# Brazilian Portuguese Lexicon Manual - LexPorBR: Psycholinguistic Corpus

*Version: Alpha2*

Lyon, 6th October 2015.

## Introduction

The main objective of the Brazilian Portuguese Lexicon - LexPorBR [1] is to deliver a psycholinguistic corpus of the Brazilian Portuguese (BP) language that provides maximum metalinguistic and psycholinguistic information about BP words. The Brazilian Portuguese Lexicon is a free and open corpus consulted in a simple and dynamic interface on the internet. The results are presented in an organized and hierarchical way, containing metalinguistic and psycholinguistic information about the words or groups of word searched.

## Psycholinguistic Corpora

Psycholinguistic corpora are used 1) to control, selection, and manipulation of words and specific criteria in psycholinguistic experiments and 2) in computational linguistic analysis on distribution and lexical behavior (Baayen, 2001). Examples of psycholinguistic corpora are: French – *Lexique*[2] (B. New, Pallier, Brysbaert, & Ferrand, 2004; B. New, Pallier, Ferrand, & Matos, 2001), Spanish – *BuscaPalabras* (Davis & Perea, 2005), English – MRC[3] (Coltheart, 1981), Dutch, English, French, German, and Spanish - ClearPOND[4] (Marian, Bartolotti, Chabal, & Shook, 2012), Cyrillic, Dutch, English, and German - CELEX [5] (Baayen, Piepenbrock, & van Rijn, 1995). For example, these corpora have been used in megastudies investigating the psycholinguistic behavior on word and pseudoword recognition in the

[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/

*English Lexicon Project* (Balota et al., 2007; Boris New, Ferrand, Pallier, & Brysbaert, 2006), *French Lexicon Project* (Ferrand et al., 2010), *Dutch Lexicon Project* (Keuleers, Diependaele, & Brysbaert, 2010), and *British Lexicon Project* (Keuleers, Lacey, Rastle, & Brysbaert, 2012). These corpora are used in the control, selection, and manipulation of words and its characteristics to create psycholinguistic experiments in a number of studies and specific researches (Gimenes & New, 2015), as well as in the development and simulation of linguistic models and computational linguistic analysis (Schreuder & Baayen, 1995).

# Brazilian Portuguese Lexicon

The Brazilian Portuguese Lexicon was born from an idea noted on a post-it in late 2012 when I was beginning my Ph.D in psycholinguistics and neurosciences, in Lyon, France. My Ph.D project investigates the morphological representation and processing of verbal inflection in BP, French, and bilinguals with BP as first language and French as second language. For the French experiments, the stimuli were selected in the *Lexique* corpus (B. New et al., 2004), which offers a range of crucial information for the experiments and result analysis (surface frequency, grammatical category, number of letters, number of neighbors, inverted form, CVCV structure, etc.). In early 2013, when we started to prepare the BP experiments, we faced the complete lack of a BP word-based psycholinguistic corpus. Looking for a corpus that feed our needs, we found the *Linguateca*[6] website (Santos & Bick, 2000) which comprises several European and Brazilian Portuguese corpora. However, we could not find any BP word-based corpus with metalinguistic and psycholinguistic information appropriated for the strict creation of psycholinguistic experiments in BP. It was when I wrote on a post-it "make the Brazilian Portuguese Lexicon". Currently, the Brazilian Portuguese Lexicon presents the homepage as shown in **Figure 1**.

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

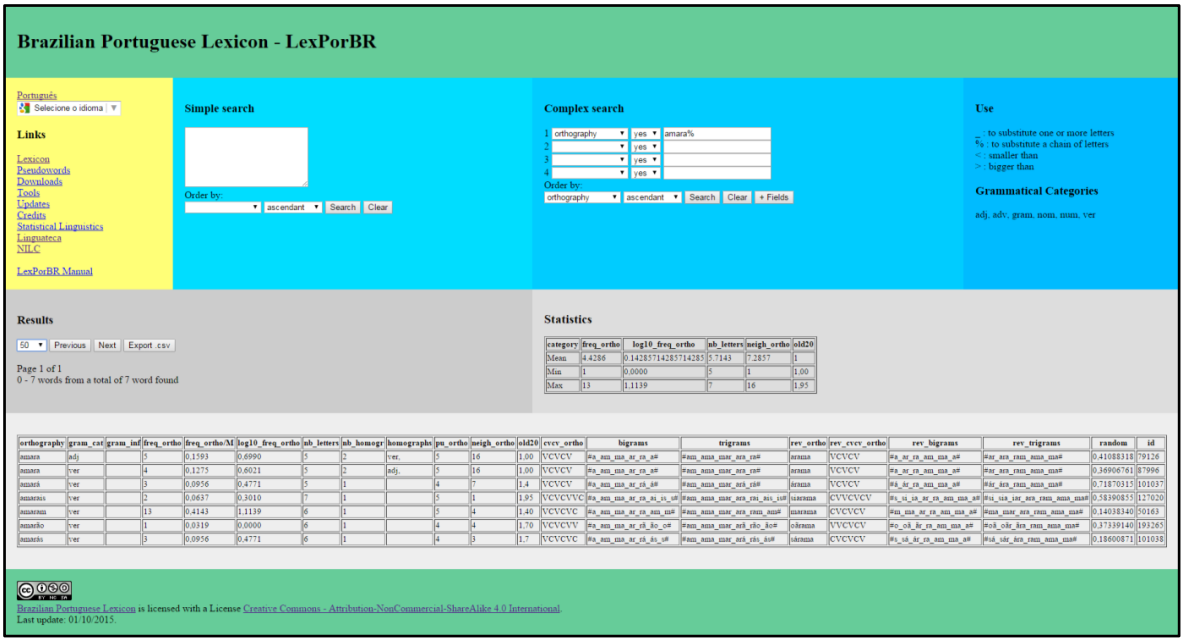



**Figure 1:** Homepage of the Brazilian Portuguese Lexicon.

## Development

In early 2014, we started the development of the Brazilian Portuguese Lexicon in four stages: 1) construction of the corpus with words, metalinguistic, and psycholinguistic information, 2) construction of the webpage in HTML, 3) importation of the corpus to a MySQL database on the internet, and 4) programming in PHP of the operation of the Brazilian Portuguese Lexicon. Additional pages were created: updates, credits, downloads, tools, and links. Then, we developed the BP pseudoword generator engine and the statistical linguistics tools.

## Updates

15/01/2013 - looking for a BP psycholinguistic corpus, we found the *Linguateca* website which concentrates a number of Portuguese corpora, but no BP psycholinguistic corpus. We decided to make the Brazilian Portuguese Lexicon as a BP word-based metalinguistic and

psycholinguistic corpus with free and open access on the internet. Skills to be developed: R[7], HTML[8], MySQL[9], PHP[10], Java[11], and CSS[12].

03/21/2013 - pre-selection in *Linguateca* website of two BP corpora: 1) Corpus Brasileiro[13] (1 billion words, 3.2 GB) and 2) *Núcleo Interdisciplinar de Linguística Computacional de São Carlos* (henceforth NILC)[14,15] (32 million words, 49 MB). After discussion with the researchers responsible for these corpora, we concluded the NILC would be the best corpus for the development of the Brazilian Portuguese Lexicon, following the criteria: 1) number of words (32 million) consistent with other psycholinguistic corpora (*Lexique*, CELEX, ClearPOND) (Brysbaert & New, 2009), 2) number and size of files (13 files, total size 49 MB), 3) organization of the corpus in .txt files separated by grammatical categories, 4) organization of files in two columns (orthography and frequency) and delimited by tabulation, and 5) publications and resources already developed by the NILC.

08/14/2013 – development of a pilot corpus with only the verbs from NILC, accounting about 80,000 forms. We used the R program for the development of 10 columns of information: 1) orthography, 2) orthographic frequency, 3) frequency per million of words, 4) log10 orthographic frequency, 5) number of letters, 6) grammatical category, 7) grammatical information, 8) reversed orthographic form, 9) CVCV structure, and 10) CVCV reversed structure. Construction of the pilot webpage in a localhost with the XAMPP[16] program containing the Apache, MySQL, PHP, and Perls preinstalled modules. Configuration of the phpMyAdmin[17] for the importation of the pilot corpus saved in .csv format to a MySQL database. We used the Notepad++[18] software to program the HTML/PHP interface page between the user and MySQL database.

10/28/2013 – pilot version of the Brazilian Portuguese Lexicon with two search engines: 1) simple search and 2) complex search. The simple search consists of a text area where users

[7] http://www.r-project.org/
[8] http://en.wikipedia.org/wiki/HTML
[9] http://www.mysql.com/
[10] http://www.php.net/
[11] http://www.java.com/
[12] http://en.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/
[17] http://www.phpmyadmin.net/home_page/index.php
[18] http://notepad-plus-plus.org/

can enter multiple words as a list. The complex search consists of four fields for criteria insertion about the words to be searched. Each search engine has a "Search" button to start the search and present the results, and a "Clear" button to delete the current data in the fields. Definition of the Brazilian Portuguese Lexicon pages: Lexicon, Pseudowords, Downloads, Tools, Updates, Credits, Statistical Linguistics, Linguateca, and NILC.

11/30/2013 - programming of Java and PHP algorithms to keep the data in the HTML fields after search. We inserted two fields for results organization, one with the arrangement criterion and another for ascendant or descendant order. We inserted the "+ Fields" button in the complex search engine for the use of eight fields. Choice of the http://www.biz.nf/ free web server host following the criteria: 1) space of 250 MB, 2) MySQL 5 database, 3) PHP 4/5 support, 4) 5,000 MB transfer, 5) free hosting, 6) free domain http://portugueselexicon.co.nf, 7) POP3/SMTP webmail, and 8) FTP control transfer. Importation of the pilot corpus in .csv format to a MySQL database and transfer of all other pages created in HTML/PHP to http://portugueselexicon.co.nf/ by FTP with the FileZilla[19] program.

12/12/2013 - given the MySQL recognizes the symbols underline "_" to replace a letter and percentage "%" to replace a chain of letters, this information was added to the tips on the homepage. Programming in PHP for the recognition of the symbols: greater than ">" and less than "<" for numerical search. The Brazilian Portuguese Lexicon development was divided in three versions: 1) Alpha, 2) Beta and 3) Delta. The Alpha version (2014) is the first version of the Brazilian Portuguese Lexicon, providing a pure orthographic corpus. The Beta version (scheduled for 2015) will provide phonological, syllabic, and lemma data. Finally, the Delta version (scheduled for 2016) will provide specific features such as: morphological information, syntactic information, age of acquisition, among others.

01/07/2014 - download on the *Linguateca* website of the 13 files of the NILC[20] in .txt format separated by grammatical categories (6 form files: adjectives, adverbs, grammatical, nouns, numerals, and verbs; 7 lemma files: adjectives, adverbs, grammatical, nouns, proper names, numerals, and verbs). We compared of the total number of words and forms to the data provided in *Linguateca*. Column with the grammatical categories (gram_cat) for each file (adjectives, adverbs, grammatical, nouns, proper names, numerals, and verbs). Column with the word type (form or lemma). Transformation of all words in lowercases and sum of

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

repeated forms. Column with an identification number (id) of the word, according to the organization of the corpus by frequency in descending order an alphabetic a-z order, so the identification number (id) also becomes the word position in the lexicon and corpus.

01/08/2014 – column with the orthographic frequency per million of words (ortho_freq/M) by calculating [1000000 * ortho_freq / total_freq]. Column with the natural log of the orthographic frequency. Column with the number of letters of the form. Exclusion of all forms with more than 30 letters and numerals, but 0-1 and 1st-9th. Column with the number of homographic forms. Column with the grammatical categories of the homographic forms.

01/10/2014 – splitting of words into letters, processing of vowels "V" and consonants "C", column with the CVCV structure (CVCV_ortho). The letters were also classified as punctuation "P", numbers "N", accents "A", and symbols "S". Column with word bigrams. Column with word trigrams.

18/01/2014 - development of an algorithm to calculate the orthographic uniqueness point (pu_ortho) and creation of a column for it. Column with the number of orthographic neighbors (ortho_neigh) (Coltheart's N) (Coltheart, Davelaar, Jonasson, & Besner, 1977) and column with the Orthographic Levenshtein Distance for the 20 closest words (old20) (Yarkoni, Balota, & Yap, 2008). These functions are available in the R package "vwr"[21] developed by Emmanuel Keuleers[22]. Column with a random number between 0 and 1 with eight digits of precision.

01/28/2014 - creation of four columns with the reversed forms from: orthography (rev_ortho), CVCV_ortho (rev_CVCV_ortho), bigrams (rev_bigrams), and trigrams (rev_trigrams). The Brazilian Portuguese Lexicon – Alpha has 21 columns of metalinguistic and psycholinguistic information: 1) orthography, 2) gram_cat, 3) gram_inf, 4) ortho_freq, 5) ortho_freq/M, 6) log10_ortho_freq, 7) nb_letters, 8) nb_homogr, 9) homographs, 10) pu_ortho, 11) ortho_neigh, 12) old20, 13) CVCV_ortho, 14) bigrams, 15) trigrams, 16) rev_ortho, 17) rev_CVCV_ortho, 18) rev_bigrams, 19) rev_trigrams, 20 ) random, and 21) id.

02/05/2014 - the word-based psycholinguistic corpus Brazilian Portuguese Lexicon has 21 columns of information and 215,175 rows of BP words. This table has a size of 45 MB in .csv format. This file was split into 36 files of about 1.5 MB in .csv format. Then the .csv files

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

were saved in UTF-8 coding (they had AINSI coding) in order to avoid problems with accents, symbols, and special characters. Each file was imported via phpMyAdmin to our internet server host in a way that each new imported file inflated the existing one.

02/12/2014 - development of a module for limitation and navigation of results. The number of words to be presented may be 50, 100, 200, or 500. It was included two buttons (“Previous” and “Next”) to navigate between the result pages. Presentation of four general information about the search: 1) total number of words found, 2) total number of pages, 3) range of words presented, and 4) page displayed. Inclusion of the button "Export .csv" to export all the results of the current search to a downloadable .csv file.

02/18/2014 - development of a basic statistical module from the current search (Davis, 2005), presenting the following information: 1) mean, 2) maximum, and 3) minimum values of the following categories: 1) ortho_freq, 2 ) log10_ortho_freq, 3) nb_letters, 4) ortho_neigh, and 5) old20.

02/23/2014 - development of a BP pseudoword generator engine from bigrams or trigrams. Unlike other pseudoword generators (Keuleers & Brysbaert, 2010; Mota & Resende, 2013), we accounted the overall frequencies bigrams and trigrams, the frequency of bigrams and trigrams according to their position in word and grammatical category. We obtained two tables as databases to generate BP pseudowords, one with the bigrams and another with the trigrams, with 1) general, 2) by position in the word, and 3) by grammatical category.

03/05/2014 – we created a BP pseudoword generator engine where the user must enter four fields: 1) number of letters of the pseudowords to be generated, 2) number of pseudowords to be generated, 3) grammatical category that these pseudowords should belong (all, adj, adv, gram, nom, num, ver), and 4) criterion for the construction of the pseudowords (bigrams or trigrams). The BP pseudoword generator engine builds pseudowords simultaneously in both directions, from left to right and from right to left, beginning with a bigram or trigram like “#xx” or #xx”. Following to the number of letters, the engine concatenates new bigrams or trigrams which share as much orthographic information as possible with the previous bigram or trigram (1 letter for bigrams and 2 letters for trigrams).

03/18/2014 - insertion of four columns in pseudoword results: 1) pseudoword grammatical category, 2) pseudoword frequency calculated as the sum of the bigrams or trigrams

frequencies that build the pseudoword, 3) log10 of the pseudoword calculated frequency, and 4) number of letters of the pseudoword. We translated the homepage with the simple and complex searches to English.

03/25/2014 - registration of a specific internet domain for the Brazilian Portuguese Lexicon (www.lexicodoportuguês.com) in HostGator[23] and redirection of this domain to the server where the Brazilian Portuguese Lexicon is hosted http://portugueselexicon.co.nf. Inauguration of the Brazilian Portuguese Lexicon on March 25th, 2014.

03/21/2015 - development of the Statistical Linguistics page with various open and free tools and resources in HTML/PHP to linguistic and statistical analysis: F1, F2, F', minF', Hartley test, normalization between 0 and 1, reverse word, Hamming distance, Levenshtein distance, orthographic neighbors (Coltheart's N), average Levenshtein distance, relative entropy, word frequency, Zipf's distribution, etc.

09/20/2015 – writing and availability of the *Manual do Léxico do Português Brasileiro - Alfa 2* in Brazilian Portuguese and writing and availability of the Brazilian Portuguese Lexicon - Alpha 2 Manual in English. We translated and implemented to English all pages of the Brazilian Portuguese Lexicon. We implemented the Google Translate gadget in all pages of the Brazilian Portuguese Lexicon for its translation to the various languages available on Google Translate. It is suggested to use the site in English in a way that the results presented in Portuguese will not be translated.

## Alpha Version

Considering the large amount of metalinguistic and psycholinguistic information that can and will be computed, implemented, and made available in the Brazilian Portuguese Lexicon, its development was divided in three versions: 1) Alpha (2014), 2) Beta (2015), and 3) Delta (2016). Currently, the Brazilian Portuguese Lexicon is in the Alpha version, inaugurated on March 25th, 2014. The main characteristic of the Brazilian Portuguese Lexicon Alpha version is that it provides a pure orthographic corpus with all information provided being strictly computed from orthographic data of the BP words. Further, the Beta version will provide

[23] http://hostgator.com/

information about: 1) phonology, 2) syllables, and 3) lemma association. Delta version will provide: 1) morphological information, 2) syntactic information, 3) allomorphy information, and as far as possible 4) Reaction Time measures from the recognition of a large number of BP words and pseudowords, following the Lexicon Projects (Balota et al., 2007; Ferrand et al., 2010; Keuleers et al., 2010, 2012).

## Creative Commons License 



# NILC/São Carlos and Linguateca

The Brazilian Portuguese Lexicon was developed from the *Núcleo Interinstitutional de Linguística Computacional de São Carlos* corpus (NILC) (Inter-Institutional Computational Linguistics Center) (Pinheiro & Aluísio, 2003) located at the *Instituto de Ciências Matemáticas e de Computação de São Carlos* (Institut of Mathematics and Computation) (ICMC/São Carlos)[26], at the University of São Paulo in São Carlos (USP/São Carlos)[27]. The lists of forms and lemmas divided in grammatical categories were downloaded from the *Linguateca* website, which also provides a series of information about the NILC, as quantitative and statistical data[28], corpus origin[29], and especially, the forms[30] and the lemmas[31] files in .txt format, separated by grammatical categories.

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

"All the material that we provide is not restricted to any group and was authorized (under the provided terms) by the respective authors or copyright holders. From resource to resource the conditions are different and are specified in its specific documentation. The tools created by the *Linguateca* are available under the GNU[32]."[33]

# Lexique

The Brazilian Portuguese Lexicon was inspired by the French psycholinguistic corpus *Lexique* (B. New et al., 2004, 2001). The *Lexique* has already offered data on French words to a number of studies and researches, being a great example of psycholinguistic corpus. This corpus exemplifies the features and utilities that a psycholinguistic corpus should and can offer as resources for research in psycholinguistics and computational linguistics. A detailed description of the *Lexique* can be found in the user's manual[34].

# R Program and Packages

The Brazilian Portuguese Lexicon was developed with the R program with the original data imported from .txt files and each new column being created and computed through certain functions and algorithms. The number of orthographic neighbors (Coltheart's N) (Coltheart et al., 1977) and the Orthographic Levenshtein Distance from de 20 closest words (OLD20) (Yarkoni et al., 2008) were calculated with the functions "coltheart.N" and "old20" available in the "vwr" R package developed by Emmanuel Keuleers. Functions from the "languageR"[35] R package developed by Harald Baayen[36] were also used in the development of the Brazilian Portuguese Lexicon.

[32] http://www.gnu.org/copyleft/gpl.html
[33] http://www.linguateca.pt/FAQ/#faq1.8
[34] http://www.lexique.org/docLexique.php
[35] https://cran.r-project.org/web/packages/languageR/index.html
[36] http://www.sfs.uni-tuebingen.de/~hbaayen/

# LexPorBR - Alpha

## Conventions

For the use of the Brazilian Portuguese Lexicon, some conventions were determined to perform the searches and understand the results.

• Grammatical categories: **adj** - adjective, **adv** – adverb, **gram** - grammatical, **nom** - noun, **num** - numeral, **prop** – proper name, **ver** - verb.

• The CVCV structures have the letters: **V** - vowel, **C** - consonant, **P** - punctuation, **N** - number, **A** - accentuation, **S** - symbol.

• The wildcard used are: **"<"** greater than, **">"** less than, **"_"** replaces a letter, **"%"** replaces a chain of letters.

• Order of the results presentation: **ascendant** - presents the results in ascending order, **descendant** - presents the results in descending order.

• Buttons: **Search** - performs the search and displays the results, **Clear** - clears all data from the fields, **+ Fields** - redirects the user to a page with more fields for complex research.

• Select yes/no: **yes** - considers the criterion, **not** – disregards the criterion.

## Columns

The Brazilian Portuguese Lexicon - Alpha has 215,175 rows with different lexical entries and 21 columns with different metalinguistic and psycholinguistic information. Thus, each row of the Brazilian Portuguese Lexicon contains a word and each column contains an information about this word.

A search example from the complex research with **cat_gram-yes-ver** and **nb_letters-yes-**
**>3<8** can be seen in **Figure 2**. This search presents all the words that have the grammatical category defined as "verb", less than 8 letters, and more than 3 letters.

| orthography | gram_cat | gram_inf | freq_ortho | freq_ortho/M | log10_freq_ortho | nb_letters | nb_homogr | homographs | pu_ortho | neigh_ortho | old20 | cvcv_ortho | bigrams | trigrams | rev_ortho | rev_cvcv_ortho | rev_bigrams | rev_trigrams | random | id |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| case | ver | | 37 | 1.1792 | 1.5682 | 4 | 1 | | 4 | 35 | 1.00 | CVCV | #c_ca_as_se# | #c_ca_as_ase_se# | esac | VCVC | #e_es_sa_ac_c# | #es_esa_sac_ac# | 0.67386608 | 28665 |
| casi | ver | | 4 | 0.1275 | 0.6021 | 4 | 1 | | 4 | 17 | 1.00 | CVCV | #c_ca_as_si_i# | #c_ca_as_asi_si# | isac | VCVC | #i_is_sa_ac_c# | #is_isa_sac_ac# | 0.94087864 | 89116 |
| caso | ver | | 84 | 2.6771 | 1.9243 | 4 | 3 | gram.nom. | 4 | 32 | 1.00 | CVCV | #c_ca_as_so_o# | #c_ca_as_aso_so# | osac | VCVC | #o_os_sa_ac_c# | #os_osa_sac_ac# | 0.66035062 | 17673 |
| cata | ver | | 8 | 0.2550 | 0.9031 | 4 | 2 | nom. | 4 | 55 | 1.00 | CVCV | #c_ca_at_ta_a# | #c_ca_at_ata_ta# | atac | VCVC | #a_at_ta_ac_c# | #at_ata_tac_ac# | 0.63842162 | 63940 |
| cato | ver | | 5 | 0.1593 | 0.6990 | 4 | 1 | | 3 | 41 | 1.00 | CVCV | #c_ca_at_to_o# | #c_ca_at_ato_to# | otac | VCVC | #o_ot_ta_ac_c# | #ot_ota_tac_ac# | 0.46417913 | 79982 |
| cava | ver | | 8 | 0.2550 | 0.9031 | 4 | 3 | adj.nom. | 4 | 47 | 1.00 | CVCV | #c_ca_av_va_a# | #c_ca_av_ava_va# | avac | VCVC | #a_av_va_ac_c# | #av_ava_vac_ac# | 0.78947595 | 63950 |
| cave | ver | | 2 | 0.0637 | 0.3010 | 4 | 2 | nom. | 4 | 32 | 1.00 | CVCV | #c_ca_av_ve_e# | #c_ca_av_ave_ve# | evac | VCVC | #e_ev_va_ac_c# | #ev_eva_vac_ac# | 0.01989358 | 130366 |
| cavo | ver | | 2 | 0.0637 | 0.3010 | 4 | 2 | adj. | 4 | 29 | 1.00 | CVCV | #c_ca_av_vo_o# | #c_ca_av_avo_vo# | ovac | VCVC | #o_ov_va_ac_c# | #ov_ova_vac_ac# | 0.73821918 | 130367 |
| cear | ver | | 8 | 0.2550 | 0.9031 | 4 | 1 | | 4 | 16 | 1.00 | CVVC | #c_ce_ea_ar_a# | #c_ce_cea_ear_a# | raec | CVVC | #r_ra_ae_ec_c# | #ra_rae_aec_ec# | 0.40265500 | 63956 |
| ceci | ver | | 1 | 0.0319 | 0.0000 | 4 | 1 | | 4 | 5 | 1.70 | CVCV | #c_ce_ec_ci_i# | #c_ce_cec_eci_ci# | icec | VCVC | #i_ic_ce_ec_c# | #ic_ice_cec_ec# | 0.42591075 | 204072 |
| ceda | ver | | 17 | 0.5418 | 1.2304 | 4 | 1 | | 4 | 26 | 1.00 | CVCV | #c_ce_ed_da_a# | #c_ce_ced_eda_da# | adec | VCVC | #a_ad_de_ec_c# | #ad_ade_dec_ec# | 0.93477629 | 44012 |
| cede | ver | | 167 | 5.3223 | 2.2227 | 4 | 1 | | 3 | 18 | 1.00 | CVCV | #c_ce_ed_de_e# | #c_ce_ced_ede_de# | edec | VCVC | #e_ed_de_ec_c# | #ed_ede_dec_ec# | 0.63774699 | 11290 |
| cedi | ver | | 10 | 0.3187 | 1.0000 | 4 | 1 | | 4 | 11 | 1.35 | CVCV | #c_ce_ed_di_i# | #c_ce_ced_edi_di# | idec | VCVC | #i_id_de_ec_c# | #id_ide_dec_ec# | 0.44348834 | 57412 |
| cedo | ver | | 23 | 0.7330 | 1.3617 | 4 | 2 | adv. | 4 | 19 | 1.00 | CVCV | #c_ce_ed_do_o# | #c_ce_ced_edo_do# | odec | VCVC | #o_od_de_ec_c# | #od_ode_dec_ec# | 0.77500498 | 37552 |
| ceei | ver | | 4 | 0.1275 | 0.6021 | 4 | 1 | | 2 | 2 | 1.85 | CVVV | #c_ce_ee_ei_i# | #c_ce_cee_eei_ei# | ieec | VVVC | #i_ie_ee_ec_c# | #ie_iee_eec_ec# | 0.89151235 | 89165 |
| cega | ver | | 13 | 0.4143 | 1.1139 | 4 | 2 | adj. | 4 | 23 | 1.00 | CVCV | #c_ce_eg_ga_a# | #c_ce_ega_ega_ga# | agec | VCVC | #a_ag_ge_ec_c# | #ag_age_gec_ec# | 0.48261355 | 50413 |
| cego | ver | | 10 | 0.3187 | 1.0000 | 4 | 3 | adj.nom. | 4 | 15 | 1.00 | CVCV | #c_ce_eg_go_o# | #c_ce_ceg_ego_go# | ogec | VCVC | #o_og_ge_ec_c# | #og_oge_gec_ec# | 0.74791504 | 57414 |
| ceia | ver | | 33 | 1.0517 | 1.5185 | 4 | 2 | nom. | 4 | 27 | 1.00 | CVVV | #c_ce_ei_ia_a# | #c_ce_cei_eia_ia# | aiec | VVVC | #a_ai_ie_ec_c# | #ai_aie_iec_ec# | 0.63512325 | 30589 |
| celo | ver | | 1 | 0.0319 | 0.0000 | 4 | 2 | nom. | 4 | 29 | 1.00 | CVCV | #c_ce_el_lo_o# | #c_ce_cel_elo_lo# | olec | VCVC | #o_ol_le_ec_c# | #ol_ole_lec_ec# | 0.13375229 | 204161 |
| cere | ver | | 1 | 0.0319 | 0.0000 | 4 | 1 | | 4 | 19 | 1.00 | CVCV | #c_ce_er_re_e# | #c_ce_cer_ere_re# | erec | VCVC | #e_er_re_ec_c# | #er_ere_rec_ec# | 0.64279096 | 204335 |
| cf | ver | | 20 | 0.6374 | 1.3010 | 3 | 1 | | 2 | 2 | 1.75 | CCP | #c_cf_f_#_ | #cf_cf_f#_ | fc | PCC | #_f_fc_c# | #_f_fc_fc# | 0.04254472 | 40456 |
| chai | ver | | 1 | 0.0319 | 0.0000 | 4 | 1 | | 4 | 4 | 1.55 | CCVV | #c_ch_ha_ai_i# | #ch_cha_hai_ai# | ihac | VVCC | #i_ia_ah_hc_c# | #ia_iah_ahc_hc# | 0.05094832 | 204309 |
| chia | ver | | 6 | 0.1912 | 0.7782 | 4 | 1 | | 4 | 12 | 1.00 | CCVV | #c_ch_hi_ia_a# | #ch_chi_hia_ia# | aihc | VVCC | #a_ai_ih_hc_c# | #ai_aih_ihc_hc# | 0.28312504 | 73356 |
| cia | ver | | 68 | 2.1671 | 1.8325 | 3 | 1 | | 3 | 30 | 1.00 | CVV | #c_ci_ia_a# | #c_ci_cia_ia# | aic | VVC | #a_ai_ic_c# | #a_aic_ic# | 0.53433187 | 20134 |
| ciam | ver | | 11 | 0.3506 | 1.0414 | 4 | 1 | | 3 | 14 | 1.00 | CVVC | #c_ci_ia_am_a# | #c_ci_cia_iam_am# | maic | CVVC | #ma_mai_ai_ic_c# | #ma_mai_aic_ic# | 0.81348379 | 54845 |
| cias | ver | | 2 | 0.0637 | 0.3010 | 4 | 1 | | 2 | 17 | 1.00 | CVVC | #c_ci_ia_as_s# | #c_ci_cia_ias_as# | saic | CVVC | #s_sa_ai_ic_c# | #sa_sai_aic_ic# | 0.28997654 | 130722 |
| cidi | ver | | 1 | 0.0319 | 0 | 4 | 1 | | 3 | 9 | 1.55 | CVCV | #c_ci_id_di_i# | #c_ci_cid_idi_di# | idic | VCVC | #i_id_di_ic_c# | #id_idi_dic_ic# | 0.41263629 | 205189 |
| cie | ver | | 5 | 0.1593 | 0.6990 | 3 | 1 | | 3 | 18 | 1.00 | CVV | #c_ci_ie_e# | #c_ci_cie_ie# | eic | VVC | #e_ei_ic_c# | #ei_eic_ic# | 0.12407612 | 80063 |
| cini | ver | | 1 | 0.0319 | 0 | 4 | 1 | | 3 | 5 | 1.7 | CVCV | #c_ci_in_ni_i# | #c_ci_cin_ini_ni# | inic | VCVC | #i_in_ni_ic_c# | #in_ini_nic_ic# | 0.16318023 | 205378 |
| cita | ver | | 618 | 19.6955 | 2.7910 | 4 | 1 | | 3 | 30 | 1.00 | CVCV | #c_ci_it_ta_a# | #c_ci_cit_ita_ta# | atic | VCVC | #a_at_ti_ic_c# | #at_ati_tic_ic# | 0.45828409 | 4745 |
| cite | ver | | 11 | 0.3506 | 1.0414 | 4 | 1 | | 4 | 14 | 1.05 | CVCV | #c_ci_it_te_e# | #c_ci_cit_ite_te# | etic | VCVC | #e_et_ti_ic_c# | #et_eti_tic_ic# | 0.23306539 | 54857 |
| cito | ver | | 55 | 1.7528 | 1.7404 | 4 | 1 | | 4 | 34 | 1.00 | CVCV | #c_ci_it_to_o# | #c_ci_cit_ito_to# | otic | VCVC | #o_ot_ti_ic_c# | #ot_oti_tic_ic# | 0.81889330 | 22891 |
| ciyo | ver | | 2 | 0.0637 | 0.3010 | 4 | 1 | | 2 | 4 | 1.80 | CCCV | #c_yi_jt_to_o# | #c_yi_jto_jto_to# | otjc | VCCC | #o_ot_tj_ic_c# | #o_oti_tjc_ic# | 0.99817464 | 130851 |
| clima | ver | | 1 | 0.0319 | 0 | 4 | 1 | | 2 | 10 | 1.1 | CCCV | #c_li_im_ma_a# | #c_lim_ima_ma# | amlc | VCCC | #a_am_ml_ic_c# | #am_aml_mic_ic# | 0.69128345 | 205808 |
| cló | ver | | 1 | 0.0319 | 0 | 3 | 1 | | 1 | 8 | 1.5 | CVV | #c_ló_ó_a# | #c_ló_óa_a# | aóc | VVC | #a_aó_óc_c# | #aó_aóc_óc# | 0.27219436 | 205862 |
| coar | ver | | 2 | 0.0637 | 0.3010 | 4 | 1 | | 4 | 11 | 1.00 | CVVC | #c_oa_ar_a# | #c_oa_oar_ar# | raoc | CVVC | #r_ra_ao_oc_c# | #ra_rao_aoc_oc# | 0.60956044 | 13093 |
| coas | ver | | 6 | 0.1912 | 0.7782 | 4 | 1 | | 4 | 15 | 1.00 | CVVC | #c_oa_as_a# | #c_oa_oas_as# | saoc | CVVC | #s_sa_ao_oc_c# | #sa_sao_aoc_oc# | 0.44505048 | 73414 |
| coça | ver | | 4 | 0.1275 | 0.6021 | 4 | 2 | nom. | 4 | 33 | 1.00 | CVCV | #c_oç_ça_a# | #c_oç_oça_ça# | açoc | VCVC | #a_aç_ço_oc_c# | #aç_açо_çoc_oc# | 0.10540561 | 89374 |
| codi | ver | | 1 | 0.0319 | 0 | 4 | 1 | | 3 | 9 | 1.5 | CVCV | #c_od_di_i# | #c_od_odi_di# | idoc | VCVC | #i_id_do_oc_c# | #id_ido_doc_oc# | 0.1638345 | 206060 |
| coe | ver | | 41 | 1.3067 | 1.6128 | 3 | 1 | | 3 | 23 | 1.00 | CVV | #c_oe_e# | #c_oe_coe_oe# | eoc | VVC | #e_eo_oc_c# | #eo_eoc_oc# | 0.63650576 | 27090 |
| cola | ver | | 39 | 1.2439 | 1.5911 | 4 | 2 | nom. | 4 | 50 | 1.00 | CVCV | #c_co_ol_la_a# | #c_co_col_ola_la# | aloc | VCVC | #a_al_lo_oc_c# | #al_alo_loc_oc# | 0.05255178 | 27863 |
| cole | ver | | 31 | 0.9880 | 1.4914 | 4 | 1 | | 4 | 36 | 1.00 | CVCV | #c_co_ol_le_e# | #c_co_col_ole_le# | eloc | VCVC | #e_el_lo_oc_c# | #el_elo_loc_oc# | 0.85920788 | 31700 |
| coli | ver | | 3 | 0.0956 | 0.4771 | 4 | 2 | nom. | 4 | 20 | 1 | CVCV | #c_co_ol_li_i# | #c_co_col_oli_li# | iloc | VCVC | #i_il_lo_oc_c# | #il_ilo_loc_oc# | 0.16136673 | 103093 |

**Figure 2**: Search result example.

It follows below the name, abbreviation, and description of each one of the 21 columns of information presented in the results obtained from a search in the Brazilian Portuguese Lexicon, Alpha.

**Orthography** (ortho): orthographic form of the word in lower case (except for proper names), respecting the specific accentuation of each word[37].

**Grammatical category** (gram_cat): grammatical category of the word (adj, adv, gram, nom, num, prop, and ver).

**Grammatical information** (gram_inf): grammatical information about the word (e.g. singular/plural, masculine/feminine, past/present/future, 1/2/3 person, etc.).

**Orthographic frequency** (ortho_freq): number of times that the word appears in NILC (around 32 million words).

**Orthographic frequency per million** (ortho_freq/M): number of times the word appears in NILC in 1 million words. Standard value used to express the frequency of a word.

**Natural logarithm of the orthographic frequency** (log10_ortho_freq): natural logarithm of the orthographic frequency. Logarithmic values are used to linearize the behavior of the frequencies of words in a corpus.

**Number of letters** (nb_letters): number of letters of the word.

**Number of homographs** (nb_homogr): number of homographic words. Words that have the same orthography or accentuation differences, but belong to different grammatical categories.

**Homographs** (homographs): grammatical categories of the homographic words.

**Orthographic uniqueness point** (pu_ortho): letter position from which the word is dissociated from the other corpus words. Direction: from left to right.

**Orthographic neighbors** (ortho_neigh): number of orthographic neighbors according to the Coltheart's N, changing only one letter at a time (Coltheart et al., 1977).

[37] We remark that the NILC was computed in 1999, before the Portuguese orthographic reform.

**Orthographic Leveinshtein Distance** (old20): average of the Orthographic Leveinshtein Distance from the 20 closest words calculated from linear regressions (Yarkoni et al., 2008).

**CVCV structure** (CVCV_ortho): word structure, where consonants are C and vowels V. Also, A for the accentuation, P for punctuation, N for numbers, and S for symbols.

**Bigrams** (bigrams): the word bigrams separated by underline "_" and bounded by hashmarks "#". The number of bigrams equals the number of letters of the word plus 1.

**Trigrams** (trigrams): the word trigrams separated by underline "_" and bounded by hashmarks "#". The number of trigrams equals the number of letters of the word.

**Reversed orthography** (rev_ortho): orthographic word reversed from (ortho).

**Reversed CVCV structure** (rev_CVCV_ortho): reversed CVCV structure of the word from (CVCV_ortho).

**Reversed bigrams** (rev_bigrams): reversed word bigrams separated by underline "_" and bounded by hashmarks "#" from (bigrams).

**Reversed trigrams** (rev_trigrams): reversed word trigrams separated by underline "_" and bounded by hshmarks "#" from (trigrams).

**Random number between 0 and 1** (random): random number between 0 and 1 with eight digits of precision.

**Identification number** (id): word identification number assigned from the corpus organized by decreasing frequency and a-z alphabetical order. The identification number is the position of the word in the lexicon and corpus.

## Simple Search

The simple search engine was developed for the search of specific words or word lists, as shown in **Figure 3**. The user must inserts the orthographic form of the words to perform a search, indeed, wildcard symbols as underline "_" for a letter and percentage "%" for a chain of letters can be used. The user can enter a list of words separated by different rows. For

example, the user can copy and paste a list of words from a spreadsheet or text editor. The user can choose the category used for the organization and presentation of words and the ascending or descending sense of organization. The "Search" button performs the search and displays the results and the "Clear" button clears the information from the fields.


**Figure 3:** Simple search.

## Complex Search

The complex search engine was developed to perform complex searches based on specific criteria of words, as the number of letters, frequency, grammatical category, orthographic neighbors, etc., as shown in **Figure 4**. In the first field, the user must choose the column of information to perform the search. In the second field, the user must choose whether to consider "yes" or disregard "no" the criteria. In the third field, the user must enter the specific criteria to be searched. The wildcard symbols underline "_" for a letter and percentage "%" for a chain of letters can be used. Still, the symbols greater than ">" and less than "<" may be used for numerical searches. The user can choose the category used for the organization and presentation of words and the ascending or descending sense of organization. The "Search" button performs the search and displays the results and the "Clear" button clears the information from the fields. Initially, the complex search presents four fields for criteria insertion, by clicking on the "+ Fields" button, the user is redirected to another page that presents eight fields for criteria insertion in complex searches.

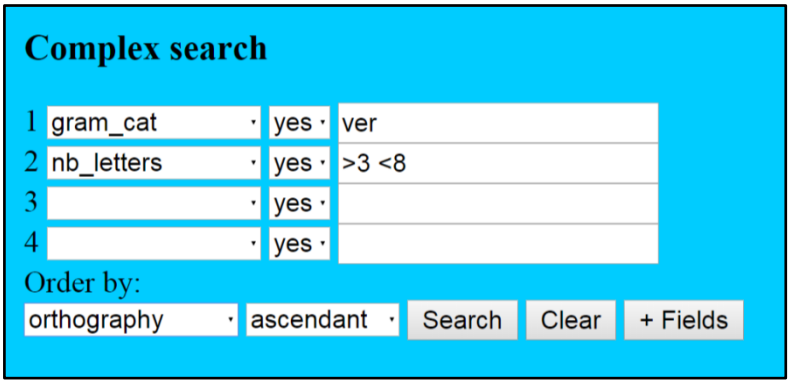



**Figure 4:** Complex search.

## Results

In the results section (**Figures 2** and **5**), the user finds the search results organized in different rows and the metalinguistic and psycholinguistic information in different columns. It can be also found specific search information, as shown in **Figure 5**: 1) total number of words found in the search, 2) range of words presented, 3) total number of pages, and 4) page number displayed. The user can choose in the upper left selection the number of words displayed on the page and navigate through the pages using the buttons "Previous" and "Next".

The top right of the results section, as shown in **Figure 5**, presents a series of statistical data established and calculated from the search (Davis & Perea, 2005; Davis, 2005): 1) mean, 2) maximum, and 3) minimum values of the following categories: 1) ortho_freq, 2) log10_ortho_freq, 3) nb_letters, d) ortho_neigh, and 4) old20. Finally, the button "Export .csv" exports all the search data to a downloadable .csv file.

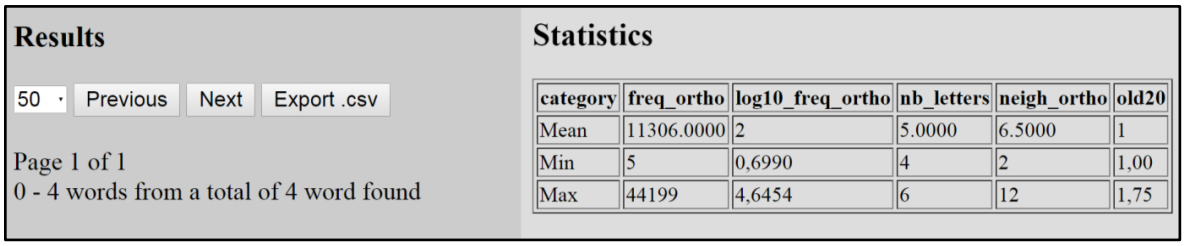



**Figure 5:** Results information and basic statistics.

## Pages

In addition to the main pages of the Brazilian Portuguese Lexicon: Lexicon [38] and Pseudowords[39], the following other pages were also created to complement the corpus and website: Downloads, Tools, Updates, Credits, and Statistical Linguistics. Downloads[40] offer for download several files from the Brazilian Portuguese Lexicon (corpus.txt, manuals, lists, conventions, bigrams, trigrams, R scripts, etc.). Tools[41] present a series of corpora links, programs, and literature on psycholinguistics, computational linguistics, and statistics. Updates[42] describe the development of the Brazilian Portuguese Lexicon and the updates made over time. Credits[43] present the proposal, the source, and the authors of the Brazilian Portuguese Lexicon; it still describes the references and relevance of the NILC/São Carlos corpus, *Linguateca* website, *Lexique* corpus, R program and packages, and Creative Commons License, ending with the acknowledgments. Finally, Statistical Linguistics is a page that provides various open and free tools and resources, as described below.

# Pseudowords

The BP pseudoword generator engine was developed to create pseudowords based on the structure and frequency of the BP words (Keuleers & Brysbaert, 2010). Unlike other pseudoword generator engines based on the syllabic structure of existing words (Keuleers & Brysbaert, 2010; Mota & Resende, 2013), the BP pseudoword generator engine of the Brazilian Portuguese Lexicon uses the bigrams and trigrams (B. New et al., 2001). All the bigrams and trigrams were computed from all the words of the Brazilian Portuguese Lexicon. The BP pseudowords are generated from the frequency and combination of bigrams or trigrams. We computed the 1) general bigram and trigram frequency, 2) bigram and trigram frequency in function of the position in the word, and 3) bigram and trigram frequency by grammatical categories.

[38] http://www.lexicodoportugues.com/index_en.php
[39] http://www.lexicodoportugues.com/pseudowords_en.php
[40] http://www.lexicodoportugues.com/downloads_en.php
[41] http://www.lexicodoportugues.com/tools_en.php
[42] http://www.lexicodoportugues.com/updates_en.php
[43] http://www.lexicodoportugues.com/credits_en.php

The user must enter four fields in the BP pseudoword generator engine: 1) number of words to be generated, 2) number of letters of the words to be generated, 3) grammatical category that these words must belong (all, adj, adv, gram, nom, num, ver), and 4) criterion for the construction of words (bigrams or trigrams). The BP pseudoword generator engine builds words simultaneously in both directions, from left to right and from right to left, beginning with a bounded bigram or trigram "#xx" or "#xx". According to the number of letters, the engine concatenates new bigrams or trigrams that share as much orthographic information as possible to the previous bigram or trigram (1 letter for bigrams and 2 letters for trigrams). The BP pseudoword generator engine has two buttons: "Submit" to generate and present the pseudoword results and "Clear" to clear all fields, as shown in **Figure 6**.

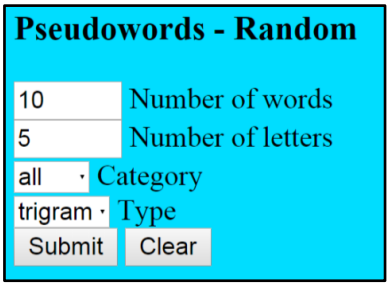
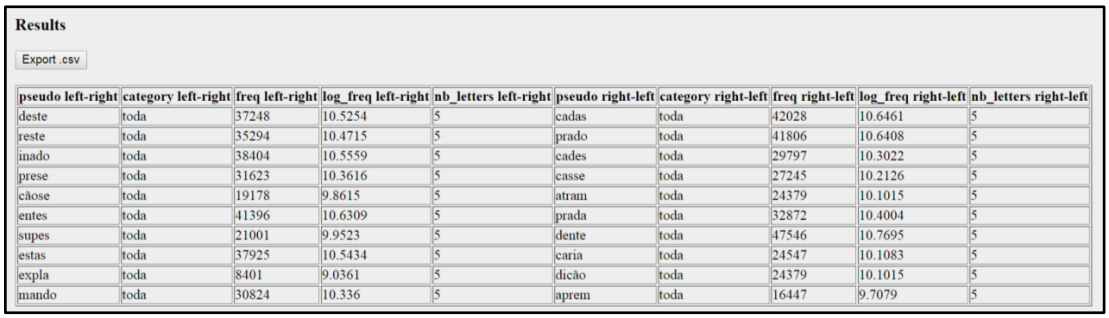



**Figure 6:** BP pseudoword generator engine.

In the pseudoword results, as shown in **Figure 7**, it is presented four columns with information about the pseudowords: 1) grammatical category defined by the user, 2) pseudoword frequency calculated from the sum of the frequencies of the pseudoword bigrams or trigrams, 3) log10 of the pseudoword calculated frequency, and 4) number of letters of the pseudoword. It is available the button "Export .csv" to export the results of the BP pseudowords generator engine to a downloadable .csv file.



Results

Export .csv

| pseudo left-right | category left-right | freq left-right | log_freq left-right | nb_letters left-right | pseudo right-left | category right-left | freq right-left | log_freq right-left | nb_letters right-left |
|---|---|---|---|---|---|---|---|---|---|
| deste | toda | 37248 | 10.5254 | 5 | cadas | toda | 42028 | 10.6461 | 5 |
| reste | toda | 35294 | 10.4715 | 5 | prado | toda | 41806 | 10.6408 | 5 |
| inado | toda | 38404 | 10.5559 | 5 | cades | toda | 29797 | 10.3022 | 5 |
| prese | toda | 31623 | 10.3616 | 5 | casse | toda | 27245 | 10.2126 | 5 |
| cãose | toda | 19178 | 9.8615 | 5 | atram | toda | 24379 | 10.1015 | 5 |
| entes | toda | 41396 | 10.6309 | 5 | prada | toda | 32872 | 10.4004 | 5 |
| supes | toda | 21001 | 9.9523 | 5 | dente | toda | 47546 | 10.7695 | 5 |
| estas | toda | 37925 | 10.5434 | 5 | caria | toda | 24547 | 10.1083 | 5 |
| expla | toda | 8401 | 9.0361 | 5 | dicão | toda | 24379 | 10.1015 | 5 |
| mando | toda | 30824 | 10.336 | 5 | aprem | toda | 16447 | 9.7079 | 5 |

**Figure 7:** BP pseudoword generator engine results.

# Statistical Linguistics

The page Statistical Linguistics[44] in the Brazilian Portuguese Lexicon provides free and open tools and resources for psycholinguistic, linguistic, and statistics that can be consulted directly on the page on the internet. These resources and tools were developed in HTML/PHP: a) F' and minF' - MS, b) minF' - F1.F2, c) Hartley test, d) normalization between 0 and 1, d) reverse words, f) Hamming distance, g) Levenshtein distance, h) orthographic neighbors (Coltheart's N), i) average Levenshtein distance, j) relative entropy, k) word frequency, and l) Zipf's distribution.

# Authors

The Brazilian Portuguese Lexicon is being developed by Gustavo Lopez Estivalet during his Ph.D financed by the program Science without Borders (CsF) [45] from the National Council for Scientific and Technological Development (CNPq) [46] scholarship, Brazil, between 2012 and 2016. Gustavo Lopez Estivalet holds his Ph.D in France in the city of Lyon, in the University Claude Barnard Lyon 1 (UCBL) [47] in the Doctoral School of Neurosciences et Cognition (ED NSCo) [48], in the Laboratory on Language, Brain, and Cognition (L2C2) [49], located at the Institute of Cognitive Sciences (ISC) [50], and being supervised by Prof. Dr. Fanny Meunier[51], which are funded by the National Council for Scientific Research (CNRS) [52].

[44] http://www.lexicodoportugues.com/stat_ling_en.php
[45] http://www.cienciasemfronteiras.gov.br/web/csf
[46] http://www.cnpq.br/
[47] http://www.univ-lyon1.fr/
[48] http://nsco.universite-lyon.fr/
[49] http://l2c2.isc.cnrs.fr/fr/
[50] http://www.isc.cnrs.fr/
[51] http://www.researchgate.net/profile/Fanny_Meunier/
[52] http://www.cnrs.fr/

# Acknowledgements

For the realization and success of the Brazilian Portuguese Lexicon, I thank the National Council for Scientific and Technological Development (CNPq) by the Ph.D scholarship from the program Science without Borders (CsF). I thank my Ph.D supervisor Prof. Dr. Fanny Meunier and Prof. Dr. Michel Hoen[53], who understood the importance of developing a BP psycholinguistic corpus. I thank the NILC Prof. Dr. Sandra M. Aluísio[54] and Prof. Dr. Maria das Graças Volpe Nunes[55] by the help in materials, information, and discussions on the NILC, as well as support, motivation and recognition of this work. I thank my work colleagues Léo Varnet[56] and Emmanuel Trouche[57] by the discussions on scripts and algorithms for the development of Brazilian Portuguese Lexicon. I thank the internet community users who work with website development and database management for the forum discussions and tutorials available. Finally, I thank Prof. Dr. Mailce Borges Mota, and the best BP Prof. Lise Lopez. Finally, I thank Luanda Lins by understanding the importance of this project and my motivation to do it. Thank you all!

# References

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

[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

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™//A lexical database for contemporary french : LEXIQUE™. *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