

Lexique 2 : A New French Lexical Database

Boris New¹, Christophe Pallier², Marc Brysbaert¹, Ludovic Ferrand³

¹Royal Holloway, University of London

²INSERM, U562 Service Hospitalier Frédéric Joliot

³CNRS and Université René Descartes, Paris, France

RUNNING HEAD: A NEW FRENCH LEXICAL DATABASE

Corresponding author :

Boris New; Laboratoire de Psychologie Expérimentale; 71 avenue Edouard Vaillant; F-92100 Boulogne-Billancourt; France

A New French Lexical Database

Abstract

In this paper, we present a new lexical database for French: *Lexique*. In addition to classical word information such as gender, number, and grammatical category, *Lexique* also includes a series of interesting new characteristics. First, word frequencies are based on two cues: a contemporary corpus of texts and the number of web pages containing the word. Second, the database is split in several tables: a graphemic table with all the relevant frequencies, a table structured around lemmas, which is particularly interesting for the study of the inflectional family, and a table about surface frequency cues. Third, *Lexique* is distributed under a GNU-like license allowing people to contribute to it. Finally, a meta search engine called Open *Lexique* has been developed so that new databases can be added very easily to the existing ones. *Lexique* can either be downloaded or interrogated freely from the web site: <http://www.lexique.org>.

Keywords: lexical database, French, frequencies

A New French Lexical Database

Psycholinguistic researchers make extensive use of word databases. These databases are particularly important because they are the foundation of most psycholinguistic studies. First the availability of a particular piece of information determines whether this factor can be studied or not. For example if frequencies of inflectional forms are given, studies on morphological processing are possible. Second, the accuracy of the measures in the database will directly influence the accuracy of the research and the statistical reliability of the experiments done.

During many years, psycholinguists studying French used *Brulex* (Content, Mousty et Radeau; 1990). As the first electronic database for psycholinguists, *Brulex* was very helpful despite the following drawbacks. *Brulex* frequencies were based on a corpus of texts published between 1919 and 1964. As frequency is one of the most important and robust factor manipulated in psycholinguistics experiments, it is important to have frequencies as reliable and recent as possible. In this respect, *Brulex* frequencies look a little bit outdated. Furthermore, *Brulex* did not include the inflectional verbal or plural forms. Thus, studies about verbal or plural forms were impossible in French. Other problems came from the fact that lemmas were not linked to inflected forms and that syllabified forms were not included. Finally the last limitation is that this database has not been updated since its publication in 1990. Other databases like *Novlex* (Lambert & Chesnet, 2001) or *Manulex* (Lété, Sprenger-Charolles, & Colé, 2004) provide more recent frequencies, but they are based on corpora for children.

For all these reasons, we decided to develop a new database. In this article we briefly describe how we created *Lexique* and how it is structured. French speakers who want more details about the structure can find them in New, Pallier, Ferrand, et Matos (2001), a paper that presents the first version of the database in detail. Here, we will mainly focus on the original features appeared after the first version such as a GNU-like license, a website, and a meta search engine. These features are particularly interesting because they can be useful for other databases in other languages as well.

In order to create a new database, our first problem was to find a corpus of texts as big and recent as possible. For this we chose the *Frantext* corpus constituted of numerous texts published between 1800 and 2000. Inside this large corpus we selected the texts published after 1950 in order to have a rather contemporary corpus. The selected 487 texts, consisting mostly of novels and essays, contained a total of 31 words.

With the use of *Frantext's* search engine, we obtained a list of 246 000 occurrences and their frequencies. Because these occurrences contained a lot of foreign and proper words, we cleaned them using *Le Grand Robert* (Robert, 1992) and the *ispell* spelling checker coupled to *Français-Gutenberg* (Pythoud, 1996). For the extraction of the morpho-syntactic information, two grammatical parsers have been used in addition to *Le Grand Robert: TreeTagger*ⁱⁱ (Schmid, 1994) and *Flemm 2*ⁱⁱⁱ (Namer, 2000).

Lexique is composed of three main databases in text format: *Graphemes*, *Lemmes* and *Surface*. *Graphemes* is the main database from which the other two are derived. *Lemmes* presents an inflectional family organization that may be useful for psycholinguists interested in lemmas or in the inflectional family. *Surface* displays information about words and their letter, bigram, trigram, phoneme, and syllable frequencies. There is also an independent archive named *Surface*, which presents detailed statistics about surface frequencies.

Graphemes and *Lemmes* are described in Table 1 and 2.

Table 1 : Graphemes fields and their description

Field Name	Description
graph	Orthographic representation
phon	Phonological representation
cgram	Grammatical category
genre	Gender
nombre	Number
lemme	Lemma
rand	Random number
frantfreqparm	Frantext frequency

A New French Lexical Database

fsfreqparm	Fastsearch frequency
nblettres	Number of letters
nbphons	Number of phonemes
cvcv	Orthographic abstract representation
pcvcv	Phonological abstract representation
puorth	Orthographic uniqueness point
puphon	Phonological uniqueness point
syll	Syllabified form
nbsyll	Number of syllables
syllcv	Syllabified abstract form
voisorth	Number of orthographic neighbours
voisphon	Number of phonological neighbours
orthrenv	Reverse orthographic representation
phonrenv	Reverse phonological representation

Table 2 : Lemmes fields and their description

Field Name	Description
lem	Orthographic representation of the lemma
graph	Inflectional family
phon	Phonological family
cgram	Grammatical class family
genre	Gender family
nombre	Number family
rand	Random number
frantfreqcum	Inflectional frantext cumulative frequency
frantfreqgraph	Inflectional frantext frequency family
fsfreqcum	Inflectional fastsearch cumulative frequency
fsfreqgraph	Inflectional fastsearch frequency family

As we wanted to have the phonological representations of the inflected forms, we could not use the ones from a dictionary like “Le Grand Robert” as the *Brulex* authors did. Thus, we used a “text to speech” application called *LAIPTTS*[®] (Keller & Zellner, 1998). Unfortunately, this application was designed for processing continuous speech. Once the first public version of *Lexique* was released, Peereman and Dufour (2003) compared phonetic notations from *Brulex* (obtained from *Le Petit Robert*) with those of *Lexique*. They detected 2500 (over the 30 000 words of *Brulex*) differences due to exceptional pronunciations or problems with the rules used by *LAIPTTS*. They also corrected the phonetic representation for the schwas positions. They suppressed the distinction between the two types of /a/, /o/, and /ɪ/. These corrections have been included in *Lexique 2* and upper versions. Recently, in the *Lexique 2.50* release we also modified the syllabification algorithm (Pallier, 1994), so that it ignores the schwa at the end of words.

Frequency is a very important factor in psycholinguistic studies. (see Monsell, 1991 for a review). As frequencies based on a corpus of texts have a certain inertia and can underestimate contemporary words like *advertisement* or *firm*, we decided to include a second frequency source based on the number of web pages written in French where the word appears.

This cue is slightly different from the standard frequency given by a corpus of texts. The standard frequency is the number of times a word appears in a text as a function of the total number of words. In contrast, web frequencies are based on the number of pages where the word appears as a function of the total number of web

A New French Lexical Database

pages. Frequencies based on web pages are interesting because:

- Web pages are more dynamic than corpora of texts as everybody can publish a web page very easily.
- Web pages exist for nearly all human activities (whereas a corpus is usually limited in the literary texts)
- Web pages are updated very regularly
- Web pages in a particular language constitute a vast corpus

We chose to use the *Fastsearch* search engine based on 15 millions French web pages for the following reasons. First, this search engine gives the precise number of pages where the word is found (contrary to *Google* which only gives approximations). Second *Fastsearch* differentiated (although this no longer seems to be true) between accentuated and non-accentuated characters. We did our research using the *SafeSearch* mode to prevent overestimates of sexually connoted words.

Recently Blair, Umland and Ma (2002) compared the frequencies of 400 English words obtained with four different search engines (*AltaVista*, *Northern Light*, *Excite* and *Yahoo!*) and the frequencies based on two different corpora of texts (Francis and Kucera, 1982; Baayen, Piepenbrock, and Van Rijn, 1993). They observed a very strong correlation between the search engines (and thus the number of hits) and the text corpora. Because the web is constantly updated Blair et al. repeated their searches 6 months later and noted that the frequencies had not changed much.

On the basis of these findings they concluded that although the two measures are different (number of pages containing the word vs. number of words), the frequencies given by the web are as representative as frequencies given by texts corpora.

Yet, it is clear that internet hit rates differ to some extent from corpus-based estimates of frequency of usage. Consider very frequent words (like the article *'the'* in English) which appear in virtually every web page: their hit rates are quite large, maybe approaching 100%, while their lexical frequencies are but a few percent. In such cases, then, hit rates overestimate the frequency of usage. On the opposite, consider now a very low frequency word, only used in certain contexts: It will occur only in a few web pages, but when it is used, it is likely to appear several times on the page, and this is not taken into account by the hit count. So, its frequency of usage could be underestimated by the hit rate.

Lexique provides text-based frequency estimates and web-based hit rates for about 129,000 distinct word forms, allowing us to examine the relationship between both variables in a very detailed way.^v All frequencies are expressed in occurrences per million (words for *Frantext* and webpages for *Fastsearch* frequencies).

Figure 1 Relationship between text-based frequencies and web-based hit rates, both expressed per million, and shown on logarithmic scales. The solid line is the linear regression line.

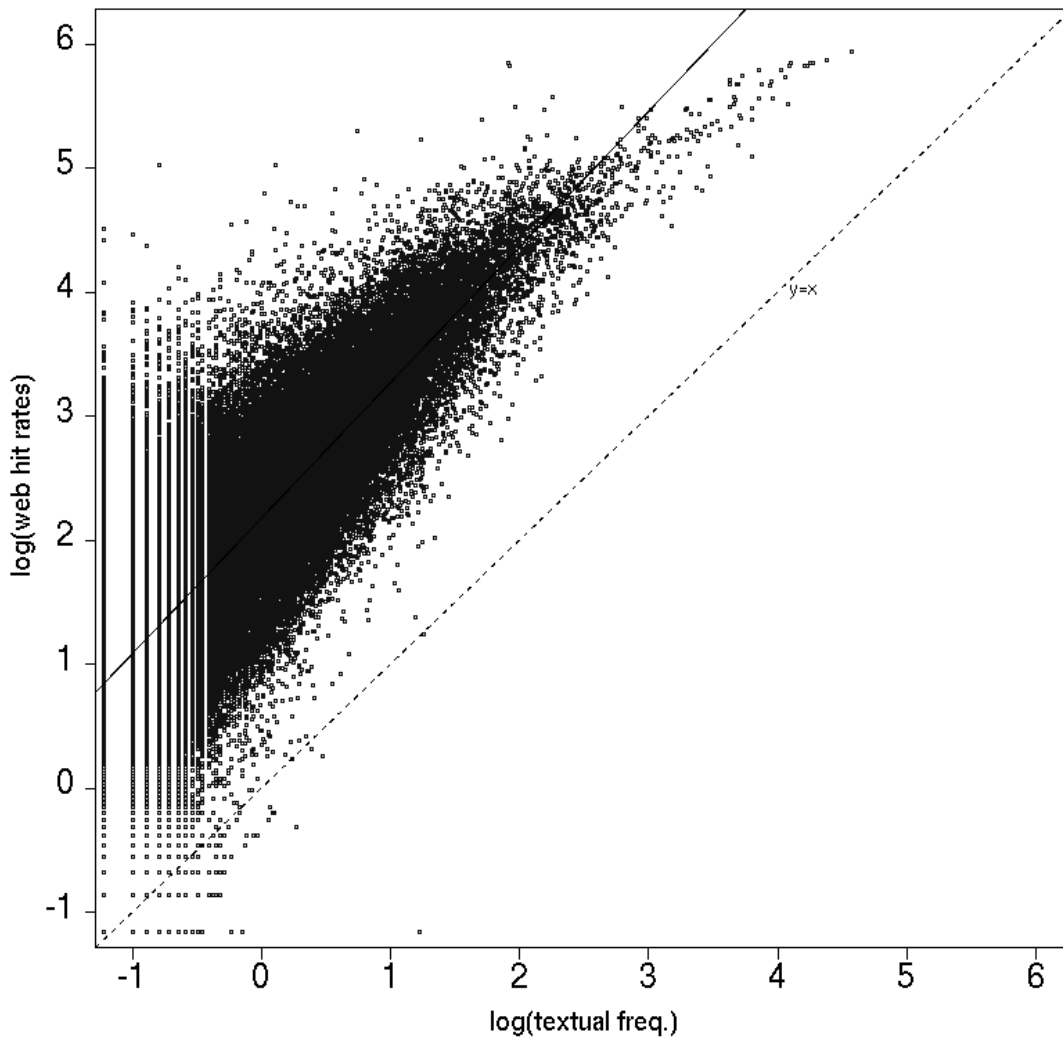


Figure 1 shows the text-based frequencies and the web-based hit rates of all the items. As expected, the hit rates are higher than the text-based frequencies, in particular for the low-frequency words. In addition, there is quite some variability among the low-frequency items. Words with a text-based frequency of 1 per million ($\log = 0$), had a web-based frequency varying from 3 per million ($\log = 0.5$) to 1,000 per million ($\log = 3$). Similarly, words with a web-based frequency of 1,000 per million ($\log = 3$), had a text-based frequency going from less than 1 per million ($\log < 0$) to more than 30 per million ($\log > 1.5$). A linear regression analysis between the two variables yielded the equation:

A New French Lexical Database

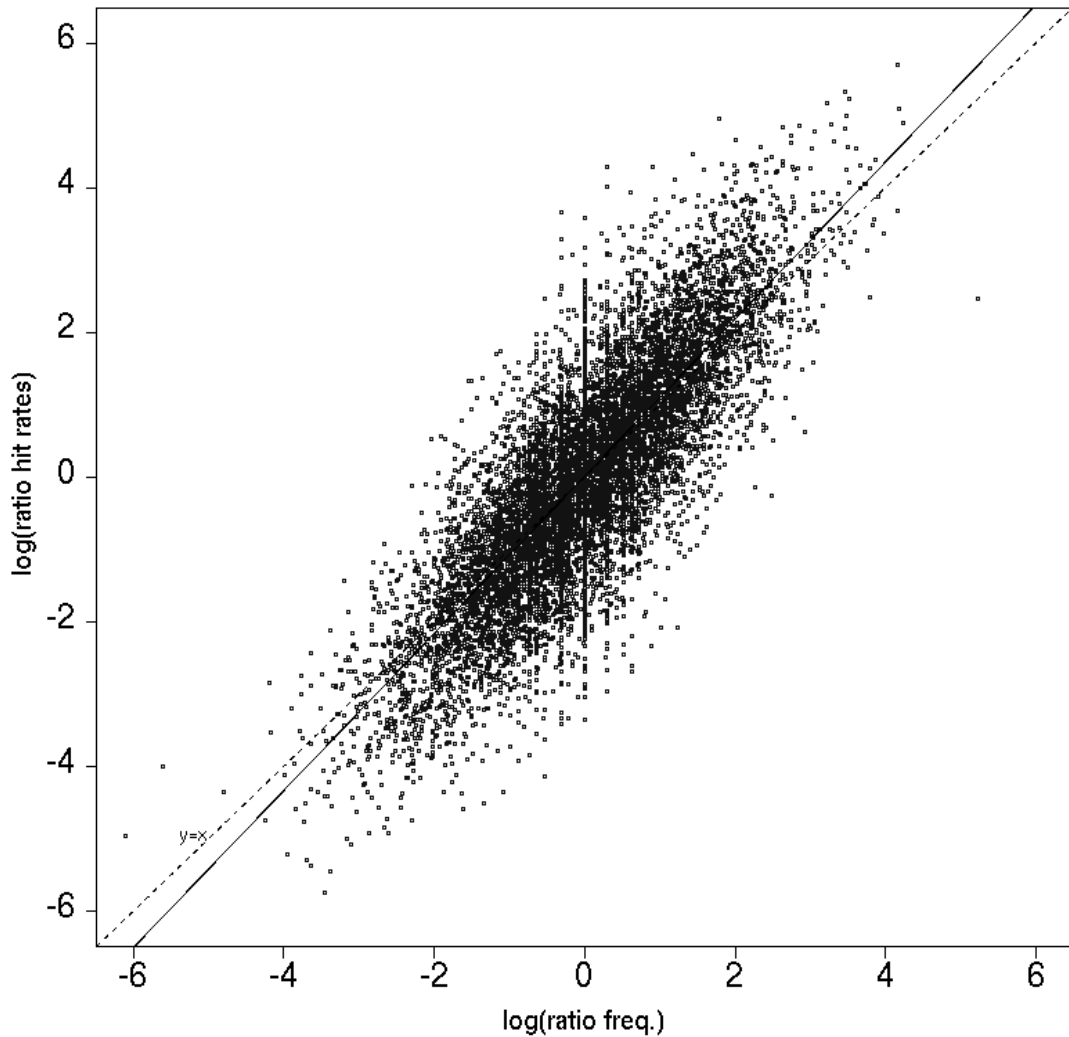
$$\log(\text{hit rate}) = 2.2 + 1.1 \log(\text{freq})$$

This equation particularly applies to words with a text-based frequency of less than 1,000 per million ($\log < 3$). The fact that the slope of the regression line (1.1) does not deviate a lot from 1 is interesting, because it means that after subtracting 2.2, the $\log(\text{hit rate})$ can be used as a rough approximation of the $\log(\text{frequency})$.

For psycholinguistic experiments, researchers are more interested in the relative position of items on the frequency continuum rather than in the absolute counts. They typically want to compare high-frequency words with low-frequency words. How good do both information sources compare in this respect? One way to test this, is to look at the frequency ratios of two randomly chosen words. If the first word is 100 times more frequent than the second on the text-based measure (i.e., $\log(\text{freq}_1 / \text{freq}_2) = 2$), then it should also be more or less 100 times more frequent on the web-based measure (i.e., $\log = 2$ as well). Similarly, if the first word is 100 times less frequent than the second on the text-based measure (i.e., $\log(\text{freq}_1 / \text{freq}_2) = -2$), then it should also be more or less 100 times less frequent on the web-based measure (i.e., $\log = -2$). In other words, we expect a one-to-one relationship between $\log(\text{freq}_1 / \text{freq}_2)$ based on the corpus of texts and $\log(\text{freq}_1 / \text{freq}_2)$ based on the web hit rates. Figure 2 shows that this is indeed the case for 10,000 randomly chosen word pairs, indicating that the relative frequencies are comparable. At the same time, the figure also gives us an idea of the divergences that can be found. If the frequencies of two words are the same in the text (i.e., $\log(\text{freq}_1 / \text{freq}_2) = 0$), on the web the frequency of the first word can vary from 100 times more frequent than the second word ($\log = 2$) to 100 times less frequent than the second word ($\log = -2$).

A New French Lexical Database

Figure 2 Ratios of frequencies and ratios of hit rates for a random sample of 10000 pairs of words. The solid line is the linear regression line.



All in all, we hope to have shown that although text-based word frequencies and web-based word frequencies in general yield comparable estimates of the familiarity of a word, there are some quite strong divergences. Needless to say, such divergences offer interesting opportunities for experimental psychologists. Do word processing times for university undergraduates (who are the usual participants in this type of experiments) agree more with the web frequencies than with the text frequencies? This can easily be checked by selecting four groups of stimuli for which text-based frequencies and web-based frequencies have been selected orthogonally.

The GNU License

The GNU^{vi} project is an effort by the Free Software Foundation (FSF) to make all the traditional UNIX utilities free for whoever wants to use them. These programs are not only free but they are also distributed with their source code under the "GNU general public license". This means that everybody can use, copy, modify, and re-

A New French Lexical Database

distribute the software, as long as the new version is distributed under the same license. This policy has led to the development of very good software able to compete with the best commercial products. Some successful examples of free software are the script languages *Php*, *Perl*, and *Awk*, the internet browser *Mozilla*, and the Office suite *Open Office*.

Lexique is distributed under a license inspired by the GNU general public license. We chose this license in order to guarantee that future *Lexique* versions will remain free, and to encourage people to contribute to future versions of *Lexique*. For the moment, the most essential contribution has been the corrections to the phonological codes made by Peereman and Dufour (2003). We hope that other contributions will follow in the future of *Lexique*.

This license also has the advantage to guarantee the continuity of *Lexique*. For example the famous database *Celex* (Baayen et al., 1993) available for English, Dutch and German has been distributed under a proprietary license. Now that funding has run out, *Celex* developments completely stopped. This should never be a problem for *Lexique* as any laboratory or person will be able to download the database, adapt it, and distribute it on their own website. This should allow *Lexique* or a derived database to live for a long time.

The Web Site

Once *Lexique* was created we wanted it to be useful for other people. Therefore, we created a web site available at <http://www.lexique.org> which consists of several sections.

Given that psycholinguistics is a very large domain and that we cannot be specialists in every aspect of it, we encourage people to contribute to *Lexique*. For this reason we made available a forum where people can ask questions, propose new features, make criticisms, etc. The website also contains a "news section" presenting announcements about *Lexique*. A hierarchical list of links presenting psycholinguistic resources is also available. Users can suggest their own links.

Classical sections as downloading, documentation, and description also appear. In the download section, you can find new databases that we made in addition to *Lexique* such as *Voisins* (which is about orthographic neighbours, see the description below) or *Frequences Frantext* (which allows users to have an overview of all the occurrences of words in *Frantext* and their frequencies (useful for the frequencies of first names for example).

Interrogating *Lexique*

There are two ways to use *Lexique*. The first is to download the database in text format (iso-8859-1) and to use a database program (for instance *Access* or *Visual Foxpro*) or some text manipulation programs (for instance *Gawk* or *Perl*). The second is to interrogate *Lexique* using the online research tools. These online research tools use *Open Lexique*, which is presented below.

Open Lexique

A problem that can arise when you constitute a database is that you would like it to be the richest possible. For this reason, there is a temptation to have an ever increasing number of fields in the database. The problem when you add too many fields, however, is that the database will become bigger and bigger and thus take more time to download, interrogate, view, or correct. This rapidly becomes a problem when you want to update the database regularly.

To solve this problem, we created *Open Lexique*: an online search engine developed in *Php* that allows users to interrogate several databases simultaneously. When we copy a new database to our server, *Open Lexique* automatically generates the web pages that are needed to interrogate this new database simultaneously with the old ones.

We give two examples to illustrate this. The first one concerns the orthographic neighbourhood. An

A New French Lexical Database

orthographic neighbour is operationally defined as a word sharing all but one letter while respecting letter position (according to Coltheart, Davelaar, Jonasson, & Besner, 1997). For instance, "roof" and "moot" are two neighbours of root). As a matter of fact, neighbours are often words with very low frequencies and researchers do not necessarily want these very rare neighbours to be included in the number of neighbours. Therefore, they need to know what the neighbours are and which frequencies are associated with these words.

Unfortunately, adding such information to *Graphemes* would make the database a lot heavier than it is. We can also imagine that in the future researchers will be interested in neighbours not only defined by substitution but also by addition or deletion of a letter (see e.g., De Moor & Brysbaert, 2000). The number of manipulations potentially interesting is unlimited and all this information cannot be placed in *Graphemes*. That's why we created *Open Lexique*. In order to be able to study characteristics of the neighbourhood family, we developed a new database called *Voisins* presenting for each word, its number of orthographic neighbours, the orthographic representations and the frequencies of these neighbours. We copied this database on our server and *Open Lexique* generated the new search engine. Now, users can for example, study what are the neighbourhood characteristics of words having more than 8 letters. Another possibility is to filter out neighbours with a frequency greater than 2 per million for instance.

Another example concerns age of acquisition (AoA). More and more studies have shown an AoA effect independent of frequency. Nevertheless, the first version of *Lexique* did not provide AoA measures. So when Ferrand, Grainger and New (2003) published their database about 400 concrete words and their AoA [in French], we found that it would be very interesting to be able to make a request on this table simultaneously to *Lexique* tables. In order to do that, we copied this new table on our server and we can, now make also request on AoA. For instance, users can select stimuli having an AoA lower than 3 (learnt before 6 years) and having a frequency lower than 10 (see Figure 3). This request will show items acquired early in childhood but having a low frequency for adults. In a similar way, we can imagine other databases that are of interest to people working on particular topics.

Figure 3 Examples of simultaneous request on *Graphemes* and *Brulex*

The image shows a search interface with two sections. The top section is titled "graphemes" and contains three rows of search criteria. The first row has a dropdown menu with "graphemes.frantfreqparm" selected, followed by an equals sign, a dropdown menu with "<" selected, and the value "<10". The second and third rows have dropdown menus with "graphemes.graph" selected, followed by equals signs and dropdown menus with "<" selected, and empty input fields. The bottom section is titled "400AoA" and contains three rows of search criteria. The first row has a dropdown menu with "400AoA.AoA" selected, followed by an equals sign, a dropdown menu with "<" selected, and the value "<3". The second and third rows have dropdown menus with "400AoA.graph" selected, followed by equals signs and dropdown menus with "<" selected, and empty input fields.

For the moment, 8 databases are available in addition to *Graphemes*, *Lemmes* and *Surface: 400 images* (Alario & Ferrand, 1999), *Brulex* (Content et al., 1990), *400 AoA* (Ferrand, et al.), *Voisins*, *Manulex Wordforms and Lemmas* (Lété et al., 2004), *Prénoms*, and *Anagrammes*. By combining these databases, users have access to the following properties: the age of acquisition of words, the number of homographs and homophones, the number and the description of anagrams, the grade-level word-frequency, the number of semantic homonyms, the imagery values of words, the neighbourhood size, the frequencies of the neighbours, etc.

Online Research Tools

Online research tools have been developed to facilitate *Lexique* queries while leaving open a large number of possibilities. They are available in French and in English. Two online tools have been created thus far. The first one allows users to ask for characteristics of a given list of words. Thus, users already having a list of words can

A New French Lexical Database

easily find their characteristics. The user selects the databases he wants to work with and then types in or copies the words list before submitting his request. His research will appear in a table that can easily be copied and pasted in a spreadsheet. Figure 4 illustrates such use.

Figure 4 Example of a request using the list of words search engine

The second search engine is complementary to the first one: it permits to find a list of words with certain characteristics. This is particularly useful when users want to select materials for an experiment. In a first time, the user selects one or several databases he wants to work with. He then accesses a second web page where he can choose the fields on which he wants a query, and he types in his request. Two types of queries exist: simple ones and those based on regular expressions.

Simple operators are presented in Table 3. They permit to make the most often used queries as "begin with", "end with", "greater than" and "lower than".

Table 3: Operators and their meanings for simple requests

Symbol	Meaning	Example	Result
*	A string or characters (in the following example it is used to request "any word beginning with an a")	a*	arbre, arbuste
.	A single character	a.o	ado, abo
<1	Lower than	<10	Words having frequency lower than 10
>1	Greater than	>30	Words having frequency greater than 30
=1	Equal to	=10	Words having frequency equal to 10
< >1 or > <1	Lower than AND upper than	<30 >10	Words having frequency lower than 30 and greater than 10

A New French Lexical Database

The second set of operators that can be used are the Regular Expressions. Regular expressions enable the user make very detailed requests. All the operators that you can use in a "Regular Expression" query are presented in Table 4.

Table 4: Operators and their meanings for "regular expressions" requests

Symbole	Meaning	Example	Result
^	Begin with	^a	arbre, arbuste
\$	End with	e\$	tente, mare
.	Any character	^a.e\$	arme, acte
[xyz]	Characters x,y or z	a[bc]	raccroché, abruti
[x-z]	All the characters from x to z	a[l-n]	amener, alourdi, anneau
[^xyz]	All the characters except xyz	[^aeiouéèïê]	All consonnants
*	Matches the preceding element zero or more times	m*	emmener, amender, entasser
+	Matches the preceding element one or more times	m+	emmener, amender
?	Matches the preceding element zero or one time	m?	amender, entasser
	Or	(buv parl)ant	buvant, parlant
{n}	Matches the preceding element n times	n{2}	patronne but not patron

Once the expression written, you can choose if you want to display items matching with it or items not matching with it, which field you want to display, and by which one you want your result to be sorted. An example of such a request is presented in Figure 5. This request uses Regular Expressions and asks for all the words beginning with an *a* followed with an *f* or a *g*, being either an adjective or a noun having a frequency greater than 10 occurrences per million and having a phonetic representation containing the fricative /f/. This request also specifies that results should be sorted according to their frequencies and that only four fields should be displayed. (the word, its phonetic representation, its grammatical category and its frequency).

Figure 5 Example of a request by properties on Graphemes

A New French Lexical Database

Simple Request
 Regular Expressions

graphemes

graphemes.graph	=	^a[fg].*
graphemes.cgram	=	NOM ADJ
graphemes.frantfreqparm	=	>10
graphemes.phon	=	.*f.*

Sort by following field graphemes.frantfreqparm Order Ascending

Display the following fields:

graphemes.graph	graphemes.cgram	graphemes.frantfreqparm
graphemes.phon	Non specified	Non specified

Display results per page

The number of results as well as the different entries are displayed in a table that can be copied and pasted in a spreadsheet. Because of necessary resource limitations, requests are limited to 2000 rows. If there are more results than displayed, users are able to navigate from one page to another. Figure 6 present the results of the Figure 5 request.

Figure 6 Results of the request presented in Figure 5

Result of the request on "graphemes"

0 - 5 results on a total of **5** words corresponding to your request

graph	cgram	frantfreqparm	phon
affirmation	NOM	12.32	afiRmasj§
affreux	ADJ	14.97	afR2
affection	NOM	23.87	afEksj§
affaires	NOM;VER:ind;pr;sub:pr	96.90	afER
affaire	NOM;VER:ind;pr;sub:pr	106.90	afER

Updates

Since the first public release in October 2000, the users community has steadily grown. Today our website sees an average of 40 different visitors per day. Since this first version, the database, its website, the online and the offline tools have been updated very regularly .

Conclusion

A New French Lexical Database

Lexique provides one of the most complete and richest databases available for the French language. This new database will be particularly interesting for researchers in psycholinguistics, in natural language processing and in linguistics.

Lexique also provides a set of interesting features in the domain of psycholinguistic resources. *Lexique's* frequencies are based on two sources: the very rich corpus of texts *Frantext* that has been developed by the ATILF, and the number of web pages containing a particular word. The corpus of texts includes 487 books published after 1950 which constitutes a total of 31 million words. *Lexique* also brings a lot of details about inflected forms that weren't available before for French. These new data are very important because they permit to study a new range of phenomena which were not possible to study before. For instance, the new features have allowed us to compare processing of French and English plurals (New, Brysbaert, Segui, Ferrand, & Rastle, in press). *Lexique* is also particular in the way it has been developed. For most psycholinguistic resources, once a public version is released, this version will be updated one or two times and then left alone. *Lexique*, being distributed under a GNU-like license permits each person who wants to participate to its development or to create a new derived base to do so. This should permit *Lexique* to live, to be corrected continuously, and to become richer. This dynamic process is encouraged by the presence of a forum where everybody can participate.

The online tools are particularly interesting because they allow to extend *Lexique*. With *Open Lexique*, new databases can be added. Users can then interrogate these new databases simultaneously to the existing ones. For example we already added *Brulex*, several databases giving measures of age of acquisition, and a table describing the orthographic neighbours.

In summary, *Lexique* is a new lexical database for French that has a lot of useful and innovative features. We hope that these features will not only be useful for *Lexique* users but will also be integrated in other projects in French or other languages.

Bibliography

- Alario, F. X., & Ferrand, L. (1999). A set of 400 pictures standardized for French : norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition, *Behavior Research Methods, Instruments, & Computers*, 31, 531-552.
- Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The Celex lexical database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.
- Blair, I., Urland, G., & Ma, J. (2002). Using Internet search engines to estimate word frequency, *Behavior Research Methods, Instruments, et Computers*, 34, 286-290.
- Coltheart, M., Davelaar, E., Jonasson, J. T. & Besner, T. (1977). Access to the internal lexicon. In S. Dornic (Ed.) *Attention and performance VI* (pp. 535-555). Hillsdale, NJ.
- Content, A., Mousty, P., & Radeau, M. (1990). BRULEX : Une base de données lexicales informatisée pour le Français écrit et parlé [A lexical computerized database for written and spoken French], *L'Année Psychologique*, 90, 551-566.
- De-Moor, W., Brysbaert, M. (2000). Neighbourhood-frequency effects when primes and targets are of different lengths. *Psychological-Research*, 63, 159-162
- Ferrand, L., Grainger, J., & New, B. (2003). Normes d'âge d'acquisition pour 400 mots monosyllabiques. [Age-of-acquisition norms for a set of 400 monosyllabic words] *L'Année Psychologique*, 104, 445-468.
- Francis, N., & Kucera, H. (1982). *Frequency analysis of English usage : Lexicon and grammar*. Boston : Houghton Mifflin
- Keller, E., & Zellner, B. (1998). Motivations for the prosodic predictive chain. *Proceedings of ESCA Symposium on Speech Synthesis*. 76, 137-141. (Also available as <http://www.unil.ch/imm/docs/LAIP/pdf.files/KellZell-98-Jenolan.zip>)
- Lambert, E., & Chesnet, D. (2001). Novlex : Une base de données lexicales pour les élèves de primaire. [A lexical database for primary school pupils] *L'Année Psychologique*, 101, 277-288.
- 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, 156-166.
- Monsell, S. (1991). The nature and locus of word frequency effects in reading, in D. Besner et G. Humphreys (Eds.), *Basic processes in reading : Visual word recognition*, (pp. 148-197). Hillsdale, NJ.
- Namer, F. (2000). Flemm : Un analyseur Flexionnel du Français à base de règles, *T.A.L.*, 41, 523-548.
- New, B., Brysbaert, M., Segui, J., Ferrand, L., & Rastle K. (in press) The processing of singular and plural nouns in French and English. *Journal of Memory and Language*.
- New, B., Pallier C., Ferrand L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet : LEXIQUE. [A lexical database on internet about contemporary French : Lexique] *L'Année Psychologique*, 101, 447-462.

A New French Lexical Database

- Pallier, C. (1994). *Rôle de la syllabe dans la perception de la parole : Etudes attentionnelles*. [Syllable role in speech perception] Thèse de doctorat, Ecole des Hautes Etudes en Sciences Sociales, Paris. (Available at <http://www.pallier.org/papers/>)
- Peereman, R., & Dufour, S. (2003). Un Correctif aux Notations Phonétiques de la Base de Données Lexique. [A corrective to the phonetic notations of the Lexique database] *L'Année Psychologique*, 103, 103-108.
- Pythoud, C. (1996). Problèmes de la correction automatique de l'orthographe lexicale du français à travers une étude de cas : le correcteur orthographique ispell et le dictionnaire Français-IREQ, [Automatic spell-checking problems: the ispell program and the Français-IREQ dictionary.] *Mémoire de licence*, Université de Lausanne.
- Robert, P. (1992). *Le Grand Robert version électronique*. Dictionnaires le Robert.
- Schmid, G. (1994). *TreeTagger - a language independent part-of-speech tagger*. Manuscript. (Available at <http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html>)

A New French Lexical Database

Author Note

This research was supported by a post-doctoral grant from the Fondation Fyssen to the first author and a British Academy Grant to the second one. We would like to thank especially Pascale Bernard and the ATILF laboratory for their help. Correspondence should be addressed to Boris New (e-mail: boris.new@univ-paris5.fr).

A New French Lexical Database

Footnotes

ⁱ <http://www.unil.ch/ling/cp/frgut.html>

ⁱⁱ <http://www.ims.uni-stuttgart.de/projekte/complex/TreeTagger/>

ⁱⁱⁱ http://www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.html

^{iv} <http://www.unil.ch/imm/docs/LAIP/LAIPPTS.html>

^v *FastSearch* is available at <http://www.alltheweb.com>

^{vi} Remember that the frequencies of occurrence were based on a corpora of written texts based on 31 millions words, and the hit rates corresponded to the number of hits per million pages returned from a Internet search engine that indexed 15 millions of French web pages.

^{vii} <http://www.gnu.org>

¹ These operators can also be used in a "Regular Expression" request