

The French Lexicon Project:
Lexical decision data for 38,840 French words
and 38,840 pseudowords

Ludovic FERRAND

CNRS and University Blaise Pascal, Clermont-Ferrand, France

Boris NEW

CNRS and University Paris Descartes, France

Marc BRYSSBAERT and Emmanuel KEULEERS

Ghent University, Ghent, Belgium

Patrick BONIN

CNRS and University of Burgundy, Dijon, France

Alain MÉOT and Maria AUGUSTINOVA

CNRS and University Blaise Pascal, Clermont-Ferrand, France

AND

Christophe PALLIER

INSERM, U562, Cognitive Neuroimaging Unit, F-91191 Gif/Yvette, France

Running-head: The French Lexicon Project

Contact address:

Ludovic FERRAND

CNRS and University Blaise Pascal

Laboratoire de Psychologie Sociale et Cognitive (LAPSCO – UMR CNRS 6024)

34, avenue Carnot

63037 Clermont-Ferrand

France

E-mail : ludovic.ferrand@univ-bpclermont.fr

Phone : + 33 473 406 257

Fax : + 33 473 406 114

Abstract

The French Lexicon Project involved the collection of lexical decision data for 38,840 French words and the same number of nonwords. It was directly inspired by the English Lexicon Project (Balota, Yap, Cortese, Hutchison, Kessler, Loftis, Neely, Nelson, Simpson, & Treiman, 2007) and produced very comparable frequency and word length effects. The present article describes the methods used to collect the data, reports analyses on the word frequency and the word length effects, and describes the Excel files that make the data freely available for research purposes.

Understanding the cognitive processes underlying visual word recognition remains a major challenge in psycholinguistics, cognitive psychology, and cognitive science. In less than 30 years, a large amount of work has identified a number of relevant variables that affect the speed and accuracy with which words can be processed (for reviews, see Balota, Yap, & Cortese, 2006; Rastle, 2007). However, at the same time it is becoming clear that the existing approach has its limitations. Nearly all research has been based on small studies involving a limited set of monosyllabic, monomorphemic words selected according to factorial designs with a limited number of independent variables matched on a series of control variables.

The emphasis on monosyllabic words can easily be understood by the facts that these words are relatively simple stimuli to work with, that researchers wanted to limit the number of words implemented in their computational models, and that for these words a lot of ratings about their lexical characteristics are available (such as subjective frequency, age of acquisition, imageability, etc.; e.g., Balota, Pilotti, & Cortese, 2001; Brysbaert & New, 2009; Cortese & Fugett, 2004; Cortese & Khanna, 2008; Desrochers & Thompson, 2009; Ferrand, Bonin, Méot, Augustinova, New, Pallier, & Brysbaert, 2008; New, Pallier, Brysbaert, & Ferrand, 2004; Stadthagen-Gonzalez & Davis, 2006).

The strong emphasis on monosyllabic monomorphemic words is a serious limitation, however, given that they represent less than 15% of the words known. As Yap and Balota (2009) recently argued, the important next step is to understand the cognitive processes underlying the visual word recognition of more complex words, i.e., polysyllabic and polymorphemic words (see also Assink & Sandra, 2003; Ferrand & New, 2003; Ferrand & Segui, 2003 for such a view).

THE ENGLISH LEXICON PROJECT

An interesting alternative approach was initiated by Balota, Yap, Cortese, Hutchison, Kessler, Loftis, Neely, Nelson, Simpson, and Treiman (2007) in what they called the “English Lexicon Project” (ELP). In this project, Balota and colleagues collected naming times and lexical decision times for over 40,000 English words from several hundreds of participants. This type of megastudy allows researchers to run large-scale regression analyses in search for the variables that influence word recognition. The following are some of the findings that have resulted from the ELP data:

1. Word frequency is the most important predictor of visual lexical decision times, accounting for up to 40% of the variance (of which 25% cannot be accounted for by other correlated variables; Baayen, Feldman, & Schreuder, 2006; Cortese & Khanna, 2007). In contrast, for word naming times, the articulatory features of the initial phoneme are the most important, explaining up to 35% of the variance (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Cortese & Khanna, 2007). In this task, word frequency explains less than 10% of the variance (of which 6% is pure), implying that for word naming it is more critical to match conditions on the first phoneme than on frequency (Kessler, Treiman, & Mullennix, 2002; Rastle, Croot, Harrington, & Coltheart, 2005; Rastle & Davis, 2002).
2. There are large quality differences between various word frequency measures. In particular, the widely used Kučera and Francis (1967; KF67) norms are a relatively poor measure of frequency. The proportion of variance explained by KF67 frequency in visual lexical decision times is more than 10% less than the variance explained by the best available frequency estimates (Balota et al., 2004; Brysbaert & New, 2009).

3. There is a quadratic effect of word length in visual lexical decision if word frequency is controlled for: RTs decrease for very short word lengths (2-4 letter), stay stable for middle word lengths (5-8 letters), and increase sharply after that (9+ letters; New, Ferrand, Pallier, & Brysbaert, 2006).
4. Many theoretically important variables account for at most 3% of the variance in lexical decision times to monosyllabic printed words, of which usually less than 1% is unquestionably due to these variables (Baayen et al., 2006).

The present study aims to supplement the ELP with a French equivalent. Having access to French data allows researchers not only to do more research on this language, but also to compare English with French and to properly chart the commonalities and differences between these languages. Although English and French are both alphabetic languages with many historical connections, there are three important differences at the word level. First, French has a much higher morphological productivity (e.g., New, Brysbaert, Segui, Ferrand, & Rastle, 2004). For instance, French adjectives exist in four forms (masculine singular, feminine singular, masculine plural, and feminine plural), and the number of different verb forms in French can easily exceed 50 (present, simple past, past imperfective, simple future, conditional, first, second, and third person singular and plural, indicative, imperative, and subjunctive, four different forms of the past and the present participles).

A second difference between English and French is that the orthography is more transparent for French words (e.g., Ziegler, Petrova, & Ferrand, 2008). Spelling-to-sound consistencies vary across orthographies (Frost, Katz, & Bentin, 1987) and in this respect French is more consistent than English but less consistent than Spanish, German, Italian or Greek (e.g., Share, 2008; Ziegler, Jacobs, & Stone, 1996; Ziegler, Stone, & Jacobs, 1997). On the other hand, French is as low in the consistency from sound to spelling as English (Ziegler

et al., 1996) and lower than many other languages. For instance, the sound /O/ (as in “oh”) can be written as *au, aud, auds, aut, auts, aux, eau, eaux, o, oc, op, ops, os, ot, ots, ôt, and ôts*. Similarly, the words *aient* and *est* are homophones.

The third main difference between French and English is that the syllabic segmentation system is more transparent in French. French has a regular syllable structure and clear syllable boundaries, whereas English has an irregular syllable structure and often unclear syllable boundaries (e.g., Ferrand, Grainger & Segui, 1996; Ferrand, Segui & Humphreys, 1997).

THE FRENCH LEXICON PROJECT

Because of financial constraints and the time intensive nature of properly measuring naming latencies (Rastle & Davis, 2002), the French Lexicon Project (FLP) only involved the collection of lexical decision times. To make the French data comparable to the English data, the FLP closely followed the design of the ELP, except for three features we thought were less desirable.

The first feature we were unhappy with was the fact that the nonwords in the ELP had been made mostly by changing a single letter of a word. Examples are “CRIP”, “YALES”, “GAINLY”, “TRINCKLE”, “PIERCELY”, “AUGMUNTED”, and “FAITHFALLY” (retrieved from <http://lexicon.wustl.edu>, October 19, 2009; stimuli were presented in capitals, although the words on the Web site are given in lowercase letters; see Balota et al., 2007, p. 447). We thought this way of constructing nonwords introduced a confounding between the length of the nonword and the wordlikeness. The longer the nonword, the more it resembled the word from which it had been derived, thereby increasing the chances of a serial spelling verification process. To take away this confound, we assembled nonwords in

such a way that their orthographic similarity to the words mimicked the orthographic similarity of the words of that length (see the method section below).

The second feature we changed was that we presented our words and nonwords in lowercase letters rather than in uppercase letters. Participants are much more used to seeing words in lowercase. Furthermore, the French language contains important diacritic marks aiding in the pronunciation (as in *élève* [pupil] or *garçon* [boy]), which are lost when capitals are used.

Finally, we were unhappy with the fact that the fixation stimulus consisted of three asterisks presented centrally for 250 msec followed by a blank interval of 250 msec and then by the centrally presented word. We feared that this sequence of events may have hindered the perception of the central letters of the words due to the spatial overlap. Therefore, we replaced the central asterisks by two vertical lines placed one above the other with a line of text between them. Participants were asked to fixate the gap between the fixation lines. The stimuli were then presented in the gap centered on the fixation point. As far as we have been able to retrace, this procedure goes back to Frederiksen and Kroll (1976).

Apart from the above changes, our study closely resembled the ELP, although we had slightly fewer stimuli (38,840 words and 38,840 nonwords vs. 40,481 words and nonwords) and slightly more participants (975 vs. 816). This was because we wanted to restrict the experiment per participant to two sessions of one hour each. Therefore, we could present only 2,000 stimuli per participant, whereas Balota et al. (2007) – based on their previous research (Balota & Spieler, 1998; Balota et al., 2004) – decided that individual participants could produce stable data up to 3,500 stimuli each (including words and nonwords). The FLP data were obtained from the Blaise Pascal University in Clermont-Ferrand, France, and the Paris Descartes University in Paris, France.

METHOD

Participants

A total of 975 university students¹ participated (mean age = 21.4 years, SD = 3.9, min = 17, max = 35). They came from the Blaise Pascal University and the Paris Descartes University. All were native speakers of French, and had normal or corrected-to-normal vision. On average, participants had 14.2 years of education (SD = 1.9). Participants took part in two different sessions that were run on different days, separated by no more than one week. They were paid 25 euros for their participation. Each session lasted about one hour.

Apparatus

The experimental software (DMDX; Forster & Forster, 2003) and testing apparatus were identical in both sites (Paris and Clermont-Ferrand). All participants were tested on the same platform. Stimuli were presented on a 17 inch Dell LCD monitor with a refresh rate of 66 MHz and a resolution of 1280 by 1024 pixels, placed at a distance of about 60 cm from the participants. The monitor was controlled by a PC Core Duo (Dell Precision 390). Stimuli were presented in lowercase in Courier New (font 12), and they appeared on the screen as white characters on a dark background. Participants responded on a Logitech Dual Action Gamepad, which is used for superfast computer games and does not have the time delays associated with keyboards (e.g., Shimizu, 2002).

Word and Nonword Stimuli

The stimuli consisted of 38,840 words and 38,840 nonwords. The words were based on Lexique 2 (New et al., 2004) and Lexique 3 (New, Brysbaert, Veronis, & Pallier, 2007). All words with a frequency of more than .1 per million words in one of the databases were initially selected. This made a list 42,136, which was screened by the first author to take out names, parts of fixed expressions, foreign words, and letter sequences unlikely to be known to

¹ Overall, 1,037 participants were tested because some had to be replaced due to technical problems (failure of DMDX to save the data ; failure of the PC and/or the screen ; failure of the gamepad; power cut; no show of participants at the second session).

the participants. Importantly, no words were deleted because they were low-frequency inflected forms (e.g., plurals, feminine forms, verb inflections, etc.).

Nonwords were formed on the basis of the words. A different procedure was used for monosyllabic and polysyllabic words. Monosyllabic words were split in onsets and rimes. Then the set of all onsets was fully recombined with the set of all rimes, resulting in a matrix of all the original words and thousands of nonword candidates. From this list, all existing words were deleted (pseudohomophones were also removed from the nonwords via an automatic text-to-speech transcription tool) and a sample of the remaining nonwords was taken such that the following distributions matched the distributions of the word samples:

- Mean log bigram frequency
- Minimal log bigram frequency
- Mean log trigram frequency
- Minimal log trigram frequency
- Mean number of neighbors (defined as words that differed by changing, adding, or deleting a letter or by swapping two adjacent letters)
- Length in number of letters and phonemes

Bigram and trigram frequencies were calculated on the basis of the 38,840 words included in the word list (type frequencies; i.e., the number of words containing the bigram or trigram). Neighbors were defined on the basis of the words in Lexique 2 and 3. Because recent research has indicated that orthographic neighbors also include words with one letter added or deleted and words with swapped letters (Davis & Taft, 2005; De Moor & Brysbaert, 2000, Perea & Lupker, 2003; Yarkoni, Balota, & Yap, 2008), we used this definition of neighborhood rather than the traditional Coltheart N (Coltheart, Davelaar, Jonasson, & Besner, 1977; see also below). The matching of words and nonwords on these features was done automatically by means of a program published by Van Casteren and Davis (2007).

In the same way, polysyllabic nonwords were created by recombining the syllables of the words (e.g., for disyllabic words: all possible first syllables of these words times all possible second syllables; for trisyllabic words: all possible first syllables times all possible second syllables times all possible third syllables; etc.) and matching their features to those of the comparison words. Because the nonword selection occurred automatically, no experimenter-based biases were present (Forster, 2000).

Creation of Sublists for Individual Participants

The procedure used by Balota et al. (2007) was adopted to create the sublists. First, 30 random permutations of all 38,840 words and 30 random permutations of all 38,840 nonwords were made and concatenated in a master list of words and a master list of nonwords. Then the first 1,000 items of each master list were given to participant 1, the second 1,000 to participant 2, and so on. Because 39 participants were necessary to go through the complete list, we tested 975 participants to obtain at least 25 observations per item ($39 \times 25 = 975$). The items were again randomly permuted for each participant, in order to have an unpredictable sequence of 1,000 words and 1,000 nonwords.

Procedure

There were 40 practice trials before each experimental session. Participants had to indicate as rapidly and as accurately as possible if the presented letter string was a French word or a nonword. The participants responded using response buttons on a Logitech Dual Action Gamepad. They answered “yes” by pressing the button corresponding to the forefinger of the preferred hand and “no” by pressing the button corresponding to the forefinger of the nonpreferred hand.

Participants received 500 words and 500 nonwords in each session. Within a session, a 5-min break was given after every 250 trials. The sequence of events was as follows: (a) two vertical lines appeared in the center of the screen for 200 msec with a gap between them wide enough to clearly present a horizontal letter string; (b) a stimulus was presented centered on the vertical lines; (c) the vertical lines remained on the screen; (d) the participant made a response; (e) the stimulus was erased from the screen. The stimulus remained on the screen until a manual response was detected or for 4 sec if no response was made. At the end of each trial there was a 1,500-msec intertrial interval with a blank, dark screen. No feedback was provided during the experiment.

RESULTS

Nearly all participants had a mean accuracy higher than 75% and a mean RT below 1,100 msec. The data of 21 participants (2.1%) who did not fulfill these criteria were dropped. For the others, following Balota et al. (2007), we used a two-step outlier procedure for the RTs of correct responses. First, all response latencies faster than 200 msec or slower than 2,000 msec were identified as outliers. Second, for the remaining RTs, the mean and standard deviation were computed and all RTs less than 3 *SDs* below the mean of the participant or greater than 3 *SDs* above the mean were considered as outliers as well. This resulted in the rejection of 3.3% of the RTs to the correct trials. Overall, the mean percentage of error was 8.9% for words (*SD* = 4.4) and 6.6% for nonwords (*SD* = 3.9). The mean RT for correct trials was 730 msec for words (*SD* = 110) and 802 msec for nonwords (*SD* = 120).

The data of the words and the nonwords are made available as two Word Excel files. These files can be found at the Web site of the Psychonomic Society (<http://brm.psychonomic-journals.org/context/supplemental.org>), at the Web site of FLP (<http://sites.google.com/site/frenchlexiconproject/>), and at the new Web site of FLP-Lexique (<http://flp.lexique.org>). The last Web site also allows researchers to correlate the FLP data

with the many word characteristics available for French words at that Web site (see also below), and to generate lists of words that correspond to certain constraints. Each Excel file contains the following columns:

- **Item:** word or nonword
- **Ntrials:** total number of observations for the item
- **Err:** percentage of errors
- **RT:** mean RT of the correct trials for the item
- **Sd:** Standard deviation of the RTs for that item
- **Rtz:** mean RT of the correct trials for the item after the RTs of the individual participants have been transformed into standardized z-scores. In this way, the item estimate is not biased by the speed and the variability of individual participants (see Faust, Balota, & Spieler, 1999). This variable has been calculated separately for the words and the pseudowords. In this way, the z-scores of the words are not influenced by the RTs to the nonwords.
- **Nused:** number of correct responses for the item

To know how useful the percentages of error (PE) and the RTs are, it is good to determine their reliability. The easiest way to do this is to calculate the split-half reliability and correct it for length using the Spearman-Brown formula. The correlation between (a) the PE calculated on the first 12 participants who saw the word and (b) the PE calculated on the remaining participants who saw the word was .76. To correct for the fact that in total we had about 25 observations per word, we applied the Spearman-Brown formula $r_{\text{corr}} = (2*r)/(1+r)$, which gives $(2*.76)/(1+.76) = .86$. For the RTs, $r = .63$ and $r_{\text{corr}} = .77$, and for RTz, $r = .72$, $r_{\text{corr}} = .84$. The reliability index r_{corr} gives an idea of how much of the variance in the variable can be explained (the remainder is noise). The lower this value, the less interesting the measure (e.g.,

intelligence tests with a reliability below .80 do not have a high status). The fact that the reliability of the z-scores is higher than the reliability of the raw RTs confirms that taking away differences in overall RT and variability between participants removes noise from the data and does not artificially reduce the variability of the items.

USE OF THE FLP DATA

As indicated above, the FLP data can be used for different types of analyses and we indeed hope that researchers will have many questions that can be investigated with the database. As an example, here are two questions that we had specifically in mind when we designed FLP: what frequency measure should we use in word recognition experiments and what is the shape of the word length effect? Thus, in the following sections we will explain how the FLP data can address these important research topics.

What frequency measure should we use in word recognition experiments?

Up to recently, the quality of frequency estimates was judged on the basis of the size of the corpus and the “representativeness” of the materials in the corpus.

A more refined measure, however, is to use human word processing data as a validation criterion and to see which frequency index explains most of the variance in the processing times of words. This procedure was initiated by Burgess and Livesay (1998) who collected lexical decision times to 240 words and correlated them with two different frequency measures: Kucera and Francis (1967) and HAL (a frequency measure Burgess and Livesay had collected themselves on the basis of Internet discussion groups). They observed a substantially larger correlation between the HAL frequencies and lexical decision times than between the KF67 frequencies and RTs.

A similar approach was followed by New et al. (2007). They correlated lexical decision times of two samples of 200 French words with word frequencies based on written sources and word frequencies based on film and television subtitles. Although the face validity of subtitle frequencies seemed lower than that of books (there are many reasons why one may assume subtitle frequencies not to be a representative sample of “normal” language or a good predictor of visual word processing), New et al. (2007) observed reliably higher correlations for subtitle frequencies than for written frequencies. As a result, they included the subtitle frequencies in Lexique 3 (a website that allows researchers to retrieve all types of information about French words; available at www.lexique.org), where it was given the name *freqfilms2*.² The alternative measure, based on books, is known in Lexique 3 as *freqlivres*.

To find out whether indeed the frequencies on the basis of subtitles are better than the frequencies based on books, all we have to do is to correlate the 38K+ PEs and RTs with the various frequency measures. Frequency measures were log transformed and 0 frequencies were given a log value of -2.5, slightly lower than the lowest value (-2.0) observed in the corpus. In addition, because Balota et al. (2004; see also Baayen et al., 2006) found that the relationship between log frequency and word processing performance is not completely linear (in particular, a floor effect seems to be reached for words with a frequency above 100 per million), we report regression analyses both for $\log(\text{frequency})$ and $\log(\text{frequency}) + \log^2(\text{frequency})$. Finally, we also investigated whether the predictive power of the word frequency measures would improve if the average of the subtitle and the book frequencies were used.

² Initially, the frequencies on the basis of subtitles in Lexique were not weighted for the origin of the films. In New et al. (2007) we reported that it was better to do so. Hence the name *freqfilms2* (subtitles weighted for origin, as described in New et al., 2007).

Table 1: Percentages of variance explained in the FLP data by the two frequency measures available in Lexique 3 (based on films and on books) and their combination. The film-based frequencies clearly outperform the book-based frequencies, even when the nonlinearity of the log frequency curve is taken into account by using a polynomial of degree 2. However, taking the average of the film and book frequencies further improves the fit, except for the percentages of error. Because of the large number of observations, differences in % variance explained of .1 are statistically significant.

	R ² (%)		
	PE	RT	RTz
Log(Freqfilms2)	14.3	32.4	35.1
Log(Freqfilms2)+Log ²	15.2	32.4	35.1
Log(Freqlivres)	10.3	26.7	28.7
Log(Freqlivres)+Log ²	10.3	27.2	29.1
Log(Freqfilms2+Freqlivres)	13.7	34.8	37.4
Log(Freqfilms2+Freqlivres)+Log ²	15.0	35.4	38.1

Table 1 shows the outcome of the analyses. As can be seen, the film-based frequencies explained 5-6% more of the variance in the PEs and RTs than the book-based frequencies, in line with the initial observation made by New et al. (2007). Interestingly, the predictive power of the frequencies further increased when the averages of the film and book frequencies were used, indicating that a combination of spoken and written frequencies may be the way forward (similar evidence was recently obtained for English by Brysbaert & New (2009) on the basis of the Elexicon Project). These analyses also confirm that less noise is present in the RTz variable than in the raw RT variable. Using standardized scores per participant takes away some of the noise introduced by differences in speed and variability between the various participants.

The search for the best word frequency measure is but one illustration of the way in which the FLP dataset can be used to validate and optimize word metrics. However, other questions that can be addressed are: Do childhood frequencies explain additional variance in adult word processing times? Do bigram and trigram frequencies matter in word recognition and is there a difference between minimal frequency and average frequency in this respect?

Which measure of orthographic and phonological similarity to other words is the best? Which variables influence nonword rejection times? How much variance is explained by semantic variables? Which semantic variables are the most important?

For a long time, researchers had to rely on the face validity of the metrics to select and match their stimulus materials. The FLP data provides an opportunity to test the importance of various variables and to make informed decisions about which word features are essential to take into account and which only have marginal effects.

What does the word length effect look like?

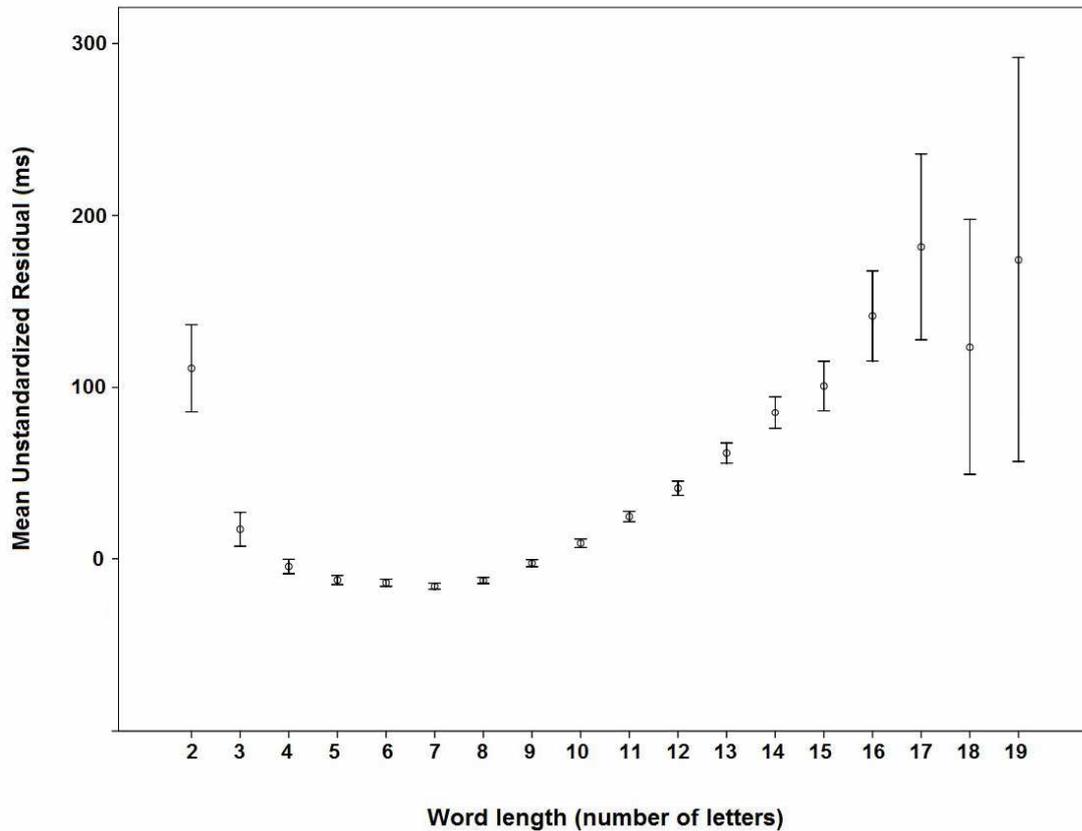
In addition to validating various word metrics, the FLP data can also be used to further examine contentious empirical findings.

A long-standing question in word recognition is whether word length is an important variable in the lexical decision task, with some authors claiming it is and other authors failing to find the effect in their studies. As mentioned in the Introduction, on the basis of an analysis of the ELP-data, New et al. (2006) proposed that word length has a quadratic effect on visual lexical decision times if word frequency is controlled for: RTs decrease for short word lengths (2-4 letter), stay stable for middle word lengths (5-8 letters), and increase sharply after that (9+ letters). However, in the section ‘The French Lexicon Project’ we saw that some features of the way in which the ELP study was run may have been responsible for the quadratic word length effect. First, the use of *** as the fixation stimulus may have made the perception of the subsequent word harder. This would have been particularly true for words of 2 and 3 letters, which overlapped completely with the asterisks. Second, the way in which the nonwords were constructed made the distinction between words and nonwords harder for long nonwords (see the examples in the Introduction). This may have been the origin of the longer RTs for words of 9+ letters. Indeed, ELP has been criticized for its long RTs to words

overall (788 msec, SD = 165; Sibley, Kello, & Seidenberg, 2009), about 50 msec longer than we found in FLP for comparable words in terms of frequency, length, and morphological complexity.

To investigate whether we could replicate the curvilinear relationship between word length and RT in the FLP data, we first partialled out the effect of log frequency. That is, the residuals were saved of a regression with $\log(\text{Freqfilms2} + \text{Freqlivres})$ and $\log^2(\text{Freqfilms2} + \text{Freqlivres})$ as predictors and RT as the dependent variable. Then we looked at the mean value of the residuals as a function of word length. Figure 1 shows the outcome. As can be seen, despite the changes we introduced, the quadratic length effect as described by New et al. (2006) still is very clearly present, indicating that it is not an artifact of the way in which the Elexicon Project was run.

Figure 1: Effect of word length when the effect of word frequency has been partialled out. As in the ELP data, we see a curvilinear effect of word length. Word processing time is shortest for words of 6-8 letters and increases for both shorter and longer words. The error bars indicate twice the SE of the means.

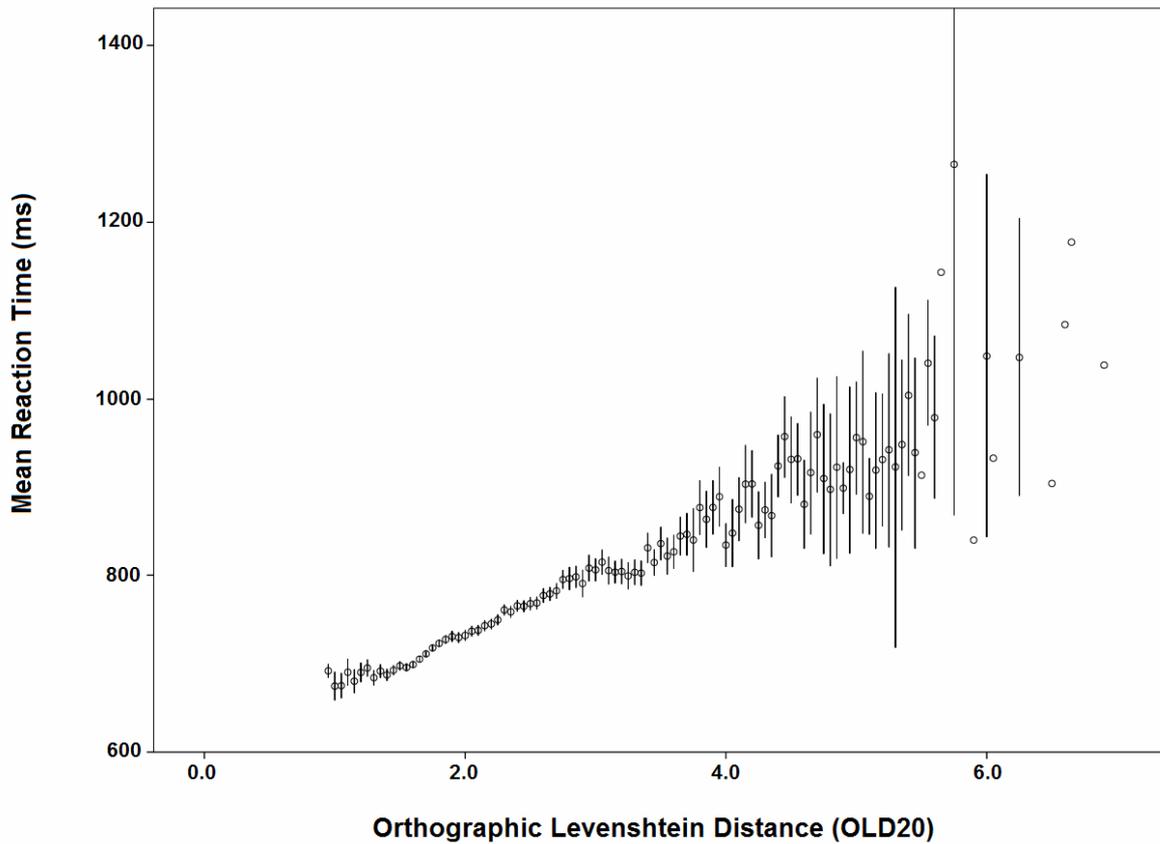


In New et al. (2006) we explained the effect by making reference to the fact that in reading saccades are typically some 8 characters long, so that the visual system may have a preference for this length. An alternative explanation was proposed by Whitney (July 31, 2008; retrieved from <http://orthoblography.blogspot.com/search/label/seriality> on October 19, 2009). According to Whitney, the quadratic effect is the outcome of two opposite factors. The first is the serial processing of the letters of a word (Whitney, 2008), resulting in a linear word length effect. The second is the time it takes for the lexicon to settle given an input. Because short words have more close neighbors than long words (i.e., words differing in only one letter), it takes longer for the lexicon to decide between the various competing representations for short words than for long words. According to Whitney, the outcome of the two opposing factors is the curvilinear function seen in Figure 1.

Still another explanation was put forward by Yarkoni et al. (2008) who suggested that researchers have been using a suboptimal measure of neighborhood density (i.e., the number of words similar to the target word). Rather than counting the number of neighbors, Yarkoni et al. argued that a better measure is the orthographic Levenshtein distance (OLD). This measure is obtained by calculating the average number of operations (letter deletion, insertion, or substitution) needed to change a word into another word. For instance, the OLD from SMILE to SIMILES is 2 (two insertions: I and S). By calculating the OLDs to the 20 closest words, Yarkoni et al. obtained a continuous variable (OLD20) that explained a substantial part of the variance in the ELP data and took away most of the word length effect. Words that were very similar to 20 other words (i.e., had a low OLD20) were responded to faster than words requiring many insertions, deletions and substitutions to be turned into other words.

To investigate whether the word length effect could be an artifact of OLD20 in French as well, we first calculated the French OLD20s. Figure 2 shows the mean lexical decision times as a function of OLD20. As can be seen, the variable had a big effect on RTs ($R^2 = .187$). When 20 or more new words could be made of the target word by one letter substitution, deletion or insertion ($OLD20 = 1$), RTs were on average faster than 700 msec. In contrast, when many changes had to be introduced to change the target word into another word ($OLD20 > 5$), RTs were around 1000 msec.

Figure 2: Effect of the orthographic similarity on lexical decision time. When 20 or more new words can be made of the target word by one letter substitution, deletion or insertion ($OLD20 = 1$), RTs are faster than 700 msec. In contrast, when many changes must be made to change the target word into another word ($OLD20 > 5$), RTs are around 1000 msec. The error bars indicate twice the SE of the means. When there is no error bar, there was only one observation in that bin.



To have a better idea of the relative importance of the different variables (frequency, frequency², length, length², and OLD20), we entered them in a stepwise regression analysis, both for reaction times (z-scores) and percentage correct. Table 2 shows the outcome. Whereas the effects of word length and word length² remained significant, their impact was much decreased: The effects specifically attributed to length on RTz decreased from $\Delta R^2 = 5.5\%$ in a regression without OLD20 to 1% in a regression with OLD20. The difference for error rate was less pronounced ($\Delta R^2 = 5.4\%$ vs. $\Delta R^2 = 4.7\%$). So, it looks indeed as if a large part of the length effect on lexical decision latencies discovered by New et al. (2006) can be explained by orthographic similarity (or orthographic distance to the nearest words).

Table 2: Percentages of variance explained in the FLP data by word frequency, word length, and similarity to other words (as measured by OLD20). Variables entered in a multiple regression analysis according to the variance they explained (stepwise function). Because of the large number of observations, all effects are significant beyond $p < .001$.

	R ² (%)	ΔR ² (%)
RTz		
Log(Freqfilms2+Freqlivres)	37.4	37.4
OLD20	44.1	6.7
Log ² (Freqfilms2+Freqlivres)	44.9	.9
Length+Length ²	45.9	1.0
PE		
Log(Freqfilms2+Freqlivres)	13.7	13.7
Length	18.2	4.5
OLD20	22.0	3.8
Log ² (Freqfilms2+Freqlivres)	22.7	.7
Length ²	22.9	.2

Table 3 shows the intercorrelation matrix of the various variables.

Table 3: Correlations between the various variables. Frequencies are log frequencies; Freqtot is the average of the book and subtitle frequencies. All correlations are significant (N = 38,335).

	1.	2.	3.	4.	5.	6.	7.	8.
1. Freqlivres	-	.657	.919	-.319	-.279	-.517	-.535	-.320
2. Freqfilms2		-	.854	-.360	-.292	-.569	-.592	-.378
3. Freqtot			-	-.376	-.322	-.590	-.612	-.370
4. Length				-	.771	.399	.393	-.057
5. OLD20					-	.433	.442	.095
6. RT						-	.956	.558
7. RTz							-	.606
8. PE								-

Again, the length issue is but an example of a large range of questions that can be addressed with the FLP dataset. The most interesting aspect of such a large scale dataset is

that one gets a panoramic overview of the impact of a variable across the entire range (Figures 1 and 2) rather than the narrow window usually offered by small-scale factorial designs. The latter are still needed, because they can give a much more detailed picture about a particular part of the problem space (Sibley, Kello, & Seidenberg, 2009). However, megastudies like FLP allow us in addition to have a look at the broader picture.

CONCLUSION

In this article we described the collection of lexical decision data for 38,000+ French words within the French Lexicon Project (FLP). These data were acquired in very much the same way as the lexical decision data of the English Lexicon Project (ELP; Balota et al., 2007). For the three variables tested, word frequency, word length, and orthographic neighborhood, we indeed found very much the same pattern of findings. For word frequency, subtitle-based frequency estimates outperformed book-based frequency estimates; and for word-length, we found a quadratic effect with the shortest lexical decision times for words of 6-8 letters when word frequency was taken into account. However, if orthographic similarity (operationalized as OLD20) was added, the word length effect on the RTs largely disappeared (though it remained significant for the accuracy data).

The FLP data are freely available for researchers who want to run other analyses either to compare the French language with the English or to address French-specific characteristics.

REFERENCES

- Assink, E. M. H., & Sandra, D. (2003) (Eds.). *Reading complex words: Cross-language studies*. New York: Kluwer Academic /Plenum Publishers.
- Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 55, 290-313.
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition for single syllable words. *Journal of Experimental Psychology: General*, 133, 283-316.
- Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, 29, 639-647.
- Balota, D. A., & Spieler, D. H. (1998). The utility of item-level analyses in model evaluation: A reply to Seidenberg and Plaut. *Psychological Science*, 9, 238-240.
- Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (a travel update). In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (2nd edition). London: Academic Press.
- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445-459.
- 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, 977-990.
- 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.

- 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). London: Academic Press.
- Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers*, *36*, 384-387.
- Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *Quarterly Journal of Experimental Psychology*, *60*, 1072-1082.
- Cortese, M. J., & Khanna, M. M. (2008). Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods*, *40*, 791-794.
- Davis, C. J., & Taft, M. (2005). More words in the neighborhood: Interference in lexical decision due to deletion neighbors. *Psychonomic Bulletin & Review*, *12*, 904-910.
- Desrochers, A., & Thompson, G. L. (2009). Subjective frequency and imageability ratings for 3,600 French nouns. *Behavior Research Methods*, *41*, 546-557.
- De Moor, W., & Brysbaert, M. (2000). Neighborhood-frequency effects when primes and targets are of different lengths. *Psychological Research*, *63*, 159-162.
- Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, *125*, 777-799.
- Ferrand, L., Bonin, P., Méot, A., Augustinova, M., New, B., Pallier, C., & Brysbaert, M. (2008). Age-of-acquisition and subjective frequency for all generally known monosyllabic French words and their relation with other psycholinguistic variables. *Behavior Research Methods*, *40*, 1049-1054.
- Ferrand, L., & New, B. (2003). Syllabic length effects in visual word recognition and naming. *Acta Psychologica*, *113*, 167-183.

- Ferrand, L., Segui, J., & Grainger, J. (1996). Masked priming of word and picture naming: The role of syllabic units. *Journal of Memory and Language*, 35, 708-723.
- Ferrand, L., Segui, J., & Humphreys, G. W. (1997). The syllable's role in word naming. *Memory & Cognition*, 25, 458-470.
- Ferrand, L., & Segui, J. (2003). Reading aloud polysyllabic words. In E. M. H. Assink & D. Sandra (Eds.), *Reading Complex Words: Cross language studies* (pp. 295-314). New York: Kluwer Academic /Plenum Publishers.
- Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, 28, 1109-1115.
- Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior, Research Methods, Instruments & Computers*, 35, 116-124.
- Frederiksen, J. R., & Kroll, J. F. (1976). Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 361-379.
- Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographic depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 104-115.
- Kessler, B., Treiman, R., & Mullenix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language*, 47, 145-171.
- Kučera, H., & Francis, W. (1967). *Computational analyses of present-day American English*. Providence, RI: Brown University Press.
- 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, 58-585.

- New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28, 661-677.
- 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.
- New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36, 516-524.
- Perea, M., & Lupker, S. J. (2003). Transposed-letter confusability effects in masked form priming. In S. Kinoshita and S. J. Lupker (Eds.), *Masked priming: State of the art* (pp. 97-120). Hove, UK: Psychology Press.
- Rastle, K. (2007). Visual word recognition. In M.G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 71-88). Oxford, UK: Oxford University Press.
- Rastle, K., Croot, K. P., Harrington, J. M., & Coltheart, M. (2005). Characterizing the motor execution stage of speech production: Consonantal effects on delayed naming latency and onset duration. *Journal of Experimental Psychology: Human Perception and Performance*, 31, 1083-1095.
- Rastle, K., & Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance*, 28, 307-314.
- Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, 134, 584-615.
- Shimizu, H. (2002). Measuring keyboard response delays by comparing keyboard and joystick inputs. *Behavior Research Methods, Instruments, & Computers*, 34, 250-256.
- Sibley, D. E., Kello, C. T., & Seidenberg, M. S. (2009). Error, error everywhere: A look at megastudies of word reading. In N. Taatgen & H. Van Rijn (Eds.), *Proceedings of the 31st*

- Annual Conference of the Cognitive Science Society* (pp. 1036-1041). Amsterdam: Cognitive Science Soc., Inc.
- Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*, 598-605.
- Van Casteren, H., & Davis, M. H. (2007). Match: a program to assist in matching the conditions of factorial experiments. *Behavior Research Methods*, *39*, 973-978.
- Whitney, C. (2008). Supporting the serial in the SERIOL model. *Language and Cognitive Processes*, *23*, 824-865.
- Yap, M., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*, 502-529.
- Yarkoni, T., Balota, D. A., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971-979.
- Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, *47*, 1-29.
- Ziegler, J. C., Jacobs, A. M., & Stone, G. O. (1996). Statistical analysis of the bidirectional inconsistency of spelling and sound in French. *Behavior Research Methods, Instruments, & Computers*, *28*, 504-515.
- Ziegler, J. C., Petrova, A., & Ferrand, L. (2008). Feedback consistency effects in visual and auditory word recognition: Where do we stand after more than a decade? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 643-661.
- Ziegler, J. C., Stone, G. O., & Jacobs, A. M. (1997). What is the pronunciation for -ough and the spelling for u/? A database for computing feedforward and feedback consistency in English. *Behavior Research Methods, Instruments & Computers*, *29*, 600-618.

Author note

This work was supported by an ANR grant n°06-CORP-00101 to Ludovic Ferrand (Agence Nationale de la Recherche, France).

We wish to thank Emmanuelle Neuville, Claire Rastoul, Valentin Flaudias, Sylvie Pires, Olivier Audebert, Eurydice Magneron and Giovanna Caris for testing the participants (from November 2007 to March 2009).

This document was created with Win2PDF available at <http://www.win2pdf.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.
This page will not be added after purchasing Win2PDF.