentas[33], Atualizações[34], Créditos[35]e Linguística Estatística[36]. ñPseudopalavrasò acessa o motor de geração de pseudopalavras do PB, conforme descrito abaixo. ñDownloadsò

[30]http://www.lexicodoportugues.com/index.php.
[31]http://www.lexicodoportugues.com/pseudowords.php.
[32]http://www.lexicodoportugues.com/downloads.php.
[33]http://www.lexicodoportugues.com/tools.php.
[34]http://www.lexicodoportugues.com/updates.php.
[35]http://www.lexicodoportugues.com/credits.php.
[36]http://www.lexicodoportugues.com/stat_ling.php.

disponibiliza uma série de arquivos pertinentes do Léxico do Português Brasileiro para *downloads* (corpus.txt, manuais, listas, convenções, bigramas, trigramas, scripts em R, entre outros). ñFerramentasò disponibiliza uma série links de corpora, programas e literatura em psicolinguística e linguística computacional. ñAtualizaçõesô descreve o desenvolvimento do Léxico do Português Brasileiro e as atualizações realizadas. ñCréditosò apresenta o objetivo, a origem do Léxico do Português Brasileiro, assim como descreve as referências e pertinência do corpus do NILC, do Linguateca, do *Lexique*, do programa e dos pacotes R e da licença *Creative Commons*, finalizando com os agradecimentos. Enfim, Linguística Estatística é uma página que disponibiliza diversos recursos e ferramentas de livre acesso, conforme descrito abaixo.

Finalmente, todas as páginas e informações do Léxico do Português Brasileiro foram traduzidas para o inglês (*Brazilian Portuguese Lexicon*)[37]. Implementou-se também o Google Tradutor em todas as páginas do Léxico do Português Brasileiro para a tradução do site para as línguas disponibilizadas nesse mecanismo. Sugere-se que o Google Tradutor seja utilizado a partir da versão inglês do *Brazilian Portuguese Lexicon*, pois assim não traduzirá os resultados das pesquisas, que por sua vez são sempre apresentados em PB.

Para terceira etapa, foi realizado o registro do domínio próprio do Léxico do Português Brasileiro (www.lexicodoportugues.com) junto ao HostGator[38] e redirecionamento deste domínio para o servidor onde o corpus foi hospedado (http://www.biz.nf/). Esse servidor foi escolhido a partir dos critérios: (1) espaço de 250 MB, (2) banco de dados MySQL 5, (3) suporte à PHP 4/5, (4) 5000 MB de transferência, (5) hospedagem gratuita, (6) domínio gratuito do tipo http://portugueselexicon.co.nf, (7) webmail POP3/SMTP e (8) controle de arquivos por FTP. Assim, os dados do Léxico do Português Brasileiro foram importados em formato .csv e configurados com a utilização do phpMyAdmin[39] para um banco de dados MySQL no servidor acima.

Na quarta etapa, foi realizada a programação de algoritmos em Java e PHP para: (1) manter os dados preenchidos nos campos da página HTML após pesquisa, (2) inserção de dois campos para organização dos resultados, um para seleção do critério de organização e outro para ordem crescente ou decrescente, (3) inserção do botão ñ+Critériosò na pesquisa

[37] http://www.lexicodoportugues.com/index_en.php.
[38] http://hostgator.com.br/.
[39] http://www.phpmyadmin.net/home_page/index.php.

complexa para disponibilização de oito campos de pesquisa, (4) reconhecimento dos símbolos maior que ñ>ò e menor que ñ<ò para as pesquisas numéricas, (5) desenvolvimento de um módulo de limitação e navegação dos resultados apresentados com o número de palavras a serem apresentadas (50, 100, 200 ou 500) e dois botões (ñAnteriorò e ñPosteriorò) para navegar entre as páginas de resultados, (6) apresentação de quatro informações gerais da pesquisa: i. total de palavras encontradas, ii. total de páginas de resultados, iii. intervalo das palavras apresentadas e iv. página apresentada, e (7) desenvolvimento do botão ñExportar .csvò para exportar o resultado da pesquisa realizada em um arquivo .csv disponibilizado para download do usuário.

Convenções

Para a utilização do Léxico do Português Brasileiro, algumas convenções foram determinadas para realização das pesquisas e apresentação dos resultados.

- Categorias gramaticais: **adj** - adjetivo, **adv** - advérbio, **gram** - gramatical, **nom** - substantivo, **num** - numeral, **prop** - nome próprio, **ver** - verbo.
- Estruturas CVCV das palavras possuem: **V** - vogais, **C** - consoantes, **P** - pontuação, **N** - números, **A** - acentos, **S** - símbolos.
- Símbolos coringas utilizados: **ñ<ò** menor que, **ñ>ò** maior que, **ñ_ò** substitui uma letra, **ñ%ò** substitui uma cadeia de letras.
- Ordem de apresentação dos resultados: **crescente** - apresenta os resultados na ordem crescente, **decrescente** - apresenta os resultados na ordem decrescente.
- Botões: **Procurar** - realiza a pesquisa e apresenta os resultados, **Limpar** -limpa os dados dos campo do formulário, **+Critérios** - direciona o usuário para uma página com mais critérios para a pesquisa complexa.
- Escolha sim/não: **sim** - considera o critério, **não** - desconsidera o critério.

Versão Alfa

Tendo em vista a enorme quantidade de informações metalinguísticas e psicolinguísticas que podem e serão computados, implementados e disponibilizados no

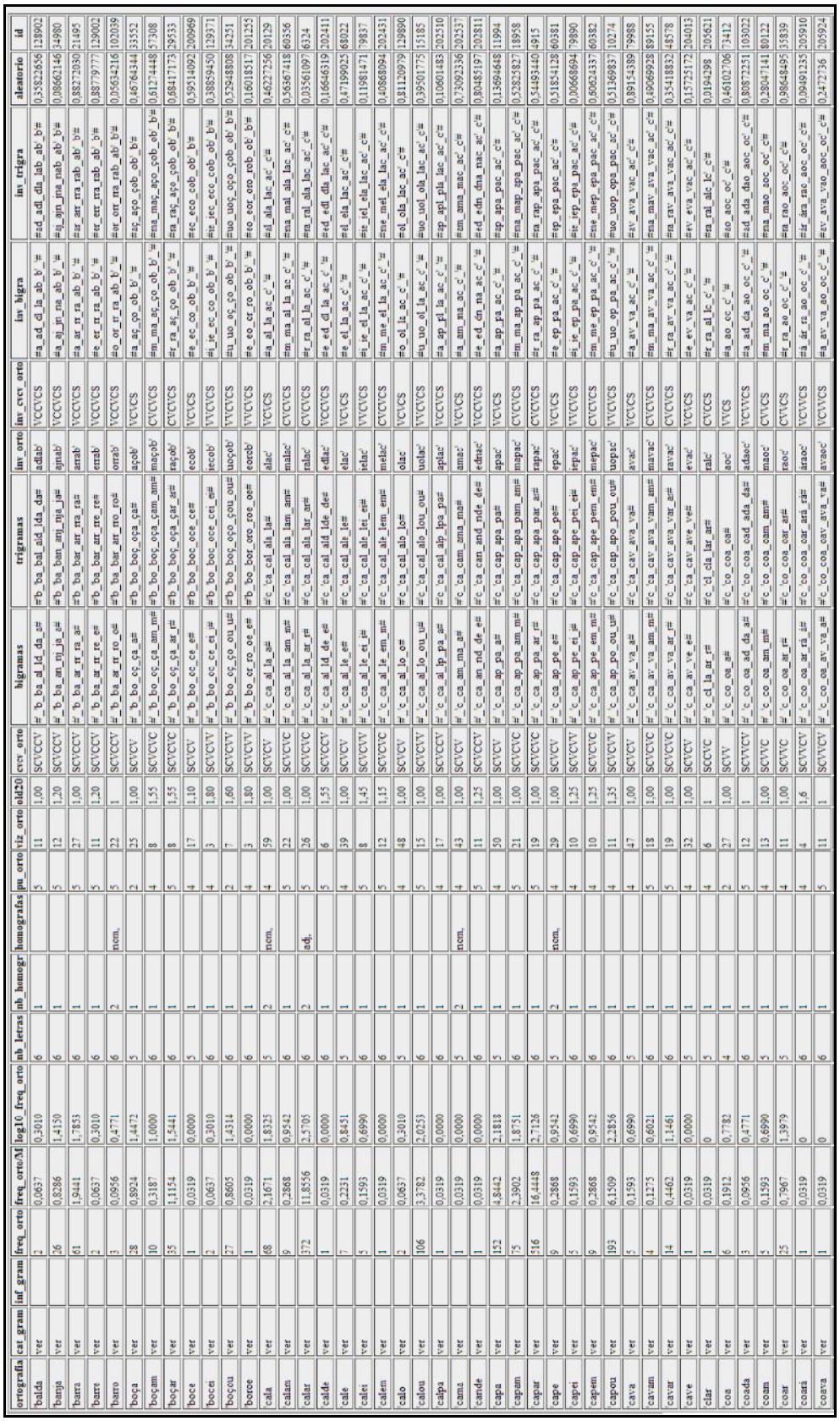
Léxico do Português Brasileiro, seu desenvolvimento foi dividido em três versões: 1) Alfa (2014), 2) Beta (2017) e 3) Delta (2019). O Léxico do Português Brasileiro versão Alfa foi inaugurada em 25 de março de 2014 e o surgimento do primeiro corpus psicolinguístico baseado em palavra do PB. A principal característica do Léxico do Português Brasileiro versão Alfa é que ele disponibiliza um *corpu*s ortográfico em que as informações foram computadas a partir de dados ortográficos das palavras do PB do NILC.

A versão Beta contará com as informações: (1) fonológicas, (2) silábicas e (3) dos lemas associados às formas. A versão Delta contará com uma série de: (1) informações morfológicas, (2) informações sintáticas e (3) medidas de tempo de reação do reconhecimento de um grande número de palavras e pseudopalavras do PB, conforme os *Lexicon Projects* (BALOTA et al., 2007; FERRAND et al., 2010; KEULEERS et al., 2012; KEULEERS; DIEPENDAELE; BRYSBAERT, 2010).

Resultados

Na seção de resultados (Figuras 3 e 4), o usuário encontra os resultados da pesquisa organizada em diferentes linhas e as informações metalinguísticas e psicolinguísticas nas diferentes colunas. Encontram-se ainda uma série de informações pertinentes à pesquisa, conforme o lado esquerdo da Figura 4: (1) número total de palavras encontradas na pesquisa, (2) intervalo de palavras apresentados, (3) número total de páginas da pesquisa e (4) número da página apresentada. Pode-se escolher no campo superior à esquerda o número de palavras apresentadas em cada página e o usuário pode navegar entre os resultados e as páginas da pesquisa através dos botões ñAnteriorò e ñPróximoò. Um exemplo de pesquisa que apresenta palavras que possuem a categoria gramatical definida como ñverboò pode ser visualizada a partir da pesquisa complexa com o critério **cat_gram - sim - ver**, conforme a Figura 3.

SOLETRAS
REVISTA
N. 33 ï 2017.1 ï GUSTAVO LOPEZ ESTIVALET
FANNY MEUNIER



ortografia	cat_gram	inf_gram	freq_orto	freq_orto/Mi	log10_freq_orto	nb_letras	nb_homogr	homografas	pu_orto	viz_orto	old20	cvcv_orto	bigramas	trigramas	inv_orto	inv_cvcv_orto	inv_bigra	inv_trigra	aleatorio	id
balda	ver		2	0,0637	0,3010	6	1		5	11	1,00	SCVCCV	#_b_ba_al_ld_da_a#	#_b_ba_bal_ald_lda_da#	adlab'	VCCVCS	#a_ad_dl_la_ab_b'_'#	#ad_adl_dla_lab_ab'_b'#	0,55823656	128902
banja	ver		26	0,8286	1,4150	6	1		5	12	1,20	SCVCCV	#_b_ba_an_nj_ja_a#	#_b_ba_ban_anj_nja_ja#	ajnab'	VCCVCS	#a_aj_jn_na_ab_b'_'#	#aj_ajn_jna_nab_ab_b'#	0,08663146	34980
barra	ver		61	1,9441	1,7853	6	1		5	27	1,00	SCVCCV	#_b_ba_ar_rr_ra_a#	#_b_ba_bar_arr_rra_ra#	arrab'	VCCVCS	#a_ar_rr_ra_ab_b'_'#	#ar_arr_rra_rab_ab'_b'#	0,88272030	21495
barre	ver		2	0,0637	0,3010	6	1		5	11	1,20	SCVCCV	#_b_ba_ar_rr_re_e#	#_b_ba_bar_arr_rre_re#	errab'	VCCVCS	#e_er_rr_ra_ab_b'_'#	#er_err_rra_rab_ab'_b'#	0,88779777	129002
barro	ver		3	0,0956	0,4771	6	2	ncm.	5	22	1	SCVCCV	#_b_ba_ar_rr_ro_o#	#_b_ba_bar_arr_rro_ro#	orrab'	VCCVCS	#o_or_rr_ra_ab_b'_'#	#or_orr_rra_rab_ab'_b'#	0,05634216	1020359
boça	ver		28	0,8924	1,4472	5	1		2	25	1,00	SCVCV	#_b_bo_oç_ça_a#	#_b_bo_boç_oça_ça#	açob'	VCVCS	#a_aç_oç_ob_b'_'#	#aç_açoç_çob_ob'_b'#	0,46764344	33552
boçam	ver		10	0,3187	1,0000	6	1		4	8	1,55	SCVCVC	#_b_bo_oç_ça_am_m#	#_b_bo_boç_oça_çam_am#	maçob'	CVCVCS	#m_ma_aç_ço_ob_b'_'#	#ma_maç_açoç_çob_ob_b'#	0,61274448	57308
boçar	ver		35	1,1154	1,5441	6	1		5	8	1,55	SCVCVC	#_b_bo_oç_ça_ar_r#	#_b_bo_boç_oça_çar_ar#	raçob'	CVCVCS	#r_ra_aç_ço_ob_b'_'#	#ra_raç_açoç_çob_ob'_b'#	0,88417173	29533
boce	ver		1	0,0319	0,0000	5	1		4	17	1,10	SCVCV	#_b_bo_oc_ce_e#	#_b_bo_boc_oce_ce#	ecob'	VCVCS	#e_ec_co_ob_b'_'#	#ec_eco_cob_ob'_b'#	0,59514092	200969
bocei	ver		2	0,0637	0,3010	6	1		4	3	1,80	SCVCVV	#_b_bo_oc_ce_ei_#_	#_b_bo_boc_oce_cei_ei#	jecob'	VVCVCS	#i_ei_ec_co_ob_b'_'#	#ei_iec_eco_cob_ob'_b'#	0,38859450	129371
boçou	ver		27	0,8605	1,4314	6	1		2	7	1,60	SCVCVV	#_b_bo_oç_ço_ou_u#	#_b_bo_boç_oço_çou_ou#	uoçob'	VVCVCS	#u_uo_oç_ço_ob_b'_'#	#uo_uoç_oço_çob_ob'_b'#	0,52948808	34251
borce	ver		1	0,0319	0,0000	6	1		4	3	1,80	SCVCVC	#_b_bo_or_ro_ce_e#	#_b_bo_bor_oro_roe_oe#	eorcb'	VVCVCS	#e_eo_or_ro_ob_b'_'#	#eo_eor_oro_rob_ob'_b'#	0,16018517	201255
cala	ver		68	2,1671	1,8325	5	2	ncm.	4	59	1,00	SCVCV	#_c_ca_al_la_a#	#_c_ca_cal_ala_la#	alac'	VCVCS	#a_al_la_ac_c'_'#	#al_ala_lac_ac'_c'#	0,46227256	20129
calam	ver		9	0,2868	0,9542	6	1		5	22	1,00	SCVCVC	#_c_ca_al_la_am_m#	#_c_ca_cal_ala_lam_am#	malac'	CVCVCS	#m_ma_al_la_ac_c'_'#	#ma_mal_ala_lac_ac'_c'#	0,56367418	60356
calar	ver		372	11,8556	2,5705	6	2	adj.	5	26	1,00	SCVCVC	#_c_ca_al_la_ar_r#	#_c_ca_cal_ala_lar_ar#	ralac'	CVCVCS	#r_ra_al_la_ac_c'_'#	#ra_ral_ala_lac_ac'_c'#	0,03561097	6324
calde	ver		1	0,0319	0,0000	6	1		5	6	1,55	SCVCCV	#_c_ca_al_ld_de_e#	#_c_ca_cal_ald_lde_de#	edlac'	VCVCS	#e_ed_dl_la_ac_c'_'#	#ed_edl_dla_lac_ac'_c'#	0,16646319	203411
cale	ver		7	0,2231	0,8451	5	1		4	39	1,00	SCVCV	#_c_ca_al_le_e#	#_c_ca_cal_ale_le#	elac'	VCVCS	#e_el_la_ac_c'_'#	#el_ela_lac_ac'_c'#	0,47199025	68022
calei	ver		5	0,1593	0,6990	6	1		5	8	1,45	SCVCVV	#_c_ca_al_le_ei_i#	#_c_ca_cal_ale_lei_ei#	ielac'	VVCVCS	#i_ei_el_la_ac_c'_'#	#ei_iel_ela_lac_ac'_c'#	0,11981471	79837
calem	ver		1	0,0319	0,0000	6	1		5	12	1,15	SCVCVC	#_c_ca_al_le_em_m#	#_c_ca_cal_ale_lem_em#	melac'	CVCVCS	#m_me_el_la_ac_c'_'#	#me_mel_ela_lac_ac'_c'#	0,40868094	202431
calo	ver		2	0,0637	0,3010	5	1		4	48	1,00	SCVCV	#_c_ca_al_lo_o#	#_c_ca_cal_alo_lo#	olac'	VCVCS	#o_ol_la_ac_c'_'#	#ol_ola_lac_ac'_c'#	0,81120979	129890
calou	ver		106	3,3782	2,0253	6	1		5	15	1,00	SCVCVV	#_c_ca_al_lo_ou_u#	#_c_ca_cal_alo_lou_ou#	uolac'	VVCVCS	#u_uo_ol_la_ac_c'_'#	#uo_uol_ola_lac_ac'_c'#	0,39501775	15185
calpa	ver		1	0,0319	0,0000	6	1		4	17	1,00	SCVCCV	#_c_ca_al_lp_pa_a#	#_c_ca_cal_alp_lpa_pa#	aplac'	VCCVCS	#a_ap_pl_la_ac_c'_'#	#ap_apl_pla_lac_ac'_c'#	0,10601485	202510
cama	ver		1	0,0319	0,0000	5	2	ncm.	4	43	1,00	SCVCV	#_c_ca_am_ma_a#	#_c_ca_cam_ama_ma#	amac'	VCVCS	#a_am_ma_ac_c'_'#	#am_ama_mac_ac'_c'#	0,73092336	202537
cande	ver		1	0,0319	0,0000	6	1		5	11	1,25	SCVCCV	#_c_ca_an_nd_de_e#	#_c_ca_can_and_nde_de#	ednac'	VCVCS	#e_ed_dn_na_ac_c'_'#	#ed_edn_dna_nac_ac'_c'#	0,80485197	202811
capa	ver		152	4,8442	2,1818	5	1		4	50	1,00	SCVCV	#_c_ca_ap_pa_a#	#_c_ca_cap_apa_pa#	apac'	VCVCS	#a_ap_pa_ac_c'_'#	#ap_apa_pac_ac'_c'#	0,13694648	11994
capam	ver		75	2,3902	1,8751	6	1		5	21	1,00	SCVCVC	#_c_ca_ap_pa_am_m#	#_c_ca_cap_apa_pam_am#	mapac'	CVCVCS	#m_ma_ap_pa_ac_c'_'#	#ma_map_apa_pac_ac'_c'#	0,52825827	18958
capar	ver		516	16,4448	2,7126	6	1		5	19	1,00	SCVCVC	#_c_ca_ap_pa_ar_r#	#_c_ca_cap_apa_par_ar#	rapac'	CVCVCS	#r_ra_ap_pa_ac_c'_'#	#ra_rap_apa_pac_ac'_c'#	0,54493440	4915
cape	ver		9	0,2868	0,9542	5	2	ncm.	4	29	1,00	SCVCV	#_c_ca_ap_pe_e#	#_c_ca_cap_ape_pe#	epac'	VCVCS	#e_ep_pa_ac_c'_'#	#ep_epa_pac_ac'_c'#	0,51854128	60381
capei	ver		5	0,1593	0,6990	6	1		4	10	1,25	SCVCVV	#_c_ca_ap_pe_ei_i#	#_c_ca_cap_ape_pei_ei#	iepac'	VVCVCS	#i_ei_ep_pa_ac_c'_'#	#ei_iep_epa_pac_ac'_c'#	0,00668694	79890
capem	ver		9	0,2868	0,9542	6	1		4	10	1,25	SCVCVC	#_c_ca_ap_pe_em_m#	#_c_ca_cap_ape_pem_em#	mepac'	CVCVCS	#m_me_ep_pa_ac_c'_'#	#me_mep_epa_pac_ac'_c'#	0,60624337	60382
capou	ver		193	6,1509	2,2856	6	1		4	11	1,35	SCVCVV	#_c_ca_ap_po_ou_u#	#_c_ca_cap_apo_pou_ou#	uopac'	VVCVCS	#u_uo_op_pa_ac_c'_'#	#uo_uop_opa_pac_ac'_c'#	0,51369837	10274
cava	ver		5	0,1593	0,6990	5	1		4	47	1,00	SCVCV	#_c_ca_av_va_a#	#_c_ca_cav_ava_va#	avac'	VCVCS	#a_av_va_ac_c'_'#	#av_ava_vac_ac'_c'#	0,89154389	79988
cavam	ver		4	0,1275	0,6021	6	1		5	18	1,00	SCVCVC	#_c_ca_av_va_am_m#	#_c_ca_cav_ava_vam_am#	mavac'	CVCVCS	#m_ma_av_va_ac_c'_'#	#ma_mav_ava_vac_ac'_c'#	0,49069928	89155
cavar	ver		14	0,4462	1,1461	6	1		5	19	1,00	SCVCVC	#_c_ca_av_va_ar_r#	#_c_ca_cav_ava_var_ar#	ravac'	CVCVCS	#r_ra_av_va_ac_c'_'#	#ra_rav_ava_vac_ac'_c'#	0,35418832	14578
cave	ver		1	0,0319	0,0000	5	1		4	32	1,00	SCVCV	#_c_ca_av_ve_e#	#_c_ca_cav_ave_ve#	evac'	VCVCS	#e_ev_va_ac_c'_'#	#ev_eva_vac_ac'_c'#	0,15725172	204013
clar	ver		1	0,0319	0	5	1		4	6	1	SCCVC	#_c_cl_la_ar_r#	#_c_cl_cla_lar_ar#	ralc'	CVCCS	#r_ra_al_lc_c'_'#	#ra_ral_alc_lc'_c'#	0,0194298	205621
coa	ver		6	0,1912	0,7782	4	1		2	27	1,00	SCVV	#_c_co_oa_a#	#_c_co_coa_oa#	aoc'	VVCS	#a_ao_oc_c'_'#	#ao_aoc_oc'_c'#	0,46102706	73412
coada	ver		3	0,0956	0,4771	6	1		5	12	1	SCVVCV	#_c_co_oa_ad_da_a#	#_c_co_coa_oad_ada_da#	adaoc'	VCVVCS	#a_ad_da_ao_oc_c'_'#	#ad_ada_dao_aoc_oc'_c'#	0,80872351	103023
coam	ver		5	0,1593	0,6990	5	1		4	13	1,00	SCVVC	#_c_co_oa_am_m#	#_c_co_coa_oam_am#	maoc'	CVVCS	#m_ma_ao_oc_c'_'#	#ma_mao_aoc_oc'_c'#	0,28047141	80122
coar	ver		25	0,7967	1,3979	5	1		4	11	1,00	SCVVC	#_c_co_oa_ar_r#	#_c_co_coa_oar_ar#	raoc'	CVVCS	#r_ra_ao_oc_c'_'#	#ra_rao_aoc_oc'_c'#	0,98648495	55839
coará	ver		1	0,0319	0	6	1		4	4	1,60	SCVVCV	#_c_co_oa_ar_rá_á#	#_c_co_coa_ará_rá#	áraoc'	VCVVCS	#á_ár_ra_ao_oc_c'_'#	#á_ára_rao_aoc_oc'_c'#	0,09491235	205910
coava	ver		1	0,0319	0	6	1		5	11	1	SCVVCV	#_c_co_oa_av_va_a#	#_c_co_coa_ava_va#	avaoc'	VCVVCS	#a_av_va_ao_oc_c'_'#	#av_ava_vao_aoc_oc'_c'#	0,24727136	205924

Figura 3. Exemplo de resultado de pesquisa do Léxico do Português Brasileiro.

Na parte superior a direita dos resultados, conforme a **Figura 4**, apresenta-se uma série de dados estatísticos estabelecidos e calculados a partir da pesquisa realizada (DAVIS, 2005; DAVIS; PEREA, 2005): 1) média, 2) valor máximo e 3) valor mínimo, das seguintes categorias: 1) freq_orto, 2) log10_freq_orto, 3) nb_letras, d) viz_orto e 4) old20. Futuramente, mais dados estatísticos serão inseridos neste módulo. Por fim, o botão ñExportar .csvò exporta todos os dados da pesquisa para um arquivo .csv disponibilizado para *download* do usuário.

Figura 4. Informações dos resultados e estatísticas básicas.

Pseudopalavras do PB

O motor gerador de pseudopalavras do PB foi desenvolvido para a criação de pseudopalavras baseadas na estrutura e frequência das palavras do PB. Diferentemente de outros motores de geração de pseudopalavras que se baseiam na estrutura silábica das palavras existentes da língua (KEULEERS; BRYSBAERT, 2010; MOTA; RESENDE, 2013), o motor de geração de pseudopalavras do PB do Léxico do Português Brasileiro utiliza os bigramas e trigramas(NEW et al., 2001). Todos os bigramas e trigramas foram contabilizados a partir de todas as palavras do Léxico do Português Brasileiro. As pseudopalavras são geradas a partir da frequência e combinação dos bigramas ou trigramas. Contabilizaram-se a (1) frequência geral dos bigramas e trigramas, (2) frequência dos bigramas e trigramas de acordo a posição na palavra e (3) frequência dos bigramas e trigramas por categoria gramatical.

No motor de geração de pseudopalavras do PB, o usuário deve inserir quatro campos: (1) número de palavras a serem geradas, (2) número de letras das palavras a serem geradas, (3) categoria gramatical que estas palavras devem pertencer (todas, adj, adv, gram, nom, num, ver) e (4) tipo de critério para a construção das palavras (bigramas ou trigramas). O motor de geração de pseudopalavras do PB constrói as palavras simultaneamente nos dois sentidos, da

esquerda para a direita e da direita para a esquerda, começando com um bigrama ou trigrama do tipo ñ#xxò ou ñxx#ò. De acordo com o número de letras, o motor vai concatenando novos bigramas ou trigramas que dividam o máximo de informação ortográfica com bigrama ou trigrama anterior (1 letra para os bigramas e 2 letras para os trigramas). O motor apresenta dois botões: ñEnviarò para gerar e apresentar os resultados das pseudopalavras e ñLimparò para limpar os dados dos campos, conforme a Figura 5.

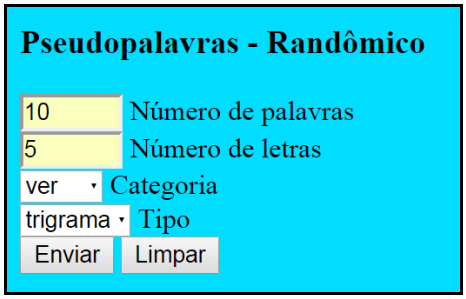


Figura 5. Motor de geração de pseudopalavras do PB.

Na tabela de resultados da geração de pseudopalavras do PB (10 pseudopalavras baseadas em verbos de 5 letras a partir de trigramas), conforme a Figura 6, quatro colunas com dados sobre as pseudopalavras são apresentadas: (1) categoria gramatical definida pelo usuário, (2) frequência da pseudopalavras calculada a partir da soma das frequências dos bigramas ou trigramas que compõem a pseudopalavra, (3) log10 da frequência calculada da pseudopalavra e (4) número de letras da pseudopalavra. Nos resultados, ainda é disponibilizado o botão ñExportar .csvò para exportar os resultados da geração de pseudopalavras do PB para um arquivo .csv disponibilizado para download do usuário.

Resultados

Exportar .csv

pseudo esq-dir	categoria esq-dir	freq esq-dir	log_freq esq-dir	nb_letras esq-dir	pseudo dir-esq	categoria dir-esq	freq dir-esq	log_freq dir-esq	nb_letras dir-esq
cosas	ver	10522	9.2612	5	cados	ver	11736	9.3704	5
desta	ver	7338	8.9008	5	atica	ver	8879	9.0914	5
presa	ver	4743	8.4644	5	atico	ver	8662	9.0667	5
resta	ver	6972	8.8497	5	inais	ver	5958	8.6925	5
antes	ver	8512	9.0492	5	apres	ver	4182	8.3385	5
cassa	ver	4555	8.424	5	prado	ver	7894	8.9739	5
supes	ver	4215	8.3464	5	dista	ver	7424	8.9125	5
manta	ver	7666	8.9446	5	prada	ver	6615	8.7971	5
extra	ver	2909	7.9756	5	dinal	ver	4247	8.354	5
estas	ver	10813	9.2885	5	cante	ver	8074	8.9964	5

Figura 6. Exemplo dos resultados da geração de pseudopalavras do PB.

Linguística Estatística

A página Linguística Estatística do Léxico do Português Brasileiro disponibiliza livremente e abertamente recursos e ferramentas psicolinguísticas e de estatística linguística que podem ser consultadas diretamente na página através da internet. Esses recursos e ferramentas foram desenvolvidos em HTML/PHP, sendo eles: (1) Fôe minFô - MS, (2) minFô - F1.F2, (3) teste de Hartley, (4) normalização entre 0 e 1, (5) inversor de palavras, (6) distância de Hamming, (7) distância de Levenshtein, (8) vizinhos ortográficos (Coltheart's N), (9) média das distâncias de Levenshtein, (10) entropia relativa, (11) frequência de palavras e (12) distribuição de Zipf, entre outros (BAAYEN, 2001, 2013; BRYSBAERT; NEW, 2009; DAVIS, 2005; DAVIS; PEREA, 2005; KEULEERS, 2013).

Licença Creative Commons

Agradecimentos

Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de Doutorado Pleno no Exterior (GDE) e ao *Centre National de La Recherche Scientifique* (CNRS) pela estrutura para o desenvolvimento do Léxico do Português Brasileiro. Agradecemos às pesquisadoras do NILC Profa. Dra. Sandra M. Aluísio e Profa. Dra. Maria das Graças Volpe Nunes pelos valiosos materiais, informações e auxílio sobre o NILC, assim como o apoio na realização deste trabalho. Agradecemos aos pesquisadores Dr. Léo Varnet e Dr. Emmanuel Trouche pelas discussões sobre os scripts e

[40]http://creativecommons.org/licenses/by-nc-sa/4.0/.

algoritmos para o desenvolvimento do Léxico do Português Brasileiro. Agradeço aos usuários dos fóruns de discussão e tutoriais da internet sobre o desenvolvimento de páginas e bancos de dados.

Referências

BAAYEN, R. H. *Word Frequency Distributions*. Dodrecht; Boston; London: Kluwer Academic Publishers, 2001.

BAAYEN, R. H. languageR: Data sets and functions with ñAnalyzing Linguistic Data: A practical introduction to statisticsò. *R Package*, p. 133, 2013.

BAAYEN, R. H.; PIEPENBROCK, R; VAN RIJN, H. *The CELEX lexical database. Release 2 [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania, 1995.

BALOTA, D. A. et al. The English Lexicon Project. *Behavior Research Methods*, v. 39, n. 3, p. 445ï 459, 2007.

BRYSBAERT, M.; NEW, B. Moving beyond Kuļera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, v. 41, n. 4, p. 977ï 990, 2009.

COLTHEART, M. et al. Access to the internal lexicon. In: DORNIC, S. (Ed.). *Attention and Performance VI*. Hillsdale, NJ: Lawrence Erlbaum Associates, p. 535ï 555, 1977.

COLTHEART, M. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, v. 33, n. 4, p. 497ï 505, 1981.

DAVIS, C. J. N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, v. 37, n. 1, p. 65ï 70, 2005.

DAVIS, C. J; PEREA, M. BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, v. 37, n. 4, p. 665ï 671, 2005.

FERRAND, L. et al. The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, v. 42, n. 2, p. 488ï 496, 2010.

GIMENES, M; NEW, B. Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, v. 48, n. 3, p. 963-972, 2015.

KEULEERS, E. et al. The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, v. 44, n. 1, p. 287ï 304, 2012.

KEULEERS, E. vwr: Useful functions for visual word recognition reserach. *R Package*, p. 19, 2013.

KEULEERS, E; BRYSBAERT, M. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, v. 42, n. 3, p. 627ï 633, 2010.

KEULEERS, E; DIEPENDAELE, K.; BRYSBAERT, M. Practice Effects in Large-Scale Visual Word Recognition Studies: A Lexical Decision Study on 14,000 Dutch Mono- and Disyllabic Words and Nonwords. *Frontiers in Psychology*, v. 1, 2010.

MARIAN, V. et al. CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE*, v. 7, n. 8, p. e43230, 2012.

MOTA, M. B.; RESENDE, N. Metodologia da pesquisa em psicolinguística: desenvolvimento de uma ferramenta para a geração automática de pseudoverbos. *Letras de Hoje*, v. 48, n. 1, p. 100ï 107, 2013.

NEW, B. et al. Une base de données lexicales du français contemporain sur internet¡ LEXIQUE^{TM}//A lexical database for contemporary french¡ LEXIQUE^{TM}. *Lôannée psychologique*, v. 101, n. 3, p. 447ï 462, 2001.

NEW, B. et al. Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, v. 36, n. 3, p. 516ï 524, 2004.

PINHEIRO, G. M.; ALUÍSIO, S. M. *Corpus NILC: descrição e análise crítica com vistas ao projeto Lacio - WebSérie de Relatórios do Núcleo Interinstitucional de Lingüística Computacional NILC - ICMC - USP*. São Carlos, SP: Universidade Federal de São Carlos - UFSCar, 2003.

SANTOS, D.; BICK, E. *Providing internet access to Portuguese corpora: the AC/DC project.* (M. Gavrilidou et al., Eds.)Proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000). *Anais*...Athens, Greece: 2000.

SCHREUDER, R; BAAYEN, R. H. Modeling Morphological Processing. In: FELDMAN, L. B. (Ed.). *Morphological Aspects of Language Processing*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers, p. 131ï 154, 1995.

YARKONI, T; BALOTA, D.; YAP, M. Moving beyond Coltheartôs N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, v. 15, n. 5, p. 971ï 979, 2008.

SOLETRAS
REVISTA
N. 33 ï 2017.1 ï GUSTAVO LOPEZ ESTIVALET
FANNY MEUNIER

The Brazilian Portuguese Lexicon Psycholinguistic Corpus

Abstract: The Brazilian Portuguese Lexicon was developed to offer a word-based corpus for psycholinguistic research in Brazilian Portuguese language. It was created from a corpus with more than 32 million words. Thus, the Brazilian Portuguese Lexicon contains more than 215 thousand lexical entries and presents 21 columns with relevant metalinguistic and psycholinguistic information, such as grammatical category, orthographic frequency, word length, and orthographic neighbors, among others. It is an open free-access corpus consulted on the Internet; it has a friendly and dynamic interface for simple and complex searches. The Brazilian Portuguese Lexicon still offers a series of data computed, a Brazilian Portuguese pseudo word generation engine, and a collection of statistical linguistic tools. Therefore, the present article aims to introduce and present the Brazilian Portuguese Lexicon, as well as to serve as a manual for its use. Further, a description of its development and creation is given. Finally, the Brazilian Portuguese Lexicon fills a huge gap in the research in psycholinguistics and computational linguistics, offering a word-based corpus with valuable metalinguistic and psycholinguistic information from Brazilian Portuguese language.

Keywords: Psycholinguistics. Computational Linguistics. Corpus. Lexicography. Linguistics. Brazilian Portuguese.

Recebido em: 29 de maio de 2017.

Aprovado em: 07 de junho de 2017.