Lexique 4: a major upgrade of the « Lexique » French Lexical Database

Boris New[1], Christophe Pallier[2], Gauvain Schalchli[3], Jessica Bourgin[1] & Manuel Gimenes[4]

[1] Univ. Savoie Mont Blanc, Univ. Grenoble Alpes, CNRS, LPNC, 38000 Grenoble, France

[2] Cognitive Neuroimaging Unit, CNRS, INSERM, CEA, Neurospin Center, 91191 Gif-sur-Yvette, France

[3] Université Bordeaux Montaigne, CLLE, UMR CNRS 5263

[4]Centre de Recherches sur la Cognition et l'Apprentissage, Université de Poitiers, Université de Tours, CNRS, Poitiers, France

Corresponding author

Boris New, boris.new@univ-smb.fr

Abstract

Lexique 4, an updated French lexical database, expands upon its predecessor, Lexique 3, by incorporating several significant improvements to enhance its utility in psycholinguistics, computational linguistics, and education. The new version is based on a larger corpus of 316 million words derived from 65,317 documents, including movies, TV shows, and documentaries' subtitles, which offers more accurate frequency estimates and includes contemporary neologisms. Lexique 4 introduces new variables, such as orthographic surface frequency, contextual diversity (CD), and detailed morphological structure, which provide a more comprehensive view of lexical properties. We find that contextual diversity is a slightly better predictor than word frequency, in line with previous work. Moreover, the integration of lexical decision times from the French Lexicon Project into Lexique 4 facilitates more in-depth linguistic research. Enhancements to the user interface, including a redesigned web platform, enable dynamic searches and sorting capabilities, increasing accessibility and usability for researchers. Statistical analyses indicate that the updated frequency measures in Lexique 4 are better predictors of lexical decision times compared to Lexique 3, supporting the value of these enhancements. Overall, Lexique 4 represents a comprehensive and flexible tool for analyzing French lexical properties, making it an essential asset for a broad range of users.

Keywords: French Lexical database, Word frequency, Contextual diversity, Orthographic surface frequency, Subtitles Frequency

Lexique 4: Enhancing French Lexical Databases

Lexical databases consist of lists of words for which different properties are provided (for example: frequency, orthography, phonology, morphology, number of letters). They are particularly useful in fields such as psycholinguistics and linguistics. By providing numerous characteristics for each word, they make it possible to address many research questions. The precision and quality of the word characteristics provided are important to run valid experiments. Lexical databases are also important tools in education and for speech and language therapists.

Lexical databases history

Existing databases vary in the number of different lexical variables they provide. Sometimes, the database is focused on one main feature such as the lexical frequency. This type of database is generally called a "frequency norm". For a long time, frequency norms were computed from book-based corpora. In English,

one of the oldest frequency norms was compiled by Ayres (1915) from a corpus of 368 000 words from private and commercial letters. The goal was educational, namely the teaching of spelling. Thorndike and Lorge (1944) published frequency norms for 30 000 words tabulated from a corpus of 18 millions of words. The goal was to help educators to teach reading. One of the most widely used frequency norms in psycholinguistics was designed by Kucera and Francis (1967) from a sample of about one million words from the Brown Corpus. More recently, frequency norms from other types of corpora were proposed. For example, New et al. (2007) showed that frequency norms based on TV subtitles corpora correlated better with human lexical decision times than book-based frequencies. Building on this finding, Brysbaert and New (2009) developed subtitle-based frequency norms for English. These subtitle-frequency norms were then computed in a lot of different languages, for example in Dutch (Keuleers et al., 2010), in Chinese (Cai & Brysbaert, 2010), in Spanish (Cuetos et al., 2012), in Polish (Mandera et al., 2015), in British (Van Heuven et al., 2014). Gimenes and New (2016) proposed new frequency norms based on Twitter, blog posts and newspapers in 66 languages. The results showed that these new frequencies predicted lexical decision reaction times similarly to the already existing frequencies (based on books or subtitles corpora), or even better than them.
Other lexical databases aim to provide more lexical characteristics than just frequency. In English, a seminal general lexical database is the MRC database (Coltheart, 1981). It contains 150 837 English words and, for each of them, 26 linguistic variables. These include semantic features (such as concreteness and imagery), developmental features (such as age of acquisition) and distributional features (such as word frequency). Another important lexical database is Celex (Baayen et al., 1995) which contains 160 594 words and includes lexical data for Dutch, English and German. It provides information on various lexical aspects, including orthography, phonology, morphology, syntax and word frequency. Since then, many lexical databases have been integrated into large-scale megastudies used in experimental psycholinguistics. Notable examples include the English Lexicon Project (Balota et al., 2007), the British Lexicon Project (Keuleers et al., 2012), and the Dutch Lexicon Project 1 (Keuleers et al., 2010) and 2 (Brysbaert et al., 2016).

French lexical databases

In French, several lexical databases exist, designed specifically for psycholinguistic research. The first widely used psycholinguistic database was Brulex (Content et al., 1990) providing information such as orthography, phonology, gender, number and word frequency count for 35 746 words. The Manulex database (Lété et al., 2004), of interest for people interested in reading acquisition, provides word frequency counts from a corpus of French elementary school readers, number of letters and syntactic category information for 48 886 words. In addition, the PsychoGLÀFF database (Calderone et al., 2014), which is derived from GLÀFF (Hathout et al., 2014), itself built from Wiktionnaire[1], the French edition of Wiktionary, extends the lexical information available in GLÀFF by including psycholinguistic variables such as word frequency, phonological transcriptions, and syllabic structure.
Finally, under various incarnations, the Lexique database (New et al., 2001, 2004) has provided a large number of lexical characteristics for about 130k French words. One of the features of Lexique is that it has always been both freely downloadable and searchableonline. The early versions contained lexical characteristics extracted from a selection of texts from the Frantext corpus (Bernard et al., 2002). Lexique 2 provided the frequency for 130 000 words, including inflected forms. Two types of lexical frequencies were available in Lexique 3: frequencies based on a corpus of 218 novels (published between 1950 and 2000) from Frantext (14.7 million words) and frequencies based on a corpus of 9 474 film and series subtitles (50 million words). Other novelties of Lexique 3 were the presence of recent and popular words at that time (i.e., internet, download), distinct frequencies for homonyms and homographs (i.e.," danse" as noun vs "danse" as a verb) and the presence of compound words.
For a long time, lexical databases included only word frequency and lexical characteristics.

[1] https://fr.wiktionary.org/

Psycholinguistic megastudies

Recently, however, researchers started to conduct extensive projects to get behavioral and chronometric measures for thousands of words. These large databases are named “megastudies” and are very useful for psycholinguists. For example, they can be used to test hypotheses on very large quantities of data using virtual experiments (see for instance: New et al., 2006; Gimenes & New, 2016) or to replicate old studies to test their robustness (Ferrand et al., 2018). The English Lexicon Project (Balota et al., 2007) provided lexical decision times (from 816 participants) and naming latencies (from 444 participants) for 40 481 words and 40 481 pseudowords. Other English megastudies were created since then, with behavioral data from different tasks: for example memory recognition (Cortese et al., 2010), eye movements in text reading (Cop et al., 2017), semantic decision (Pexman et al., 2017), and auditory lexical decision (Tucker et al., 2019). Megastudies were developed in a lot of different languages as, for example, in Chinese (Tse et al., 2017, 2023), Dutch (Brysbaert et al., 2016; Keuleers et al., 2010), Korean (Siew et al., 2021), Malay (Yap et al., 2010), Spanish (Aguasvivas & Sainz, 2018) and British (Keuleers et al., 2012). In French, the first megastudy (the French Lexical Project) was performed by Ferrand et al. (2010) who collected lexical decision times for 38,840 words. A more recent French megastudy, called Megalex, was designed by Ferrand et al. (2018), which provides visual lexical decision times and accuracy rates for 28 466 words and auditory lexical decision data for 17,876 words.

In summary, general lexical databases offer a broad range of lexical information, typically including frequency measures, while recent megastudies provide extensive behavioral data. In this paper, we present the latest version of Lexique, Lexique 4.

Lexique 4: main characteristics

Lexique 4 introduces significant innovations compared to Lexique 3. First, the corpus is much larger and relies on more recent documents (65,317 movies, series, and documentaries’ subtitles, comprising a total of 316 million words), whereas Lexique 3 was based on a corpus of books totaling 14.8 million words and a corpus of subtitles totaling 50 million words. Therefore, the frequency estimates derived from the Lexique 4 corpus are potentially more accurate than those from Lexique 3, even though the Lexique 4 corpus is largely a superset of the Lexique 3 subtitle corpus, as many movies and TV shows from the subtitle corpus in Lexique 3 are also included in Lexique 4. Some recent neologisms are included in Lexique 4 such as “covoiturer” (carshare), “selfie” (selfie), “vegan” (vegan), “chouille” (party), “youtubeur” (youtuber). The number of word types (168 536) is much greater than in Lexique 3 (125 654) which represents an increase of 34%.

Lexique 4 now lists orthographic surface frequency, which is the sum of the frequencies of the different parts of speech corresponding to the same orthographic form.

Contextual diversity is included for each word type. This variable represents the number of distinct documents in which a specific word appears within the corpus. It has been shown to be a better predictor of lexical decision times than lexical frequency (Adelman et al., 2006).

The morphological structure for all the lemmas available in Lexique 3 was also added. We used the Dictionnaire étymologique de la langue française (Bloch et al., 1964) as a reference source. The process involved: 1) Identifying the morphological base (i.e. *composer* for *décomposable*) of each lemma 2) Segmenting the word into its morphological structure in terms of affixation and composition (i.e. *1-1-1*, meaning it involves a prefix, base, and suffix) 3) Providing a detailed decomposition indicating the boundaries of affixes and bases. (i.e. *_dé/compos(er).able*).

As breaking down the morphological structure of all French words is a complex task, we are also making available the original file made by the linguist, which contains much more morphological information than that

included in the main table of Lexique 4 (type of affixes and suffixes, etc.) and could be useful to psycholinguists working on morphology.
The lexical decision times provided from the French Lexicon Project (Ferrand et al., 2010) have been directly integrated into Lexique 4 to facilitate the work of the researchers. We chose to use reaction times from the French Lexicon Project (FLP) rather than those from Megalex because FLP provides a much larger corpus of approximately 38,336 words, compared to around 25,777 words in Megalex. Additionally, word length in the FLP is not limited to 9 letters or fewer, but extends up to 19 letters. This broader range of word lengths is particularly valuable for researchers focusing on morphology (morphologically complex words generally being longer), orthographic neighborhood effects, syllabic units, or word-length effects.
Finally, the www.lexique.org website was redesigned from scratch and improved to facilitate online searches. For example, search results are now presented in a dynamic table, making it possible, for example, to dynamically sort rows according to any field.

Method

To construct our corpus, we downloaded the subtitles of 65,317 documents, movies, TV shows, documentaries, or anime, representing a total of 316 million words. These documents were obtained from the OpenSubtitles sub-corpus (Lison & Tiedemann, 2016: this is a collection of translated movie subtitles from http://www.opensubtitles.org/) belonging to OPUS 2018[2] which is part of a growing collection of translated texts from the web. First we checked that there were no duplicate files. As many of the subtitles had been generated using automatic optical character recognition (OCR), we implemented a rule-based procedure to identify files containing OCR-related errors (e.g., the letter i frequently misrecognized as l). This process led us to exclude 1,357 files. Then, to tokenize and grammatically annotate our corpus, we relied on Cordial Analyseur 8.13. We evaluated several POS taggers, including TreeTagger (Schmid, 1999), the Stanford Tagger (Toutanova et al., 2003), and Cordial Analyseur itself, using a set of ambiguous sentences containing homograph words[3]. Our results showed that Cordial Analyseur performed better in correctly disambiguating POS tags compared to the other tools. We obtained a list of items, including compound words and their frequencies. These items included symbols (including punctuation), abbreviations, foreign words, and proper nouns. To "clean up" this list[4], we employed Lexique 3, Dicollecte 6.4.1 which is a dictionary for spell-checkers and Lefff 3.4 (Sagot, 2010) which is a lexicon of inflected forms of French. We provide a file on OSF containing all words along with their frequency and raw count data.
The problem of obtaining Lexique's phonological codes is a tricky one, since Lexique has always included a large number of words. In Lexique 2 and 3, these phonological forms have been progressively corrected thanks to extensive manual work by the Lexique user community, notably Ronald Peereman, Christian Lachaud and Jean-Philippe Goldman. Thus, for words that were already in Lexique 3.83, we have adopted the phonology of Lexique 3.83. For words in Lexique 4 that were not present in Lexique 3, we used Glaff 1.2.2 which is a lexicon based on the French Wiktionary. If Glaff's phonology was absent, the phonology was derived from a text-to-speech tool Multitel Elite 2.0.1 (Pagel et al., 1998).

Description of Lexique 4 variables

Mot: the orthographic form of the word. All words in this column appear at least once in the corpus. As explained above, two identical orthographic forms can appear on different rows if they exist as different parts of speech. For instance, “lit” (bed) as a noun and “lit” (reads) as a verb are two different records.
Phono: the phonological form of the word, encoded with the same phonological codes as in the previous versions of Lexique (see Table A1 in Appendix). We kept this system both for reasons of backward

[2] https://opus.nlpl.eu/
[3] https://fr.wikipedia.org/wiki/Homographe
[4] This list is available in the Lexique 4 OSF project

compatibility and because these codes are easy to type directly from any standard keyboard, including non-French ones.

Phono_IPA : the phonological form of the word encoded with the International Phonetic Alphabet (IPA), which provides a standardized and widely recognized representation of phonemes.

Lemme: the lemma of the word. A lemma is an entry in a dictionary that is used to represent all the other possible inflected forms. For example, the word “cats” has for lemma “cat”.

CGram: the grammatical category of the word. The different codes used are presented in the appendix (Table A2). Some words belong to multiple grammatical categories (e.g., the word danse can function as either a noun or a verb). Such words are therefore listed as separate entries.

CGramOrtho: the grammatical categories of the orthographic form of the word. For instance the orthographic form “danse” in French can be both a noun and a verb so this variable will indicate: NOM, VER. This variable can be useful if a user wants to select words that can only be nouns for example.

Genre: the grammatical gender of a word indicates whether it is feminine (f), masculine (m) or epicene (e, a noun that can be feminine and masculine).

Nombre: the grammatical number of a word indicates whether it is singular (s) or plural (p) or invariant (i).

InfoVER: verb properties (mood, tense and person). Table A3 in the appendix represents the different codes used in this column.

FreqOrtho: when a word belongs to multiple grammatical categories (and is on multiple lines with different POS [Part Of Speech]), FreqOrtho is the sum of the different word frequencies. It is expressed as the number of occurrences per million. Please note that the absence of a form from Lexique 4 should not be interpreted as evidence that its frequency is zero; it only indicates that the form does not appear in our 316-million-word subtitle corpus.

FreqLemme: the sum of the frequencies of all the words having the same lemma. For instance, FreqLemme of “cat” will be the sum of the frequencies of “cat” and “cats”. It is expressed as the number of occurrences per million.

FreqMot: number of occurrences of a word per million in our 316 million words subtitle corpus.

CDOrtho: contextual diversity of an orthographic form. It is the proportion of documents in the corpus where an orthographic form occurs. For example, CDOrtho of “danse” is the proportion of documents where “danse” (both as a noun and as a verb) occurs.

IsLem: indicates whether a word is a lemma (coded 1) or not (coded 0).

NbLettres: number of letters in a word.

NbPhons: number of phonemes in a word.

OLD20: Orthographical Levenstein Distance 20 as proposed by Yarkoni et al. (2008). It is a measure of lexical similarity. It calculates the average number of single-character edits (insertions, deletions, or substitutions) needed to change a word into its 20 nearest neighbors in a given lexicon. A lower OLD20 score indicates greater similarity to other words, suggesting a higher orthographic neighborhood density.

PLD20: Phonological Levenstein Distance 20. It is the same measure as OLD20 but it is based on the phonological form.

CVOrtho: Orthographical Structure of a word. Consonants are coded “C” and vowels “V”. For example, the word “chien” is noted “CCVVC”.

CVPhono: phonological structure of a word, with the same coding of consonants and vowels as in OrthoCV and semi-vowels that are coded “Y”. For example, the word “chien” (/ʃjɛ̃/) is noted “CYV”.

VoisOrtho: number of orthographical neighbours of a word. Orthographical neighbours (it corresponds to the N from Coltheart et al., 1977) are all the words with the same letters at the same position as the target word, but with only one different letter. For example, “char” and “chut” are orthographical neighbours of the word “chat”.

VoisPhono: number of phonological neighbours. Phonological neighbours are computed on the phonological form.

Lexique 4

NbHomog: number of homographs. Number of words having the same spelling in the database.
NbHomoph: number of homophones. Here, homophones are words with the same pronunciation but with different spellings or different part of speech.
SyllPhono: The phonological forms were syllabified according to a syllabification algorithm described in Dufour et al. (2002). The syllabification is calculated on the phonological representation present in Lexique, with final schwas removed. This syllabification is based on the general principle of syllabic segmentation between two consonants except in the cases of stops + liquids or a a labiodental fricative followed by a liquid consonant.
SyllNb: number of syllables.
SyllCV: Syllabic phonological structure. Consonants are coded as “C”, vowels as “V” and semivowels as “Y”.
PUOrtho: orthographical uniqueness point. It is the first letter (from left to right) where the word can be identified without ambiguity. We have calculated the uniqueness points on the basis of the lemmas so that plural forms do not interfere with the calculations (otherwise all forms with a plural have a uniqueness point equal to their length). For orthographic forms that are not lemmas, the orthographic uniqueness point is 0.
PUPhon: phonological uniqueness point. It is the first phoneme (from left to right) where the word can be identified without ambiguity
The next 3 variables concern morphology. As these fields were generated manually, inflected forms were not taken into account and only lemmas that were not too rare and were not foreign borrowings were taken into account (46,305 words were taken into account). These columns will be improved and added to by the community (as it has been the case, for example for the phonology in previous versions of Lexique).
MorphoBase: the base morpheme of the word.
MorphoStruct: number of affixes in a word. It is composed of three numbers separated by a hyphen. The first number indicates the number of prefixes, the second number indicates the number of roots, and the third number indicates the number of suffixes.
MorphoDecomp: it shows the different morphemes in the word. Three symbols are placed in front of the different types of morphemes. Prefixes are preceded with an underscore, roots are preceded with a slash and suffixes are preceded with a dot (i.e. décomposer -> _dé/compos(er).er). Other symbols are used:

- [] which is added to the base: baisse -> ba[i]sse -> basse; actionner -> action[n]-er -> action
- () what is removed from the base: abasourdi -> a-basourdi(r) ->abasourdir

Preval: Percentage of participants who know the lemma of the word. The data were collected from the website “Combien de mots connaissez-vous ?” active from 2010 to 2014 where everyone could indicate whether they know or not some words mixed with pseudowords. The website randomly selected 100 stimuli, 67 real French words taken from the Lexique database and 33 pseudowords, which were initially displayed in red. Users were instructed to click on the stimuli they believed were real words. Once clicked, the selected stimuli turned green, indicating that the user endorsed them as known words. After submitting the page, the website estimated the user’s vocabulary size by inferring it from the proportion of real words endorsed among the 100 stimuli. An archive of this page can still be accessed thanks to the “Wayback Machine”[5].
PrevalNb: Number of participants from which the previous variable (Preval) was generated.
RT_FLP: Visual lexical decision times based on the French megastudy FLP (Ferrand et al., 2018).
zRT_FLP: Standardized reaction times from FLP.
Err_FLP: Error rates from FLP.

Results

Univariate analyses

[5] https://web.archive.org/web/20140217163018/https://www.abyssum.com/Mesmots/

Lexique 4

Some univariate statistical analyzes were first carried out. In Lexique 4, there are 170 782 word types, and 36 179 hapax (words that occur only once in the corpus). Table 1 presents the number of words having n different parts of speech (n is a number between 1 and 5).

Table 1 *Number of words having 1 to 5 parts of speech*

1	2	3	4	5
153 652	15 222	1 871	31	6

Concerning gender, there are 23 401 feminine nouns, 33 851 masculine words and 2 820 epicenes. Concerning number, there are 36 347 singular nouns, 22 615 plural nouns and 989 invariant words. Table 2 presents the number of words for different frequency and CDOrtho intervals.

Table 2 *Number of words for different frequency and CDOrtho intervals*

	Range	Number of words
FreqOrtho	Freq = 0	0
	Freq > 0 and <= 0.1	126 917
	Freq > 0.1 and <= 1	41 285
	Freq > 1 and <= 10	16 679
	Freq > 10 and <= 100	4 202
	Freq > 100 and <= 1000	656
	Freq > 1000 and above	124
CDOrtho	CD >= 0 and <= 0.5	165 406
	CD > 0.5 and <= 1	7 870
	CD > 1 and <= 2	5 762
	CD > 2 and <= 4	4 165
	CD > 4 and <= 8	2 762
	CD > 8 and <= 16	1 693
	CD > 16 and <= 32	1030
	CD > 32 and <= 64	638
	CD > 64 and <=100	537

As has already been observed in the literature (Brysbaert & New, 2009), we can see that rare words (both in terms of frequency and CD) make up the vast majority of the lexicon. This observed increase in the number of words as frequency or CDOrtho decreases is consistent with Zipf's law (Zipf, 1935).
Table 3 shows statistical indices to describe some variables in Lexique 4.

Table 3 *Descriptive statistics for different variables from Lexique 4*

Variable	min	max	mean	median	sd
N Letters	1	25	9.17	9	2.63
N Phonemes	1	31	6.86	7	2.23
OLD20	1	14.6	2.63	2.45	0.988
PLD20	1	21.6	2.29	2	0.943
N Ortho Neighbors'	0	33	1.37	1	2.49
N Phono Neighbours	0	377	5.34	2	14.1
N Homophons	0	23	1.94	1	2.54
N Syllables	1	18	2.99	3	1.06

Lexique 4

Ortho UP	0	25	2.65	0	3.85
Phono UP	0	19	3.44	4	3.20
Prevalence	0	100	89.2	95	18.0

Correlation analyses

A correlation analysis was conducted to compare frequencies between Lexique 3 and Lexique 4. The correlation between word frequencies in Lexique 3 and Lexique 4 is exceptionally high at 0.96 (N = 124 391). This strong correlation is primarily driven by very high-frequency words. When considering only words with frequencies below 10, it decreases to 0.91 (N = 119 337). The correlations are therefore generally very strong between Lexique 3 and Lexique 4, but this correlation decreases for infrequent words, which indicates that these frequencies may differ significantly. This observation is a well-known characteristic of Zipfian distributions (Zipf, 1949): the reliability of frequency as an estimator depends on its absolute value, improving as the frequency increases.

We have also carried out global correlations on the variables in Lexique 4 that appear most relevant for the majority of psycholinguistic analyses. The correlations are presented in Table 4. For all analyses, logarithmic transformations were carried out for FreqOrtho, FreqLemme, CDOrtho and prevalence variables.

Table 4 *Correlation matrix of lexical and orthographic variables associated with word processing.*

	FLP zRT	*FreqOrtho*	*FreqLemmes*	*CDOrtho*	*Preval*	*N Letters*	*OLD20*
FLP zRT	*1*	*-0.59*	*-0.43*	*-0.6*	*-0.37*	*0.37*	*0.4*
FreqOrtho		*1*	*0.51*	*0.99*	*0.29*	*-0.31*	*-0.28*
FreqLemmes			*1*	*0.53*	*0.53*	*-0.11*	*-0.37*
CDOrtho				*1*	*0.32*	*-0.29*	*-0.28*
Preval					*1*	*0.13*	*-0.11*

Lexique 4

	FLP zRT	*FreqOrtho*	*FreqLemmes*	*CDOrtho*	*Preval*	*N Letters*	*OLD20*
FLP zRT	*1*	*-0.59*	*-0.43*	*-0.6*	*-0.37*	*0.37*	*0.4*
N Letters						*1*	*0.72*
OLD20							*1*

We observe that FLP zRTs are most strongly correlated with orthographic frequency and orthographic CD. Next, we find a moderate correlation with lemma frequency. Finally, we observe weaker correlations with prevalence, number of letters, and OLD20. Additionally, we note a strong correlation between frequency and CD, as well as between word length and OLD20, which aligns with previous findings in the literature (Hollis, 2020; Yarkoni et al., 2008).

Regression analyses

To validate the new frequencies in Lexique 4, we checked if the frequencies from Lexique 4 could better predict lexical decision reaction times than those from Lexique 3. For this purpose, we used reaction times provided by the French Lexicon Project (FLP) database (Ferrand et al., 2010). To capture the non-linearity of frequency, we used a third-degree polynomial function. Number of syllables and letters, OLD20 and number of phonemes were also entered as predictors in the multiple regressions. Below we list the percentage of variance explained in the lexical decision times (adjusted R^2) by each of the frequency measures.

We first compared FLP RTs and FLP standardized RTs. In order to compare them, we used FreqOrtho because, in lexical decision tasks, a presented word lacks grammatical category information. Table 5 presents the number of observations and the adjusted R^2 for both analyzes.

Table 5 *Adjusted R^2 for regression analyses comparing FLP RT and FLP zRT as dependent variables (N = 37,559 words)*

Dependent Variable	Predictors	Adjusted R^2
FLP RT	Lex4 FreqOrtho	0.3863
FLP zRT	Lex4 FreqOrtho	0.4125

Since our frequencies predict standardized FLP RTs better than raw RTs, we will conduct subsequent analyses exclusively on standardized RTs. It is unsurprising that the zRTs do better than RTs, because centering subject RTs on zero means that a large part of the by subject variance is taken out. In a second

analysis we wanted to know whether the new Lexique 4 frequencies predicted reaction times better than the Lexique 3 frequencies. Table 6 presents the results of the regression analyses.

Table 6 *Adjusted R^2 for regression analyses comparing Lexique 3 FreqOrtho and Lexique 4 FreqOrtho as predictors (N = 37,499 words)*

Dependent Variable	Predictors	Adjusted R^2
FLP zRT	Lex3 FreqOrtho	0.4172
FLP zRT	Lex4 FreqOrtho	0.4329

As we hypothesized, the analysis using Lexique 4 frequencies explains more variance than the analysis using Lexique 3 frequencies.

A potential question that may arise concerns orthographic forms that can correspond to multiple grammatical categories (i.e. “danse” [“dance”] as a noun and “danse” [“dance”]) as verb): for such a form, is it better to use the sum of the frequencies of the orthographic forms or the highest frequency? We compared the variable FreqOrtho and another variable which corresponds to the frequency of the most frequent grammatical category of an orthographic form (named “FreqMax” in Table 7).

Table 7 *Adjusted R^2 for regression analyses comparing Lexique 4 FreqOrtho and Lexique 4 FreqMax as predictors (N = 37,559 words)*

Dependent Variable	Predictors	Adjusted R^2
FLP zRT	Lex4 FreqOrtho	0.4125
FLP zRT	Lex4 FreqMax	0.4079

Our analyses show that it is better to use the sum of the frequencies of the orthographic forms rather than the frequency of the most frequent grammatical category at least in a lexical decision task (conclusions may differ for other tasks, particularly those involving words in context.). These results are consistent with those observed in morphology (Baayen et al., 1997; New et al., 2004), where a low-frequency singular form (i.e. "nuage" [“cloud”]) benefits from the existence of a high-frequency plural form (i.e. "nuages" [“clouds”])) because all the letters of “nuage” are present in “nuages”. So the idea is that, whenever “nuages” is presented it activates the singular form “nuage”. In this case, "danse" presented as a noun will benefit from the activation of the word "danse" presented as a verb. From this result, and for the sake of simplicity, we decided to include in Lexique 4 only FreqOrtho (and not FreqMax).

A new variable available in Lexique 4 is contextual diversity. We compared this indicator to sum frequency to see the variance explained by each of these indicators. The results are presented in Table 8.

Table 8 *Adjusted R^2 for regression analyses comparing lex4 FreqOrtho and Lex4 CDOrtho as predictors (N = 37,559 words)*

Dependent Variable	Predictors	Adjusted R^2
FLP zRT	Lex4 FreqOrtho	0.4125
FLP zRT	Lex4 CDOrtho	0.4198

According to the literature (Adelman et al., 2006), we observe that the Contextual Diversity of the orthographic form predicts reaction times slightly better than the frequency of the orthographic form.

Another question concerns the different frequency variables available in Lexique 4. Which variable best predicts reaction times: FreqMot, FreqLemme or FreqOrtho? Table 9 shows the results of the different regression analyses.

Lexique 4

Table 9 *Adjusted R^2 for regression analyses comparing lex4 FreqMot, Lex4 FreqLemme and Lex4 FreqOrtho as predictors (N = 37,559 words)*

Dependent Variable	Predictors	Adjusted R^2
FLP zRT	Lex4 FreqMot	0.3616
FLP zRT	Lex4 FreqLemme	0.3294
FLP zRT	Lex4 FreqOrtho	0.4125

Interestingly, word frequency and lemma frequency predict reaction times less well than orthographic word frequency. Lemma frequency predicts reaction times less accurately than word frequency[6].
Other regression analyses were performed to test whether the 3 frequency measures in Lexique 4 are complementary. In other words, we tested whether FreqLemme or FreqMot improve the model when controlling for FreqOrtho. Results are presented in Table 10.

Table 10 *Adjusted R^2 for regression analyses comparing the contribution of Lex4 FreqLemme and Lex4 FreqMot as predictors (N = 37,559 words)*

Dependent Variable	Predictors	Adjusted R^2
FLP zRT	Lex4 FreqOrtho	0.4125
FLP zRT	Lex4FreqOrtho + Lex4 FreqLemme	0.4466
FLP zRT	Lex4 FreqOrtho + Lex4 FreqMot	0.4131

The results indicate that FreqMot adds very little to the model. However, FreqLemme explained an important part of variance, even when FreqOrtho is already entered in the model, confirming the importance of morphology during lexical access.
Finally, another variable proposed in Lexique 4 is *Prevalence*. We ran analyses to see if this variable explained variance when the frequency of the orthographic form is already entered in the regression model (see Table 11). For these analyses, we used the probit transformation of prevalence scores, computed using the formula proposed by Keuleers et al. (2015)[7].

Table 11 *Adjusted R^2 for regression analyses to see the contribution of Prevalence as predictor (N = 36,478 words)*

Dependent Variable	Predictors	Adjusted R^2
FLP zRT	Lex4 FreqOrtho	0.4104
FLP zRT	Lex4 FreqOrtho + Lex4 FreqLemme	0.4456
FLP zRT	Lex4 FreqOrtho + Lex4 FreqLemme + Preval	0.4737

Our new prevalence indicator for French contributes additional explained variance, consistent with previous findings (Brysbaert et al., 2019).

[6] However, in a certain number of analyses which we have not been able to report here so as not to make the article too long, lemma frequency predicts reaction times in lexical decision making better than word frequency: this is the case in particular if only nouns or uninflected verbs are selected. But in all cases, orthographic frequency predicts reaction times better than lemma frequency.

[7] $\text{Prevalence}=\Phi-1(0.005+\text{Pknown}\times 0.99)$

Lexique 4

We also wanted to know the part of explained variance for each important variables individually. The results are presented in Table 12.

Table 12 *Adjusted R^2 for regression models predicting standardized lexical decision latencies (FLP zRT). The full model ("All") included all lexical predictors. To assess the unique contribution of each predictor, adjusted R^2 was recomputed after removing one predictor at a time. The last column reports the decrease in adjusted R^2 relative to the full model, with larger values indicating predictors with greater explanatory importance (N = 36,478 words)*

Dependent Variable	Predictors	Adjusted R^2	Decrease in Adj. R^2
FLP zRT	All	0.4933	-
FLP zRT	All - FreqOrtho	0.3922	0.1011
FLP zRT	All - Prevalence	0.4614	0.0319
FLP zRT	All - FreqLemmas	0.4758	0.0175
FLP zRT	All - OLD20	0.4811	0.0122
FLP zRT	All - N letters	0.4815	0.0118

Table 12 summarizes the contribution of several key lexical variables to the prediction of standardized lexical decision times (zRTs) from the FLP. The full model, which includes all predictors, accounts for approximately 49.3% of the variance in zRTs. To assess the unique contribution of each variable, we removed them one at a time and examined the drop in adjusted R^2.

The largest decrease is observed when orthographic frequency (FreqOrtho) is removed from the model (10.11%), highlighting its strongest predictive power among all variables. This is consistent with our earlier findings showing that orthographic frequency is the most robust predictor in lexical decision tasks.

Removing prevalence also leads to a substantial drop (3.19%), indicating that this variable captures a meaningful portion of the variance, likely reflecting participants' familiarity with the words beyond pure exposure frequency.

Lemma frequency (FreqLemmas) also contributes significantly (1.75%), supporting the idea that morphological representations play a role during lexical access.

The contribution of word length and orthographic neighborhood density (OLD20) is more modest but non-negligible (1.22% and 1.18%, respectively), confirming that both variables play an independent role in lexical processing, though to a lesser extent than frequency-based measures.

In summary, orthographic frequency, prevalence, and lemma frequency are the main contributors to lexical decision times, with smaller but consistent effects for word length and OLD20.

To visualize the influence of lexical variables word length on lexical decision latencies, we plotted standardized reaction times (zRTs) as a function of the number of letters (Figure 1). In this analysis, we limited the plot to word lengths ranging from 2 to 15 letters for which a sufficient number of observations were available, ensuring robust estimates at each length level. Standardized lexical decision latencies were modeled as a function of word length using ordinary least squares regression. To allow for potential non-linear associations without imposing a priori functional constraints, word length was entered as a restricted cubic spline as implemented in the rms package. The Figure shows that for words shorter than 9 letters, there is no length effect, whereas from 8 letters onward, we observe a slight, linear inhibitory effect. Overall, these results align well with those observed by New et al. (2006) in the English Lexicon Project. In contrast, we do not observe the slowdown reported by New et al. (2006) for 3- and 4-letter words. As the authors

themselves suggested, this effect in the ELP data may be due to the use of a unusually large fixation point ("* * *") prior to stimulus presentation.

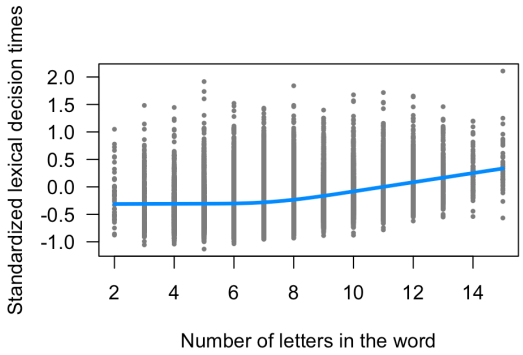

Figure 1 Effect of word length on standardized lexical decision times (zRTs) once other effects have been partialled out. Each point represents a word from the Lexique 4 dataset.

We used the same approach (using rms package) to examine the effect of other lexical variables. To investigate the role of orthographic similarity in word recognition, we examined the relationship between OLD20 and standardized lexical decision times. Figure 2 shows that for words with OLD20 values below approximately 2.5, there is no observable effect of orthographic similarity. However, for higher OLD20 values—i.e., for words that are more orthographically distinct—lexical decision times tend to increase. This inhibitory effect is consistent with previous findings and suggests that lower neighborhood density hampers visual word recognition, possibly because they display unusual or less typical letter patterns, which could make them harder to recognize.

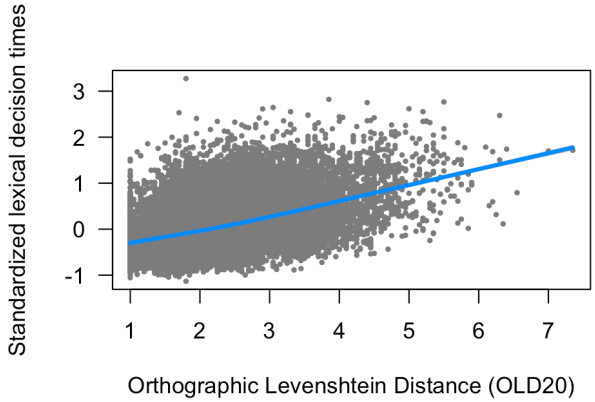

Figure 2 Effect of OLD20 on standardized lexical decision times (zRTs) once other effects have been partialled out.

To investigate the role of prevalence in word recognition, we examined the relationship between prevalence and standardized lexical decision times.
Figure 4 illustrates the relationship between word prevalence and standardized lexical decision times. As expected, words known by a larger proportion of participants tend to be recognized more quickly.

Lexique 4

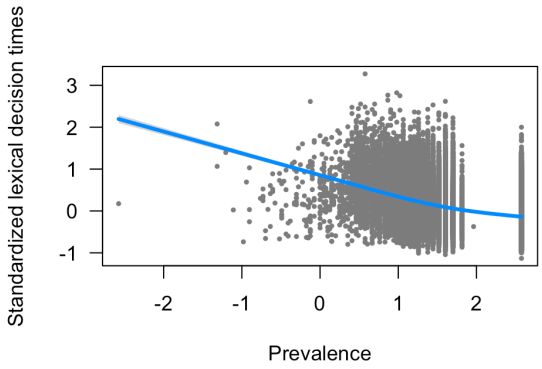

Figure 3 Effect of Prevalence on standardized lexical decision times (zRTs) once other effects have been partialled out.

Figure 4 illustrates the relationship between lemma frequency and standardized lexical decision reaction times from the French Lexical Project (FLP).

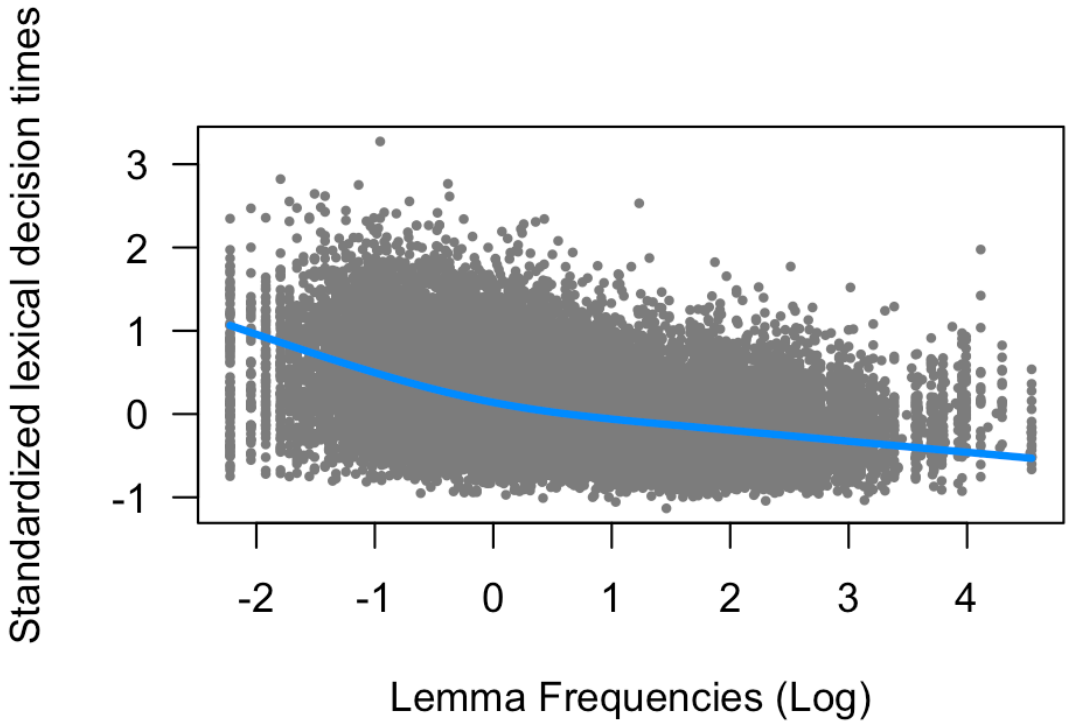

Figure 4. Effect of Lemma Frequencies on standardized lexical decision times (zRTs) once other effects have been partialled out.

To investigate the role of orthographic frequency in word recognition, we examined the relationship between orthographic frequency and standardized lexical decision times. Figure 5 illustrates the relationship between word frequency and standardized lexical decision reaction times from the French Lexical Project (FLP). This

result confirms the well-established frequency effect in lexical decision tasks

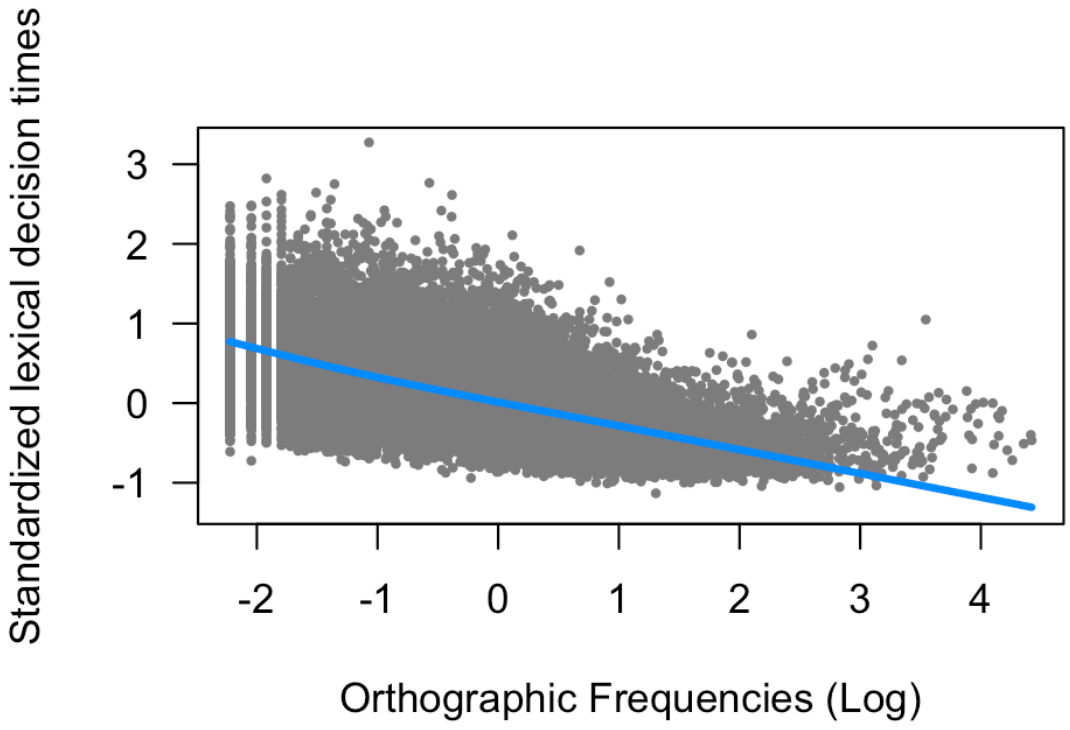

Figure 5. Effect of Orthographic Frequencies on standardized lexical decision times (zRTs) once other effects have been partialled out.

Discussion

The development and release of Lexique 4 represent a significant advancement in the field of psycholinguistics and computational linguistics. Building on the foundations laid by Lexique 3, this new iteration introduces several critical improvements that enhance both the breadth and depth of lexical data available for the French language. The expanded corpus size, updated to 316 million words from a diverse array of 65,317 documents, marks a substantial increase from the previous version, providing more robust and accurate frequency estimates. The inclusion of recent neologisms ensures that the database remains relevant in the context of contemporary language use.

Another feature of Lexique 4 is the incorporation of contextual diversity (CD) as a variable. This addition aligns with findings from previous research (Adelman et al., 2006), which demonstrated the superior predictive power of contextual diversity over simple frequency counts for certain linguistic tasks.

The detailed morphological annotations provided in Lexique 4 offer a granular view of the morphological structure of French words. This feature is particularly useful for research focusing on morphological processing and language acquisition. This morphological information for a large number of words provides a promising basis for future developments.

The new web interface of Lexique 4, Open Lexicon, significantly enhances accessibility and usability. The dynamic search functionality and the ability to sort and filter results in real-time facilitate more efficient data retrieval and analysis. This enhancement is likely to encourage broader use of the database in educational and therapeutic contexts, as well as in linguistic research. Moreover, Open Lexicon allows users to merge and integrate Lexique with other lexical databases seamlessly. The main advantage of Open Lexicon is its flexibility and user-friendliness, enabling researchers to tailor their data queries and combine diverse

linguistic variables (e.g. age of acquisition or imageability) from various sources in a single, cohesive platform. We hope that this tool will encourage researchers to develop linguistic norms for as many words as possible. Currently, there is a lack of French norms for age of acquisition, subjective frequency, imageability, and semantic richness for tens of thousands of words. This integrative approach expands the range of potential research questions and applications, making it an invaluable resource for psychologists, psycholinguists, linguists, educators, and language therapists.

Our statistical analyses highlight several important findings. The high correlation between word frequencies in Lexique 3 and Lexique 4 underscores the continuity and reliability of the data, while the increased corpus size and updated content in Lexique 4 provide even more precise frequency measures. The regression analyses indicate that Lexique 4 frequency measures are slightly better predictors of lexical decision times from the FLP project than those of Lexique 3, supporting the value of the updates and expansions made in this version.

Moreover, the comparison between summed and maximum frequencies reveals that *FreqOrtho* tends to be a better predictor of lexical decision times. This finding suggests that cumulative exposure to different forms of a word contributes more significantly to lexical access than exposure to the most frequent form alone.

We have chosen to propose three frequency indices in Lexique 4. *FreqMot* because it is the basic frequency from which other types of frequencies can be derived. *FreqOrtho* because it is the frequency that best predicts reaction times in a lexical decision task. *FreqLemme* because it explains a relatively large proportion of variance (almost 5%, which is considerable in psycholinguistics) in addition to *FreqOrtho*. To sum up, if a researcher is conducting a lexical decision task with isolated words, FreqOrtho appears to be the most appropriate measure. Ideally, care should also be taken when using words such as ‘pianos’, which will have a very low orthographic frequency but a much higher lemma frequency. Finally, these results do not allow us to say which frequency will be the most relevant in another task than a lexical decision task with isolated words. Alternatively, researchers may consider using the contextual diversity (CD) of the orthographic form. However, its psychological relevance remains debated in the literature (Hollis, 2020).

In conclusion, Lexique 4 is currently the most comprehensive lexical database for the French language. Its rich set of linguistic variables, updated corpus, and improved user interface make it an invaluable tool for researchers, educators, and clinicians. The ongoing updates and user feedback integration will ensure that Lexique 4 continues to evolve, maintaining its relevance and utility in the study of French psycholinguistics.

Availability

It is possible to download the last version of Lexique 4 or to query it online at the website http://www.lexique.org. Lexique 4 is also downloadable from this OSF repository https://osf.io/arjcz/?view_only=d17fab25993f44dfa657d0b79f8556f0.

Declarations

Acknowledgments

We would like to thank the linguist Danielle Béchennec, who carried out considerable manual work to generate the morphological information included in Lexique 4.

Funding This research did not receive any specific grant.

Conflict of interest We have no conflicts of interest to disclose.

Ethics approval Our research adheres to the ethical guidelines established by the **Oslo** Declaration on Ethical Standards, ensuring that all processes meet the highest standards of transparency, participant rights, and integrity in scientific inquiry.

Consent to participate Not applicable

Consent for publication Not applicable

Lexique 4

Data Availability Lexique 4 used for the analyses is available in an OSF repository, https://osf.io/arjcz/?view_only=d17fab25993f44dfa657d0b79f8556f0.

Code Availability The script used for the analyses is available in an OSF repository, https://osf.io/arjcz/?view_only=d17fab25993f44dfa657d0b79f8556f0.

Open Practices

Materials and analysis code are available at https://osf.io/arjcz/?view_only=d17fab25993f44dfa657d0b79f8556f0. None of the reported studies were preregistered.

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814-823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Aguasvivas, J. A., & Sainz, M. A. (2018). The Spanish Lexicon Project: A database of lexical decision times for 40,481 Spanish words. *Behavior Research Methods, 50*(1), 174-185. https://doi.org/10.3758/s13428-017-0976-6

Ayres, L. P. (1915). *The spelling vocabularies of personal and business letters.* The Arthur H. Thomas Company.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language, 37*, 94–117. https://doi.org/10.1006/jmla.1997.2509

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

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445-459. https://doi.org/10.3758/BF03193014

Bernard, J. M., Chaffin, R., Daniel, T., Hirsh, R., & Giguère, G. (2002). *Frantext: A French electronic text database*. [Database]. ATILF.

Bloch, O., Von Wartburg, W., & Meillet, A. (1964). *Dictionnaire étymologique de la langue française* (4th ed.), Presses Universitaires de France.

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods, 51,* 467-479. https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., & New, B. (2009). Moving beyond Kucera 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

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441. https://doi.org/10.1037/xhp0000159

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLOS ONE, 5*(6), e10729. https://doi.org/10.1371/journal.pone.0010729

Lexique 4

Calderone, B., Hathout, N., & Sajous, F. (2014). From GLÀFF to PsychoGLÀFF: A large psycholinguistics-oriented French lexical resource. In *Proceedings of the 16th EURALEX International Congress*, 431–446.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33*(4), 497-505. https://doi.org/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). Academic Press. https://doi.org/10.4324/9781003309734-29

Content, A., Mousty, P., & Radeau, M. (1990). Brulex: Une base de données lexicales informatisée pour le français écrit et parlé. *L'année Psychologique, 90*(4), 551-566. https://doi.org/10.3406/psy.1990.2946

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods, 49*(2), 602-615. https://doi.org/10.3758/s13428-016-0734-0

Cortese, M. J., Khanna, M. M., & Hacker, S. (2010). Recognition memory for 40,000 words: Data from the English Lexicon Project. *Behavior Research Methods, 42*(1), 149-154. https://doi.org/10.3758/BRM.42.1.149

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2012). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica, 33*(2), 133-143. https://doi.org/10.2118/0103023

Dufour, S., Peereman, R., Pallier, C., & Radeau, M. (2002). VoColex: A lexical database on phonological similarity between French words. *L'Année Psychologique*, 102, 725-746. https://doi.org/10.3406/psy.2002.29616

Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., Augustinova, M., & 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. https://doi.org/10.3758/BRM.42.2.488

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2018). Megalex: A megastudy of visual and auditory word recognition. *Behavior Research Methods, 50*(6), 2506-2521. https://doi.org/10.3758/s13428-018-1036-1

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

Hathout, N., Tanguy, L., & Sajous, F. (2014). GLÀFF, a large versatile French lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)* (pp. 728-734). Reykjavik, Iceland: European Language Resources Association (ELRA).

Hollis, G. (2020). Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language*, *114*, 104146. https://doi.org/10.1016/j.jml.2020.104146

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*, 174. https://doi.org/10.3389/fpsyg.2010.00174

Lexique 4

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. https://doi.org/10.3758/s13428-011-0118-4

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology, 68*(8), 1665–1692. https://doi.org/10.1080/17470218.2015.1022560

Kucera, H., & Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

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. https://doi.org/10.3758/BF03195560

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016)* (pp. 923-929). Portorož, Slovenia: European Language Resources Association (ELRA).

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). SUBTLEX-PL: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods, 47*(2), 471-483. https://doi.org/10.3758/s13428-014-0489-4

New, B., Brysbaert, M., Segui, J., Ferrand, L., & Rastle, K. (2004). The processing of singular and plural nouns in French and English. *Journal of Memory and Language, 51*(4), 568-585. https://doi.org/10.1016/j.jml.2004.06.010

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4), 661-677. https://doi.org/10.1017/S014271640707035X

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*, 45-52. https://doi.org/10.3758/BF03193811

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: Lexique. *L'Année Psychologique, 101*(3), 447-462. https://doi.org/10.3406/psy.2001.1341

Pagel, V., Lenzo, K., & Black, A.W. (1998) Letter to sound rules for accented lexicon compression. *Proc. 5th International Conference on Spoken Language Processing (ICSLP 1998)*, paper 0561. https://doi.org/10.21437/icslp.1998-39.

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: concrete/abstract decision data for 10,000 English words. *Behavior Research Methods*, *49*, 407-417. https://doi.org/10.3758/s13428-016-0720-6

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010).*

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to German. In Natural language processing using very large corpora (pp. 13-25). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-2390-9_2

Siew, C. S., Yi, K., & Lee, C. H. (2021). Syllable and letter similarity effects in Korean: Insights from the Korean Lexicon Project. *Journal of Memory and Language*, *116*, 104170. https://doi.org/10.1016/j.jml.2020.104170

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words.* Bureau of Publications, Teachers Co.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 human language technology conference of the north american chapter of the association for computational linguistics*, 252-259. https://doi.org/10.3115/1073445.1073478

Tse, C. S., Chan, Y. L., Yap, M. J., & Tsang, H. C. (2023). The Chinese Lexicon Project II: A megastudy of speeded naming performance for 25,000+ traditional Chinese two-character words. *Behavior Research Methods*, *55*(8), 4382-4402. https://doi.org/10.3758/s13428-022-02022-z

Tse, C. S., Yap, M. J., Chan, Y. L., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*, *49*, 1503-1519. https://doi.org/10.3758/s13428-016-0810-5

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The massive auditory lexical decision (MALD) database. *Behavior Research Methods*, *51*, 1187-1204. https://doi.org/10.3758/s13428-018-1056-1

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

Yap, M. J., Liow, S. J. R., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, *42*, 992-1003. https://doi.org/10.3758/BRM.42.4.992

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

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology.* Boston, MA: Houghton Mifflin.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley.

APPENDIX

Table A1 Phonological codes used in Lexique 4

Lexique 4 code	example	named sound	IPA symbol
a	b**a**t, pl**a**t	A	a
i	l**i**t, ém**i**s	I	i
y	l**u**	U	y
u	r**ou**e	Ou	u
o	p**eau**, m**o**t	o (close)	o
O	él**o**ge, f**o**rt	o (open)	ɔ
e	**été**	e (close)	e
E	p**ai**re, tr**ei**ze	e (open)	ɛ
°	abord**e**ra	elidable schwa	ə
2	d**eu**x	e (fermé)	ø
9	**oeu**f, p**eu**r	e (ouvert)	œ
5	c**in**q, l**in**ge	in (nasal)	ɛ̃
1	**un**, parf**um**	un (nasal)	œ̃
@	**an**ge	an (nasal)	ɑ̃
§	**on**, sav**on**	on (nasal)	ɔ̃

Lexique 4 code	example	named sound	IPA symbol
j	**y**eux, pa**ill**e	y (semi-vowel)	j
8	**hu**it, l**u**i	ui (semi-vowel)	ɥ
w	**ou**i, n**ou**er	w (semi-vowel)	w

Lexique 4 code	example	named sound	IPA symbol
p	**p**ère, sou**p**e	p (plosive)	p
b	**b**on, ro**b**e	b (plosive)	b
t	**t**erre, vi**t**e	t (plosive)	t
d	**d**ans, ai**d**e	d (plosive)	d
k	**c**arré, la**qu**e	k (plosive)	k
g	**g**are, ba**gu**e	g (plosive)	g
f	**f**eu, neu**f**	f (fricative)	f
v	**v**ous, rê**v**e	v (fricative)	v
s	**s**ale, de**ss**ous	s (fricative)	s
z	**z**éro, mai**s**on	z (fricative)	z
S	**ch**at, tâ**ch**e	ch (fricative)	ʃ
Z	**g**ilet, mi**j**oter	ge (fricative)	ʒ
m	**m**ain, fe**mm**e	m (nasal)	m
n	**n**ous, to**nn**e	n (nasal)	n
N	a**gn**eau, vi**gn**e	gn (nasal palatal)	ɲ
l	**l**ent, so**l**	l (liquid)	l
R	**r**ue, veni**r**	R	ʁ
x	**j**ota	jota (Spanish borrowing)	x

Lexique 4

G	camping	ng (English borrowing)

Table A2 Codes used to describe the different part of speech

Abbreviation	Part of speech	
ADJ	Adjective	30874
ADJ:dem	Demonstrative adjective	4
ADJ:ind	Indefinite adjective	34
AD:int	Interrogative adjective	4
ADJ:num	Numeral adjective	158
ADJ:pos	Possessive adjective	31
ADV	Adverb	1870
ART:def	Definite article	8
ART:inf	Indefinite article	3
AUX	Auxiliary	85
CON	Conjunction	36
LIA	Euphonic liaison	1
NOM	Common noun	66638
ONO	Onomatopoeia	192
PRE	Preposition	83
PRO:dem	Demonstrative pronoun	19
PRO:ind	Indefinite pronoun	44
PRO:int	Interrogative pronoun	19
PRO:per	Personal pronoun	48
PRO:pos	Possessive pronoun	24
PRO:rel	Relative pronoun	19
VER	Verb	87424

Table A3 Codes used to describe the different tenses, moods, and persons a verb can take.

Mood		Person		Tense	
ind	indicative	1s	1st person singular	pre	present
cnd	conditional	2s	2nd person singular	fut	future
sub	subjunctive	3s	3rd person singular	imp	imperfect

Lexique 4

par	participle	1p	1st person plural	pas	past
inf	infinitive	2p	2nd person plural		
imp	imperative	3p	3rd person plural		