The French Lexicon Project

Ludovic Ferrand[1], Boris New[2], Marc Brysbaert[3], Emmanuel Keuleers[3], Patrick Bonin[4], Maria Augustinova[1], Alain Méot[1], Christophe Pallier[5]

[1]CNRS and Blaise Pascal University, France, [2]CNRS and Paris Descartes University, France, [3]Ghent University, Belgium, [4]CNRS and University of Burgundy, France, [5]INSERM, Cognitive Neuroimaging Unit, France

Introduction

The French Lexicon Project (FLP) involved the collection of lexical decision data for 38,840 French words and 38,840 pseudowords. It was directly inspired by the English Lexicon Project (ELP; Balota et al., 2007).

Our aim was to supplement the ELP with a French equivalent.

This type of megastudy allows psycholinguists to run large-scale regression analyses in search for the variables that influence word recognition (Yap & Balota, 2009).

Methods: Lexical decision task

- 975 university students (ages 17-35)
- 38,840 French words taken from *Lexique* (including plurals, feminine forms and verb inflections)
- 38,840 pseudowords constructed by recombining subsyllabic elements
- 2 sessions of 1 hour (500 words + 500 pseudowords per session)

	FLP	ELP
Number of words	38 840	40 481
Length in letters	8.5 (2-20)	8.0 (1-21)
Length in syllables	2.5 (1-7)	2.5 (1-8)
Subtitle frequency per million	21.13 (0-25 988)	25.2 (.02-41 857)
Accuracy words	91% (8-100)	84%(0-100)
RT words	740 ms (515-1464)	784 ms (415-1755)
Accuracy pseudowords	93% (8-100)	88% (0-100)
RT pseudowords	807 ms (519-1604)	856 ms (589-1814)

Table 1. Main statistics of the FLP and the ELP.

Availability of FLP data

The FLP data (RT, RTz, PE) can be found at https://sites.google.com/site/frenchlexicon/home

Use of the FLP data

The FLP data can be used for different types of analyses. As an example, here are two questions we had in mind when we designed FLP.

(1) What frequency measure should we use?

We correlated the 38K+ PEs and RTs with various frequency measures (taken from *Lexique*): one based on written sources (Freqbooks) and one based on film subtitles (Freqfilms).

- Film-based frequencies explained 5%-6% more of the variance in the PEs and RTs than did the book-

based frequencies (see Table 2), in line with the demonstration of Brysbaert and New (2009) for ELP.

	R^2(%)		
	PE	RT	RTz
Log(Freqfilms)	14.3	32.4	35.1
Log(Freqfilms) + Log²	15.2	32.4	35.1
Log(Freqbooks)	10.3	26.7	28.7
Log(Freqbooks) + Log²	10.3	27.2	29.1
Log(Freqfilms + Freqbooks)	13.7	34.8	37.4
Log(Freqfilms + Freqbooks) + Log²	15.0	35.4	38.1

Table 2. Percentages of variance explained in the FLP data by various frequency measures.

- A combination of film-based and written frequencies also increased the predictive power of frequencies.

(2) What is the shape of the word length effect?

Using ELP data, New et al. (2006) found that word length has a quadratic effect on visual lexical decision times if frequency is controlled for.

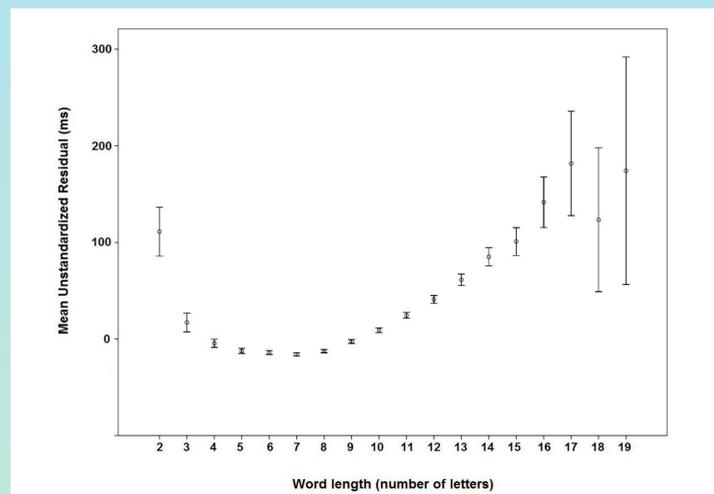We replicated this quadratic length effect (see Figure 1) with the FLP data.

Figure 1. Quadratic effect of word length (when frequency has been partialed out).

Word frequency, word length and orthographic similarity

Yarkoni et al. (2008) tested a new measure of orthographic similarity (OLD20) that explained a substantial part of the variance in the ELP and took away most of the word length.

We tested whether the word length effect could be an artefact of OLD20 in French as well.

- A large part of the quadratic length effect can be explained by orthographic similarity (Table 3).

RTz	R^2 (%)	ΔR^2 (%)
Log(Freqfilms + Freqbooks)	37.4	37.4
OLD20	44.1	6.7
Log²(Freqfilms + Freqbooks)	44.9	0.9
Length + Length²	45.9	1.0
PE		
Log(Freqfilms + Freqbooks)	13.7	13.7
Length	18.2	4.5
OLD20	22.0	3.8
Log²(Freqfilms + Freqbooks)	22.7	0.7
Length²	22.9	0.2

Table 3. Percentages of variance explained in the FLP data. For all effects, p < .001

- Significant effects of word length, but decreased impact: from ΔR^2 = 5.5% in a regression without OLD20 to 1% in a regression with OLD20 (less pronounced for PEs: 5.4% to 4.5%).

Comparison with the ELP and DLP data

Keuleers et al. (2010) compared the FLP, ELP and DLP (Dutch Lexicon Project) dataset.

Monosyllabic and disyllabic only ΔR^2 (%)	FLP	ELP	DLP
Subtitle frequency	30.7	42.7	34.4
OLD20	0.7	3.4	ns
Length (letters)	0.1	0.2	0.1
Length (syllables)	0.3	0.3	ns

Table 4. Percentages of variance explained in the FLP, ELP and DLP for mono- and disyllabic words only.

- Higher impact on OLD20 in ELP than DLP and FLP (see Table 4).

New questions to address

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

References

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.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., 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*, 488-496.

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 Language Sciences, Psychology, 1:174*.

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.

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.