Behavior Research Methods, Instruments, & Computers
2004, 36 (3), 516-524

Lexique 2 : A new French lexical database

BORIS NEW
Royal Holloway, University of London, London, England
and CNRS and Université René Descartes, Paris, France

CHRISTOPHE PALLIER
INSERM, U562 Service Hospitalier Frédéric Joliot, Paris, France

MARC BRYSBAERT
Royal Holloway, University of London, London, England

and

LUDOVIC FERRAND
CNRS and Université René Descartes, Paris, France

In this article, we present a new lexical database for French: *Lexique*. In addition to classical word information such as gender, number, and grammatical category, *Lexique* 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 into a graphemic table with all the relevant frequencies, a table structured around lemmas (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 metasearch engine, *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 http://www.lexique.org.

Psycholinguistics researchers make extensive use of word databases. These databases are particularly important because they are the foundation of most psycholinguistics 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.

For many years, psycholinguists studying French used *Brulex* (Content, Mousty, & Radeau, 1990). As the first electronic database for psycholinguists, *Brulex* was very helpful despite the following drawbacks. Its frequencies were based on a corpus of texts published between 1919 and 1964. Because frequency is one of the most important and robust factors manipulated in psycholinguistics experiments, it is important to have frequencies as reliable and recent as possible. In this respect, *Brulex* frequencies look rather outdated.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, 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, and Matos (2001), an article that presents the first version of the database in detail. Here, we will mainly focus on the original features that appeared after the first version, such as a GNU-like license, a Web site, and a metasearch engine. These features are particularly interesting because they can be useful for databases in other languages, as well.

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

With the use of *Frantext's* search engine, we obtained a list of 246,000 occurrences and their frequencies. Because these occurrences contained many foreign and proper words, we removed the words not referenced in a

This research was supported by a postdoctoral grant from the Fondation Fyssen to B.N. and a British Academy grant to M.B. We thank Pascale Bernard and the ATILF laboratory for their help. Correspondence should be addressed to B. New, Laboratoire de Psychologie Expérimentale, 71 avenue Edouard Vaillant, F-92100 Boulogne-Billancourt, France (e-mail: boris.new@univ-paris5.fr).

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
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

standard French Dictionary (Robert, 1992), using the *ispell* spelling checker coupled to *Français-Gutenberg* (Pythoud, 1996). For the extraction of morphosyntactic information, two grammatical parsers have been used in addition to *Le Grand Robert: TreeTagger* (Schmid, 1994; available at http://www.ims.uni-stuttgart.de/projekte/ corplex/treetagger/) and *Flemm 2* (Namer, 2000; available at http://www.univ-nancy2.fr/pers/uaner/ Telecharger_ Flemm. html).

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 the inflectional family. Surface displays information about words and their letter, bigram, trigram, phoneme, and syllable frequencies. An independent archive, also named *Surface*, presents detailed statistics about surface frequencies.

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

Because 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 2,500 (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 schwa positions and suppressed the distinction between the two types of /a/, /o/, and /r/. These corrections have been included in *Lexique* 2 and subsequent 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). Since 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 in which 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 in which the word appears as a function of the total number of Web pages. Frequencies based on Web pages are interesting because (1) Web pages are more dynamic than corpora of texts because Web pages are easily published; (2) Web pages exist for nearly all human activities, whereas a corpus is usually limited to literary texts; (3) Web pages are updated very regularly; and (4) Web pages in a particular language constitute a vast corpus.

We chose to use the *Fastsearch* (available at http://www.alltheweb.com) search engine based on 15 million French Web pages for the following reasons. First, this search engine gives the precise number of pages in which the word is found, whereas Google gives only approximations. Second, *Fastsearch* differentiated (although this no longer seems to be true) between accentuated and nonaccentuated characters. We did our research with the *SafeSearch* mode to prevent overestimates of words with sexual connotations.

Recently, Blair, Urland, 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 (Baayen, Piepenbrock, and van Rijn, 1993; Francis and Kučera, 1982). 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 substantially.

Table 2
***Lemmes* Fields and Their Descriptions**

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

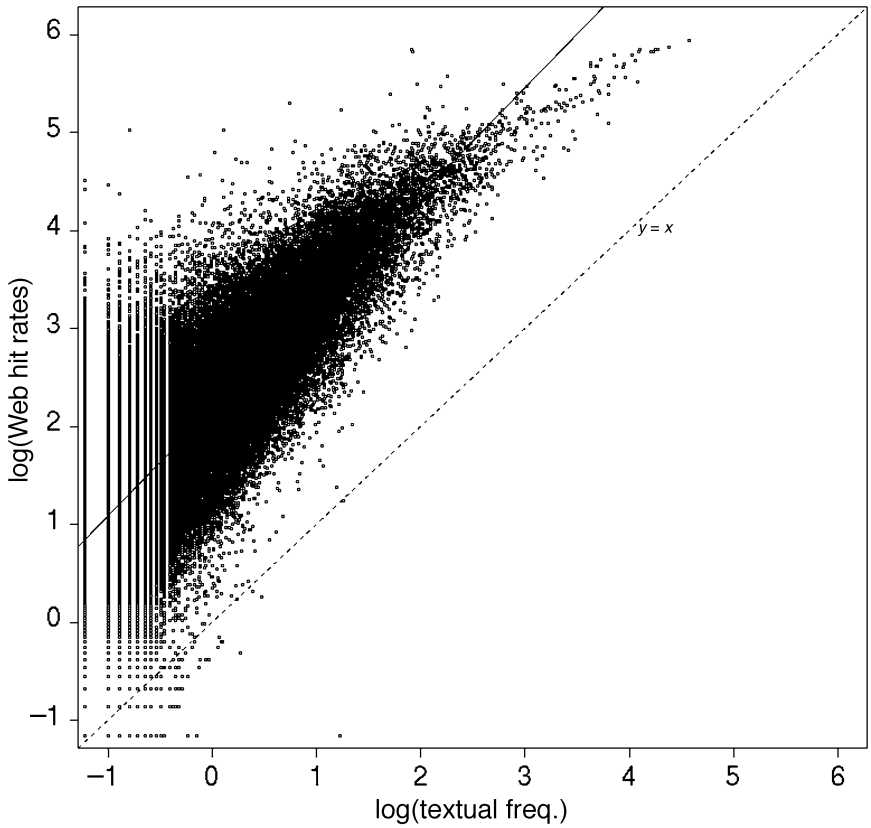


Figure 1. Relationship between text-based frequencies and Web-based hit rates, both expressed per million, and shown on logarithmic scales. The solid line is the linear regression line.

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 those given by corpora of texts.

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%, whereas their lexical frequencies are but a few percent. In such cases, then, hit rates overestimate the frequency of usage. On the other hand, consider a very low-frequency word, used only in certain contexts: It will occur in only a few Web pages, but when it is used, it is likely to appear several times on the page, a fact not considered 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.[1] All frequencies are expressed in occurrences per million (words for *Frantext* and Web pages for *Fastsearch* frequencies).

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, especially for the low-frequency words. In addition, there is considerable 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 ranging 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 following equation:

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

This equation applies particularly 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 much from 1 is interesting, because it means that after subtraction of 2.2, the log(hit rate) can be used as a rough approximation of the log(frequency).

For psycholinguistics experiments, researchers are more interested in the relative position of items on the frequency continuum than in the absolute counts. They

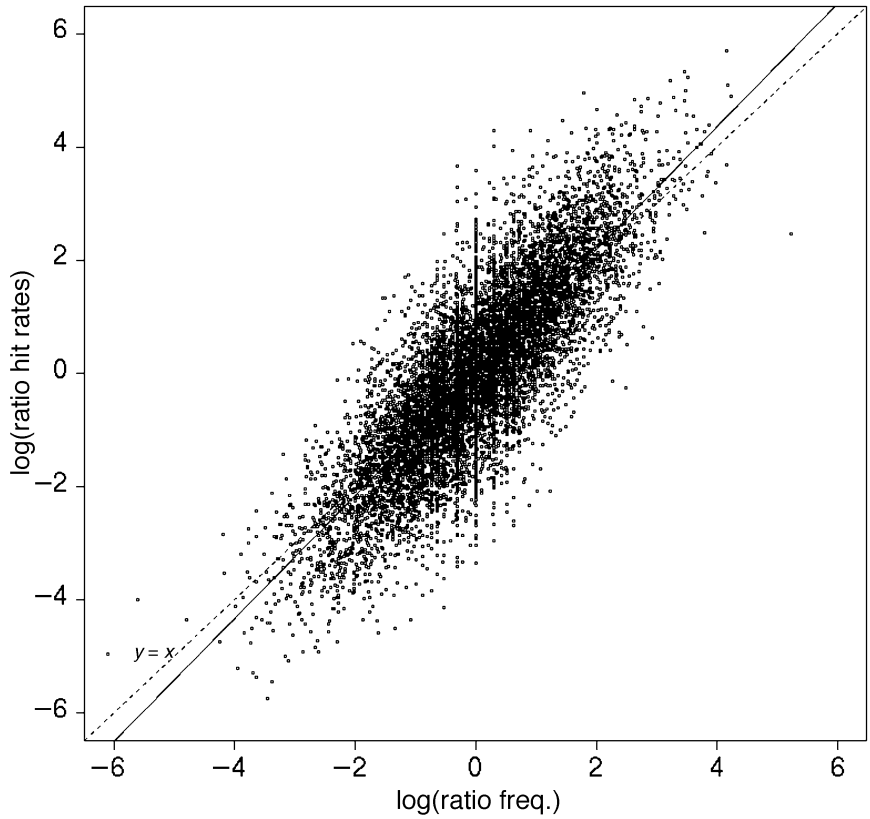


Figure 2. Ratios of frequencies and ratios of hit rates for a random sample of 10,000 pairs of words. The solid line is the linear regression line.

typically want to compare high-frequency words with low-frequency words. But one might ask the question, how well do the text-based and Web-based 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 approximately 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., $\log(\text{freq}_1/\text{freq}_2) = -2$], it should also be approximately 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 shows 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).

All in all, we hope to have demonstrated 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 college undergraduates (who are the usual participants in this type of experiment) 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 (available at http://www.gnu.org) 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 redistribute the software, as long as the new version is distributed under the same license. This policy has lead 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

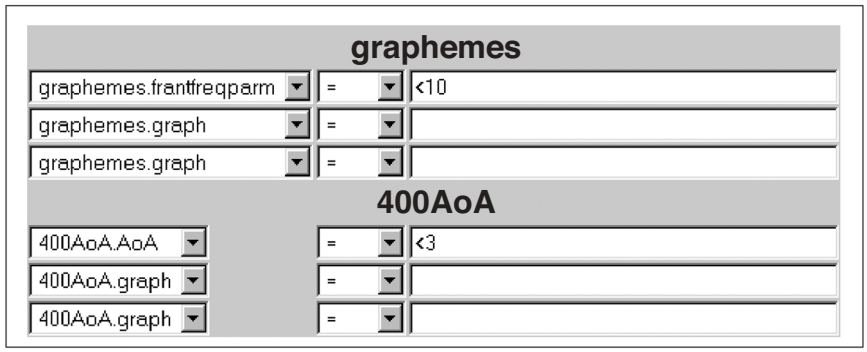


Figure 3. Examples of simultaneous request on *Graphemes* and *Brulex*.

to guarantee that future versions of *Lexique* will remain free and to encourage people to contribute. 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.

This license also has the advantage of guaranteeing 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. Since funding ran out, *Celex* developments have completely stopped. This should never be a problem for *Lexique* because any institution or individual will be able to download the database, adapt it, and distribute it on their own Web sites. 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, and so forth. The Web site also contains a "news section" presenting announcements about *Lexique*. A hierarchical list of links presenting psycholinguistics resources is also available, and users can suggest their own links.

Lexique contains standard sections such as downloading, documentation, and description. In the download section, one can find new databases that we made, such as *Voisins* (which is about orthographic neighbors; 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* with on-line research tools, using *Open Lexique*, which is presented below.

Open *Lexique*

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

To solve this problem, we created *Open Lexique*, an on-line 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 along with the old ones.

We give two examples to illustrate this. The first concerns the orthographic neighborhood. An orthographic neighbor is operationally defined as a word sharing all but one letter while respecting letter position (see Coltheart, Davelaar, Jonasson, & Besner, 1997). For instance,

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

Table 3
Operators and Their Meanings for Simple Requests

Symbol	Meaning	Example	Result
*	a string of 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 greater than	<30 >10	words having frequency lower than 30 and greater than 10

roof and *moot* are two neighbors of *root*). As a matter of fact, neighbors are often words with very low frequencies. Researchers do not necessarily want these very rare words to be included in the number of neighbors that they manipulate. Therefore, they need to know what the neighbors are, as well as the frequencies associated with these words.

Unfortunately, adding such information to *Graphemes* would make the database too heavy to be handled easily. We can also imagine that future researchers will be interested in neighbors defined not only by substitution but by addition or deletion of a letter (see, e.g., De Moor & Brysbaert, 2000). The number of potentially interesting manipulations is unlimited, and all this information cannot be placed in *Graphemes*, so we created *Open Lexique*. In order to be able to study characteristics of the neighborhood family, we developed a new database called *Voisins* presenting each word's number of orthographic neighbors, the orthographic representations, and the frequencies of these neighbors. We copied this database on our server, and *Open Lexique* generated the new search engine. Now, users can, for example, study the neighborhood characteristics of words having more than eight letters. Another possibility is to filter out neighbors with a frequency greater than two 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 of 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 with *Lexique* tables. In order to do that, we copied this new table on our server and we can now also make requests on AoA. For instance, users can select stimuli having an AoA lower than 3 (learned before age 6) 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.

For the moment, eight databases are available in addition to *Graphemes*, *Lemmes*, and *Surface*: *400 Images* (Alario & Ferrand, 1999); *Brulex* (Content et al., 1990);

Table 4
Operators and Their Meanings for Regular Expressions Requests

Symbol	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	[^aeiyouéèïê]	all consonants
*	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

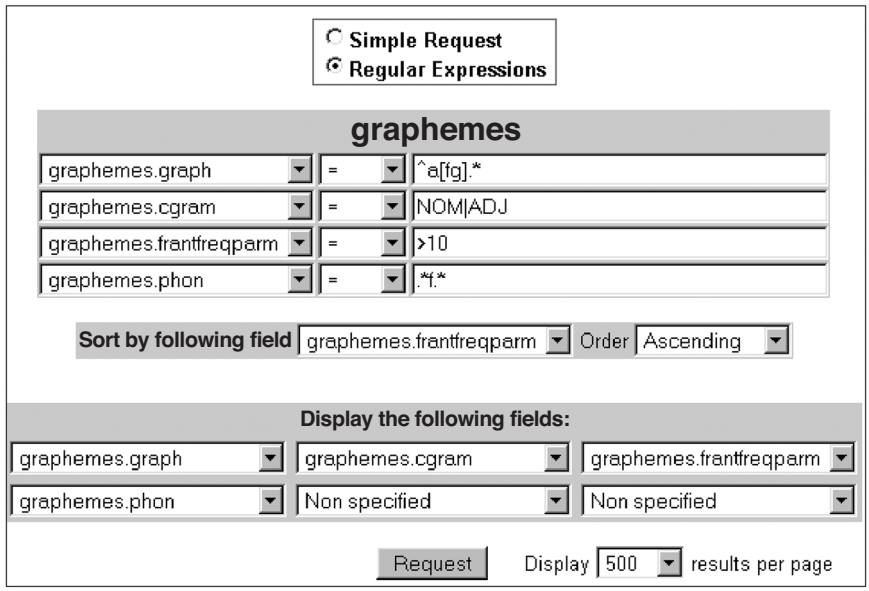


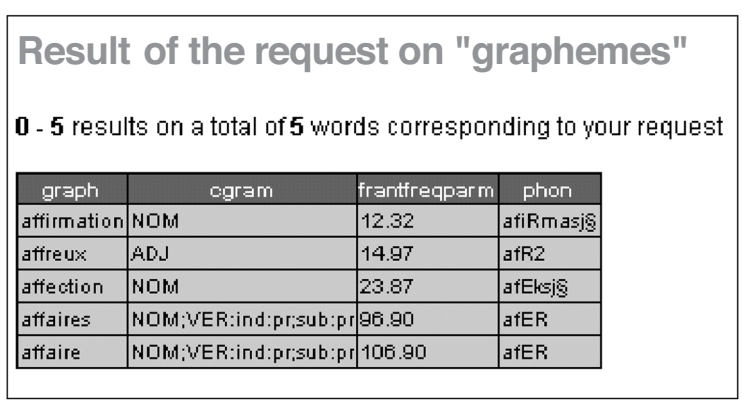
Figure 5. Example of a request by properties on *Graphemes*.

400 AoA (Ferrand et al., 2003); *Voisins*; *Manulex Word-forms*; and *Lemmas* (Lété et al., 2004); *Prénoms*; and *Anagrammes*. By combining these databases, users have access to the following properties: AoA of words, number of homographs and homophones, number and description of anagrams, grade-level word frequency, number of semantic homonyms, imagery values of words, neighborhood size, frequencies of neighbors, and so forth.

On-line Research Tools

French and English on-line research tools have been developed to facilitate *Lexique* queries while leaving open a large number of possibilities. Two on-line tools have been created thus far. The first allows users to ask for characteristics of a given list of words. Thus, users already having a list of words can easily find their characteristics. Users select the databases they want to work with and then type in or copy the word list before submitting their request. Their research will appear in a table that can easily be copied and pasted in a spreadsheet. Figure 4 illustrates such use.

The second search engine is complementary to the first: It permits users to find a list of words with certain characteristics. This is particularly useful when users want to select materials for an experiment. Initially, users select one or several databases they want to work with. They then access a second Web page where they can choose the fields on which they want a query, typing in their request. Two types of queries exist: Simple Requests and Regular Expressions.



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

Figure 6. Results of the request presented in Figure 5.

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

The second set of operators that can be used are the Regular Expressions, which enable users to make very detailed requests. All the operators that one can use in a Regular Expression query are presented in Table 4.

Once the expression is written, users can choose if they want to display items matching with it or items not matching it, which field they want their display, and by which one they want their 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 by an *f* or a *g*, being either an adjective or a noun with a frequency greater than 10 occurrences per million and with 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).

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 2,000 rows. If there are more results than can be displayed, users can navigate from one page to another. Figure 6 presents the results of the Figure 5 request.

Updates

Since the first public release of *Lexique* in October 2000, the community of users has steadily grown. Today, our Web site sees an average of 40 different visitors per day. The database, its Web site, and the on-line and offline tools have been updated regularly.

Conclusion

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

Lexique also provides a set of interesting features in the domain of psycholinguistics resources. *Lexique*'s frequencies are based on two sources: *Frantext*, the rich corpus of texts 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 wealth of details about inflected forms previously unavailable for French. These new data are very important because they permit users to study a new range of phenomena that could not be studied 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 psycholinguistics resources, once a public version is released, this version is updated once or twice and then left alone. *Lexique*, which is distributed under a GNU-like license, permits each person who wants to participate in its development or to create a new derived base to do so. This should permit *Lexique* evolve, to be corrected continuously, and to become of ever greater value. This dynamic process is encouraged by the presence of a forum in which everyone can participate.

The on-line tools are particularly interesting because they allow users to extend *Lexique*. With *Open Lexique*, new databases can be added. Users can then interrogate these new databases simultaneously with existing ones. For example, we have already added *Brulex*, several databases giving measures of AoA, and a table describing orthographic neighbors.

In summary, *Lexique* is a new lexical database for French that has many 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.

REFERENCES

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)*. University of Pennsylvania, Philadelphia: Linguistic Data Consortium.

Blair, I. V., Urland, G. R., & Ma, J. E. (2002). Using Internet search engines to estimate word frequency. *Behavior Research Methods, Instruments, & 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: Erlbaum.

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). Neighborhood-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., & Kučera, 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.

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 & G. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 148-197). Hillsdale, NJ: Erlbaum.

Namer, F. (2000). Flemm: Un analyseur flexionnel du Français à base de règles [Flemm: Inflectional analyzer for French with rules]. *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 & 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 the Internet about contemporary French: Lexique]. *L'Année Psychologique*, **101**, 447-462.

Pallier, C. (1994). *Rôle de la syllabe dans la perception de la parole: Études attentionelles* [Syllable role in speech perception]. Thèse de doctorat. Paris: École des Hautes Études en Sciences Sociales. (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'orthographie 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 French–IREQ dictionary] available at http://www.vuil.ch/ling/frgvt.html. *Mémoire de licence*, Université de Lausanne.

Robert, P. (1992). *Le Grand Robert version électronique*. Paris: Dictionnaires le Robert.

Schmid, G. (1994). *TreeTagger—A language-independent part-of-speech tagger*. Available at http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html.

NOTES

1. Remember that the frequencies of occurrence were based on a corpora of written texts based on 31 million words, and the hit rates corresponded to the number of hits per million pages returned from an Internet search engine that indexed 15 million French Web pages.

2. These operators can also be used in a Regular Expression request.

(Manuscript received July 23, 2003;
accepted for publication May 16, 2004.)