Manual do Léxico do Português:
corpus psicolinguístico do português brasileiro

versão Alfa1

Lyon, 23 de abril de 2014.

Introdução

O principal objetivo do Léxico do Português [1] é oferece 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 foi construído com o intuito de ser um corpus aberto, com acesso livre, consultado em uma plataforma simples, intuitiva e dinâmica através da internet. A partir de uma determinada 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 têm sido utilizados principalmente para 1) controle, seleção e manipulação de palavras e critérios específicos para a criação de experiências psicolinguísticas e 2) análise em linguística computacional da distribuição e do comportamento de um determinado léxico (R. Harald Baayen, 2008). Alguns exemplos de corpora psicolinguísticos são: francês - *Lexique*[2] (Matos, Ferrand, Pallier, & New, 2001; New, Pallier, Brysbaert, & Ferrand, 2004), espanhol – *BuscaPalabras* (Davis & Perea, 2005), inglês – MRC[3] (Max Coltheart, 1981), alemão, espanhol, francês, holandês e inglês - ClearPOND[4]

[1] http://www.lexicodoportugues.com/
[2] http://www.lexique.org/
[3] http://www.psych.rl.ac.uk/

(Marian, Bartolotti, Chabal, & Shook, 2012), alemão, cirílico, holandês e inglês - CELEX[5] (R. H. Baayen, Piepenbrock, & van Rijn, 1995). Estes corpora foram utilizados, por exemplo, em megaestudos psicolinguísticos que investigam o comportamento humano no reconhecimento e processamento de palavras e pseudopalavras, no *English Language Project* (Balota et al., 2007; New, Ferrand, Pallier, & Brysbaert, 2006) e no *French Language Project* (Ferrand et al., 2010). Estes corpora são largamente utilizados na seleção, controle e manipulação de palavras para a criação de experiências psicolinguísticas em inúmeros estudos e pesquisas específicas (R. Harald Baayen, 2008), assim como no desenvolvimento e simulação de modelos linguísticos (Schreuder & Baayen, 1995).

Léxico do Português

O Léxico do Português nasceu de uma ideia anotada em um *postit* no final de 2012 quando estava começando meu Doutorado em Lyon, na França. Meu projeto de Doutorado está compreendido nas áreas de psicolinguística e neurociências, tendo como objetivo investigar a representação e o processamento morfológico flexional verbal no PB, no francês e em bilíngues com o PB como língua materna e o francês como língua estrangeira. A preparação de experiências demanda um enorme controle e manipulação das características metalinguísticas e psicolinguísticas das palavras utilizadas como estímulos. Para as experiências em francês, os estímulos foram selecionados através do corpus *Lexique* (New et al., 2004), que oferece uma série de informações indispensáveis e outras muito interessantes para a criação das experiências e análise dos resultados (tais como: 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 uma situação completamente inesperada da completa falta de existência de um corpus psicolinguístico do PB. Procurando por um corpus que suprisse às nossas necessidades, tivemos acesso ao Linguateca[6] (Santos & Bick, 2000), que por sua vez, é um site que reúne vários corpora do português europeu e brasileiro. Entretanto, mesmo no Linguateca, não foi possível encontrar nenhum corpus do português brasileiro com dados metalinguísticos e psicolinguísticos apropriados para a criação rigorosa de experiências

[4] http://clearpond.northwestern.edu/
[5] http://celex.mpi.nl/
[6] http://www.linguateca.pt/

psicolinguísticas em PB. Foi neste momento que anotei em um *postit*: "fazer o léxico do português". Após todo o desenvolvimento do Léxico do Português descrito abaixo neste manual, o Léxico do Português apresenta a página principal conforme a Figura 1.

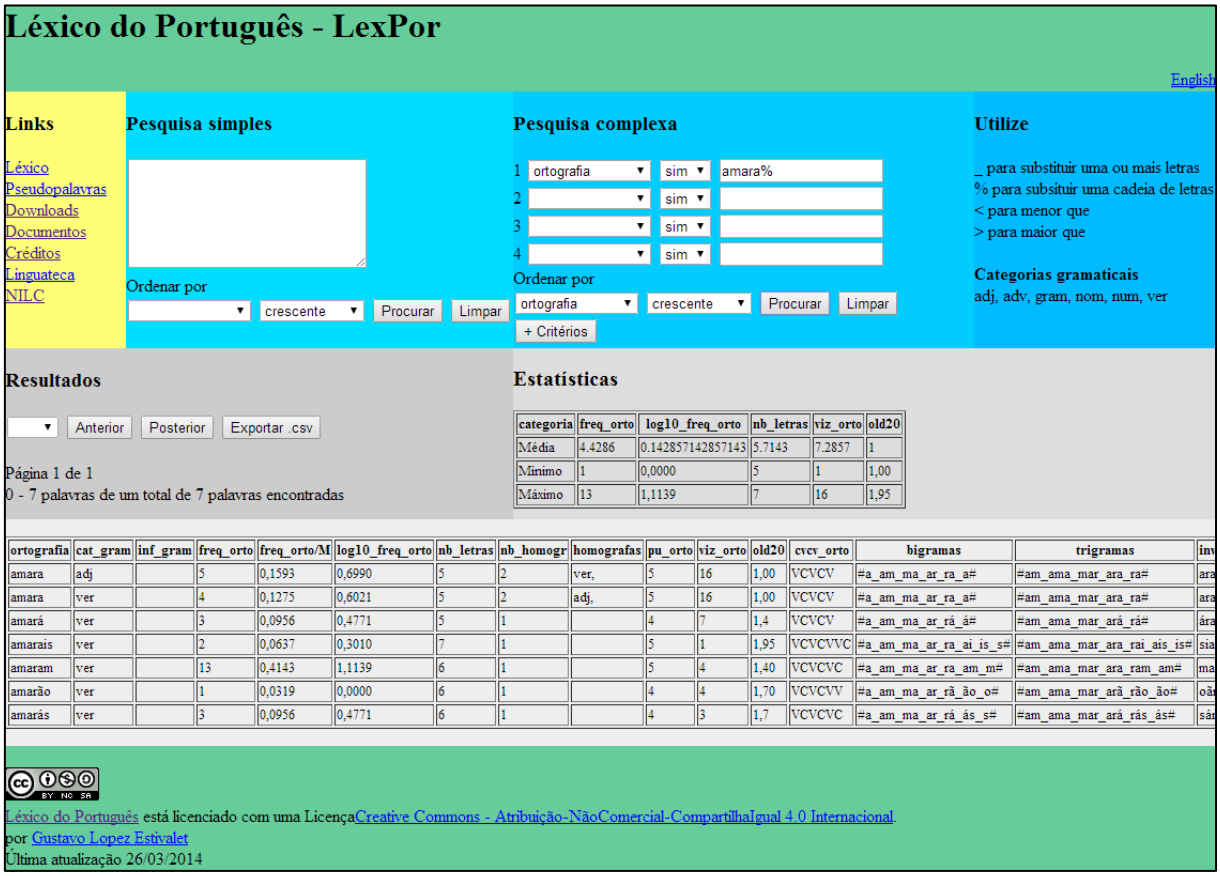


Figura 1: Página inicial do Léxico do Português.

Desenvolvimento do Léxico do Português

No começo de 2014, começou o desenvolvimento do Léxico do Português, sendo dividido em quatro etapas: 1) construção do corpus com as listas de palavras contendo todas as informações metalinguísticas e psicolinguísticas possíveis de serem computadas, 2) construção da página na internet em HTML para a interface entre o usuário e o banco de dados, 3) importação do corpus para um banco de dados MySQL em um servidor de internet e 4) programação lógica em PHP do funcionamento do Léxico do Português. Além disso, foram criadas as demais páginas do site: créditos, downloads, documentos, links, etc. Ao final, foi desenvolvido o motor de geração de pseudopalavras do PB.

História

15/01/2013 – procura de um corpus psicolinguístico do PB para a construção de experiências psicolinguísticas. 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 como um corpus metalinguístico e psicolinguístico do PB baseado em palavras e de acesso livre através da internet. Conhecimentos necessários para o desenvolvimento do mesmo: R[7], HTML[8], MySQL[9], PHP[10], Java[11] e CSS[12].

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

14/08/2013 – desenvolvimento do corpus piloto do Léxico do Português apenas com os verbos do CN, 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 através da utilização de servidor local com o programa XAMPP[16], contendo os

[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

módulos Apache, MySQL, PHP e Perls pré-instalados. Configuração e utilização do phpMyAdmin[17] para importação do corpus piloto salvo 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 de programação lógica em PHP.

28/10/2013 – versão piloto do Léxico do Português com dois motores de pesquisa: 1) pesquisa simples e 2) pesquisa complexa. A pesquisa simples foi constituída de uma área de texto onde se podem inserir múltiplas palavras em forma de lista. A pesquisa complexa foi constituída de 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 todos os dados presentes nos campos. Definição das páginas do Léxico do Português: Léxico – pesquisa lexical, Pseudopalavras - geração de pseudopalavras do PB, Downloads – arquivos para download, Documentos – documentos do desenvolvimento do Léxico do Português, Créditos – créditos, Linguateca - Linguateca e NILC - NILC/São Carlos.

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 de apresentação. Inserção do botão "+ Critérios" na pesquisa complexa para disposição 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 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 por FTP com o programa FileZilla[19] de todas as páginas criadas em HTML e PHP para http://portugueselexicon.co.nf/. Bom funcionamento geral de todas as funcionalidades do site.

12/12/2013 – tendo em vista que o próprio 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:

[17] http://www.phpmyadmin.net/home_page/index.php
[18] http://notepad-plus-plus.org/
[19] https://filezilla-project.org/

maior que “>” e menor que “<” para as pesquisas numéricas. Bom funcionamento do reconhecimento de "<" e ">" para pesquisas por conjuntos de características numéricas. Limitação e divisão do Léxico do Português em três versões: 1) Alfa, 2) Beta e 3) Delta. A versão Alfa (2014) marca a primeira versão do Léxico do Português, 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 uma série de funcionalidades específicas como classificação e decomposição morfológica, posição e papel temático sintático, entre outras.

07/01/2014 – desenvolvimento do Léxico do Português versão Alfa. Download dos 13 arquivos em formato .txt do corpus do CN [20] no Linguateca 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). Comparação do número total de palavras e formas de todos os arquivos 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 o número da posição da palavra no léxico.

08/01/2014 – criação de uma coluna com a frequência por milhão (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, para ser mais fácil identificar-se palavras que contenham esses caracteres. Criação de uma coluna

[20] http://www.linguateca.pt/acesso/contabilizacao.php

com os bigramas das palavras concatenando as letras duas a duas. Criação de uma coluna com os trigramas das palavras concatenando as letras três a três. Criação de duas novas colunas com os números de bigramas (nb_bigramas) e trigramas (nb_trigramas).

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) (M. 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 para cada uma das palavras do corpus.

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 conta com 21 colunas de informações metalinguísticas e psicolinguísticas: 1) ortografia, 2) cat_gram, 3) inf_gram, 4) freq_orto, 5) freq/M_orto, 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. Futuramente, a versão Beta do Léxico do português conterá ainda 1) forma fonológica e categorias derivadas, 2) silabificação e categorias derivadas e 3) associação forma e lema com frequência do lema.

05/02/2014 – ao final, o corpus psicolinguístico Léxico do Português possui 21 colunas de informações e 215.175 linhas com diferentes palavras do PB. Esta tabela em formato .csv ficou com um tamanho de 45 MB. Este arquivo foi dividido em 36 arquivos de aproximadamente 1,5 MB no formato .csv. Em seguida, os arquivos .csv foram salvos novamente no Bloco de Notas com codificação UTF-8 (pois antes os arquivos possuíam a codificação AINSI), afim de evitar-se problemas com acentos e símbolos especiais. Cada um dos arquivos foi importado através do phpMyAdmin de nosso servidor, de forma que cada novo arquivo importado ia inflando o arquivo já existente. Teste geral e verificação do perfeito funcionamento do Léxico do Português.

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

12/02/2014 – desenvolvimento de um módulo de limitação dos dados apresentados com botões de navegação das páginas de resultados e dados gerais da pesquisa realizada. 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 junto a esse módulo: 1) total de palavras encontradas, 2) total de páginas de resultados, 3) intervalo de palavras apresentadas e 4) página apresentada. Desenvolvimento do botão "Exportar .csv" para exportar todo 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. Futuramente, mais dados estatístico serão inseridos neste módulo. Sendo assim, considerou-se encerrado o desenvolvimento da versão Alfa do Léxico do Português em 27 de fevereiro de 2014, inaugurado no site http://portugueselexicon.co.nf/.

23/02/2014 – desenvolvimento de um motor de geração de pseudopalavras do PB a partir de todos os dados acumulados no desenvolvimento e construção do Léxico do Português. Diferentemente de outros motores geradores de pseudopalavras (Keuleers & Brysbaert, 2010; Mota & Resende, 2013), utilizamos todos os bigramas e trigramas para a geração de pseudopalavras. Contabilizamos a frequência geral de cada bigrama e trigrama. Mais do que isto, contabilizamos a frequência de cada bigrama e trigrama de acordo à sua posição na palavra e por categoria gramatical. 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 categorias gramaticais.

05/03/2014 – criação de um motor de geração de pseudopalavras onde o usuário pode inserir quatro campos: 1) número de letras das palavras a serem geradas, 2) número de palavras 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 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 o bigrama ou o trigrama anterior (1 letra para os bigramas e 2 letras para os trigramas).

18/03/2014 – inserção de quatro colunas com dados sobre as pseudopalavras: 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 das pseudopalavras. Tradução da página principal de pesquisa simples e pesquisa complexa para o inglês.

25/03/2014 – desenvolvimento das páginas: documentos e créditos. Registro do domínio próprio do Léxico do Português (www.lexicodoportuguês.com) junto ao HostGator[23] e redirecionamento deste domínio para o servidor onde o Léxico do português está hospedado (portugueselexicon.co.nf/). Inauguração oficial do Léxico do Português em 25 de março de 2014 no site www.lexicodoportuguês.com.

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, o desenvolvimento do mesmo foi dividido em três versões: 1) Alfa (2014), 2) Beta (2015) e 3) Delta (2016). Atualmente o Léxico do Português está na versão Alfa, inaugurada em 25/03/2014. A versão Alfa marca a criação do Léxico do Português e o surgimento do primeiro corpus psicolinguístico do PB. A principal característica da versão Alfa do Léxico do Português é que ela disponibiliza um corpus ortográfico, ou seja, todas as informações disponibilizadas na versão Alfa foram estritamente computadas a partir de dados ortográficos das palavras do PB. Futuramente, a versão Beta contará com todas as informações: 1) fonológicas das formas, 2) silábicas das formas e 3) dos lemas associados às formas. A versão final Delta contará também com uma série de: 1) informações morfológicas das palavras, 2) informações sintáticas das palavras, 3) informações de alomorfia e, na medida do possível, 4) medidas de tempo de reação do reconhecimento de um grande número de palavras e pseudopalavras do PB, seguindo os modelos do *English Lexicon Project* (ELP) (Balota et al., 2007) e do *French Lexicon Project* (FLP) (Ferrand et al., 2010).

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

Corpus do NILC/São Carlos e Linguateca

O Léxico do Português foi desenvolvido a partir do corpus do Núcleo Interinstitucional de Linguística Computacional (NILC/São Carlos) (Pinheiro & Aluísio, 2003), aqui chamado apenas de CN, que por sua vez, é sediado no Instituto de Ciências Matemáticas e de Computação ICMC/São Carlos (ICMC/São Carlos)[24], da Universidade de São Paulo em São Carlos (USP/São Carlos)[25]. As listas de formas e lemas divididas em categorias gramaticais foram baixadas do Linguateca, onde também se encontra uma série de informações do CN, como dados quantitativos e estatísticos[26], descendência do corpus[27] e, principalmente, os arquivos de formas[28] e lemas[29] 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[30]. Há que ter, no entanto, em atenção a diferença fundamental entre o que de facto disponibilizamos e o que é apenas por nós catalogado. No primeiro caso na secção "Acesso a Recursos", no segundo caso na secção "Catálogo de Recursos". As condições de utilização destes últimos devem ser confirmadas junto dos respectivos autores." [31]

Lexique

Inicialmente, a criação e o desenvolvimento do Léxico do Português foram inspirados no corpus psicolinguístico do francês *Lexique* (Matos et al., 2001; New et al., 2004). O *Lexique* já ofereceu dados sobre as palavras do francês a uma série de estudos e pesquisas, servindo como um ótimo exemplo de corpus psicolinguístico completo, simples e de fácil utilização.

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

Ainda, este 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 do *Lexique* pode ser encontrada no manual do *Lexique*[32].

Programa R e pacotes

O Léxico do Português foi desenvolvido com a utilização do programa R, sendo que os dados originais foram importados a partir de arquivos .txt e cada nova coluna foi criada e computada através de determinadas funções e algoritmos. O número de vizinhos ortográficos (Coltheart's N) (M. Coltheart et al., 1977) e a distância de Levenshtein ortográfica (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" desenvolvido por Harald Baayen[33] também foram utilizadas no desenvolvimento do Léxico do Português.

Licença Creative Commons

O Léxico do Português de Gustavo Lopez Estivalet[34] está licenciado com uma Licença Creative Commons - Atribuição-NãoComercial-CompartilhaIgual 4.0 Internacional[35]. Baseado no trabalho disponível em http://www.linguateca.pt/acesso/contabilizacao.php. Podem estar disponíveis autorizações adicionais às concedidas no âmbito desta licença em http://www.lexicodoportugues.com/creditos.php.

[32] http://www.lexique.org/docLexique.php
[33] http://www.sfs.uni-tuebingen.de/~hbaayen/
[34] http://www.researchgate.net/profile/Gustavo_Estivalet
[35] http://creativecommons.org/licenses/by-nc-sa/4.0/

Convenções

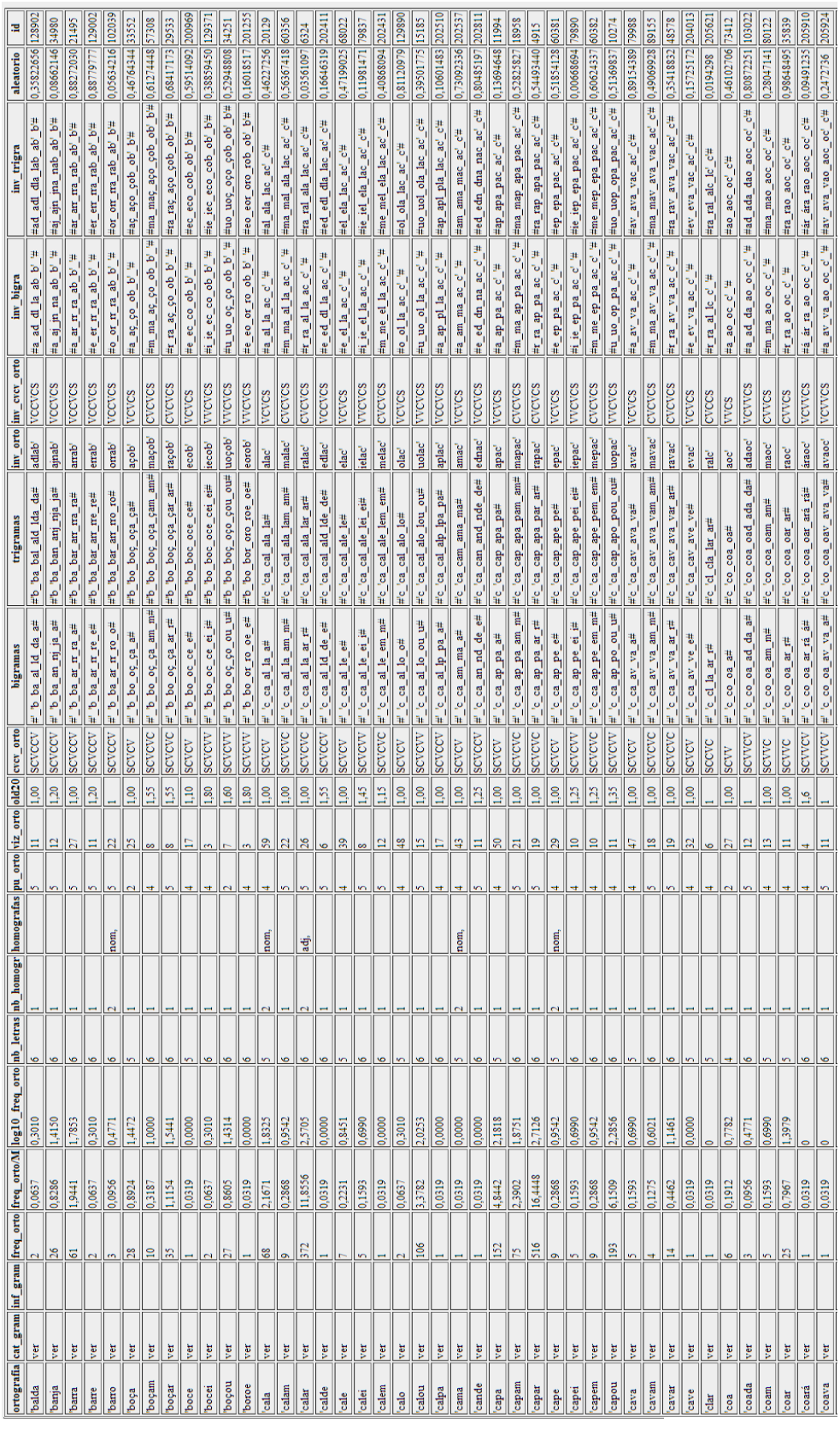
Para a utilização do Léxico do Português, algumas convenções foram determinadas para realização das pesquisas e apresentação dos resultados.

- As categorias gramaticais são: **adj** – adjetivo, **adv** – advérbio, **gram** – gramatical, **nom** – substantivo, **num** – numeral, **prop** – nome próprio, **ver** – verbo.
- As estruturas CVCV das palavras possuem as letras: **V** – vogais, **C** – consoantes, **P** - pontuação, **N** - números, **A** – acentos, **S** – símbolos.
- Os símbolos coringas utilizados são: **"<"** 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.
- Botões: **Procurar** – realiza a pesquisa e apresenta os resultados, **Limpar** – limpa todos os dados de todos os 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.

Categorias

O Léxico do Português 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 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. Enfim, esta pesquisa apresenta todas as 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	#'_'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	0,8286	1,4150	6	1		5	12	1,20	SCVCCV	#'_'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'#	0,08662146	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_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	nom.	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	102039
'boça	ver		28	0,8924	1,4472	5	1		2	25	1,00	SCVCV	#'_'b_bo_o'_ça_a#	#'b_'bo_bo'_oça_ça#	'açob'	VCVCS	#a_aç_'_co_ob_'_b'#	#a_aç_'oço_ob_ob'_b'#	0,46764344	33552
'boçam	ver		10	0,3187	1,0000	6	1		4	8	1,55	SCVCVC	#'_'b_bo_o'_ça_am_m#	#'b_'bo_bo'_oça_çam_am#	'maçob'	CVCVCS	#m_ma_aç_'_co_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_o'_ça_ar_r#	#'b_'bo_bo'_oça_çar_ar#	'raçob'	CVCVCS	#r_ra_aç_'_co_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'#	#e_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'	VVCVCS	#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_o'_ço_ou_u#	#'b_'bo_bo'_oço_çou_ou#	'uoçob'	VVCVCS	#u_uo_o'_ço_ob_'_b'#	#u_uoç_o'ço_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'	VVCVCS	#e_eo_or_ro_ob_'_b'#	#e_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'	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	60336
'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	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'#	#o_ol_la_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'#	#u_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'	VCVCCS	#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'#	#a_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'#	#a_apa_pac_ac'_c'#	0,11694648	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'#	#m_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'#	#r_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'#	#e_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'#	#u_uop_opa_pac_ac'_c'#	0,51369837	10274
'cara	ver		5	0,1593	0,6990	5	1		4	47	1,00	SCVCV	#'_'c_ca_ar_ra_a#	#'c_'ca_car_ara_ra#	'arac'	VCVCS	#a_ar_ra_ac_'_c'#	#a_ara_rac_ac'_c'#	0,89154389	79988
'caram	ver		4	0,1275	0,6021	6	1		5	18	1,00	SCVCVC	#'_'c_ca_ar_ra_am_m#	#'c_'ca_car_ara_ram_am#	'marac'	CVCVCS	#m_ma_ar_ra_ac_'_c'#	#ma_mar_ara_rac_ac'_c'#	0,49069928	89155
'carar	ver		14	0,4462	1,1461	6	1		5	19	1,00	SCVCVC	#'_'c_ca_ar_ra_ar_r#	#'c_'ca_car_ara_rar_ar#	'rarac'	CVCVCS	#r_ra_ar_ra_ac_'_c'#	#r_rar_ara_rac_ac'_c'#	0,35418832	48578
'carve	ver		1	0,0319	0,0000	5	1		4	32	1,00	SCVCV	#'_'c_ca_ar_ve_e#	#'c_'ca_car_arv_ve#	'evrac'	VCVCS	#e_ev_ra_ac_'_c'#	#e_eva_rac_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'#	#r_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'#	#a_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'#	#m_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'#	#r_rao_aoc_oc'_c'#	0,98648495	35839
'coará	ver		1	0,0319	0	6	1		4	4	1,6	SCVVCV	#'_'c_co_oa_ar_rá_á#	#'c_'co_coa_oar_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_oav_ava_va#	'avaoc'	VCVVCS	#a_av_va_ao_oc_'_c'#	#av_ava_vao_aoc_oc'_c'#	0,2472736	205924

Figura 2: Exemplo de pesquisa.

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

Ortografia (orto): forma ortográfica da palavra em letras minúsculas (com exceção dos nomes próprios), respeitando os acentos específicos de cada palavra[36].

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 CN (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 comumente utilizado para expressar a frequência de uma palavra.

Logaritmo natural da frequência ortográfica (log10_freq_orto): logarítmico natural da frequência ortográfica. Os valores logarítmicos são utilizados principalmente 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.

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

Vizinhos ortográficos (viz_orto): número de vizinhos ortográficos a partir do N de Coltheart, ou seja, alterando-se apenas uma letra por vez, quantos vizinhos uma palavra pode ter (M. Coltheart et al., 1977).

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

Estrutura consoante-vogal CVCV (CVCV_orto): estrutura da palavra na forma CV, onde todas as consoantes são C e todas as vogais V. Ainda possui a codificação A para acentos, P para pontuação, N para números e S para símbolos.

Bigramas (bigramas): bigramas que constituem a palavra separados por "_", onde "#" indica as fronteiras da palavra. O numero de bigramas é igual ao número de letras da palavra mais 1.

Trigramas (trigramas): trigramas que constituem a palavra separados por "_", onde "#" indica as fronteiras da palavra. O numero de trigramas é igual ao número de letras da palavra.

Ortografia invertida (inv_orto): forma ortográfica da palavra invertida a partir de (orto).

Estrutura consoante-vogal CVCV invertida (inv_CVCV_orto): estrutura da palavra na forma CVCV invertida a partir de (CVCV_orto).

Bigramas invertidos (inv_bigra): bigramas que constituem a palavra separados por "_", onde "#" indica as fronteiras da palavra invertido a partir de (bigramas).

Trigramas invertidos (inv_trigra): trigramas que constituem a palavra separados por "_", onde "#" indica as fronteiras da palavra invertido 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. Automaticamente, o número de identificação é a posição da palavra no corpus e no léxico.

Pesquisa simples

O motor de pesquisa simples foi desenvolvido exclusivamente para a pesquisa de palavras específicas ou listas de palavras, conforme a Figura 3. Neste motor, o usuário pode somente realizar uma pesquisa a partir da inserção 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 todas 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 as colunas de informações pelas quais deseja realizar sua pesquisa. No segundo campo, ele deve escolher se deseja considerar "sim" ou desconsiderar "não". No terceiro campo, o usuário deve efetivamente 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 que contenham determinados critérios. 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 todas as informações dos campos. Inicialmente, a pesquisa complexa apresenta quatro campos de critérios para a pesquisa, clicando-se no botão "+ Critérios", o usuário é submetido a outra página que apresenta oito campos de critérios para a pesquisa.

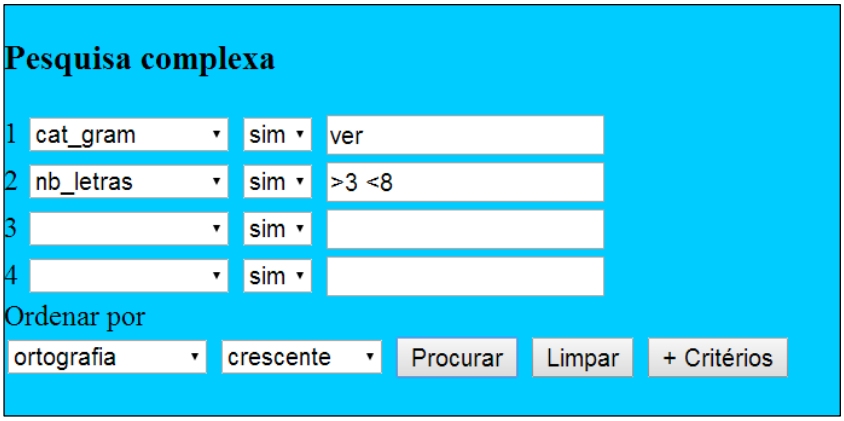


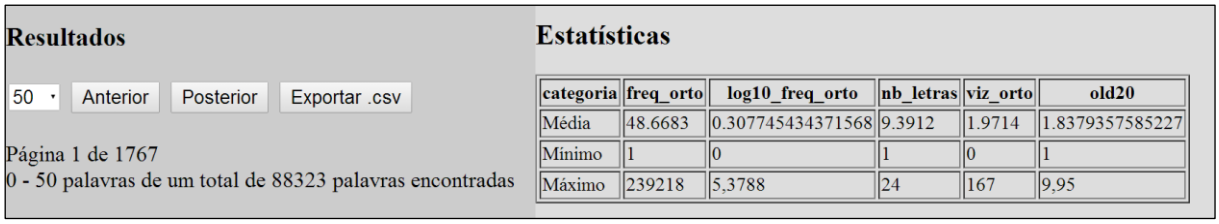
Figura 4: Pesquisa complexa.

Resultados

Na seção de resultados, conforme a Figura 2, o usuário encontra os resultados de sua pesquisa organizada em diferentes linhas e com as informações metalinguísticas e psicolinguísticas nas diferentes colunas. Além do resultado da pesquisa, pode-se ainda encontrar uma série de informações pertinentes a pesquisa, conforme a Figura 5. As seguintes informações são disponibilizadas: 1) número total de palavras encontradas na pesquisa, 2) intervalo de palavras apresentados na tela, 3) número total de páginas de acordo com o número de palavras encontradas e apresentadas na pesquisa e 4) número da página apresentada entre todas as páginas. Ainda, pode-se escolher no campo superior à esquerda o número de palavras apresentadas em cada página. Através dos botões "Anterior" e "Próximo", o usuário pode navegar entre os resultados e as páginas da pesquisa.

Na parte superior a direita dos resultados, conforme a Figura 5, apresenta-se uma série de dados estatísticos estabelecidos e calculados a partir da pesquisa realizada (Davis, 2005;

Davis & Perea, 2005). As informações estatísticas apresentadas são as seguintes: 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.



categoria	freq_orto	log10_freq_orto	nb_letras	viz_orto	old20
Média	48.6683	0.307745434371568	9.3912	1.9714	1.8379357585227
Mínimo	1	0	1	0	1
Máximo	239218	5,3788	24	167	9,95

Figura 5: Resultados e estatísticas básicas.

Pseudopalavras

O motor gerador de pseudopalavras foi desenvolvido com o objetivo de obter-se um dispositivo para a criação de pseudopalavras baseadas na estrutura e frequência das palavras (Keuleers & Brysbaert, 2010) do PB. Porém, diferentemente de outros motores de geração de pseudopalavras que se baseiam na estrutura silábica das palavras existentes (Keuleers & Brysbaert, 2010; Mota & Resende, 2013), o motor de geração de pseudopalavras do Léxico do Português utiliza os bigramas e trigramas para a construção das pseudopalavras do PB (New et al., 2004). Para tanto, utilizou-se todos os bigramas e trigramas contabilizados a partir de todas as palavras do Léxico do Português. Assim, as pseudopalavras podem ser geradas a partir da frequência e combinação dos bigramas ou trigramas. Contabilizou-se a 1) frequência geral, 2) frequência de acordo com a posição na palavra e 3) frequência por categoria gramatical, de cada bigrama e trigrama.

No motor de geração de pseudopalavras do PB, o usuário pode inserir quatro campos: 1) número de palavras a serem geradas, 2) número de letras das palavras a serem geradas, 3) categoria gramatical de base 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 de geração de pseudopalavras do PB apresenta dois botões: “Criar” para gerar e apresentar os resultados das pseudopalavras criadas e “Limpar” para limpar todos os dados dos campos, conforme a Figura 6.

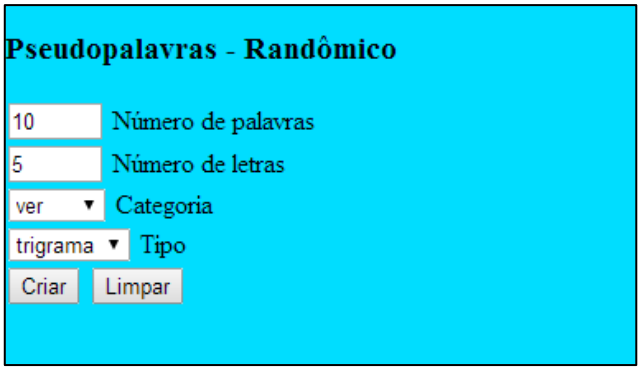


Figura 6: Motor de geração de pseudopalavras.

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 as pseudopalavras, 3) log10 da frequência calculada das pseudopalavras e 4) número de letras das pseudopalavras. Nos resultados, ainda é disponibilizado o botão “Exportar .csv” para exportar os resultados da geração de pseudopalavras para um arquivo .csv encaminhado 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.

Corpora

Léxico do Português[37] - Português Brasileiro

Lexique[38] – Francês

MRC[39] - Inglês

CELEX[40] - Alemão, Cirílico, Holandês e Inglês

ClearPOND[41] - Alemão, Espanhol, Francês, Holandês e Inglês

Recursos psicolinguísticos

Programa R[42] – (R. Harald Baayen, 2008)

languageR[43] package for R – (R. Harald Baayen, 2013)

vwr[44] package for R -(Keuleers, 2013)

Wuggy: a multilingual pseudoword generator[45] - (Keuleers & Brysbaert, 2010)

Autores

O Léxico do Português 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) [46] do Conselho Nacional de Desenvolvimento

[37] http://www.lexicodoportugues.com/
[38] http://www.lexique.org/
[39] http://www.psych.rl.ac.uk/
[40] http://celex.mpi.nl/
[41] http://clearpond.northwestern.edu/
[42] http://www.r-project.org/
[43] http://cran.r-project.org/web/packages/languageR/languageR.pdf
[44] http://cran.r-project.org/web/packages/vwr/vwr.pdf
[45] http://crr.ugent.be/programs-data/wuggy
[46] http://www.cienciasemfronteiras.gov.br/web/csf

Científico e Tecnológico (CNPq)[47], Brasil, entre 2012 e 2015, e, sua Orientadora Prof. Dr. Fanny Meunier[48], financiada pelo *Conseil National de la Recherche Scientifique* (CNRS)[49], França. Os dois pesquisadores desenvolvem atualmente seus trabalhos de pesquisa na França, na cidade de Lyon, na *Université Claude Barnard Lyon 1* (UCBL)[50] na *École Doctorale de Neurosciences et Cognition* (ED NSCo)[51] junto ao *Laboratoire sur le Langage, le Cerveau et la Cognition* (L2C2)[52], localizado no *Institute de Sciences Cognitives* (ISC) [53].

Agradecimentos

Para a realização e êxito do Léxico do Português, 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 pesquisador Prof. Dr. Michel Hoen[54] pela compreensão da importância do desenvolvimento de um corpus psicolinguístico do PB e por acreditaram e apoiaram a realização do mesmo com sugestões e críticas construtivas durante o processo. Agradeço às professores do NILC/São Carlos Prof. Dr. Sandra M. Aluísio[55] e Prof. Dr. Maria das Graças Volpe Nunes[56] pelos valiosos materiais, informações e auxílio sobre o CN, assim como o apoio e a motivação na realização deste trabalho. Agradeço aos colegas Léo Varnet [57] e Emmanuel Trouche [58] pelas discussões sobre os scripts e algoritmos para o desenvolvimento do Léxico do Português. Agradeço os usuários dos fóruns de discussão e tutoriais da internet sobre o desenvolvimento de páginas e bancos de dados. Agradeço minha mentora em psicolinguística que sempre apoiou meu trabalho Prof. Dr. Mailce Borges Mota e a melhor professora de português do mundo Prof. Lise Lopez. Finalmente, agradeço Luanda Lins por compreender a importância deste projeto para mim e minha motivação em fazê-lo, trocando horas de sono por elaborações de algoritmos e processamento computacional.

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

Referências

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics*. Cambridge: Cambridge University Press.

Baayen, R. H. (2013). R Package 'languageR': Data sets and functions with 'Analyzing Linguistic Data: A practical introduction to statistics'. (pp. 133).

Baayen, R. H., 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

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

Keuleers, E. (2013). R Package 'vwr' (pp. 19).

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

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

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

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

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* (pp. 60). São Carlos, SP: Universidade Federal de São Carlos - UFSCar.

Santos, D., & Bick, E. (2000, 31 May - 2 June 2000). *Providing internet access to Portuguese corpora: the AC/DC project.* Paper presented at the Proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000), 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