MULTILEXICON: A Psycholinguistic Database Lexicon from Multiple Languages

Gustavo Estivalet, Maylton Fernandes, Márcio Leitão

Federal University of Paraiba, Laboratory of Language Processing, Brazil

Joao Pessoa, 10th October 2020.

Abstract

A common challenge in psycholinguistic researcher is the use of reliable lexical norms of frequency and neighborhood for experimental stimuli selection and control. In multilingualism investigation, this challenge is multiplied by lexicon differences and interlanguage influences. Therefore, the use of a multilingual database aligned by word translation with intra- and inter-language norms would be an optimal scenario and a useful tool for stimuli selection and control in psycholinguistic experiments (Marian et al., 2012).

The primary objective of this work was the development of the MULTILEXICON, a word-based database with multiple languages aligned by translation containing many intra- and inter-language lexical norms (Davis, 2005). The secondary objectives were: i. construct a friendly open-access multilingual lexicon aligned by word translation; ii. deliver the main standardized norms of frequency, neighborhood, and orthographic similarity among languages; and iii. provide an analysis of these norms distribution for English, French, and Portuguese .

Therefore, two questions guided this work: How neighborhood and orthographic similarity might be approached in multilingualism research? How the word norms from the languages above interact among each other in the lexicon? The main hypothesis is that typological characteristics may define the neighborhood and similarity norms with strong correlation between French and Portuguese, moderate between French and English, and weak between English and Portuguese. Our hypothesis is that beyond typological characteristics, there is a common large etymological convergence among these languages regarding neighborhood and formal similarity (Davis et al., 2009).

Method

For the database construction, we used as word-based lexicon sources the Brazilian Portuguese Lexicon for Portuguese (Estivalet et al., 2019; Estivalet & Meunier, 2015), the SUBTLEX-UK for English (van Heuven et al., 2014), and the *Lexique* for French (New et al., 2004, 2007). All words checked by the Hunspell dictionaries from these languages were translated using the Google Sheets and Google Translate. Compound translations in two or more words were deleted from this version of the lexicon; further, compound translations will be simplified to ensure a one-to-one translation pair for each word. Afterwards, intra- and inter-language neighborhoods were calculated for each word by means of Coltheart's N (Coltheart et al., 1977), Levenshtein Distance (Levenshtein, 1966), OLD20 (Yarkoni et al., 2008), and uniqueness point. Finally, orthographic similarity among language pairs were calculated by the Relative Levenshtein Distance (Schepens et al., 2012).

Results

The result was a friendly multilingual lexicon with more than 120K word-types from each language with aligned translations, neighborhood, orthographic similarity, as well as complementary norms, such as frequency, word length, and cvcv structure for English, French, and Portuguese (Brysbaert & New, 2009).

We hope the MULTILEXICON become a useful and reference tool in future psycholinguistic and translation work. It will be extended to other well studied languages in psycholinguistics, such as German, Dutch, Spanish, and Italian (Gimenes & New, 2016). Finally, the MULTILEXICON is an open-access database available in the Internet: http://www.lexicodoportugues.com/multilexicon/index.html.

Nb	Category	Description	Nb	Category	Description
1	ortho_en	Orthography – EM	46	old20_en_fr	OLD20 – EN in FR
2	ortho_fr	Orthography – FR	47	old20_en_bp	OLD20 – EN in BP
3	ortho_bp	Orthography – BP	48	old20_fr_en	OLD20 – FR in EN
4	freqM_en	Frequency per million – EN	49	old20_fr_bp	OLD20 – FR in BP
5	freqM_fr	Frequency per million – FR	50	old20_bp_en	OLD20 – BP in EN
6	freqM_bp	Frequency per million – BP	51	old20_bp_fr	OLD20 – BP in FR
7	zs_en	Zipf scale – EN	52	old20_en_all	OLD20 – EN in all
8	zs_fr	Zipf scale – FR	53	old20_fr_all	OLD20 – FR in all
9	zs_bp	Zipf scale – BP	54	old20_bp_all	OLD20 – BP in all
10	zr_en	Zipf rank – EN	55	up_en_fr	Uniqueness point – EN in FR
11	zr_fr	Zipf rank – FR	56	up_en_bp	Uniqueness point – EN in BP
12	zr_bp	Zipf rank – BP	57	up_fr_en	Uniqueness point – FR in EN
13	nchar_en	Number of characters – EN	58	up_fr_bp	Uniqueness point – FR in BP
14	nchar_fr	Number of characters – FR	59	up_bp_en	Uniqueness point – BP in EN
15	nchar_bp	Number of characters – BP	60	up_bp_fr	Uniqueness point – BP in FR
16	neigh_en	Coltheart’s N – EN	61	up_en_all	Uniqueness point – EN in all
17	neigh_fr	Coltheart’s N – FR	62	up_fr_all	Uniqueness point – FR in all
18	neigh_bp	Coltheart’s N – BP	63	up_bp_all	Uniqueness point – BP in all
19	ld_en	Levenshtein Distance – EN	64	wld_en_fr	Word Levenshtein Distance – EN to FR
20	ld_fr	Levenshtein Distance – FR	65	wld_en_bp	Word Levenshtein Distance – EN to BP
21	ld_bp	Levenshtein Distance – BP	66	wld_fr_en	Word Levenshtein Distance – FR to EN
22	old20_en	OLD20 – EN	67	wld_fr_bp	Word Levenshtein Distance – FR to BP
23	old20_fr	OLD20 – FR	68	wld_bp_en	Word Levenshtein Distance – BP to EN
24	old20_bp	OLD20 – BP	69	wld_bp_fr	Word Levenshtein Distance – BP to FR
25	up_en	Uniqueness point – EN	70	wrld_en_fr	Word Relative Levenshtein Distance – EN to FR

26	up_fr	Uniqueness point – FR	**71**	wrld_fr_bp	Word Relative Levenshtein Distance – FR to BP
27	up_bp	Uniqueness point – BP	**72**	wrld_bp_en	Word Relative Levenshtein Distance – BP to EN
28	neigh_en_fr	Coltheart's N – EN in FR	**73**	wup_en_fr	Word uniqueness point – EN to FR
29	neigh_en_bp	Coltheart's N – EN in BP	**74**	wup_fr_bp	Word uniqueness point – FR to BP
30	neigh_fr_en	Coltheart's N – FR in EN	**75**	wup_bp_en	Word uniqueness point – BP to EN
31	neigh_fr_bp	Coltheart's N – FR in BP	**76**	flo_en_fr	First letter overlap – EN to FR
32	neigh_bp_en	Coltheart's N – BP in EN	**77**	flo_fr_bp	First letter overlap – FR to BP
33	neigh_bp_fr	Coltheart's N – BP in FR	**78**	flo_bp_en	First letter overlap – BP to EN
34	neigh_en_all	Coltheart's N – EN in all	**79**	cvcv_en	CVCV structure – EN
35	neigh_fr_all	Coltheart's N – FR in all	**80**	cvcv_fr	CVCV structure – FR
36	neigh_bp_all	Coltheart's N – BP in all	**81**	cvcv_bp	CVCV structure – BP
37	ld_en_fr	Levenshtein Distance – EN in FR	**82**	cvcvo_en_fr	CVCV overlap – EN to FR
38	ld_en_bp	Levenshtein Distance – EN in BP	**83**	cvcvo_fr_bp	CVCV overlap – FR to BP
39	ld_fr_en	Levenshtein Distance – FR in EN	**84**	cvcvo_bp_en	CVCV overlap – BP to EN
40	ld_fr_bp	Levenshtein Distance – FR in BP	**85**	rev_en	Reverse word – EN
41	ld_bp_en	Levenshtein Distance – BP in EN	**86**	rev_fr	Reverse word – FR
42	ld_bp_fr	Levenshtein Distance – BP in FR	**87**	rev_bp	Reverse word - BP
43	ld_en_all	Levenshtein Distance – EN in all	**88**	random	Random number
44	ld_fr_all	Levenshtein Distance – FR in all	**89**	id	Identity
45	ld_bp_all	**Levenshtein Distance – BP in all**			

Table 1: Categories of the MULTILEXICON.

References

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. https://doi.org/10.3758/BRM.41.4.977

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). 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. https://doi.org/10.3758/BF03206399

Davis, C. J., Perea, M., & Acha, J. (2009). Re(de)fining the orthographic neighborhood: The role of addition and deletion neighbors in lexical decision and reading. *Journal of Experimental Psychology: Human Perception and Performance, 35*(5), 1550–1570. https://doi.org/10.1037/a0014253

Estivalet, G. L., Hartmann, N. S., Marquiafavel, V., Lukasova, K., Carthery-goulart, M. T., & Aluisio, S. M. (2019). LexPorBR Infantil: Uma base lexical tripartida e com interface Web de textos ouvidos, produzidos e lidos por crianças. In C. A. Prolo & L. H. M. de Oliveira

(Eds.), *Proceedings of the XII Symposium in Information and Human Language Technology (STIL2019)* (pp. 190–199). http://comissoes.sbc.org.br/ce-pln/stil2019/proceedings-stil-2019-Final-Publicacao.pdf

Estivalet, G. L., & Meunier, F. (2015). The Brazilian Portuguese Lexicon: An Instrument for Psycholinguistic Research. *PLOS ONE*, *10*(12), e0144016. https://doi.org/10.1371/journal.pone.0144016

Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, *48*(3), 963–972. https://doi.org/10.3758/s13428-015-0621-0

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and reversals. *Soviet Physics*, *10*(8), 707–710.

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. https://doi.org/10.1371/journal.pone.0043230

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(04). https://doi.org/10.1017/S014271640707035X

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. https://doi.org/10.3758/BF03195598

Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, *15*(1), 157–166. https://doi.org/10.1017/S1366728910000623

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

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. https://doi.org/10.3758/PBR.15.5.971