Multi-LEX: a database of multi-word frequencies for French and English

Marjorie Armando, Jonathan Grainger, Stephane Dufau

Accepted author version, Behavior Research Methods, 2022

Multi-LEX: a database of multi-word frequencies for French and English

Marjorie ARMANDO[1,2,3], Jonathan GRAINGER[1], Stephane DUFAU[1,4]

1. Laboratoire de psychologie cognitive (UMR7290), CNRS & Aix-Marseille University, France
2. Aix Marseille Univ, CNRS, LIS, Marseille, France
3. Pôle pilote Ampiric, Institut National Supérieur du Professorat et de l'Éducation, Aix-Marseille Université, Marseille, France
4. Queensland Brain Institute, The University of Queensland, Brisbane, QLD, Australia

CORRESPONDING AUTHOR: Stephane DUFAU, Laboratoire de psychologie cognitive UMR7290, Aix-Marseille Université Case D, 3, place Victor HUGO, 13331 MARSEILLE CEDEX 3, France. Email : stephane.dufau@univ-amu.fr, Telephone : +33 4 13 55 09 77

ABSTRACT

Written word frequency is a key variable used in many psycholinguistic studies and is central in explaining visual word recognition. Indeed, methodological advances on single word frequency estimates have helped to uncover novel language-related cognitive processes, fostering new ideas and studies. In an attempt to support and promote research on a related emerging topic, visual multi-word recognition, we extracted from the exhaustive Google Ngram datasets a selection of millions of multi-word sequences and computed their associated frequency estimate. Such sequences are presented with Part-of-Speech information for each individual word. An online behavioral investigation making use of the French 4-gram lexicon in a grammatical decision task was carried out. The results show an item-level frequency effect of word sequences. Moreover, the proposed datasets were found useful during the stimulus selection phase, allowing more precise control of the multi-word characteristics.

INTRODUCTION

Written word frequency is a key variable used in many psycholinguistic studies and is central in explaining visual word recognition when the subject performs a lexical decision or naming task. Indeed, word frequency is intrinsically linked to the level of activation of words in computational models that make use of a lexical access component (the original Interactive-Activation model, McClelland & Rumelhart, 1981, and subsequent models such as that of Davis, 2010, Grainger & Jacobs, 1996, or Perry, Ziegler & Zorzi, 2007). In the original Interactive-Activation model of single word recognition, McClelland and Rumelhart employed word frequency to determine the resting-level activation of word nodes prior to stimulus onset. Whether the aim is to select a confound-free linguistic material, to accurately assess word frequency effect in empirical research or to disprove a hypothesis in modeling a particular psycholinguistic phenomenon, it is crucial to have the best possible estimates of word counts (see Zevin & Seindenberg, 2002, for a seminal study comparing the influence of a particular frequency metric, and Brysbaert, Mandera & Keuleers, 2018, for a more recent review). A variable of such importance has unsurprisingly stimulated many methodological proposals. To name a few, Kucera and Francis (1967) relied on the 1-million-word Brown corpus to compute individual word counts; in the 1990's, more than 3000 different Usenet newsgroup conversations were aggregated, providing a corpus of 131 million words (HAL-corpus; 70000 unique words; Burgess & Livesay,1998); and a corpus of 16.6 million words from British and American English texts was used in CELEX (Baayen, Piepenbrock &., & Gulikers, 1996).

More recent methodological advances were made with the use of movie subtitles (New et al., 2007), which are thought to better reflect actual language usage in the general population. In controlled experiments employing a lexical decision task for example, such a word count explains the greatest variance in participants' performance (Brysbaert & New, 2009), even though a composite variable made up of the best metrics have been used (see for example Ferrand et al., 2018). Advances have also been made in building corpora and word frequency estimates in languages other than English (Chinese, Dutch, French, Greek, Portuguese or Spanish; see Boada et al., 2019, for the most recent proposal in Catalan) or in targeting a population other than adults (see Zeno et al., 1995, Lete et al., 2004, or Terzopoulos et al., 2017, for child-based material). Databases including a word frequency estimate are proposed in a variety of supports, the oldest ones being proposed in book format (e.g. Zeno) or on CD-ROM (e.g. CELEX) and the most recent being freely downloadable online as single files (SUBTLEX family). To facilitate research, many resources associate these files with an online search engine capability, such as the English Corpora[1] (English) or Lexique[2] (French).

Methodological advances in single word frequency measures are directly linked to the level of research interest in single word recognition. Such advances have helped in uncovering novel cognitive processes and foster new ideas and studies. The success achieved by the word-based psycholinguist community in understanding cognitive phenomena in this domain could not have been achieved without normative databases. It is essential for any scientific discipline to generate and facilitate the use of such tools, and this is especially true for a domain that is emerging.

MULTI-WORDS

Recent advances in the study of reading, embedding the single word recognition processes in a more global sentence processing context has opened up new research directions. Indeed, Grainger, Dufau and Ziegler (2016) proposed a new framework in which orthographic

processing that spans multiple words connects word reading to sentence reading (see also Snell et al., 2018, for computational implementation of such a mechanism). Amongst other advances building on this framework, Snell and Grainger (2017) revisited the sentence superiority effect in the light of the hypothesized parallel multi-word orthographic processing. In this study, participants had to identify a word within a 4-word sequence displayed for 200 ms, a sequence being grammatical or ungrammatical. Results showed enhanced word identification performance within grammatical sequences, thus pointing to a level of parallel processing across multiple words that enables rapid extraction of their syntactic categories.

Researchers interested in uncovering cognitive processes behind multiple word recognition need a reliable and easy-to-access resource to select and control the linguistic multi-word material and to analyze participant data. However, little has so far been published in the psycholinguistic literature on sequences of multiple words and their associated frequencies, and experimenters are left with relatively few practical resources. For instance, in English, the British National Corpus can be used to search for particular word combinations (pre-selected by the experimenter) and to measure their associated frequencies. In an eye-movement study where participants had to read a single sentence presented on a screen, Siyanova-Chanturia et al. (2011) used frequent triplets of words like "knife and fork" and reversed the word order ("fork and knife") to generate less frequent forms of word combinations. The material selection and phrasal frequency measurement were carried out through the British National Corpus in addition to a human-based completion test, to ensure that the normal and reversed combinations were distinct in frequency. In a same-different matching task in French, Pegado and Grainger (2020) created their material by assembling 5 different words and manipulating their order to create ungrammatical sequences without further control over the frequency of the combinations. Even if a corpus is used to measure frequency, stimulus creation and/or selection is still achieved by the experimenter with unavoidable individual biases as exposed in Kuperman (2015): "compilation of lists of stimuli with required characteristics may be non-random and critically depends on experimenters' intuitions and experience, leaving the door open to experimenter bias". In other words, the development of multi-word databases could be the key to scientific progress in the domain of multi-word processing. On the data analysis side, a multi-word frequency effect could arise from a grammatical decision task where participants have to distinguish grammatical word sequences from ungrammatical ones (e.g. see Mirault, Snell & Grainger, 2018, where speeded grammaticality judgments revealed a transposed-word effect). A sentence database with frequency norms would be an ideal tool for revealing effects of multi-word frequency in grammatical decisions – just like lexical decision has been for word frequency. To reveal such an effect on the item level (therefore at a finer grain than the high- Vs. low-frequent group level as in Siyanova-Chanturia et al., 2011), we conducted a tentative experiment with a grammatical decision task that made use of our list of French 4-grams and their associated frequencies.

The Google Books corpus and the Ngram datasets.

Since 2002, in order to archive and reference human knowledge, Google has scanned an estimated 25 million books published in 430 different languages and has performed automatic optical character recognition to transform printed texts to machine-encoded material. Results of this initiative,

Google Books, are available online[3]. The Google Books initiative gave birth to two side projects, Google Scholar, which indexes full text and/or metadata of scholarly literature, and Ngram, which reports frequencies of any set of words and charts their values (Michel et al., 2011). The latest version of the Google Books Ngram datasets (2019) consists of a list of combinations of words, the "n-grams", and their occurrences over five centuries. The letter "n" in n-grams stands for the number of words in any given combination, going from n=1 (single words; unigram) to n=5 (5-word sequence; 5-gram), "gram" being used in the sense of a unit. Such a dataset follows a series of works by the same company that measured the frequency of n-grams in web pages, first in English (Brants, 2006) then in 10 other European languages (Brants & Franz, 2009). The more recent database on published books was used for example to perform a quantitative analysis of affectionate communication in the past 50 years in China (Wu et al., 2019) or to analyze how rational versus reasonableness judgments are associated with different contexts (Grossmann et al., 2020). The current Ngram corpus is based on 8 million books randomly selected from the Google Books corpus (including 4.5 million books in English) and represents an estimate of 6% of all the books ever published at the time of the corpus publication. To provide a language-independent interface, n-grams were tokenized and annotated for syntax with 12 language universal part-of-speech (PoS) tags introduced in Petrov, Das & McDonald (2011): NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuation marks) and X (a catch-all for other categories such as foreign words). The part-of-speech tagging procedure is described in detail in Lin et al. (2012). An example of an English 4-gram is "the_DET house_NOUN is_VERB red_ADJ" where PoS are attached to their referring word with an underscore.

The Ngram datasets are available in eight languages (English, Spanish, French, German, Russian, Italian, Chinese and Hebrew), all of them being proposed with an online and graphical interactive search engine[4] (see Figure 1). In addition to this viewer, the compressed text files containing both the word-based- and PoS-related-raw data are available for direct download.

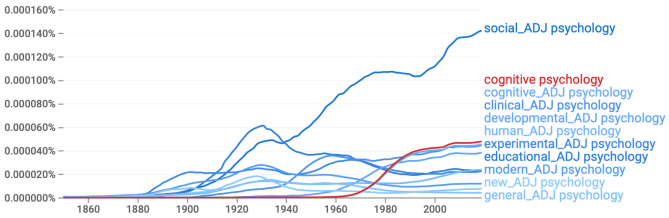

Figure 1. Written frequency of the most frequent 2-word associations with a generic adjective as the first member and "psychology" as the second member, for the period 1850-2019 (x-axis; smoothing of 8 years). Results are from the Google Ngram viewer for the dual search "*_ADJ psychology" and "cognitive psychology" (plotted in red). Frequency is given in percentage (y-axis).

The n-gram lexicons tailored for psycholinguistics.

The current proposal aims at making the handling of the n-gram lexicon databases more suitable for psycholinguistic research. More specifically, from the Ngram datasets, we propose a selection of 2- to 5-word sequences in French and English along with their associated frequency estimate. Lists of word sequences are available for e.g. material selection based on word identity, part-of-speech information and occurrences. In addition, an online interface is proposed for database exploration and material selection with ready-to-use filters, a handy tool for researchers not willing or able to manipulate lists of millions of sequences on their computer.

SELECTION OF NGRAMS

The Google Ngram database can be manipulated in two ways. The first method utilizes the dedicated online search engine and display tool developed by Google (Figure 1). It can be used as a set of predefined

word sequences to be analyzed. Although not intended to be manipulated in this way, the online tool displays frequency values expressed in percentages that can be automatically extracted from the source code of the web page. The second method consists of starting from raw data and extracting word sequences that are suitable for psycholinguistic research, a method employed and described here. The current version of the Ngram database raw data is composed of hundreds to thousands language-specific files, each of them with millions of word sequences associated to their number of occurrences per year (one thousand 5-gram files of 4 million sequences each leading to 4 billion sequences listed over 5 centuries; one hundred 2-gram files leading to 400 million sequences). At this stage, the n-gram lists are composed of PoS-tagged sequences ("the_DET table_NOUN"), untagged sequences ("the table") and partially tagged sequences ("the_DET table"), these three forms having the same number of occurrences. Concerning the number of occurrences, in recent years, approx. 5 billion n-gram occurrences per year are listed by Google for each language and each N.

Given the mixed PoS composition of the n-gram lists, the necessary step for usability was therefore to select the meaningful sequences that could potentially be useful for psycholinguists. To do so, we downloaded the different 2019 Ngram data files on SSD drives for efficient data reading and writing (2- to 5-grams, French and British English file sets, available in a compressed format). We wrote computer programs handling parallel processing to read and process data and prepared a 40-core workstation to work on these n-gram lexicons. We applied to the initial n-gram lists the following treatments. We first selected the n-grams that had no punctuation signs, that were fully PoS-tagged and that had at least two records in the time period from year 2012 to year 2019. We then lowercased the sequences and summed the number occurrences in which the n-gram appeared in over these years. In the online Ngram Viewer, this simple selection corresponds to a search with the case-insensitive option activated (meaning that the occurrences of "The table" and "the table" are added), years set to 2012 and 2019, and a smoothing option set to zero (then summing the last 8-year output values makes up for the summation over the period of interest). The next step was to match individual words in the word sequences to a list of the most frequent words extracted from single word databases: 37,559 unique English words from the SUBTLEX-UK and SUBTLEX-US word lexicons (Van Heuven et al., 2014; Brysbaert & New, 2009), and 47,707 unique French words from the LEXIQUE-3 lexicon (New et al., 2001; French having more inflected forms than English). Those n-grams that did not have at least one match were discarded, meaning that selected n-grams had at least one of their unigrams on the single word lexicons. These two selection steps discarded most of the n-grams that were of low frequency or ill-formed, leading to eight lists of 50 million (2-grams) to 500 million (5-grams) unique n-grams per language. The last selection process was more quantitative and was based on the frequency of the n-grams. Indeed, to be manageable, these lists had to be reduced: for each single word in the word lexicons, we selected the n-grams containing this word in their unigrams, e.g. selecting "american society" or "american revolution" by searching for the single word "american" in the English 2-gram lexicon. We then computed the minimum between the set size and the square root of the set size multiplied by 10. The impact of such formulae on a set size follows. Say a frequent word was present in 100,000 sequences. The square root of 100,000 is about 316. We selected the 10 x 316 = 3160 most frequent sequences corresponding to this word (as the minimum between 3160 and 100,000 is 3160), thus discarding 90k+ less frequent sequences. For a less frequent word appearing in say ten n-grams, we selected all of the ten n-grams as 10 x 3.16 (3.16 being the square root of 10) is higher than 10. As this example shows, this simple formula ensures that (i) the single-word-based n-gram sets of 100 or less n-grams were not affected by this last selection process, and (ii) the single-word-based n-gram sets of more than 100 n-grams are limited by the square root of their sizes. This step selected approx. ten million unique word sequences per language and per N. At this stage, the n-gram lexicons have misspelled unigrams or unigrams of foreign origin, and some of the n-grams are non-phrases like "root of the". To overcome this situation, an additional filter was applied to these n-gram lexicons: we selected the 50,000 most frequent word sequences, in parallel to a random selection of 950,000 sentence-like sequences (e.g. the sequence "an additional filter" was selected and not the sequence "additional filter was"). At the very end of the selection process, each dataset of French and English 2- to 5-grams is therefore proposed in three lengths: the approx. ten-million-sequence full selection, the one-million-sequence selection composed of sentence-like sequences, and the most-frequent 50,000-sequence selection. The ten-million-sequence selection is only available in a compressed csv format (for more details, see the open practice section).

SELECTION RESULTS

Eight lists of 2- to 5-word sequences were generated for both English and French. They contained approx. 10 million sequences each, each of the lists being composed of between 86,393 and 130,703 unique words. An overview of the n-gram lexicons is provided in Table 1. We can see that the size of the single word lexicons as well as the n-gram lexicons is a bit larger for French, such difference being probably linked to a larger number of derivatives word forms, especially for the verb category (e.g. the family size of "abandonner" in the French 5-gram lexicon is 27 while the family size of the English cognate "to abandon" in the English 5-gram lexicon is 5). In the "examples" column of Table 1, generic selection rows, it can be noted that not all of the n-grams form stand-alone phrases: some appear to be missing extra words to be complete as in "did not transform into" or "on the marsh and the". Such *incomplete* sequences are less present in the sentence-like selection rows. Nevertheless, it can be anticipated that most psycholinguists selecting their material would probably need extra selection steps. The online web application provides just this sort of tool.

Table1. Characteristics of the n-gram lexicons after selection. Three selections are proposed per language and number of grams. Unique sequences are referred to as types in the single word literature.

Selection	Language	Lexicon	Number of unique sequences	Number of unique words within sequences	Examples
Generic	English	2-grams	8,161,430	90,835	instincts_NOUN dulled_VERB he_PRON daring_VERB shining_VERB countenance_NOUN some_DET toffee_NOUN cool_ADJ depths_NOUN
		3-grams	15,448,922	96,390	may_VERB also_ADV learn_VERB where_ADV the_DET friendlies_NOUN gallery_NOUN in_ADP chelsea_NOUN on_ADP a_DET plane_NOUN concierge_NOUN to_PRT have_VERB
		4-grams	15,116,507	90,139	they_PRON can_VERB not_ADV abide_VERB drained_VERB out_ADP of_ADP his_PRON the_DET smallest_ADJ possible_ADJ volume_NOUN did_VERB not_ADV transform_VERB into_ADP was_VERB an_DET early_ADJ opponent_NOUN
		5-grams	10,602,231	86,393	for_ADP the_DET typing_NOUN of_ADP the_DET the_DET heligoland_NOUN bight_NOUN in_ADP the_DET the_DET glories_NOUN of_ADP the_DET forest_NOUN in_ADP medicine_NOUN and_CONJ honorary_NOUN consultant_NOUN on_ADP the_DET marsh_NOUN and_CONJ the_DET
	French	2-grams	7,548,063	128,832	colonel_NOUN lui_PRON salaires_NOUN pour_ADP espérons_VERB partager_VERB militaires_ADJ exceptionnels_ADJ comme_ADP inquiétants_ADJ
		3-grams	15,425,851	129,688	soirée_NOUN dans_ADP le_DET libres_ADJ et_CONJ sur_ADP être_VERB admise_VERB ici_ADV bien_ADV fonctionné_VERB pour_ADP jouerait_VERB sur_ADP la_DET
		4-grams	16,313,733	129,770	je_PRON ne_ADV vous_PRON dise_VERB un_DET remède_NOUN et_CONJ un_DET de_ADP faire_VERB la_DET leçon_NOUN dont_PRON les_DET autorités_NOUN allemandes_ADJ d'_ADJ esprit_NOUN à_ADP la_DET
		5-grams	12,412,134	130,703	car_CONJ tu_VERB ne_ADV pourras_VERB plus_ADV des_ADP modes_NOUN de_ADP recrutement_NOUN des_ADP craqua_VERB une_DET allumette_NOUN et_CONJ mit_VERB de_ADP l'_ADV intendante_ADJ de_ADP la_DET des_ADP moindres_ADJ potins_NOUN de_ADP salon_NOUN
Sentence-like	English	2-grams	1,000,000	47,458	same_ADJ sentiment_NOUN pipes_NOUN placed_VERB small_ADJ humming_NOUN successful_ADJ modern_ADJ gay_ADJ filmmaker_NOUN
		3-grams	1,000,000	44,231	with_ADP its_PRON folds_NOUN hard_ADJ to_PRT wring_VERB we_PRON were_VERB lazing_VERB as_ADP the_DET capitalist_ADJ colours_NOUN were_VERB different_ADJ
		4-grams	1,000,000	42,488	the_DET prescribed_VERB time_NOUN limit_NOUN and_CONJ the_DET structural_ADJ funds_NOUN result_NOUN of_ADP a_DET dispute_NOUN turned_VERB pink_ADJ with_ADP pleasure_NOUN some_DET of_ADP the_DET cocaine_NOUN
		5-grams	1,000,000	43,345	an_DET empty_ADJ chair_NOUN beside_ADP her_PRON the_DET executor_NOUN of_ADP the_DET estate_NOUN the_DET life_NOUN and_CONJ death_NOUN struggle_NOUN similarity_NOUN between_ADP the_DET two_num_cases_NOUN been_VERB doing_VERB his_PRON best_ADJ to_PRT
	French	2-grams	1,000,000	55,942	meilleure_ADJ négociatrice_NOUN il_PRON saboté_VERB moins_ADV honnêtement_ADV devenait_VERB idiote_ADJ distincte_ADJ spéciale_ADJ
		3-grams	1,000,000	58,357	de_ADP télépathie_NOUN est_VERB je_PRON faillis_VERB pousser_VERB est_VERB un_DET moyen_NOUN la_DET messe_NOUN ésotérique_ADJ

					ont_VERB été_VERB approfondis_VERB
		4-grams	1,000,000	57,062	et_CONJ les_DET oignons_NOUN hachés_ADJ pour_ADP apaiser_VERB son_DET chagrin_NOUN mais_CONJ aussi_ADV les_DET particuliers_NOUN fit_VERB sursauter_VERB l'_ADV artilleur_NOUN une_DET bouffée_NOUN de_ADP printemps_NOUN
		5-grams	1,000,000	59,415	d'_ADP avoir_VERB accepté_VERB de_ADP travailler_VERB les_DET étoiles_NOUN et_CONJ les_DET planètes_NOUN qui_PRON ne_ADV coûtent_VERB pas_ADV trop_ADV construira_VERB une_DET gare_NOUN tout_ADV près_ADV entre_ADP services_NOUN centraux_ADJ et_CONJ services_NOUN
Most frequent	English	2-grams	50,000	10,090	he_PRON actually_ADV he_PRON understands_VERB government_NOUN spending_NOUN the_DET instrument_NOUN the_DET boundary_NOUN
		3-grams	50,000	5,663	do_VERB we_PRON not_ADV these_DET last_ADJ few_ADJ many_ADJ of_ADP its_PRON of_ADP the_DET program_NOUN and_CONJ the_DET earth_NOUN
		4-grams	50,000	5,037	true_ADJ to_PRT his_PRON word_NOUN we_PRON had_VERB come_VERB to_PRT is_VERB waiting_VERB for_ADP you_PRON through_ADP thick_ADJ and_CONJ thin_ADJ in_ADP spite_NOUN of_ADP having_VERB
		5-grams	50,000	5,492	you_PRON should_VERB not_ADV be_VERB here_ADV and_CONJ see_VERB if_ADP he_PRON could_VERB and_CONJ it_PRON might_VERB have_VERB been_VERB which_DET can_VERB be_VERB used_VERB to_PRT had_VERB tried_VERB so_ADV hard_ADJ to_PRT
	French	2-grams	50,000	15,483	un_DET jury_NOUN de_ADP distance_NOUN de_ADP pascal_NOUN nous_PRON suit_VERB les_DET perspectives_NOUN
		3-grams	50,000	9,514	par_ADP un_DET personnel_NOUN ne_ADV sommes_VERB plus_ADV avant_ADP de_ADP s'installer_VERB qui_PRON est_VERB assez_ADV dont_PRON les_DET contours_NOUN
		4-grams	50,000	8,926	sais_VERB pas_ADV qui_PRON vous_PRON installé_VERB dans_ADP un_DET ancien_ADJ nous_PRON ne_ADV pouvions_VERB plus_ADV de_ADP ne_ADV pas_ADV subir_VERB il_PRON fut_VERB d'_ADV abord_ADV
		5-grams	50,000	9,018	en_ADP moins_ADV de_ADP trois_DET ans_NOUN sous_ADP peine_NOUN de_ADP se_PRON voir_VERB de_ADP colère_NOUN et_CONJ d'_DET indignation_NOUN certain_ADJ nombre_NOUN d'_ADP entre_ADP elles_PRON le_DET droit_NOUN de_ADP se_PRON faire_VERB

FREQUENCY COMPUTATION

Each of the word sequences in the Ngram database is associated with a yearly-based number of occurrences which we separately aggregated over the period 2012 to 2019 (8 years). The frequency of each unique n-gram was computed by dividing its occurrence number by the sum of occurrences of all the different n-grams over the same period (a number available in the raw data). The total occurrences per N over the period were the following: 3.6783×10^{10}, 3.8615×10^{10}, 3.6783×10^{10}, 3.4950×10^{10} for English (from 2- to 5-ngrams resp.) and 3.3125×10^{10}, 3.0693×10^{10}, 3.2249×10^{10}, 3.0693×10^{10} for French (also from 2- to 5-ngrams resp.). We can note two things. First, within each language, the total occurrence counts are similar. This is not surprising given that a word is quasi-systematically embedded within a sentence, thus generating similar counts. Second, the total counts are similar between languages. This reflects a similarity between the corpus sizes; roughly a similar number of pages that were scanned and processed for the two languages. These frequency measures were further normalized by applying a 10^6 factor multiplication to match frequency norms used in single word recognition (frequency per million; fpm or FPM). A final transformation consisted in normalizing the fpm by performing successively a log10 transform (the distributions follow a normal shape) and a z-score (the distributions are centered at zero with a unit-based standard deviation). The histograms of the standardized frequency index are displayed in Figure 2. In addition to the number of occurrences, the databases contain the two frequency measures presented above, FPM and ZFI, without any other transformations.

Why FPM and ZFI? The lowest log10 transformed frequency per million value in our databases is -4.26 (cf. Table 2, English 2-grams). This corresponds to the log10 of the minimal number of occurrences in our database, 2, multiplied by a million, and divided by the total number of occurrences in the English 2-gram database (3.6783×10^{10}). Such computation giving negative values is due to the high value of the denominator (corresponding to the corpus size) that is typical of modern lexicology where corpora are gigantic compared to historical ones, e.g. the word-based Brown corpus (Kucera & Francis, 1967), that has 1 million tokens. For this specific corpus, words that occur once have a frequency per million equal to 1 and therefore a log10 transform of 0. In order to investigate word-based language usage with positive values, lexicographers came to the idea of employing a simple additional transform, i.e. going from the frequency transform log10(fpm) to log10(fpm)+4, corresponding to the log10 of the occurrences per 10 billion words (Carroll 1970, 1971). Carroll called this new transform the "Standard Frequency Index" (SFI), an index used in corpus analyses such as Manulex or HelexKids, two scholarly book-based corpora (Lété, Sprenger-Charolles & Colé, 2004; Terzopoulos et al., 2017). When such an index is used (i.e. in a relatively small corpus in token count compared to the corpus reported in this article), the SFI ranges between 10 and 90. This scale is therefore convenient enough to compute the distribution once the mean is known, the less positive values (10-50) corresponding to low frequency words and the high positive values to high frequency words (50 to 90). More recently, van Heuven et al. (2014) developed a similar index called Zipf that corresponds to the log10(fpm)+3, a log10 transform of the occurrence per billion words. Applied to a corpora of movie subtitles, the Zipf index ranges from 1 to 7. Going back to the Ngram corpus, transformations manipulating an additional term would naturally lead us to use an index corresponding to log10(fpm)+5, i.e. the occurrence per 100 billion sequences. We have not proceeded to this step as the corpus sizes in the Google Ngram databases differ a lot. The size of the American English corpus is for example twice the size of the British one (the one that we used). The Russian corpus in comparison is of a smaller size. To account for future work on these corpora, we let the FPM measure without any further transformation, such that researchers using Multi-LEX will be able to apply the transformation of their choice. We preferred to introduce a standardized score, ZFI, that put all the different measures from the different languages and N on a common ground. Indeed, as seen in the Figure 2, histograms of ZFI are quasi-normal, and thanks to the z-scoring, the mean of ZFIs have a zero value and their standard deviation is one. Unfortunately, such measures have negative values. To overcome this, one can apply an additional transformation as is done for intelligence quotient calculation, i.e. proceed to ZFI x 15 + 100, a distribution centered at 100 with a standard deviation of 15 (a value of 130 would be at 2 standard deviations away from the mean). Z-scoring a frequency is particularly useful when one works with several n-gram databases, e.g. controlling for the 2-gram inner frequencies (gram1/gram2, gram2/gram3, gram3/gram4) of a sequence of 4 grams. In such a context, standardized frequency offers a common ground in which frequencies are comparable *between* distinct n-gram databases.

FREQUENCY RESULTS

Word sequence frequencies were computed using standardized frequency formula. An overview of the frequency index is provided in Table 2 and Figure 2. They show the distributions' Gaussian shape ranging from -4 to 8. The standardized frequency index for word sequences has a minimum for French of 4-grams (-3.07) and a maximum for English 2-grams (8.17). Means and standard deviations of the distributions are the ones from a z-score transformation. The medians and means are quite close together, suggesting that the distributions are normal.

Table 2. Descriptive statistics of the Frequency per million, log10(FPM), and standardized frequency index, z(log10(fpm)). Min: minimum, Max: maximum, Mean: average, STD: standard deviation, Q25: 25th percentile, Q75: 75th percentile.

		English				French			
		2-gram	3-gram	4-gram	5-gram	2-gram	3-gram	4-gram	5-gram
FPM	Min	-4.26	-4.29	-4.26	-4.24	-4.22	-4.19	-4.21	-4.19
	Max	3.80	2.37	1.83	1.49	3.12	2.25	1.94	1.37
	Mean	-2.30	-2.37	-2.54	-2.72	-2.37	-2.33	-2.52	-2.68
	STD	0.75	0.64	0.58	0.54	0.74	0.62	0.55	0.50
	Q25	-2.81	-2.80	-2.93	-3.07	-2.88	-2.76	-2.89	-3.00
	median	-2.39	-2.41	-2.57	-2.74	-2.48	-2.39	-2.56	-2.72
	Q75	-1.89	-1.99	-2.19	-2.39	-1.97	-1.97	-2.19	-2.38
ZFI	Min	-2.63	-3.00	-2.98	-2.82	-2.49	-2.99	-3.07	-3.03
	Max	8.17	7.44	7.54	7.80	7.39	7.37	8.09	8.12
	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	STD	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Q25	-0.68	-0.68	-0.68	-0.64	-0.68	-0.70	-0.67	-0.64
	median	-0.11	-0.07	-0.05	-0.03	-0.15	-0.10	-0.08	-0.09
	Q75	0.55	0.59	0.61	0.62	0.53	0.58	0.59	0.59

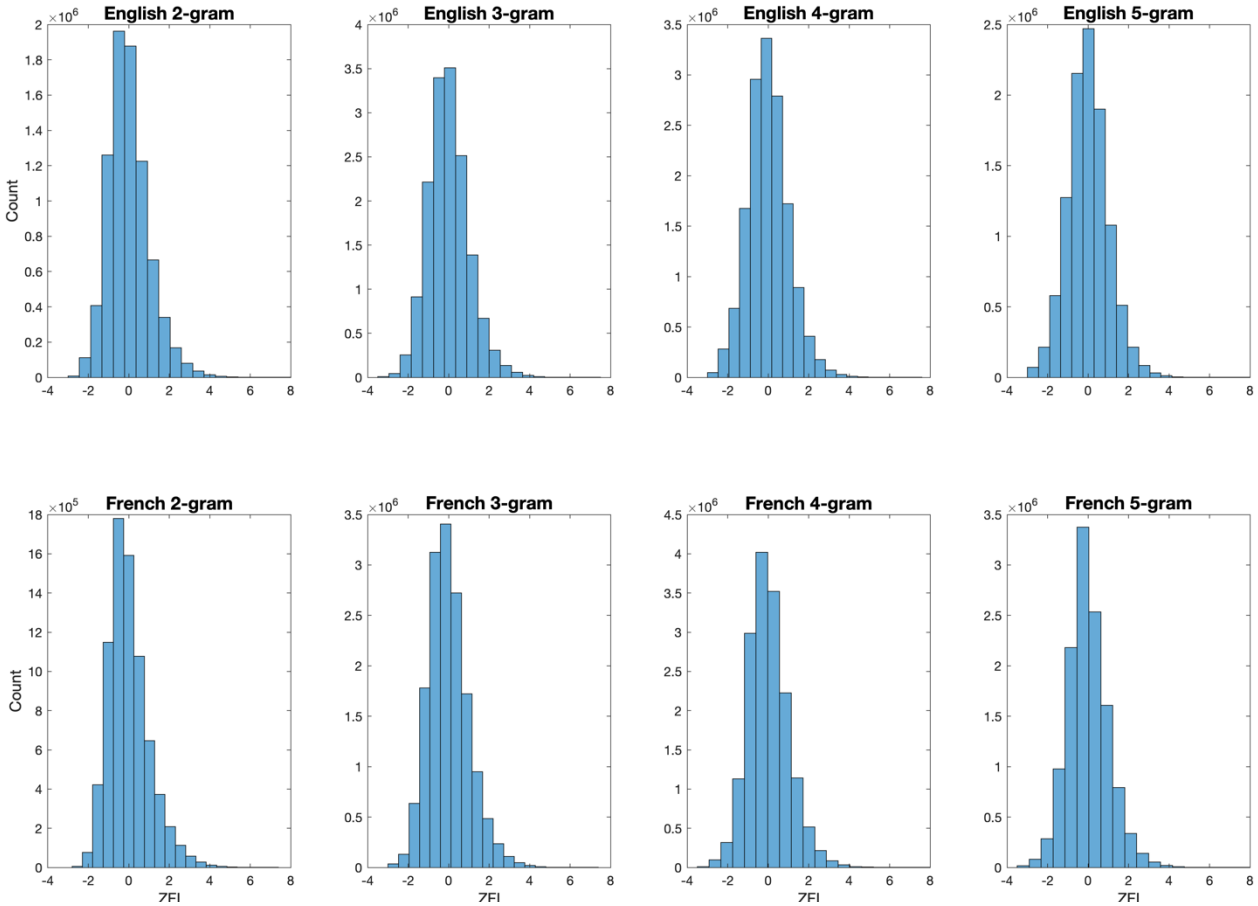

Figure 2. Distribution of the z-scored frequency index (ZFI) for English and French 2- to 5-grams. ZFI corresponds to the standardized log10 of the individual n-gram occurrence per million n-grams.

Frequency of the Part-of-Speech sequences is easily computable from the general n-gram lexicon using some standard routines in R (e.g. "group_by(PoS)" with tidyverse) or Matlab ("tabulate(PoS)"). For the English 5-grams for example, there is 41,422 combinations of PoS, the most frequent being DET NOUN ADP DET NOUN as in "the turn of the century" (N= 311,066; 2.93% of all the 5-PoS sequences).

ASSESSMENT OF THE DATABASE IN AN ONLINE EXPERIMENT

We wanted to know whether the frequency of an n-gram played a role in its recognition, just as a lexical frequency modulates the time it takes to recognize a particular word. As exposed in a previous section, such effect has been shown between groups of n-grams, i.e. between low- and high-frequency 3- grams (Siyanova-Chanturia et al., 2011), or in other context such as auditory sentence recognition (Arnon and Cohen Priva, 2013) or language production (Janssen and Barber, 2012). It would be reassuring if an experiment using Multi-LEX were to show a similar effect. Indeed, finding an effect of sequence frequency would both show the utility of Multi-LEX, our n-gram frequency databases, and validate the frequency measures per se. Moreover, we designed the study to analyze a putative frequency effect at the item-level, meaning that we were interested in regressing the participants' performance (response times, in milliseconds) to the frequencies of the n-grams. To do so, we performed an online psychology study consisting of categorizing 200 French 4-word sequences and 200 shuffled word sequences as grammatical or not (grammatical decision task).

Participants. Announcements posted in various social media led to the participation of 183 persons of whom 123 completed the full experiment. A further selection based on participant's general accuracy resulted in a dataset of 119 unique responders. A questionnaire at the beginning of the test asking for age, gender, mother tongue and handiness was proposed. Self-report for age gave a median value of 28 years (range [18; 69]). Eighty-eight participants were female and 119 had French as their mother tongue (1 Portuguese, 1 Russian, 1 Spanish, 1 Turkish). One hundred participants reported being right-handed. Additionally, participants were informed that their browser language was monitored once at the start of the experiment. All participants had their browser language set to French.

Stimuli selection. Stimuli consisted of 200 French 4-word sequences each forming a grammatically coherent structure such as "debout dans le wagon" ("standing in the wagon"; grammatical sequences compared to ungrammatical sequences). The sequences were taken from the Google 4-gram French database (2019 edition; Michel 2011). From this list, we conducted a double selection to make sure that sequences were homogenous and that no word within the sequences differs in frequency of use. First, concerning the adjective-, noun- and verb-tagged grams for part-of-speech (PoS), (i) words were 3- to 6-letter long, (ii) the log10 word frequency fell between -1 and 1 standard deviation of the whole word population, as well as (iii) the log10 of the orthographic distance. Second, other PoS-tagged grams were selected on their number of letters (between 2 and 6 that included determiner "le" or "la" for example, "the" in English) irrespective of any other criterion. Word sequences were first chosen randomly then hand-picked for sentence likeliness, thus selecting "the sky is clear" for example and not the part of sentence-like "sky is clear and". We made sure that the final 200 sentences followed the two conditions: (i) the mean word's lemma log10 frequency fell within the [-1; 1] standard deviation interval; and (ii) the 4-gram log10 frequency fell within the [-2; 2] standard deviation interval. No selection criterion was applied to the 2-gram or 3-gram frequencies. Ungrammatical sequences were built by shuffling the 4-grams of each Grammatical sequence. We made sure that the result could not form a part of a sentence. Following the grammatical sequence construction, no criterion was applied to the 2-gram or 3-gram frequencies. Stimuli from both Grammaticality conditions were divided in 2 different lists. Grammatical sequences were randomly assigned to one of two lists (list A and list B), and their associated Ungrammatical sequences were assigned to the other list (list B and List A respectively). This led to each list having 100 Grammatical and 100 unrelated Ungrammatical sequences. Each participant was randomly assigned to one of the two experimental lists. For the analysis, the log10 of the 4-gram frequency was used as the main factor of analysis (Frequency), as well as the number of letters of each sequence (NbLetter).

Experimental procedure. Prior to the experiment, visual instructions were provided about the grammatical decision task. Participants were asked to categorize the stimuli presented (Grammatical or Ungrammatical word sequences) as rapidly and accurately as possible. Following instructions, five practice trials were presented. The main experiment consisted of 200 trials. A trial was defined in a simple form: a fixation cross presented in the center of the screen for 500 milliseconds (ms) followed by a stimulus that was printed on screen until a response was given. The participant could either answer on Correct and Incorrect screen buttons displayed below the stimuli (mobile devices) or on the S (Correct) and L (Incorrect) keys of a computer keyboard (desktops and laptops). Stimuli were displayed in black Courier New letters on a light grey background and disappeared after a choice was made. Inter-trial interval was set at 1500 ms following answers. Trials were grouped in 4 blocks allowing three self-paced pauses during the test. On average, the test lasted 15 minutes. All data was recorded anonymously, complying to the General Data Protection Regulation section of the European Research Council research program POP-R (grant ERC742141). This study was ethically approved by the French IRB "Comité de Protection des Personnes SUD-EST IV" (No. 17/051). All participants gave their informed consent before the experiment started. The task was programmed in javascript / PHP and hosted on a standard Apache web server (https://ilcb-online-test.net).

Data processing and statistical analysis. Data from 123 participants who completed the full study were considered. First, trials with response times (RT) below 300 ms and above 6000 ms were excluded from further processing as well as items with mean accuracy over participants below 75% (2 items). Second, participants having their mean accuracy above the participant-based general mean accuracy minus 2.5 standard deviations were retained (119 participants). Third, trials having response times being above or below the overall participant-based mean response time +/- 2.5 standard deviation were discarded. Overall, 8.32% of the whole data set was not included for the statistical analysis.

Raw response times were further log10 transformed as in a standard lexical decision analysis.

Accuracies and response times were respectively analyzed using logistic and linear mixed-effects regression modeling (Baayen et al., 2008; Jaeger, 2008). In such analyses, participants and items were considered as crossed random factors. Following Baayen et al. (2008), |t|- and |z|-values larger or equal to 1.96 were deemed significant.

Results. Frequency and NbLetter influenced the grammatical decision response times (b= -0.03, SD=0.004, t=-7.5, and b= -0.007, SD=0.002, t=-3.7 resp.), but not accuracy (|t| and |z| inferior to 1.96).

Response times were therefore negatively influenced by Frequency (the more frequent a sequence is, the less time it takes to categorize it as being grammatically correct) and positively influenced by the number of letters (it takes more time to read a long sequence than a short one). Figure 3 shows the consolidated RTs for each sequence across the participants. We can clearly see the influence of Frequency on RTs, a linear regression on these sets of points giving an explain variance of 28.1%. Details of the analysis are given as supplementary material.

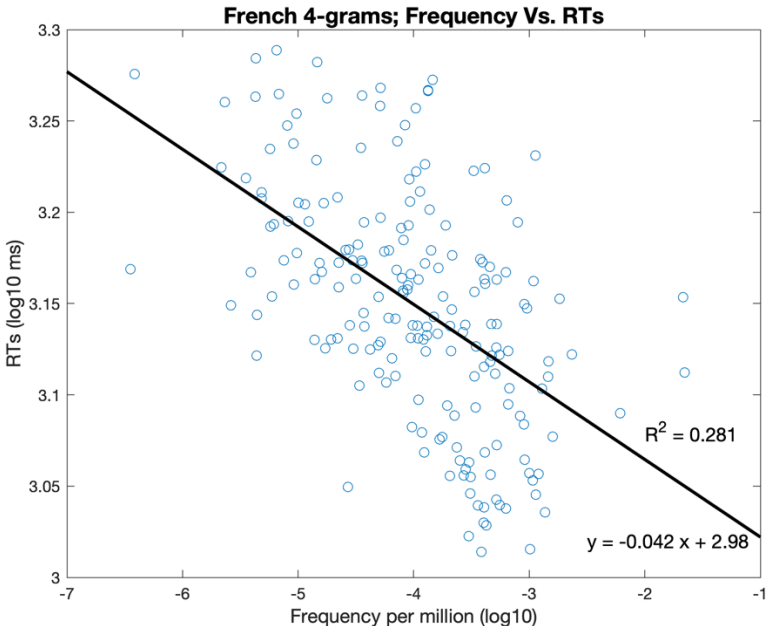

Figure 3. Grammatical decision RTs as a function of sequence frequency. The black line is the regression line between these two variables.

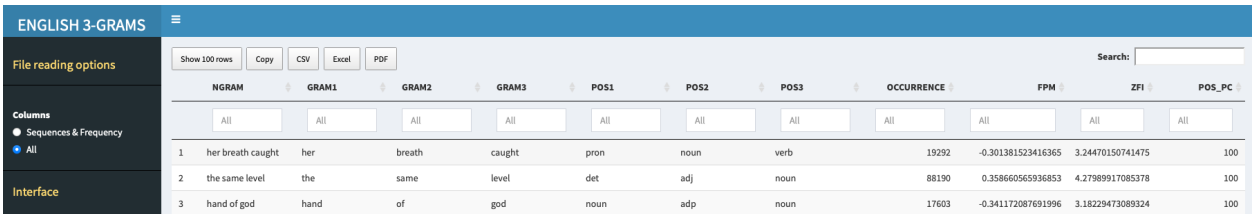

Figure 4. Screenshot of the web interface for the English 3-gram lexicon. The database can be searched for a unique word, a combination of words, a unique PoS or a combination of PoS. For example, a search for "PRON VERB ADJ" in POS1, POS2 and POS3 search fields led to the selection of 304 entries within the 50k lexicon (e.g., "it is important", "it felt good", "it becomes possible" and so on).

Discussion. The result of the online experiment (mixed model on RTs) clearly shows the expected effect of the n-gram frequency. Although small, the effect of frequency is sufficiently consistent across items to generate a strong t-value. The size of the effect in terms of R^2 in the regression analysis is somehow less than the one typically found in the lexical decision literature (frequency explains from 30 to 40 percent of the variance, depending on the RT dataset and the lexical frequency database). What is more surprising is the shape of the relation between Frequency and RTs, which was found to be linear. Actually, the single word literature always gives a relation in a form of a banana-shape, a negative relationship similar to that shown in Figure 2 but with a floor effect that mainly affects the more frequent words: frequent words all have the same RT, breaking the negative relationship with Frequency (i.e. the frequency effect lies in the lower frequency bands). Even though the primary goal of the experiment is met in demonstrating the usefulness of Multi-LEX, such n-gram frequency effects should be investigated further to either confirm or infirm the results exposed here.

ASSOCIATED VARIABLES AND COMPUTER PROGRAMS

A set of computer programs is provided, both to compute tailored lists of sentences and to manipulate the selections in the form of a web application. The programs for lists mainly read the Google Ngram files in a parallel fashion, decompress them in real time and generate lists of N-word sequences according to particular criteria such as excluding the Part-Of-Speech tags. The programs perform additional work on frequency computation and can further select sequences within the generated lists. Such programs are solely dedicated to reading the format employed by Google in generating the Ngram database and cannot be used to read other kinds of corpora without modifications. The second program, the web application, lets the user read and display in a table either part of or whole lists of word sequences. Each sequence record is associated with the individual words composing the sequence, as well as the individual word PoS tags, number of occurrences and the standardized frequency index. One can search for a particular PoS in a particular position or search partial matches of individual words (see Figure 4). We also provide within-list selection tools based on individual words, PoS tags and frequency values. Whole lists of word sequences in tabulated format are also directly downloadable from the web application. Those files have the following columns: the NGRAM (e.g. “one of the most important”), GRAM1 (“one”) to GRAM5 (“important”), POS1 (“num”) to POS5 (“adj”), the number of occurrences summed over 2012-2019, the Frequency Per Million, the standardized frequency index (ZFI), and finally the dominant PoS sequence associated to the NGRAM, expressed in percentage (POS_PC; 100% means that the NGRAM was only found in one PoS combination). A lower percentage means that either several forms of PoS exist for a particular word sequence, or that the linguistic processing in recognizing PoS gave some inconsistent results.

DISCUSSION

From the Google Book Ngram corpus, we selected a set of Part-Of-Speech tagged 2- to 5-grams in English and French. These selections of a few million word sequences, presented as lists in compressed text files, are intended to help psycholinguists optimally choose their material for studying written word sequences or sentence processing. We also provide the source code that helped us generate such lists, allowing interested researchers to start to generate their own lists (noting that the selection process is quite time- and resource-consuming). Finally, we propose an online graphical application that makes the within-list material selection more convenient and user-friendly.

To assess the usefulness of the database, the 4-gram list in French was tested in an online experiment. The 200 word sequences selected from this n-gram lexicon were categorized more quickly when the sequence was more frequent.

Our proposed selection process and computation comes with several pitfalls. First, our selection process depends on the way in which Google has digitized and processed books. On the one hand, Pechenick, Danforth and Dodds (2015) identified several limitations concerning the frequency count, including a divergence between years for identical word sequences and a bias toward the inclusion of scientific literature. Indeed, Brysbaert, Keuleers and New (2011) further noted that, for unigrams, the Ngram database is poor in correlating human performance to frequency estimates compared to standard lexical databases such as the SUBTLEX family (11% drop in the explained variance). On the other hand, the Google Ngram entries are automatically tagged with Part-Of-Speech and such a process is necessarily error-prone (Lin et al., 2012). Mis-tagged sequences might be rejected by the selection process on the basis of a false PoS assignment.

Second, and finally, we present a frequency measure summed over the years 2012-2019. This single information per n-gram is to be taken at face value while entries of historical corpora such as Google Ngrams are meant to be analyzed in terms of trends. The coherence of frequencies along several periods gives an indication of a certain reliability that a single point does not offer.

To overcome some of the Google Ngram limitations, Younes and Reips (2019) proposed a set of strategies to be used in researching n-grams. Amongst the different proposals, two of them are relevant to our lists: investigating several corpora of different languages and cross-checking different corpora from the same language. Applied to psycholinguistics, readers working in English can refer to other tools including the British National Corpus (Leech & Rayson, 2014), the Corpus of Contemporary American English (Davies, 2008), the Corpus of Historical American English (Davies, 2012), or the Global Web-based English Corpus (Davies & Fuchs, 2015).

Readers interested in conducting research in multi-word processing should be aware of an extra potential pitfall intrinsic to n-grams. As we saw in the Selection Results section, some n-grams are not self-contained phrases. Selecting such stimuli could be problematic in a grammatical decision task, for example, “of the process of” could generate more processing time to be classified compared to “at the same time”. “Of the process of” could even be misclassified as an ungrammatical sequence. In such cases, a preliminary rating of the material with an independent group of participants could be of use. On a side note, some authors reported the classification between grammatical and ungrammatical word sequences as “phrasal decision”. Even though correct when idiomatic expressions or figures of speech are used as stimuli, phrasal decision refers to an overly precise concept in which a sequence of words must be, by definition, a phrase or a sentence. In a grammatical decision task, instructions are given to the participant to classify word sequences as being correct or not, in the sense that a group of words that is syntactically correct should be classified as grammatical. The “of the process of” example shows why the grammatical decision task is more general, in that the sequence just has to be syntactically correct and not necessarily a phrase or sentence.

CONCLUSION

There is little material currently available for psycholinguists interested in multi-word sequences and syntax and this is particularly true for languages other than English. To provide the community with relevant material, we took advantage of the Google Ngram database to produce lists of n-grams taken from millions of books. As a first initiative, we proposed a selection of n-grams (2-word to 5-word sequences) in French and English along with scripts to compute

additional or custom-made n-gram selections. Each of the lists' entries are associated with Part-Of-Speech tags for individual words, counts for occurrences, as well as a standardized frequency estimate.

FOOTNOTES

1. Available at https://www.english-corpora.org
2. Available at http://www.lexique.org/
3. books.google.com
4. https://books.google.com/n-grams

OPEN PRACTICES STATEMENT

The files containing the English sequences are available at https://zenodo.org/record/7214223 (DOI: 10.5281/zenodo.7214223) and the files containing the French sequences are available at https://zenodo.org/record/7214248 (DOI: 10.5281/zenodo.7214248). For each language and each number of words, three types of files are downloadable: a set of the 50,000 most frequent sequences, a set of a million sequences including the former set, and finally the whole sequence set. Such sets are compressed (standard gzip algorithm) and easily readable by standard software like R. The code necessary for running the n-gram computational selection processes is available at https://github.com/lpc-cnrs-amu/N-word-frequency, so that researchers can generate their own personal lists of items. Languages available in the Google Ngram database are English, Spanish, French, German, Russian, Italian, Chinese and Hebrew. The code of the R Shiny application is available at https://github.com/lpc-cnrs-amu/Multi-LEX. This repository contains some additional code to load and search the dataset with R scripts. The R Shiny application is available https://analytics.huma-num.fr/popr-ngram/. Supplementary material for the grammatical decision experiment is available at https://osf.io/6wq8y/ (see https://osf.io/524sv for the statistical results in a web-like page mixing the R code and its result).

ACKNOWLEDGMENTS AND FUNDING INFORMATION

This study was supported by grants ANR-11-LABX-0036, ANR-15-CE33-0002-01 and ANR-16-CONV-0002 from the French Agence Nationale de la Recherche (ANR), and ERC advanced grant 742141. This work was carried out within the Pôle Pilote Ampiric, funded by the French State's Future Investment Program (PIA3/France Relance) as part of the "Territories of Educational Innovation" action. We thank A. McGonigal for proofreading this article.

REFERENCES

Arnon, I., & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. Language and speech, 56(3), 349-371.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of memory and language, 59(4), 390-412.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). The CELEX lexical database (cd-rom).

Boada, R., Guasch, M., Haro, J., Demestre, J., & Ferré, P. (2019). SUBTLEX-CAT: Subtitle word frequencies and contextual diversity for Catalan. Behavior Research Methods, 1-16.

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. Behavior Research Methods, Instruments, & Computers, 30, 272-277.

Brants, T. (2006). Web 1T 5-gram Version 1. http://www. ldc. upenn. edu/Catalog/CatalogEntry. jsp? catalogId= LDC2006T13.

Brants, T., & Franz, A. (2009). Web 1T 5-gram, 10 European languages version 1. Linguistic Data Consortium.

Brysbaert, M., Keuleers, E., & New, B. (2011). Assessing the usefulness of google books' word frequencies for psycholinguistic research on word processing. Frontiers in Psychology, 2, 27. https://doi.org/10.3389/fpsyg.2011.00027

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.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. Current Directions in Psychological Science, 27(1), 45-50.

Carroll, J. B. (1970). An alternative to Juilland's usage coefficient for lexical frequencies, and a proposal for a standard frequency index (SFI). Computer studies in the humanities and verbal behavior, 3(2), 61-65.

Carroll, J. B. (1971). Measurement properties of subjective magnitude estimates of word frequency. Journal of Verbal Learning and Verbal Behavior, 10(6), 722-729.

Davies, M. (2008). The corpus of contemporary American English (COCA): 560 million words, 1990-present. https://www.english-corpora.org/coca/ (retrieved from 2020).

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. Corpora, 7(2), 121-157.

Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). English World-Wide, 36(1), 1-28.

Davis, C. J. (2010). The spatial coding model of visual word identification. Psychological Review, 117(3), 713–758.

Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., Dufau, S., Mathôt, S. & Grainger, J. (2018). MEGALEX: A megastudy of visual and auditory word recognition. Behavior Research Methods, 50(3), 1285-1307.

Grainger, J., Dufau, S., & Ziegler, J. C. (2016). A vision of reading. Trends in Cognitive Sciences, 20(3), 171-179.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. Psychological Review, 103, 518–565.

Grossmann, I., Eibach, R. P., Koyama, J., & Sahi, Q. B. (2020). Folk standards of sound judgment: Rationality Versus Reasonableness. Science Advances, 6(2), eaaz0289.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of memory and language, 59(4), 434-446.

Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production. PloS one, 7(3), e33202.

Kucera, H. & Francis, W.N. (1967). Computational analysis of present-day English. Providence, RI: Brown University.

Kuperman, V. (2015). Virtual experiments in megastudies: A case study of language and emotion. The Quarterly Journal of Experimental Psychology, 68(8), 1693-1710.

Leech, G., & Rayson, P. (2014). Word frequencies in written and spoken English: Based on the British National Corpus. Routledge.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. Behavior Research Methods, Instruments, & Computers, 36(1), 156-166.

Lin, Y., Michel, J. B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the google books n-gram corpus. In Proceedings of the ACL 2012 system demonstrations (pp. 169-174). Association for Computational Linguistics.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: 1. An account of basic findings. Psychological Review, 88, 375–407.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. Science, 331(6014), 176-182.

Mirault, J., Snell, J., & Grainger, J. (2018). You that read wrong again! A transposed-word effect in grammaticality judgments. Psychological Science, 29(12), 1922-1929.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. Applied Psycholinguistics, 28(4), 661-677.

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.

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. PloS One, 10(10), e0137041.

Pegado, F., & Grainger, J. (2020). A transposed-word effect in same-different judgments to sequences of words. Journal of Experimental Psychology: Learning, Memory, and Cognition.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. Psychological Review, 114(2), 273.

Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. arXiv preprint arXiv:1104.2086.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(3), 776.

Snell, J., & Grainger, J. (2017). The sentence superiority effect revisited. Cognition, 168, 217-221.

Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. Psychological Review, 125(6), 969.

Terzopoulos, A. R., Duncan, L. G., Wilson, M. A., Niolaki, G. Z., & Masterson, J. (2017). HelexKids: A word frequency database for Greek and Cypriot primary school children. Behavior Research Methods, 49(1), 83-96.

van Heuven, W. J., 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.

Wu, M. S., Li, B., Zhu, L., & Zhou, C. (2019). Culture Change and Affectionate Communication in China and the United States: Evidence From Google Digitized Books 1960–2008. Frontiers in Psychology, 10, 1110.

Younes, N., & Reips, U. D. (2019). Guideline for improving the reliability of Google Ngram studies: Evidence from religious terms. PloS One, 14(3), e0213554.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duwuri, R. (1995). The educator's word frequency guide. Brewster, NY: Touchstone Applied Sciences.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. Journal of Memory and Language, 47(1), 1-29.