Manual do Léxico do Português Brasileiro - LexPorBR: corpus psicolinguístico

versão Alfa2

Lyon, 6 de outubro de 2015.

Introdução

O principal objetivo do Léxico do Português Brasileiro - LexPorBR[1] é oferecer um corpus psicolinguístico 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

Corpora psicolinguísticos são utilizados 1) no controle, seleção e manipulação de palavras e critérios específicos para a criação de experiências psicolinguísticas e 2) em análises em linguística computacional da distribuição e do comportamento lexical (Baayen, 2001). Exemplos de corpora psicolinguísticos são: francês - *Lexique*[2](B. New, Pallier, Brysbaert, & Ferrand, 2004; B. New, Pallier, Ferrand, & Matos, 2001), espanhol – *BuscaPalabras* (Davis & Perea, 2005), inglês – MRC[3] (Coltheart, 1981), alemão, espanhol, francês, holandês e inglês - ClearPOND[4] (Marian, Bartolotti, Chabal, & Shook, 2012), alemão, cirílico, holandês e inglês - CELEX[5] (Baayen, Piepenbrock, & van Rijn, 1995). Esses corpora foram utilizados,

[1] http://www.lexicodoportugues.com/
[2] http://www.lexique.org/
[3] http://www.psych.rl.ac.uk/
[4] http://clearpond.northwestern.edu/
[5] http://celex.mpi.nl/

por exemplo, em megaestudos que investigam o comportamento psicolinguísticos no processamento de palavras e pseudopalavras, no *English Lexicon Project* (Balota et al., 2007; Boris New, Ferrand, Pallier, & Brysbaert, 2006), no *French Lexicon Project* (Ferrand et al., 2010), no *Dutch Lexicon Project* (Keuleers, Diependaele, & Brysbaert, 2010), e no *British Lexicon Project* (Keuleers, Lacey, Rastle, & Brysbaert, 2012). Esses corpora 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 modelos linguísticos(Schreuder & Baayen, 1995).

Léxico do Português Brasileiro

O Léxico do Português Brasileiro nasceu de uma ideia anotada em um *postit* no final de 2012 quando estava começando meu Doutorado em psicolinguística e neurociências, em Lyon, na França. Meu projeto de Doutorado tem como objetivo 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 através do corpus *Lexique* (B. New et al., 2004), que oferece uma série de informações indispensáveis para criação das experiências e análise dos resultados (frequência da forma, categoria gramatical, número de letras, número de vizinhos, forma invertida, estrutura CVCV, entre outras). No começo de 2013, quando começamos a preparar as experiências em PB, deparamo-nos com a completa falta de um corpus psicolinguístico do PB. Procurando por um corpus que suprisse nossas necessidades, tivemos acesso ao site do Linguateca[6] (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 neste momento que anotei em um *postit*: "fazer o léxico do português brasileiro". Atualmente, o Léxico do Português Brasileiro apresenta a página principal conforme a **Figura 1**.

[6] http://www.linguateca.pt/

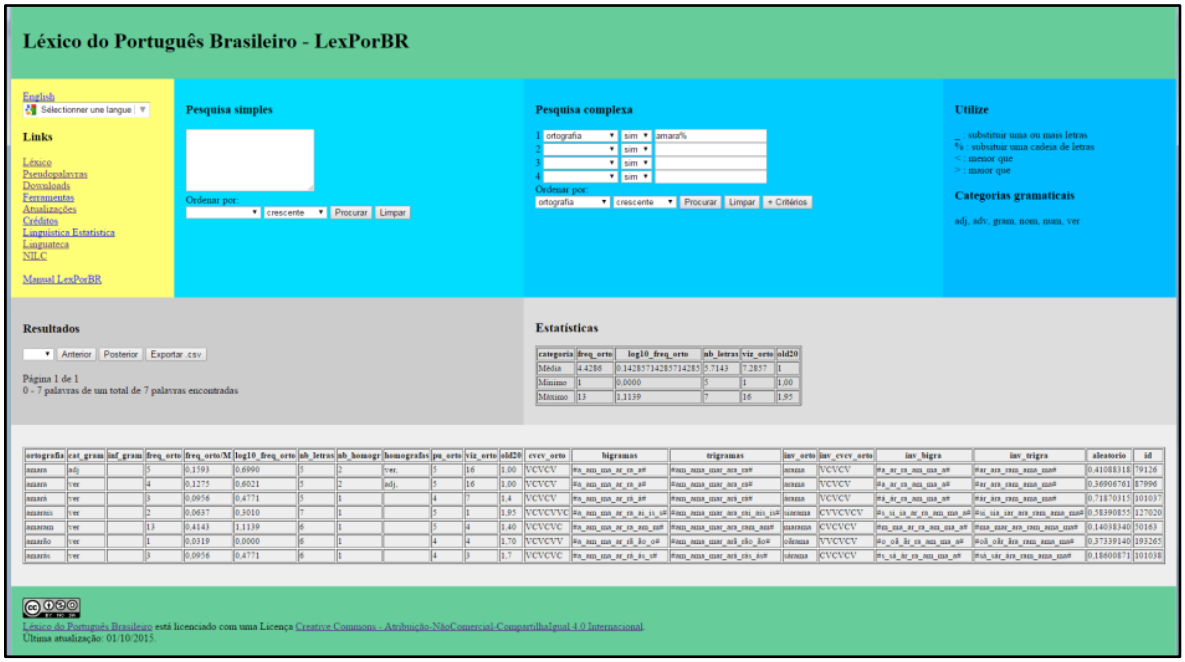

Figura 1: Página inicial do Léxico do Português Brasileiro.

Desenvolvimento

No começo de 2014, começou o desenvolvimento do Léxico do Português Brasileiro 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 Léxico do Português Brasileiro. Além disso, foram criadas as demais páginas do site: atualizações, créditos, downloads, ferramentas e links. Em seguida, foi desenvolvido o motor de geração de pseudopalavras do PB e os recursos e ferramentas de estatística linguística.

Atualizações

15/01/2013 – procura de um corpus psicolinguístico do PB. Conhecimento do Linguateca que hospeda uma série de corpora do português, porém nenhum corpus psicolinguístico do PB. Decisão de criar-se o Léxico do Português Brasileiro como um corpus metalinguístico e

psicolinguístico do PB baseado em palavras e de acesso livre e aberto através da internet. Conhecimentos necessários: R[7], HTML[8], MySQL[9], PHP[10], Java[11] e CSS[12].

21/03/2013 – pré-seleção de dois corpora do PB no Linguateca para o desenvolvimento do Léxico do Português Brasileiro: 1) Corpus Brasileiro[13] (1 bilhão de palavras, 3,2 GB) e 2) corpus do Núcleo Interdisciplinar de Linguística Computacional (NILC) de São Carlos (doravante NILC)[14,15] (32 milhões de palavras, 49 MB). Após discussão com os pesquisadores responsáveis desses corpora, chegamos à conclusão que o NILC seria mais pertinente para o desenvolvimento do Léxico do Português Brasileiro pelos seguintes critérios: 1) número 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.

14/08/2013 – desenvolvimento do corpus piloto do Léxico do Português Brasileiro com apenas os verbos do NILC, contabilizando cerca de 80 mil formas. Utilização do programa R para o desenvolvimento de 10 colunas de informações: 1) ortografia, 2) frequência da forma, 3) frequência por milhão de palavras, 4) log10 da frequência da forma, 5) número de letras, 6) categoria gramatical, 7) informações gramaticais, 8) forma ortográfica invertida, 9) estrutura CVCV e 10) estrutura CVCV invertida. Construção do site piloto do Léxico do Português Brasileiro através da utilização de servidor local com o programa XAMPP[16], contendo os módulos Apache, MySQL, PHP e Perls pré-instalados. Configuração e utilização do phpMyAdmin[17] para importação do corpus piloto em formato .csv para um banco de dados MySQL. Utilização do programa Notepad++[18] para a programação da página HTML piloto de interface entre usuário e banco de dados MySQL e da programação lógica em PHP.

[7] http://www.r-project.org/
[8] http://pt.wikipedia.org/wiki/HTML
[9] http://www.mysql.com/
[10] http://www.php.net/
[11] http://www.java.com/pt_BR/
[12] http://pt.wikipedia.org/wiki/Cascading_Style_Sheets
[13] http://corpusbrasileiro.pucsp.br/cb/Inicial.html
[14] http://www.nilc.icmc.usp.br/nilc/index.php
[15] http://www.linguateca.pt/acesso/corpus.php?corpus=SAOCARLOS
[16] http://www.apachefriends.org/pt_br/index.html
[17] http://www.phpmyadmin.net/home_page/index.php
[18] http://notepad-plus-plus.org/

28/10/2013 – versão piloto do site do Léxico do Português Brasileiro com dois motores de pesquisa: 1) pesquisa simples e 2) pesquisa complexa. A pesquisa simples contém uma área de texto onde se podem inserir múltiplas palavras em forma de lista. A pesquisa complexa contém quatro campos de inserção de critérios das palavras a serem pesquisadas. Cada motor de pesquisa foi desenvolvido com um botão "Procurar" para iniciar a pesquisa e apresentar os resultados e um botão "Limpar" para apagar os dados presentes nos campos. Definição das páginas do Léxico do Português Brasileiro: Léxico, Pseudopalavras, Downloads, Ferramentas, Atualizações, Créditos, Linguística Estatística, Linguateca e NILC.

30/11/2013 – programação de algoritmos em Java e PHP para manter os dados preenchidos nos campos da página HTML após pesquisa. Inserção de dois campos para organização dos resultados, um para seleção do critério de organização e outro para ordem crescente ou decrescente. Inserção do botão "+ Critérios" na pesquisa complexa para disponibilização de oito campos de pesquisa. Escolha do servidor de internet gratuito http://www.biz.nf/, pelos seguintes 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. Importação do corpus piloto no formato .csv para um banco de dados MySQL e envio dos arquivos por FTP com o programa FileZilla[19] para http://portugueselexicon.co.nf/.

12/12/2013 – tendo em vista que o MySQL reconhece os símbolos "_" para substituir uma letra e "%" para substituir uma cadeia de letras, esta informação foi acrescentada às instruções na página principal. Programação em PHP para reconhecimento dos símbolos: maior que ">" e menor que "<" para as pesquisas numéricas. Divisão do Léxico do Português Brasileiro em três versões: 1) Alfa, 2) Beta e 3) Delta. A versão Alfa (2014) é a primeira versão do Léxico do Português Brasileiro, disponibilizando um corpus puramente ortográfico. A versão Beta (prevista para 2015) conterá os dados fonológicos, silábicos e de lema. Finalmente, a versão Delta (prevista para 2016) apresentará informações específicas das palavras, como informações morfológicas e sintáticas, entre outras.

07/01/2014 – download dos 13 arquivos em formato .txt do corpus do NILC[20] no sitre do Linguateca separados por categorias gramaticais (6 arquivos de formas: adjetivos, advérbios, gramaticais, nomes, numerais e verbos; 7 arquivos de lemas: adjetivos, advérbios,

[19] https://filezilla-project.org/
[20] http://www.linguateca.pt/acesso/contabilizacao.php

gramaticais, nomes, nomes próprios, numerais e verbos). Comparação do número total de palavras e formas com os dados fornecidos no Linguateca. Criação de uma coluna com as respectivas categorias gramaticais (cat_gram) de cada arquivo (adjetivo, advérbio, gramatical, nome, nome próprio, numeral e verbo). Criação de uma coluna com o tipo de palavra (forma ou lema). Transformação de todas as palavras em letras minúsculas e soma de todas as formas repetidas. Criação de uma coluna com um número de identificação (id) da palavra de acordo com a organização do corpus por frequência em ordem decrescente, logo, o número de identificação (id) passou a ser também a da posição da palavra no léxico e no corpus.

08/01/2014 – criação de uma coluna com a frequência por milhão de palavras (freq_orto/M) através do cálculo [1000000 * freq_orto / freq_total]. Criação de uma coluna com o log natural da frequência. Criação de uma coluna com o número de letras das formas. Exclusão de todas as formas com mais de 30 letras e numerais, salvo 0-1, 1º-9º e 1ª-9ª. Criação de uma coluna com o número de formas homógrafas. Criação de uma coluna com as categorias gramaticais das formas homógrafas.

10/01/2014 - separação das palavras em letras, transformação das vogais em V e das consoantes em C, criação de uma coluna com a estrutura CVCV (CVCV_orto). Ainda, foram utilizadas as letras P para pontuação, N para números, A para acentos e S para símbolos. Criação de uma coluna com os bigramas das palavras. Criação de uma coluna com os trigramas das palavras.

18/01/2014 - desenvolvimento do algoritmo para o cálculo do ponto de unicidade ortográfico (pu_orto) e criação de uma coluna para o mesmo. Criação de uma coluna com o número de vizinhos ortográficos (viz_orto) (Coltheart, Davelaar, Jonasson, & Besner, 1977). Criação de uma coluna com a distância de Levenshtein ortográfica (old20) (Yarkoni, Balota, & Yap, 2008). Estas funções são disponibilizadas no pacote "vwr"[21] desenvolvido por Emmanuel Keuleers[22] para o programa R. Criação de uma coluna com um número aleatório entre 0 e 1 com oito dígitos de precisão.

28/01/2014 - criação de quatro colunas com as formas invertidas de: ortografia (inv_orto), CVCV_orto (inv_CVCV_orto), bigramas (inv_bigramas) e trigramas (inv_trigramas). Sendo assim, a versão Alfa do Léxico do Português Brasileiro conta com 21 colunas de informações

[21] http://cran.r-project.org/web/packages/vwr/index.html
[22] http://crr.ugent.be/members/emmanuel-keuleers

metalinguísticas e psicolinguísticas: 1) ortografia, 2) cat_gram, 3) inf_gram, 4) freq_orto, 5) freq_orto/Mo, 6) log10_freq_orto, 7) nb_letras, 8) nb_homogr, 9) homografas, 10) pu_orto, 11) viz_orto, 12) old20, 13) CVCV_orto, 14) bigramas, 15) trigramas, 16) inv_orto, 17) inv_CVCV_orto, 18) inv_bigra, 19) inv_trigra, 20) aleatorio e 21) id.

05/02/2014 – o corpus psicolinguístico Léxico do Português Brasileiro possui 21 colunas de informações e 215.175 linhas com diferentes palavras do PB. Essa tabela em formato .csv ficou com um tamanho de 45 MB. Esse arquivo foi dividido em 36 arquivos de aproximadamente 1,5 MB no formato .csv. Os arquivos .csv foram salvos novamente com codificação UTF-8 (pois possuíam codificação AINSI) afim de evitarem-se problemas com acentos e símbolos especiais. Cada um dos arquivos foi importado através do phpMyAdmin para nosso servidor, de forma que cada arquivo importado inflava o arquivo já existente.

12/02/2014 – desenvolvimento de um módulo de limitação e navegação dos resultados apresentados. O número de palavras a serem apresentadas pode ser 50, 100, 200 ou 500. Dois botões (“Anterior” e “Posterior”) para navegar entre as páginas de resultados. Apresentação de quatro informações gerais da pesquisa: 1) total de palavras encontradas, 2) total de páginas de resultados, 3) intervalo das palavras apresentadas e 4) página apresentada. Desenvolvimento do botão “Exportar .csv” para exportar o resultado da pesquisa realizada em um arquivo .csv disponibilizado para download do usuário.

18/02/2014 - desenvolvimento de um módulo de estatística básica do resultado da pesquisa efetuada (Davis, 2005) apresentando as seguintes informações: 1) média, 2) valor máximo e 3) valor mínimo das seguintes categorias: 1) freq_orto, 2) log10_freq_orto, 3) nb_letras, 4) viz_orto e 5) old20.

23/02/2014 – desenvolvimento de um motor de geração de pseudopalavras do PB a partir dos bigramas ou trigramas. Diferentemente de outros motores geradores de pseudopalavras (Keuleers & Brysbaert, 2010; Mota & Resende, 2013), contabilizamos a frequência geral dos bigrama e trigrama, a frequência de cada bigrama e trigrama de acordo à posição na palavra e por categoria gramatical, para a geração de pseudopalavras do PB. Obtenção de duas tabelas para o banco de dados de geração de pseudopalavras do PB, uma com os bigramas e outra com os trigramas 1) gerais, 2) por posição na palavra e 3) por categoria gramatical.

05/03/2014 – criação de um motor de geração de pseudopalavras onde o usuário deve inserir quatro campos: 1) número de letras das pseudopalavras a serem geradas, 2) número de pseudopalavras a serem geradas, 3) categoria gramatical de base que estas pseudopalavras devem pertencer (todas, adj, adv, gram, nom, num, ver) e 4) tipo de critério para a construção das pseudopalavras (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 "#xx" ou "xx#". De acordo com o número de letras, o motor vai concatenando novos bigramas ou trigramas que dividam o máximo de informação ortográfica com o bigrama ou trigrama anterior (1 letra para os bigramas e 2 letras para os trigramas).

18/03/2014 – apresentação de quatro colunas com dados sobre os resultados das pseudopalavras: 1) categoria gramatical, 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 das pseudopalavras. Tradução da página principal de pesquisa simples e pesquisa complexa para o inglês.

25/03/2014 – registro do domínio próprio do Léxico do Português Brasileiro (www.lexicodoportuguês.com) junto ao HostGator[23] e redirecionamento deste domínio para o servidor onde o Léxico do Português Brasileiro está hospedado http://portugueselexicon.co.nf. Inauguração oficial do Léxico do Português Brasileiro em 25 de março de 2014.

21/03/2015 - desenvolvimento da página Linguística Estatística com diversas ferramentas e recursos abertos e gratuítos em HTML/PHP para análise linguística e estatística: F1, F2, F', minF', teste de Hartley, normalização entre 0-1, inverter palavra, distância de Hamming, distância de Levenshtein, vizinhos ortográficos (Coltheart's N), média das distâncias de Levenshtein, entropia relativa, frequência de palavras, distribuição Zipf, etc.

20/09/2015 - escritura e disponibilização do Manual do Léxico do Português Brasileiro - Alfa 2 em português brasileiro e escritura e disponibilização do Brazilain Portuguese Lexicon - Alpha 2 Manual em inglês. Tradução e implementação de todas as páginas e informações do Léxico do Português Brasileiro em inglês (Brazilian Portuguese Lexicon). Implementação do Google Tradutor em todas as páginas do Léxico do Português Brasileiro para a tradução do site para as línguas disponibilizadas no Google Tradutor. Sugere-se que o Google Tradutor

[23] http://hostgator.com.br/

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 apresentadoe em PB.

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 (2015) e 3) Delta (2016). Atualmente o Léxico do Português Brasileiro está na versão Alfa, inaugurada em 25 de março de 2014. A versão Alfa marca a criação do Léxico do Português Brasileiro e o surgimento do primeiro corpus psicolinguístico do PB. A principal característica da versão Alfa do Léxico do Português Brasileiro é que ela disponibiliza um corpus ortográfico em que as informações disponibilizadas foram computadas a partir de dados ortográficos das palavras do PB. Futuramente, a versão Beta contará com as informações: 1) fonológicas das formas, 2) silábicas das formas e 3) dos lemas associados às formas. A versão Delta contará também com uma série de: 1) informações morfológicas, 2) informações sintáticas, e na medida do possível, 3) medidas de tempo de reação do reconhecimento de um grande número de palavras e pseudopalavras do PB, seguindo os modelos dos *Lexicon Projects* (Balota et al., 2007; Ferrand et al., 2010; Keuleers et al., 2010, 2012).

Licença Creative Commons

[24] http://www.researchgate.net/profile/Gustavo_Estivalet
[25] http://creativecommons.org/licenses/by-nc-sa/4.0/

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) (Pinheiro & Aluísio, 2003) sediado no Instituto de Ciências Matemáticas e de Computação de São Carlos (ICMC/São Carlos)[26], da Universidade de São Paulo em São Carlos (USP/São Carlos)[27]. As listas de formas e lemas divididas em categorias gramaticais foram baixadas do site do Linguateca, onde encontra-se uma série de informações do NILC, como dados quantitativos e estatísticos[28], descendência do corpus[29] e, principalmente, os arquivos de formas[30] e lemas[31] no formato .txt, separados por categorias gramaticais.

"Todo o material que disponibilizamos não é restrito a nenhum grupo e foi autorizado (nos termos em que o disponibilizamos) pelos respectivos autores ou detentores de direitos de autor. De recurso para recurso as condições são diferentes, estando especificadas na documentação de cada um deles. As ferramentas criadas pela Linguateca são disponibilizadas nos termos da Licença pública geral GNU[32]."[33]

Lexique

A criação e o desenvolvimento do Léxico do Português Brasileiro foram inspirados no corpus psicolinguístico do francês *Lexique* (B. New et al., 2004, 2001). O *Lexique* tem oferecido dados sobre as palavras do francês a uma série de estudos e pesquisas, sendo um ótimo exemplo de corpus psicolinguístico. Esse corpus exemplifica as funcionalidades e utilidades que um corpus psicolinguístico deve e pode oferecer como recursos para a pesquisa em psicolinguística e linguística computacional. Uma descrição detalhada desse corpus é encontrada no manual do *Lexique*[34].

[26] http://www.icmc.usp.br/Portal/
[27] http://www.saocarlos.usp.br/
[28] http://www.linguateca.pt/acesso/desc_corpus.php?corpus=SAOCARLOS
[29] http://www.linguateca.pt/acesso/NILCsaocarlos.html
[30] http://www.linguateca.pt/acesso/contabilizacao.php#listaPosSAOCARLOS
[31] http://www.linguateca.pt/acesso/contabilizacao.php#listaLemasSAOCARLOS
[32] http://www.gnu.org/copyleft/gpl.html
[33] http://www.linguateca.pt/FAQ/#faq1.8
[34] http://www.lexique.org/docLexique.php

Programa e pacotes R

O Léxico do Português Brasileiro foi desenvolvido com o programa R, com os dados originais importados a partir de arquivos .txt e cada coluna sendo criada e computada através de determinadas funções e algorítmos. O número de vizinhos ortográficos (Coltheart's N) (Coltheart et al., 1977) e a distância de Levenshtein ortográfica das 20 palavars mais próximas (OLD20) (Yarkoni et al., 2008) foram calculados a partir das funções "coltheart.N" e "old20" disponibilizadas no pacote "vwr" desenvolvido por Emmanuel Keuleers. Uma série de funções do pacote "languageR"[35] desenvolvido por Harald Baayen[36] também foram utilizadas no desenvolvimento do Léxico do Português Brasileiro.

LexPorBR - Alfa

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: **"<"** menor que, **">"** maior que, **"_"** substitui uma letra, **"%"** substitui uma cadeia de letras.

- Ordem de apresentação dos resultados: **crescente** – apresenta os resultados na ordem crescente, **decrescente** – apresenta os resultados na ordem decrescente.

[35] https://cran.r-project.org/web/packages/languageR/index.html
[36] http://www.sfs.uni-tuebingen.de/~hbaayen/

- 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.

Colunas

O Léxico do Português Brasileiro versão Alfa apresenta 215.175 linhas com diferentes entradas lexicais e 21 colunas com diferentes informações metalinguísticas e psicolinguísticas. Sendo assim, cada linha do Léxico do Português Brasileiro contém uma palavra e cada coluna uma determinada informação sobre esta palavra.

Um exemplo de pesquisa a partir da pesquisa complexa com o critério **cat_gram – sim - ver** pode ser visualizado na **Figura 2**. Essa pesquisa apresenta palavras que possuem a categoria gramatical definida como “verbo”.

ortografia	cat_gram	inf_gram	freq_orto	freq_orto/M	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'#	0,35822656	128902
'banja	ver		26	0,8386	1,4150	6	1		5	12	1,20	SCVCCV	#'_b_ba_an_nj_ja_a#	#'_b_ba_ban_anj_nja_a#	ajnab'	VCCVCS	#a_aj_jn_na_ab_b'_'#	#aj_ajn_jna_nab_ab'_b'#	0,08662146	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_a#	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_e#	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	nom.	5	22	1	SCVCCV	#'_b_ba_ar_rr_ro_o#	#'_b_ba_bar_arr_rro_o#	orrab'	VCCVCS	#o_or_rr_ra_ab_b'_'#	#or_orr_rra_rab_ab'_b'#	0,05634216	102039
'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_ço_ob_b'_'#	#aç_aço_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_aço_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_aço_çob_ob'_b'#	0,68417173	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_i#	#'_b_bo_boc_oce_cei_ei#	iecob'	VCVCVS	#i_ie_ec_co_ob_b'_'#	#ie_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'	VCVCVS	#u_uo_oç_ço_ob_b'_'#	#uo_uoç_oço_çob_ob'_b'#	0,52948808	34251
'boroe	ver		1	0,0319	0,0000	6	1		4	3	1,80	SCVCVV	#'_b_bo_or_ro_oe_e#	#'_b_bo_bor_oro_roe_oe#	eorob'	VCVCVS	#e_eo_or_ro_ob_b'_'#	#eo_eor_oro_rob_ob'_b'#	0,16018517	201255
'cala	ver		68	2,1671	1,8325	5	2	nom.	4	59	1,00	SCVCV	#'_c_ca_al_la_a#	#'_c_ca_cal_ala_la#	alac'	CVCVCS	#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	60556
'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'	VCCVCS	#e_ed_dl_la_ac_c'_'#	#ed_edl_dla_lac_ac'_c'#	0,16646319	202411
'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_ie_el_la_ac_c'_'#	#ie_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'#	0,10601483	202510
'cama	ver		1	0,0319	0,0000	5	2	nom.	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'	VCCVCS	#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'#	0,52822827	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	nom.	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_ie_ep_pa_ac_c'_'#	#ie_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'#	0,49069938	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	148578
'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'	CCVCS	#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'#	0,80872251	103022
'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	35839
'coará	ver		1	0,0319	0	6	1		4	4	1,6	SCVCCV	#'_c_co_oa_ar_rá_á#	#'_c_co_coa_oar_ará_á#	áraoc'	VCCVCS	#á_ár_ra_ao_oc_c'_'#	#ár_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_oav_ava_va#	avaoc'	VCVVCS	#a_av_va_ao_oc_c'_'#	#av_ava_vao_aoc_oc'_c'#	0,2472736	205924

Figura 2: Exemplo de resultado de pesquisa.

Segue abaixo o nome, a abreviação e a descrição das 21 colunas de informações apresentadas nos resultados de uma pesquisa no Léxico do Português Brasileiro, versão Alfa.

Ortografia (orto): forma ortográfica da palavra em letras minúsculas (com exceção dos nomes próprios), respeitando os acentos específicos de cada palavra[37].

Categoria gramatical (cat_gram): categorial gramatical da palavra (adj, adv, gram, nom, num, prop, ver).

Informação gramatical (inf_gram): informações gramaticais sobre a palavra (ex. singular/plural, masculino/feminino, passado/presente/futuro, 1/2/3 pessoas, etc.).

Frequência ortográfica (freq_orto): número de vezes que a palavra aparece no NILC (cerca de 32 milhões de palavras).

Frequência ortográfica por milhão (freq_orto/M): número de vezes que a palavra aparece entre 1 milhão de palavras. Valor padrão para frequência de palavras.

Logaritmo natural da frequência ortográfica (log10_freq_orto): logarítmico natural da frequência ortográfica. Os valores logarítmicos são utilizados para linearizar-se o comportamento das frequências das palavras no corpus.

Número de letras (nb_letras): número de letras da palavra.

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.

Homógrafas (homografas): categorias gramaticais das palavras homógrafas.

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.

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).

[37] Destaca-se que o NILC foi realizado em 1999, antes da reforma ortográfica do português.

Distância de Leveinshtein ortográfica (old20): distância ortográfica de Leveinshtein das 20 palavras mais póximas calculadas a partir de regressões lineares (Yarkoni et al., 2008).

Estrutura CVCV (CVCV_orto): estrutura CVCV da palavra, onde consoantes são C e vogais são V. Ainda, A para acentos, P para pontuação, N para números e S para símbolos.

Bigramas (bigramas): bigramas que constituem a palavra separados por "_" e limitados por "#". O número de bigramas é igual ao número de letras da palavra mais 1.

Trigramas (trigramas): trigramas que constituem a palavra separados por "_" e limitados por "#". O número de trigramas é igual ao número de letras da palavra.

Ortografia invertida (inv_orto): forma invertida da ortografia (orto).

Estrutura CVCV invertida (inv_CVCV_orto): estrutura CVCV da palavra invertida a partir de (CVCV_orto).

Bigramas invertidos (inv_bigra): bigramas que constituem a palavra separados por "_" e limitados por "#" invertidos a partir de (bigramas).

Trigramas invertidos (inv_trigra): trigramas que constituem a palavra separados por "_" e limitados por "#" invertidos a partir de (trigramas).

Número aleatório entre 0 e 1 (aleatorio): número aleatório entre 0 e 1 com oito algarismos de precisão.

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 e no léxico.

Pesquisa simples

O motor de pesquisa simples foi desenvolvido para a pesquisa de palavras específicas ou listas de palavras, conforme a **Figura 3**. O usuário deve realizar a pesquisa a partir da forma ortográfica das palavras, contudo, os símbolos coringas "_" para uma letra e "%" para uma cadeia de letras podem ser utilizados. O usuário pode inserir uma lista de palavras separadas

em diferentes linhas. Por exemplo, pode-se copiar e colar uma lista de palavras de uma planilha ou editor de texto. O usuário pode escolher a categoria utilizada para a organização e apresentação das palavras e o sentido de organização crescente ou decrescente. O botão "Procurar" realiza a pesquisa e apresenta os resultados e o botão "Limpar" limpa as informações dos campos.

Figura 3: Pesquisa simples.

Pesquisa complexa

O motor de pesquisa complexa foi desenvolvido para a realização de pesquisas complexas a partir de critérios específicos das palavras, como número de letras, frequência, categoria gramatical, vizinhos ortográficos, etc., conforme a **Figura 4**. 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 "sim" ou desconsiderar "não" o critério. No terceiro campo, o usuário deve inserir os critérios específicos de sua pesquisa. Os símbolos coringas "_" para uma letra e "%" para uma cadeia de letras podem ser utilizados. Ainda, os símbolos maior que ">" e menor que "<" podem ser utilizados para pesquisas numéricas de grupos de palavras. O usuário pode escolher a categoria utilizada para a organização e apresentação das palavras e o sentido de organização crescente ou decrescente. O botão "Procurar" realiza a pesquisa e apresenta os resultados e o botão "Limpar" limpa as informações dos campos. Inicialmente, a pesquisa complexa apresenta quatro campos de critérios para a pesquisa, clicando-se no botão "+ Critérios", o usuário é enviado a uma página que apresenta oito campos de critérios para a pesquisa.

Figura 4: Pesquisa complexa.

Resultados

Na seção de resultados (**Figuras 2** e **5**), o usuário encontra os resultados da pesquisa organizada em diferentes linhas e com as informações metalinguísticas e psicolinguísticas nas diferentes colunas. Encontram-se ainda uma série de informações pertinentes à pesquisa, conforme a **Figura 5**: 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 "Anterior" e "Próximo".

Na parte superior a direita dos resultados, conforme **Figura 5**, apresenta-se uma série de dados estatísticos estabelecidos e calculados a partir da pesquisa realizada (Davis & Perea, 2005; Davis, 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 estatístico serão inseridos neste módulo. Por fim, o botão "Exportar .csv" exporta todos os dados da pesquisa para um arquivo .csv disponibilizado para download do usuário.

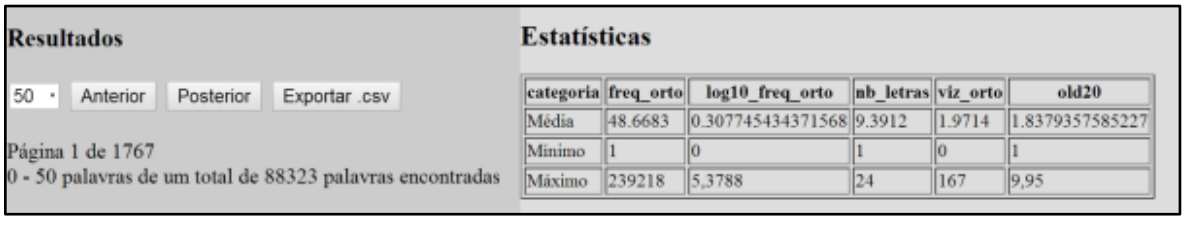

Figura 5: Informações dos resultados e estatísticas básicas.

Páginas

Além das páginas principais do Léxico do Português Brasileiro: Léxico[38] e Pseudopalavras[39], as seguintes páginas ainda foram criadas para complementar o website: Downloads, Ferramentas, Atualizações, Créditos e Linguística estatística. Downloads[40] disponibiliza uma série de arquivos pertinentes do Léxico do Português Brasileiro para downloads (como: corpus.txt, manuais, listas, convenções, bigramas, trigramas, scripts em R, etc.). Ferramentas [41] disponibiliza uma série links de corpora, programas e literatura em psicolinguística e linguística computacional. Atualizações[42] descreve o desenvolvimento do Léxico do Português Brasileiro e as atualizações que são realizadas com o tempo. Créditos[43] apresenta o objetivo, a origem e os autores do Léxico do Português Brasileiro; ainda descreve as referências e pertinência do corpus do NILC/São Carlos, 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 abertos e livre, conforme descritos abaixo.

Pseudopalavras

O motor gerador de pseudopalavras do PB foi desenvolvido para a criação de pseudopalavras baseadas na estrutura e frequência das palavras (Keuleers & Brysbaert, 2010) 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 (B. 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.

[38] http://www.lexicodoportugues.com/index.php
[39] http://www.lexicodoportugues.com/pseudowords.php
[40] http://www.lexicodoportugues.com/downloads.php
[41] http://www.lexicodoportugues.com/tools.php
[42] http://www.lexicodoportugues.com/updates.php
[43] http://www.lexicodoportugues.com/credits.php

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 "#xx" ou "xx#". 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: "Enviar" para gerar e apresentar os resultados das pseudopalavras e "Limpar" para limpar os dados dos campos, conforme a **Figura 6**.

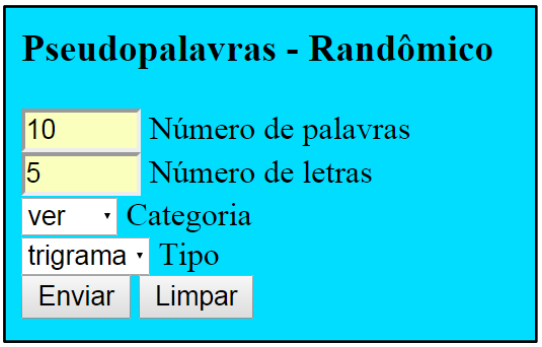

Figura 6: Motor de geração de pseudopalavras do PB.

Na tabela de resultados da geração de pseudopalavras do PB, conforme a **Figura 7**, 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 "Exportar .csv" 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 7: Resultados da geração de pseudopalavras do PB.

Linguística Estatística

A página Linguística Estatística[44] 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: a) F' e minF' – MS, b) minF' – F1.F2, c) teste de Hartley, d) normalização entre 0 e 1, e) inversor de palavras, f) distância de Hamming, g) distância de Levenshtein, h) vizinhos ortográficos (Coltheart's N), i) média das distâncias de Levenshtein, j) entropia relativa, k) frequência de palavras e l) distribuição de Zipf.

Autores

O Léxico do Português Brasileiro foi e está sendo desenvolvido por Gustavo Lopez Estivalet durante a realização de seu Doutorado, financiado com bolsa de Doutorado Pleno no Exterior (GDE) do Programa Ciências sem Fronteiras (CsF)[45] do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)[46], Brasil, entre 2012 e 2016, e, sua

[44] http://www.lexicodoportugues.com/stat_ling.php
[45] http://www.cienciasemfronteiras.gov.br/web/csf
[46] http://www.cnpq.br/

Orientadora Prof. Dr. Fanny Meunier[47], financiada pelo *Conseil National de la Recherche Scientifique* (CNRS)[48], França. Os dois pesquisadores desenvolvem atualmente seus trabalhos de pesquisa na França, na cidade de Lyon, na *Université Claude Barnard Lyon 1* (UCBL) [49] junto a *École Doctorale de Neurosciences et Cognition* (ED NSCo) [50] no *Laboratoire sur le Langage, le Cerveau et la Cognition* (L2C2)[51], localizado no *Institut de Sciences Cognitives* (ISC) [52].

Agradecimentos

Para a realização e êxito do Léxico do Português Brasileiro, agradeço o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de Doutorado Pleno no Exterior (GDE) do Programa Ciências sem Fronteiras (CsF). Agradeço minha Orientadora Prof. Dr. Fanny Meunier e o Prof. Dr. Michel Hoen[53] pela compreensão da importância de um corpus psicolinguístico do PB. Agradeço às pesquisadoras do NILC/São Carlos Prof. Dr. Sandra M. Aluísio[54] e Prof. Dr. Maria das Graças Volpe Nunes[55] pelos valiosos materiais, informações e auxílio sobre o NILC, assim como o apoio na realização deste trabalho. Agradeço aos colegas Léo Varnet[56] e Emmanuel Trouche[57] pelas discussões sobre os scripts e 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. Agradeço à Prof. Dr. Mailce Borges Mota e a melhor professora de PB Prof. Lise Lopez. Finalmente, agradeço à Luanda Lins por compreender a importância deste projeto para mim e minha motivação em fazê-lo.

[47] http://www.researchgate.net/profile/Fanny_Meunier/
[48] http://www.cnrs.fr/
[49] http://www.univ-lyon1.fr/
[50] http://nsco.universite-lyon.fr/
[51] http://l2c2.isc.cnrs.fr/fr/
[52] http://www.isc.cnrs.fr/
[53] http://www.researchgate.net/profile/Michel_Hoen/
[54] http://www.researchgate.net/profile/Sandra_Aluisio/
[55] http://www.researchgate.net/profile/Maria_Nunes10/
[56] http://www.researchgate.net/profile/Leo_Varnet/
[57] http://cnrs.academia.edu/EmmanuelTrouche

Referências

Baayen, H. R. (2001). *Word Frequency Distributions* (Vol. 18). Dodrecht; Boston; London: Kluwer Academic Publishers.

Baayen, H. R., Piepenbrock, R., & van Rijn, H. (1995). *The CELEX lexical database. Release 2 [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. doi:10.3758/BF03193014

Brysbaert, M., & New, B. (2009). Moving beyond Kučera 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*, *41*(4), 977–990. doi:10.3758/BRM.41.4.977

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. doi:10.1080/14640748108400805

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI* (pp. 535–555). Hillsdale, NJ: Lawrence Erlbaum Associates.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, *37*(1), 65–70. doi:10.3758/BF03206399

Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, *37*(4), 665–671. doi:10.3758/BF03192738

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., … Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488–496. doi:10.3758/BRM.42.2.488

Gimenes, M., & New, B. (2015). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*. doi:10.3758/s13428-015-0621-0

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. doi:10.3758/BRM.42.3.627

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). 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*, *1*. doi:10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. doi:10.3758/s13428-011-0118-4

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE*, *7*(8), e43230. doi:10.1371/journal.pone.0043230

Mota, M. B., & Resende, N. (2013). Metodologia da pesquisa em psicolinguística: desenvolvimento de uma ferramenta para a geração automática de pseudoverbos. *Letras de Hoje*, *48*(1), 100–107.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*(1), 45–52. doi:10.3758/BF03193811

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516–524. doi:10.3758/BF03195598

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet : LEXIQUE^{TM}//A lexical database for contemporary french : LEXIQUE^{TM}. *L'année Psychologique*, *101*(3), 447–462. doi:10.3406/psy.2001.1341

Pinheiro, G. M., & Aluísio, S. M. (2003). *Corpus NILC: descrição e análise crítica com vistas ao projeto Lacio - Web. Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional NILC - ICMC - USP*. São Carlos, SP: Universidade Federal de São Carlos - UFSCar.

Santos, D., & Bick, E. (2000). Providing internet access to Portuguese corpora: the AC/DC project. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, & G. Stainhauer (Eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000)* (pp. 205–210). Athens, Greece.

Schreuder, R., & Baayen, R. H. (1995). Modeling Morphological Processing. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing* (pp. 131–154). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. doi:10.3758/PBR.15.5.971