Une ensemble de ressources concernant la langue française sur internet : LEXIQUE ™

Document version 2.64

Boris New[1], Christophe Pallier[2]

[1]Laboratoire de Psychologie expérimentale
UMR 8581 CNRS, Université René Descartes, Paris V
71, avenue Edouard Vaillant, 92774 Boulogne Billancourt Cedex, France

[2]Laboratoire de Sciences Cognitives et Psycholinguistique,
UMR 8554, CNRS, Ecole des Hautes Etudes en Sciences Sociales (EHESS),
54 Boulevard Raspail,75270 Paris CEDEX 06,

E-mail :new@psycho.univ-paris5.fr

Remerciements: Nous tenons à remercier Pascale Bernard de l'Inalf pour ses précieux renseignements, ainsi que Ray Sydney et l'équipe de FastSearch pour leurs moteurs de recherche Internet, Helmut Schmid pour son excellent lemmatiseur et Sid Kouider pour son aide et son programme permettant le calcul des voisins.

Mots clés : Reconnaissance de mots, Fréquence, Base de donnée

Historique de cette documentation

2.62c Correction des champs frantfreqparm qui devient freqfrant et fsfreqparm qui devient freqweb
2.62b Noms des champs pour toutes les bases d'Open Lexique
Tableau avec la répartition du nombre de mots de Graphemes par nombre de syllabes et nombre de lettres
Histogramme des fréquences
2.62a Correction des catégories grammaticales
Précisions sur le champs « Syllabation »
2.62 Corrections de mise en formes (liens hypertextes, moins de notes de bas de page, en-têtes corrigés)
2.61 Introduction rapide pour le nouveau venu
Correction de la description de BigrMoy.txt selon Surface 2.10
Rajout de la table des tableaux (permettant un accès direct aux codes grammaticaux ou phonologiques)
Petites corrections pour les outils hors-ligne (Undows)
Création de la section « Les autres bases »
2.60 Ajout d'une description précise du point d'unicité phonologique et orthographique dans Lexique 2.60
2.50 Ajout de la description des champs concernant le nombre de voisins (orthographiques et phonologiques) et des représentations inversées
Amélioration de la description de la table Surface.txt
2.02 Correction mineure du tableau présentant les formes phonologiques
Ajout de cet historique
2.01 Corrections du tableau présentant les formes phonologiques
2.00 Ajout/Modifications des sections: -2.2 Fréquences à partir des pages web
-3.2 Acquisition de la forme phonologique
-4.4 Organisation du dossier Surface
6. Licence
7 Les outils
1.00 Naissance de cette documentation

Introduction rapide pour le nouveau venu

Si vous cherchez une information particulière et ne connaissez rien à Lexique, nous vous conseillons de procéder de la façon suivante :

- lisez ce manuel (dans les grandes lignes) afin de
    - déterminer dans quelle base se trouve l'information que vous cherchez (le plus souvent c'est la base Graphemes)
    - comprendre comment cette base est structurée (quel sont le ou les champs dont vous avez besoin)
    - déterminer quelle recherche vous allez utiliser (online ou offline). Essayez d'abord la recherche online et si vous ne pouvez utiliser celle-ci pour avoir l'information qui vous intéresse, essayez alors l'interrogation offline. (Undows)

Si vous avez un problème, faites d'abord une recherche sur le forum. Si vous ne trouvez pas de réponse à votre question, n'hésitez pas à la poster.

TABLE DES MATIERES

TABLE DES TABLEAUX

TABLE DES FIGURES

Une base de données lexicales pour la langue française: Lexique 2

Pendant longtemps, les psycholinguistes ont sélectionné manuellement le matériel verbal dans le *Trésor de la Langue Française* (Imbs, 1971). Leur travail a été grandement facilité quand Content, Mousty et Radeau (1990) ont mis à leur disposition *Brulex*, une base de données informatisée regroupant les 35 746 entrées lexicales du *Petit Robert* et leurs fréquences selon le *TLF*. Ces fréquences étaient estimées sur un corpus de textes littéraires datant de 1919 à 1964 et comprenant 26 millions de mots. Une limitation notable de *Brulex* était l'absence des formes fléchies telles que les verbes conjugués ou certaines formes écrites plurielles ou féminines. Cela pose problème par exemple pour toutes les études concernant les formes fléchies en français ou pour estimer des fréquences d'unités telles que les syllabes. *NOVLEX*, une base de données plus récente (Lambert et Chesnet, 2001), fournit les formes fléchies mais se fonde sur un corpus spécialisé de textes pour enfants de 417 000 mots. C'est pourquoi nous avons entrepris de construire une nouvelle base de données avec des estimations de fréquences plus complètes, plus actuelles, et comprenant les formes fléchies.

1 Description du corpus original

Afin de constituer la base initiale de mots, nous avons sélectionné dans la base *Frantext* les textes publiés entre 1950 et 2000 : cela représentait un corpus de 31 millions d'items. *Frantext* est une base de données textuelles regroupant 3 200 textes représentatifs du français des 19e et 20e siècle, développée par l'INALF-Nancy, devenu aujourd'hui l'ATILF. Ces textes étaient essentiellement des romans, mais comprenaient également quelques recueils de poésie, des essais et des traités scientifiques ou techniques. Nous avons obtenu une liste de 246 000 items distincts ainsi que leur fréquence (Le logiciel d'interrogation ne traitait malheureusement pas correctement les noms composés : un mot comme « garde-manger » était identifié comme deux items distincts « garde » et « manger »). Ces items comprenaient des symboles (dont la ponctuation), des abréviations, des mots étrangers et des noms propres. Pour "nettoyer" cette liste, nous avons employé le dictionnaire *Francais-Gutenberg 1.0* (Pythoud, 1996) et le dictionnaire *Le Grand Robert* (Robert, 1996). Le résultat de ce filtrage a produit une liste de 130 000 items ayant des formes orthographiques distinctes.

2 Calcul des fréquences

2.1 Fréquences à partir d'un corpus de textes

La fréquence des mots joue un rôle fondamental dans la plupart des tâches psycholinguistiques (voir Monsell, 1991 pour une synthèse). De nombreuses études ont montré que les performances étaient meilleures pour les

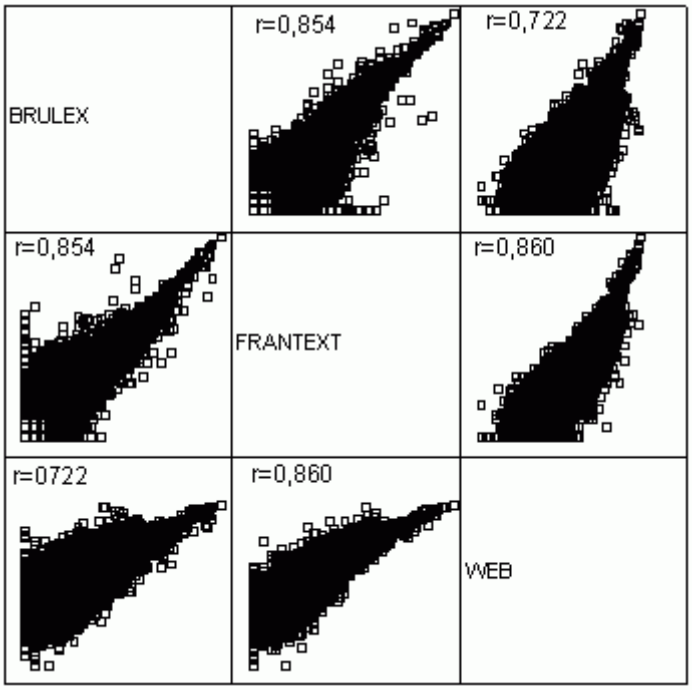
mots de haute fréquence que pour les mots de basse fréquence, que cela soit en terme de nombre d'erreurs ou de temps de réaction. Cependant, d'autres facteurs comme l'âge d'acquisition, ou la familiarité, généralement très corrélés avec la fréquence d'usage, interviennent (Morrison et Ellis, 1995 ; Connine, Mullenix, Shernoff et Yelen, 1990). Pour décorréler ces différents facteurs, il est primordial d'avoir de bonnes estimations de chacun d'entre eux.

Dans *Lexique*, nous proposons deux estimateurs des fréquences d'usage : le premier est fondé sur le corpus initial de *Frantext*, constitué de textes littéraires ; le second repose sur le nombre de pages web françaises contenant un mot donné. Ce deuxième estimateur, fondé sur quinze millions de pages web, nous a paru constituer une source d'information supplémentaire sur l'usage du français.

2.2 Fréquences à partir des pages web

Plus précisément, nous avons soumis au moteur de recherche *FastSearch*, les 130 000 formes orthographiques obtenues à partir du corpus *Frantext*. Nous avons choisi ce moteur en raison de son indication précise du nombre de pages contenant le mot recherché (*Google* p. ex. ne donne que des approximations), et du fait qu'il différencie les caractères accentués des caractères non accentués. En revanche, il n'effectue pas de différenciation entre les majuscules et les minuscules tout comme nos fréquences basées sur *Frantext*. L'interrogation était effectuée sur les 15 millions de pages françaises répertoriées, en mode *SafeSearch* pour éviter la sur-représentation des mots à connotation sexuelle. Pour chaque mot a été obtenu le nombre de pages dans lesquelles celui-ci apparaissait ; il ne s'agit donc pas exactement de la fréquence lexicale de la forme, mais néanmoins d'un estimateur de l'usage de ce mot. Par exemple, des mots tels que *publicité*, *entreprise* ou *télévision* se retrouvent avec des fréquences comparables à celles de mots tels que *champ*, *arbre* ou *chaise* selon *FastSearch*, mais avec des fréquences très divergentes selon *Frantext*. D'autres items tels que *kiwi* sont extrêmement rares selon *Brulex* ou *Frantext* alors que *FastSearch* les considère, de façon plus réaliste, comme "plutôt rares". Pour comparer ces deux estimations de fréquence entre elles et par rapport aux fréquences du TLF, nous avons construit le diagramme de corrélation de la Figure 1 à partir du logarithme des fréquences de 23 440 items selon le *TLF*, *Frantext* et *FastSearch*.

Figure 1 Nuages de points présentant les corrélations entre les fréquences basées sur *Brulex*, *Frantext* et les fréquences basées sur le web



Plus récemment, Blair, Urland et Ma (2002) ont effectué une comparaison sur 400 mots anglais entre les fréquences obtenues en nombre de hits de 4 moteurs de recherche (*AltaVista*, *Northern Light*, *Excite* et *Yahoo!*), et les fréquences fondées sur des bases de textes (Francis et Kucera, 1982; et Baayen, Piepenbrock, et van Rijn, 1993). Ces auteurs observent une forte corrélation entre les différents moteurs et les bases de textes. Le web étant plus versatile que les bases de textes, ils vont aussi interroger à nouveau ces moteurs 6 mois plus tard et constater que les fréquences n'ont pas significativement changé. En revanche, ils constatent une corrélation moyenne (entre 0,45 et 0,49) entre ces différentes bases et l'indice de familiarité donné par les sujets. Cette corrélation est tout aussi importante pour les fréquences estimées par les moteurs de recherche que pour celles données par les bases de textes.

Ils en concluent que même si l'indice fréquentiel (nombre de pages contenant ce mot) donné par les moteurs de recherche n'est pas le même que celui donné par les bases de textes (nombre de mots apparaissant dans le corpus), cet indice semble tout aussi représentatif que celui donné par les corpus de textes.

3 Obtention des autres descripteurs

3.1 Catégorie grammaticale, genre et nombre

Pour obtenir la catégorie grammaticale, le genre, le nombre et le lemme des mots (un lemme est le mot choisi pour représenter toute une famille de formes apparentées. Par exemple: *manger* est le lemme de *mangea*, *mangeait*, …etc.), nous avons utilisé conjointement le *Grand Robert*, et les deux lemmatiseurs: *Tree Tagger* de

Helmut Schmid et *Flemm 2.0* (Namer, 2000). En effet, aucune de ces sources seules ne permettait d'avoir une information suffisamment complète.

3.2 Acquisition de la forme phonologique

Dans une troisième étape, nous avons dérivé la forme phonologique de nos entrées grâce au logiciel *LAIPTTS 1.13*. Ce logiciel utilise un noyau de 500 règles de conversion graphème-phonème rendant compte de plus de 86% des prononciations. Afin de traiter les exceptions, il dispose aussi d'un dictionnaire composé de 6 000 mots ayant des prononciations exceptionnelles. Sur 4 000 phrases du quotidien *Le Monde*, l'auteur rapporte que son logiciel a un taux d'erreur de 0,001 %.

Or ce logiciel (*LAIPTTS*) était un logiciel prévu pour générer de la parole à partir de textes continus et non de mots isolés. Peereman et Dufour (sous presse) ont examiné, une fois la première version de *Lexique* rendue publique, les codes phonémiques donnés par *Lexique* en les comparant aux notations phonémiques données par *Brulex* (elles-mêmes basées sur le dictionnaire *Le Petit Robert*). Ils ont ainsi détecté 2 500 différences (sur les 30 000 entrées que contient *Brulex*) de codifications phonémiques entre *Lexique* et *Brulex*. Ces 2 500 différences relevaient soit de mots à prononciation exceptionnelle, soit de problèmes de règles de conversion utilisées par le logiciel. Ils ont donc corrigé ces entrées. Ils ont aussi retraité l'ensemble des codes phonémiques pour le positionnement des schwas. Afin de rendre les codes phonémiques les plus cohérents possibles, les auteurs de ces corrections ont aussi supprimé la distinction entre les deux types de "a" et les deux types de "o", les deux types de "r", l'arrêt glottique, ainsi que la marque d'aspiration "h".

Le site *http://leadserv.u-bourgogne.fr/bases/lexiquecorr/* met à disposition un document décrivant les corrections réalisées, les scripts de correction utilisés ainsi que l'ensemble des correctifs. Ces corrections ont été intégrées à la version 2 de *Lexique*.

4 Organisation de la base

Etant donné le grand nombre d'informations disponibles, nous avons choisi pour des raisons d'accessibilité et de lisibilité de diviser notre base en trois tables principales :

- ***Graphemes.txt*** : une base organisée à partir des formes orthographiques qui comprend environ 129 000 entrées.

- ***Lemmes.txt*** : une base organisée à partir des lemmes qui comprend environ 54 000 entrées. Nous avons choisi la forme "infinitif" pour les verbes et la forme "masculin singulier" pour les participes passés, adjectifs et noms.

- ***Surface.txt***: une base qui résume les statistiques fréquentielles concernant les lettres, bigrammes, trigrammes, phonèmes et syllabes pour chaque mot. Elle comprend 129 000 entrées tout comme *graphemes.txt*.

Ces tables sont fournies sous forme de fichiers textes, les champs étant séparés par des tabulations. Cela permet de les importer facilement avec la plupart des logiciels. Deux dossiers supplémentaires, *Surface* et *Outils*,

contiennent respectivement des informations fréquentielles détaillées à propos des lettres, bigrammes, trigrammes, phonèmes et syllabes, et des outils facilitant l'utilisation des tables.

4.1 Organisation de la table *Graphemes*

La table *Graphemes* est présentée dans le fichier *graphemes.txt*. C'est la base à partir de laquelle nous avons créé les autres bases (*Lemmes* et *Surface* p.ex.). Nous allons présenter dans cette partie une description des différents champs qui constituent cette base.

La Tableau 1 présente les différents champs de cette table pour quelques items.

Tableau 1 Présentation d'un extrait de *Graphemes.txt*

Légende: **graph**: le mot; **phon**: les formes phonologiques du mot; **cgram**: les catégories grammaticales de ce mot; **genre**: le genre; **nombre**: le nombre; **lemme**: les lemmes de ce mot;

graph	phon	cgram	genre	nombre	lemme	freqfran	freqweb	nblettres	nbphons	cvcv	p_cvcv	puorth	puphon	syll	nbsyll	cv-cv
danse	d@s	NOM;VER:imp:p	f	s;2s;1s;3s	danse;danser	49.71	10745.56	5	3	CVCCV	CVC	5	3	d@s	1	CVC
dansent	d@s	VER:ind:pr;sub:pr		3p	danser	5.29	546.01	7	3	CVCCVCC	CVC	6	3	d@s	1	CVC
danser	d@se	NOM;VER:infi	m	s	danser	21.26	2320.22	6	4	CVCCVC	CVCV	6	4	d@-se	2	CV-CV
dansera	d@s*Ra	VER:ind:futu		3s	danser	0.16	40.91	7	6	CVCCVCV	CVCVCV	7	6	d@-s*-R	3	CV-CV-
danserai	d@s*RE	VER:ind:futu		1s	danser	0.10	10.51	8	6	CVCCVCV	CVCVCV	8	6	d@-s*-R	3	CV-CV-
danseraient	d@sRE	VER:cond:pr		3p	danser	0.13	3.36	11	5	CVCCVCV	CVCCV	9	4	d@-sRE	2	CV-CCV
danserais	d@s*RE	VER:cond:pr		1s;2s	danser	0.06	4.27	9	6	CVCCVCV	CVCVCV	9	6	d@-s*-R	3	CV-CV-
danserait	d@s*RE	VER:cond:pr		3s	danser	0.23	5.88	9	6	CVCCVCV	CVCVCV	9	6	d@-s*-R	3	CV-CV-
danseras	d@s*Ra	VER:ind:futu		2s	danser	0.13	5.95	8	6	CVCCVCV	CVCVCV	8	6	d@-s*-R	3	CV-CV-
danserez	d@s*Re	VER:ind:futu		2p	danser	0.03	9.81	8	6	CVCCVCV	CVCVCV	7	6	d@-s*-R	3	CV-CV-
danserons	d@sR§	VER:ind:futu		1p	danser	0.13	12.26	9	5	CVCCVCV	CVCCV	9	5	d@-sR§	2	CV-CCV
danseront	d@sR§	VER:ind:futu		3p	danser	0.19	29.84	9	5	CVCCVCV	CVCCV	9	5	d@-sR§	2	CV-CCV
danses	d@s	NOM;VER:ind:p	f	2s	danse;danser	14.19	2402.67	6	3	CVCCVC	CVC	6	3	d@s	1	CVC
danseur	d@s9R	NOM	m	s	danseur	6.94	602.54	7	5	CVCCVVC	CVCVC	7	5	d@-s9R	2	CV-CVC
danseurs	d@s9R	NOM	m	(p)	danseur	7.87	1440.37	8	5	CVCCVVC	CVCVC	8	5	d@-s9R	2	CV-CVC
danseuse	d@s2z	NOM	f	s	danseur	6.58	674.34	8	5	CVCCVVC	CVCVC	8	5	d@-s2z	2	CV-CVC
danseuses	d@s2z	NOM	f	(p)	danseur	5.74	521.15	9	5	CVCCVVC	CVCVC	9	5	d@-s2z	2	CV-CVC
dansez	d@se	VER:imp:pr;ind:pr		2p	danser	0.55	129.24	6	4	CVCCVC	CVCV	6	4	d@-se	2	CV-CV
dansiez	d@sje	VER:ind:impf;sub:pr		2p	danser	0.06	6.23	7	5	CVCCVVC	CVCYV	6	5	d@-sje	2	CV-CYV
dansions	d@sj§	VER:ind:impf;sub:pr		1p	danser	0.32	12.26	8	5	CVCCVVC	CVCYV	6	5	d@-sj§	2	CV-CYV

freqfrant: les fréquences de *frantext* par million d'occurences; **freqweb**: les fréquences de *fastsearch* (web) par million de pages; **nblettres**: le nombre de lettres; **nbphons**: nombre de phonèmes; **cvcv**: la structure orthographique; **p-cvcv**: la structure phonologique; **puorth**: point d'unicité orthographique; **puphon**: point d'unicité phonologique; **syll**: forme phonologique syllabée; **nbsyll**: nombre de syllabes **; cv-cv** : structure phonologique syllabée

-Graphie (*graph*): La graphie est la forme orthographique du mot (p. ex.*chienne*)

-Phonie (*phon*): Les codes phonémiques utilisés sont présentés dans le Tableau 2

Tableau 2 Codes phonémiques

Codes Lexique	Exemples	Sons nommés
Voyelles		
a	bat, plat	a
i	lit, émis	i
y	lu	u
u	roue	ou
O	éloge, peau	o (fermé ou ouvert)
e	été	e-fermé
E	paire, treize	e-ouvert
*	premier, abattre	schwa
2	deux	e-fermé
9	œuf, peur	e-ouvert
5	cinq, linge	in (voy. nasale)
1	un, parfum	un (voy. nasale)
@	ange	an (voy. nasale)
§	on, savon	on (voy. nasale)
o	minoen	o d'origine étrangère
Semi-Voyelles		
j	yeux, paille	y (semi-voyelle)
8	huit, lui	ui (semi-voyelle)
w	oui, nouer	w (semi-voyelle)

Codes Lexique	Exemples	Sons nommés
Consonnes		
p	père, soupe	p (occlusive)
b	bon, robe	b (occlusive)
t	terre, vite	t (occlusive)
d	dans, aide	d (occlusive)
k	carré, laque	k (occlusive)
g	gare, bague	g (occlusive)
f	feu, neuf	f (fricative)
v	vous, rêve	v (fricative)
s	sale, dessous	s (fricative)
z	zéro, maison	z (fricative)
S	chat, tâche	ch (fricative)
Z	gilet, mijoter	ge (fricative)
m	main, femme	m (cons. nasale)
n	nous, tonne	n (cons. nasale)
N	agneau, vigne	gn (c. nasale palat.)
l	lent, sol	l (liquide)
R	rue, venir	R
x	jota	jota (emprunt espagn.)
G	camping	ng (emprunt angl.)
h	hachoir	h aspiré

- Classe grammaticale (*cgram*) : Si une même entrée peut appartenir à plusieurs classes grammaticales différentes, celles-ci sont séparées par un point-virgule. Les différents codes utilisés pour représenter les catégories grammaticales sont présentés dans le Tableau 3.

Tableau 3 Codes des catégories grammaticales

Abréviations	Signification
ABR	Abréviations
ADJ	Adjectif
ADV	Adverbe
Cond	Conditionnel
CONJ	Conjonction
demo	démonstratif
DET	Déterminant
EXCLAM	Exclamation
Futu	Futur
Imp	Impératif
Impf	Imparfait
Ind	Indicatif
Indef	indéfini
Infi	Infinitif
INT	Interjection
invar	Une des formes est invariable
LOC	Locution
NOM	Nom
ONOMAT	Onomatopée
pers	Pronom personnel
poss	Pronom possessif
Pper	Participe passé
Ppre	Participe présent
Pr	Présent
PRE	Préposition
PREF	Préfixe
PRO	Pronom
Ps	Passé simple
Sub	Subjonctif
SUFFIXE	Suffixe
SYM	Symbole
VER	Verbe

- Genre (*genre*) : Les différents codes utilisés pour représenter le genre sont présentés dans le Tableau 4.

Tableau 4 Codes utilisés pour le genre

m	masculin
f	féminin
é	épicène

(Un épicène est un mot dont la forme ne varie pas avec le genre (p. ex. *pianiste*))

- Nombre (*nombre*) : Les codes utilisés pour représenter le singulier, le pluriel, etc. sont indiqués dans le Tableau 5.

Tableau 5 Codes du champ nombre

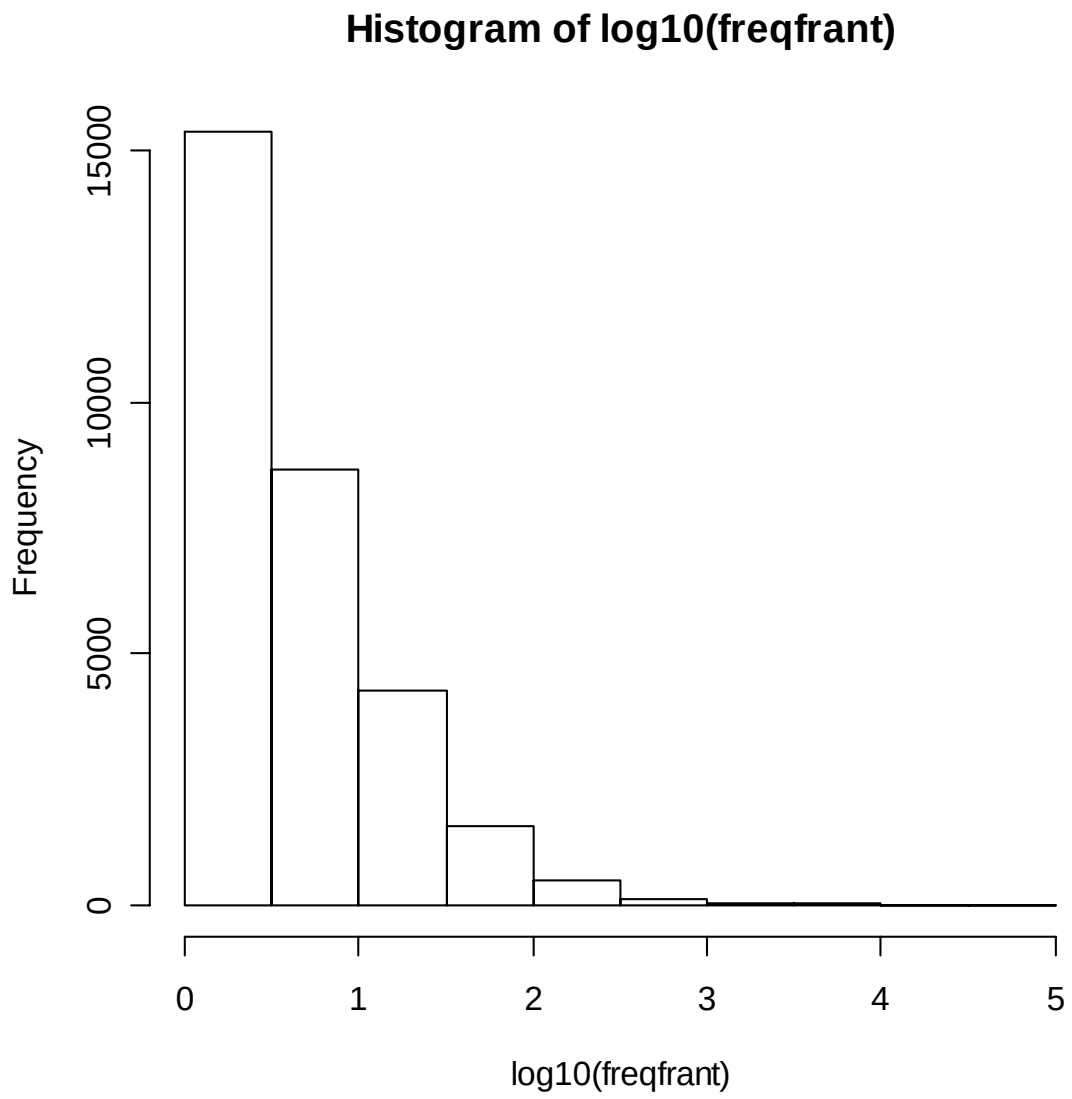
s	Singulier
p	Pluriel
(p)	probablement pluriel mais peut aussi être pluriel ou singulier (vieux)
1s	1ère personne du singulier
2s	2ème personne du singulier
3s	3ème personne du singulier
1p	1ère personne du pluriel
2p	2ème personne du pluriel
3p	3ème personne du pluriel

- Lemme (*lemme*) : Le lemme est la forme canonique, c'est à dire l'infinitif pour un verbe, la masculin singulier pour un nom ou un adjectif. Par exemple, l'item *chienne* a pour lemme *chien.*

- Nombre aléatoire (*rand*) : Un nombre aléatoire tiré entre 1 et 1 000 000. Si vous utilisez cette colonne afin de trier les résultats obtenus, vous pouvez ainsi obtenir des items dont les premières lettres sont distribuées dans la totalité de l'alphabet (ce peut être très utile lors de la constitution du matériel d'une expérience).

- Fréquence par million selon *Frantext* (*freqfrant*) : Elle correspond à la fréquence fournie par *Frantext*, normalisée par une division par 31 (le corpus original comprenant 31 millions d'occurences). La somme des fréquences de ce champs ne fait pas un million en raison du premier filtrage effectué. En effet, après avoir collecté toutes les formes orthographiquement distinctes présentes dans la base de textes Frantext, nous avons dû enlever de cette liste toutes les formes étrangères, noms propres, etc.

Figure 2 : Histogramme des fréquences Frantext pour les mots de fréquence supérieure à 1



Ce graphique indique qu'entre : (les fréquences sont données en occurrences par millions)

0 et 1: 98 000 mots

1 et 3 : 15 000 mots

3 et 10 : 9 000 mots

10 et 31 : 4000 mots

31 et 100 : 1600 mots

100 et plus : 700

-Fréquence par million de pages selon *FastSearch* (*freqweb*) : Le nombre de pages web par million où ce mot apparaît, selon *FastSearch* (sur un corpus de 14,27 millions de pages).

- Nombre de lettres (*nblettres*)

Tableau 6 : Nombre de mots dans Graphemes en fonction du nombre de syllabes et du nombre de lettres

Nombre de caractères \ Nombre de syllabes	1	2	3	4	5	6	7	8	9
1	43	2	2	0	0	0	0	0	0
2	100	20	0	0	0	0	0	0	0
3	533	50	3	1	0	0	0	0	0
4	1328	807	5	0	0	0	0	0	0
5	2055	3398	185	0	0	0	0	0	0
6	1957	7471	1563	5	0	0	0	0	0
7	1108	10025	5057	128	0	0	0	0	0
8	398	9175	9784	762	2	0	0	0	0
9	65	5808	12366	2055	43	1	0	0	0
10	5	2769	11155	3648	213	0	0	0	0
11	0	792	7625	4684	581	12	0	0	0
12	0	197	3786	4346	910	63	0	0	0
13	0	18	1457	2950	1104	135	8	0	0
14	0	2	384	1576	971	201	15	0	0
15	0	0	62	641	639	254	26	0	0
16	0	0	7	204	331	187	44	0	0
17	0	0	0	24	125	131	29	5	0
18	0	0	0	5	54	73	27	7	1
19	0	0	0	0	17	39	26	6	0
20	0	0	0	0	4	10	12	13	0
21	0	0	0	0	0	1	1	1	0

- Nombre de phonèmes (*nbphons*) : C'est le nombre de phonèmes d'après la représentation phonologique présentée dans le champ *phon.*

- Structure orthographique (*cvcv*) : Elle décrit la structure orthographique. Les voyelles sont notées *V*, les consonnes sont notées par *C*. Ainsi *chienne* est représentée par *ccvvccv*.

-Structure de la forme phonologique (*p-cvcv*) : C'est un découpage du mot en voyelles (*V*) et consonnes (*C*) selon sa représentation phonologique.

-Point d'unicité orthographique (*puorth*) : Le point d'unicité orthographique correspond au rang de la lettre en partant de la gauche à partir duquel le mot peut être identifié sans ambiguïté. Nous avons calculé les points d'unicité pour les sur la base des lemmes pour que les formes plurielles ne parasitent pas les calculs (sinon toutes les formes ayant un pluriel ont un point d'unicité égale à leur longueur).Pour les formes orthographiques n'étant pas lemme, le point d'unicté est de 0. [avant la version 2.60 les voisins n'étaient pas calculés sur les lemmes mais sur toutes les entrées de *graphemes*]

- Point d'unicité phonologique (*puphon*) : Le point d'unicité phonologique correspond au rang du phonème en partant de la gauche à partir duquel le mot peut être identifié sans ambiguïté. Le point d'unicité phonologique a aussi été calculé sur la base des lemmes. Pour certains lemmes très rares nous n'avions pas leurs représentations phonologiques (les représentations phonologiques ont été calculées sur les formes orthographiques). Pour les formes orthographiques n'étant pas lemmes, ou pour les formes dont le lemme dont nous n'avons pas de représentation phonologique, le point d'unicité phonologique est donc de 0.

- Syllabation (*syll*) : Les formes phonologiques ont été syllabées selon un algorithme de syllabation décrit dans Dufour, Peereman, Pallier et Radeau (sous presse). Une version mise à jour de l'article décrivant l'algorithme utilisé est présente à l'adresse suivante En résumé, nous avons retenu la syllabation adoptée par Pallier (1994). La syllabation est calculée sur la représentation phonologique présente dans Lexique **dont on a enlevé les schwas finaux**. Cette syllabation est basée sur le principe général d'une segmentation syllabique entre deux consonnes sauf dans les cas des occlusives + liquides ou d'une fricative labio-dentale suivie d'une liquide. Le script de syllabation (syllabation.awk) est distribué avec lexique.

- Nombre de syllabes (*nbsyll*)

- Structure phonologique syllabique (*cv-cv*) : Elle décrit la structure phonologique du mot syllabé. Les consonnes sont notées *C*, les voyelles sont notées *V* et les semi-voyelles *Y*

- Nombre de voisins orthographiques (*voisorth*) : Le nombre de voisins orthographiques calculés selon toutes les entrées de la base *Graphemes*. Les voisins orthographiques d'un mot sont les mots qui peuvent être créés en changeant une lettre sans modifier pour autant la position des autres lettres (Coltheart, Davelaar, Jonasson et Besner, 1977). Par exemple, les mots *vidé*, et *aidé* sont tous des voisins orthographiques du mot aidé. Les différents voisins de chaque mot sont présentés dans la table Voisins (que l'on peut télécharger sur http://www.lexique.org).

- Nombre de voisins phonologiques (*voisphon*) : Les voisins phonologiques d'un mot sont des mots qui peuvent être créés en changeant un phonème sans modifier les autres. Ils ont aussi été calculés à partir des entrées phonologiques de la base *Graphemes*.

- Représentation orthographique inversée (*orthrenv*) : Ex: *erbra* (arbre). Ce type de champs, une fois trié, est très utile pour les personnes travaillant sur les terminaisons (p.ex. en morphologie)

- Représentation phonologique inversée (*phonrenv*) : Ex: *RbRa* (aRbR). Même champs que précédemment mais pour la représentation phonologique.

4.2 Organisation de la table *Lemmes*

La table Lemmes est présentée dans le fichier *Lemmes.txt*. La base *Lemmes* a été créée à partir de *Graphemes*. Nous allons présenter dans cette partie une description des différents champs qui constituent cette base.

Le Tableau 7 présente les différents champs de cette table pour quelques items.

Tableau 7: Présentation d'un extrait de *Lemmes.txt*

lem	graph	phon	cgram	genre	nombre	freqfrantcum	freqfrantgraph	freqwebcum	freqwebgraph
danse	danse;danses	d@s;d@s	NOM;VER:imp:pr;ind:p	f	s;1s;3s;2s	63.9	49.71;14.19	13148.23	10745.56;240
danser	dansa;dansai;dansaient;dans	d@sa;d@sE;d@s	ADJ;NOM;VER:cond:pr	f;m	1s;3s;2s;2	116	0.84;0.06;4.90;	18057	51.34;2.31;15
danseur	danseur;danseurs;danseuse;	d@s9R;d@s9R;d(	NOM	m;f	s;(p)	27.13	6.94;7.87;6.58;	3238.4	602.54;1440.3
dansoter	dansota;dansotait;dansotter	d@sOta;d@sOtE;	VER:ind:impf;ind:ps;infi		3s	0.12	0.03;0.06;0.03	0.21	0;0.14;0.07
dansé	dansé;dansée;dansées;dans	d@se;d@se;d@s	ADJ;VER:pper	f;m	s;(p);p	4.06	3.16;0.35;0.10;	488.44	367.81;53.31;
dantesque	dantesque;dantesques	d@tEsk;d@tEsk	ADJ	é	(p)	0.25	0.19;0.06	83.99	55.69;28.30
danubien	danubien;danubienne;danubie	danybj5;danybjEn;	ADJ	f;m	(p)	0.39	0.10;0.13;0.10;	63.11	23.05;19.26;1
daphnie	daphnies	dafni	NOM	f	(p)	0.06	0.06	23.75	23.75
daphné	daphné	dafne	NOM	m	s	0.06	0.06	153.26	153.26
dard	dard;dards	daR;daR	NOM	m	s;(p)	2.03	1.35;0.68	393.45	304.35;89.10
dardant	dardant;dardantes	daRd@;daRd@t	ADJ;VER:ppre	m;f	3p;p;(p)	0.68	0.65;0.03	22.2	21.99;0.21
darder	darda;dardaient;dardait;darda	daRda;daRdE;daR	ADJ;VER:imp:pr;ind:im	é;f;m	2s;1s;3s;3	2.5	0.10;0.06;0.32;	163.55	9.60;4.97;18.2
dardillon	dardillon	daRdij§	NOM	m	s	0.03	0.03	0.56	0.56
dardé	dardé;dardée;dardées;dardés	daRde;daRde;daR	ADJ;VER:pper	é;f;m	s;(p);p	0.96	0.32;0.35;0.19;	26.05	17.02;3.64;1.7
dargeot	dargif	daRZif	NOM	m	s	0.03	0.03	0.7	0.70
dariole	darioles	daRjOl	NOM	f	(p)	0.06	0.06	6.51	6.51
darique	darique;dariques	daRik;daRik	NOM	f	s;(p)	0.22	0.06;0.16	5.95	1.26;4.69
darne	darne	daRn	ADJ;NOM	é;f	s	0.06	0.06	95.89	95.89
daron	daron;daronne;daronnes;daro	daR§;daROn;daR(	NOM	f;m	s;(p)	1.22	0.42;0.48;0.06;	15.27	12.75;2.03;0.(
darse	darse;darses	daRs;daRs	NOM	f	s;(p)	0.58	0.32;0.26	58.42	45.53;12.89
dartre	dartres	daRtR	NOM	f	(p)	0.13	0.13	17.86	17.86

Légende: lem: le lemme; graph: les formes fléchies du lemme; phon: les formes phonologiques des formes fléchies; cgram: les catégories grammaticales auxquelles appartiennent les formes fléchies; genre: le genre des formes fléchies; nombre: le nombre des formes fléchies; freqfrantcum : la fréquence du lemme selon *Frantext* (en tant que somme des fréquences des formes fléchies associées);); freqfrantgraph: les fréquences des formes fléchies selon *Frantext* freqwebcum la fréquence du lemme du web (en tant que somme des fréquences des formes fléchies associées); freqwebgraph: les fréquences des formes fléchies du web.

Tableau 8: Gros plan sur un verbe:"abaisser"

Abaisser	abaissa;abaissai;abaissaient;abaissait;abaissant;abaisse;abaissent;abaisser;abaissera;abaisserai;abaisseraient;abaisserait;abaisses;abaissez;abaissons;abaissât;abaissèrent;abaissé;abaissée;abaissées;abaissés	abEsa;abEsE;abEsE;abEsE;abEs@;abEs;abEs;abese;abEsRa;abEsRE;abEsRE;abEsRE;abEs;abEse;abEs§;abEsA;abEsER;abese;abese;abese;abese

ADJ;NOM;VER:cond:pr;imp:pr;ind:futu;ind:impf;ind:pr;ind:ps;infi;pper;ppre;sub:impf;sub:pr	F;m	1s;2s;2p;1p;3s;3p;s;(p);p	658	45;2;8;40;74;167;42;138;3;3;4;6;2;3;1;1;7;66;24;4;18	45732	761;62;259;625;3560;7960;1730;16800;576;72;66;258;332;1190;120;13;143;6100;2820;855;1430

- Lemme (*lem*) : Cette base est organisée selon ce champs qui est le lemme.

- Graphies (*graph*) : Ce champs présente les graphies des formes fléchies associées à ce lemme. Ainsi pour le lemme *chien*, les graphies sont *chien*, *chienne*, *chiens* et *chiennes*.

Les champs qui suivent présentent l'information de Graphemes.txt pour chacune des formes fléchies.

- Phonies (*phon*)

- Classes grammaticales (*cgram*)

- Genre (*genre*)

- Nombre (*nombre*)

- La fréquence cumulée du lemme selon *Frantext* (*frantfreqcum*) : C'est la somme des fréquences des formes orthographiques (calculées ci-dessous).

- La fréquence des formes orthographiques selon *Frantext* (*frantfreqgraph*) : Ce sont les fréquences des formes fléchies du lemme. Ainsi le lemme *arbre* ayant deux formes fléchies *arbre* et *arbres*, nous affichons *8 004.64;8 448.17*

- La fréquence cumulée du lemme selon *FastSearch* (*fsfreqcum*)

- La fréquence des formes orthographiques selon *FastSearch* (*fsfreqgraph*)

4.3 Organisation de la table *Surface*

Le fichier *Surface.txt* résume l'information concernant les fréquences des lettres, bigrammes, trigrammes, phonèmes et syllabes pour chaque item de *Graphemes.txt*.

Afin d'effectuer ce résumé, nous avons tout d'abord calculé la fréquence cumulée de chaque unité (lettre, bigramme, etc.) pour chaque position. Pour ce faire, nous avons sommé la fréquence du mot où cette lettre apparaissait à telle ou telle position. Une fois obtenues ces fréquences par position, la fréquence d'un mot présentée dans la base *Surface* correspond à la moyenne de la fréquence des unités le composant.

Par exemple, la fréquence du champs GrTok pour *abaissa* correspond à la moyenne des fréquences de *a* en première position, *b* en deuxième, etc.

Tableau 9 Présentation du mot *abaissa* dans la table *Surface*

Graph	GrTok	GrTokEt	BigrTok	BigrTokEt	TrigrTok	TrigrTokEt	PhonTok	PhonTokEt	SyllTok	SyllTokE
abaissa	28950.19	16528.04	3399.29	3468.15	587.64	521.15	16491.51	23064.28	22815.54	17605.3

4.4 Organisation du dossier *Surface*

Le dossier *Surface* comprend des fichiers donnant des statistiques sur les lettres, bigrammes, trigrammes, phonèmes et syllabes calculées à partir de la table *Graphemes*.

Il comprend 5 sous-dossiers correspondant chacun à une unité d'analyse: lettre, bigramme, trigramme, phonème et syllabe.

Chaque dossier est organisé de la même façon et comprend le même type de fichiers. Nous allons ici décrire le dossier qui contient les informations à propos des bigrammes mais l'organisation des autres dossiers est en tout point similaire à celui-ci. Les calculs sont réalisés soit à partir des fréquences lexicales (ou fréquence de type), soit en fonction des fréquences textuelles (ou fréquence de token). La fréquence lexicale d'une lettre, par exemple, correspond au nombre de mots dans laquelle elle apparaît. La fréquence textuelle d'une lettre correspond à la somme des fréquences de tous les mots où cette lettre est apparue.

4.4.1 Bigr.txt

Ce fichier décrit pour chaque bigramme, sa fréquence de type et sa fréquence textuelle pour chacune des positions qu'il peut prendre.

Le Tableau 10 présente un extrait de la table B*igr.txt* pour le bigramme *ab*

Tableau 10: Présentation de la table Bigr.txt

Bigr	Pos1BigrType;Pos1BigrTok	Pos2BigrType;Pos2BigrTok	Pos3BigrType;Pos3BigrTok
ab	734;1308.86	736;1375.05	223;522.24

Par exemple, la deuxième colonne du Tableau 10 montre que le bigramme *ab* en première position a une fréquence lexicale de 734 et une fréquence textuelle de 1 308. La deuxième colonne montre la fréquence lexicale et textuelle de *ab* en deuxième position, etc.

4.4.2 BigrMoy.txt

Ce fichier indique pour chaque bigramme sans position particulière sa fréquence. C'est la somme de ses fréquences par position. (données dans Bigr.txt).

Le Tableau 11 présente un extrait de ce tableau pour le bigramme *ab*

Tableau 11: Présentation du bigramme *ab* dans la table BigrMoy.txt

Graph	MoyFreqType	MoyFreqToken
ab	3573	5643.7

La deuxième colonne du Tableau 11 indique que le bigramme *ab* apparaît toutes positions confondues 5643,7 fois (par million de mots) dans le corpus de Frantext (fréquence textuelle ou de token). La première colonne indique que *ab* apparaît dans 3573 mots différents (fréquence lexicale ou de type).

4.4.3 BigrMots.txt

Ce fichier indique pour chaque mot, la fréquence lexicale et la fréquence textuelle de chacun de ses bigrammes selon sa position.

Le Tableau 10 présente un extrait de la table *BigrMots.txt* pour le moy *abaissa*

Tableau 12: Présentation de la table BigrMots.txt

Graph	Pos1BigrType;Pos1BigrTok	Pos2BigrType;Pos2BigrTok
ab-ba-ai-is-ss-sa	734;1308.86	169;249.63

La deuxième colonne nous indique que le bigramme *ba* en deuxième position a une fréquence lexicale de 169 et une fréquence textuelle de 249.

4.4.4 BigrMotsMoy.txt

Ce fichier indique pour chaque mot, la moyenne de la fréquence lexicale, la moyenne de la fréquence textuelle, la moyenne des écart-types de la fréquence lexicale et enfin la moyenne des écart-types de la fréquence textuelle de tous les bigrammes le composant.

Le Tableau 13 présente un extrait de la table *BigrMotsMoy.txt* pour l'entrée *abaissa*.

Tableau 13: Présentation du mot *abaissa* dans la table *BigrMotsMoy.txt*

Graph	MoyFreqType	MoyFreqToken	EtTypeEtType	EtToken	Nb
ab-ba-ai-is-ss-sa	926.17	3399.29	533.52	3468.15	6

Ainsi la deuxième colonne du Tableau 13 donne la moyenne des fréquences textuelle de tous les bigrammes dont *abaissa* est composé.

4.4.5 Calculs à partir de la dernière position

Les dossiers lettres, bigrammes, trigrammes, phonèmes et syllabes contiennent tous un dossier DER (*bigrder* pour le dossier bigrammes p. ex.) où se trouvent les mêmes fichiers mais avec un calcul commençant par la dernière unité utilisée. Ainsi *freqbigrder.txt* présente la même information que *freqbigr.txt*; il est à noter que la première colonne correspond aux statistiques de la lettre en dernière position, la deuxième colonne à l'avant-dernière position, etc.

5 Les autres bases

Au fur et à mesure, nous avons créé d'autres bases de données. Vous pouvez cliquer sur les liens afin de disposer d'une explication plus détaillée.

- Fréquence Frantext : la base avec les fréquences brutes (mots et nonmots)
- Voisins : une base de voisins orthographiques
- Anagrammes : une base d'anagrammes
- Prenoms : une base de prénoms
- Corpatext : un corpus de textes

6 Disponibilité et site web

Afin de faciliter l'accès à *Lexique*, nous avons créé un site web disponible à l'adresse suivante: *http://www.lexique.org*

Ce site web s'organise autour de plusieurs pôles :
-une description de *Lexique*;
-des nouvelles permettant de connaître les dernières modifications apportées soit au site soit à la base elle-même;
-un forum où les utilisateurs peuvent poser des questions, faire des propositions d'évolutions futures, etc. Quand une question nous est posée, nous essayons d'y répondre le plus rapidement possible. En général, la réponse est postée le lendemain. D'autre part, les propositions faites par les utilisateurs sont prises en compte dans le développement de *Lexique*.
-une liste catégorisée de liens vers d'autres bases de données ou outils en rapport avec le domaine de l'ingénierie linguistique;
-des outils "en ligne" permettant l'interrogation de *Lexique*;
-le téléchargement des bases et des outils "hors ligne" constituant *Lexique*. Les outils "hors ligne" sont des outils ne nécessitant pas de connexion à internet pour être utilisables.

La base de données *Lexique* est disponible sous la forme d'une application que l'on peut installer facilement sur un ordinateur équipé du système *Windows*. Pour les utilisateurs de système *Macintosh* ou *UNIX*, il est aussi possible de télécharger directement les bases sous forme compressée (zip). Une mise à jour des bases ainsi que des outils "en ligne" et "hors ligne" est effectuée régulièrement.

7 Licence

Un des objectifs de *Lexique* est de rendre disponible publiquement une base de données qui soit la plus grande et la plus fiable possible. Pour cela *Lexique* est publié sous une licence qui autorise toute personne à utiliser, copier, et même modifier la base, du moment que celle-ci reste sous cette même licence.

Cette licence correspond à la "Licence Publique Générale" existant dans le monde des logiciels libres. Nous avons choisi cette licence afin de garantir la gratuité des futures versions de *Lexique*, ainsi que pour encourager les différents utilisateurs à participer à l'élaboration de cette base, ce qui a déjà été le cas avec la collaboration de Peereman et Dufour (sous presse) pour ne citer qu'un exemple.

Cette licence présente aussi l'avantage de garantir une certaine pérennité à cette base. En effet, la célèbre base de données développée par l'Institut de Nimejgen, *Celex* a toujours été distribuée sous une licence propriétaire. Maintenant que les sources de financement de ce projet ont été coupées, le développement de *Celex* semble définitivement arrêté. C'est un problème auquel ne sera pas confronté *Lexique*. Cette licence garantit que si un jour le projet ne devait plus être soutenu par les auteurs à l'origine du projet, un autre laboratoire pourrait tout à fait télécharger la base, la modifier et la redistribuer.

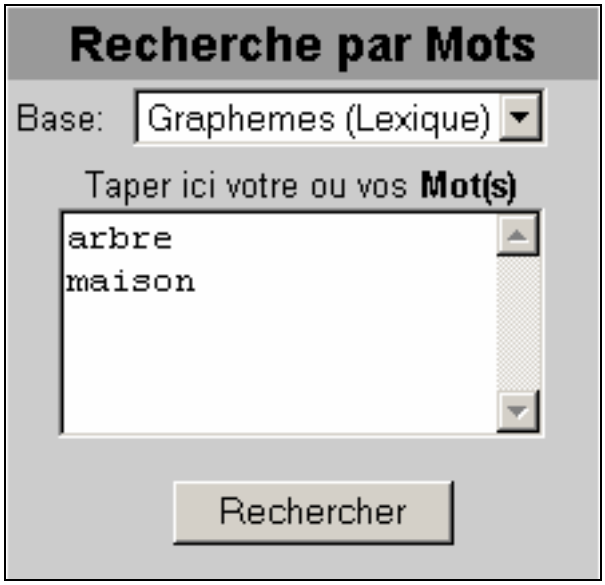
8 Les outils

Afin de rendre *Lexique* disponible au plus grand public, nous avons mis à disposition plusieurs outils gratuits permettant de l'interroger.

8.1 Les outils "en ligne"

Il existe deux moteurs de recherche "en ligne" : un moteur permettant de faire des requêtes à partir d'une simple liste de mots. Cela permet aux personnes désirant obtenir une certaine caractéristique donnée pour une liste de mots de l'obtenir instantanément. Ce moteur permet à l'utilisateur de choisir sa base, taper son ou ses mots et de lancer sa recherche. Celle-ci apparaît alors dans un tableau qu'il peut par exemple copier et coller dans un tableur tel qu'Excel. La figure Figure 3 présente un exemple d'un tel type de requête.

Figure 3 Exemple de requête de type "Recherche par Mots"



Le deuxième moteur de recherche permet d'effectuer des recherches par propriétés sur *Lexique* et d'autres bases simultanément.

Pour cela, l'utilisateur choisit la ou les bases sur lesquelles il désire procéder à son interrogation. Dans un deuxième temps, il choisit le type de recherche qu'il désire effectuer : il peut effectuer : 1) soit une recherche simple permettant d'utiliser quelques opérateurs basiques Ces opérateurs sont présentés dans le tableau ci-dessous.

Tableau 14 Présentation des opérateurs utilisés dans recherches simples

Symbole	Signification	Exemple	Résultat
*	Toute chaîne de caractères	a*	arbre, arbuste
.	Tout caractère	a.o	ado
<	Inférieur à	<10	Mots fréquence inférieure à 10
>	Supérieur à	>30	Mots de fréquence supérieure 30
=	Egal à	=10	Mots de fréquence égale à 10
< > ou > <	Inférieur et Supérieur à	<10 >30	Mots de fréquence inférieure à 10 et supérieure à 30

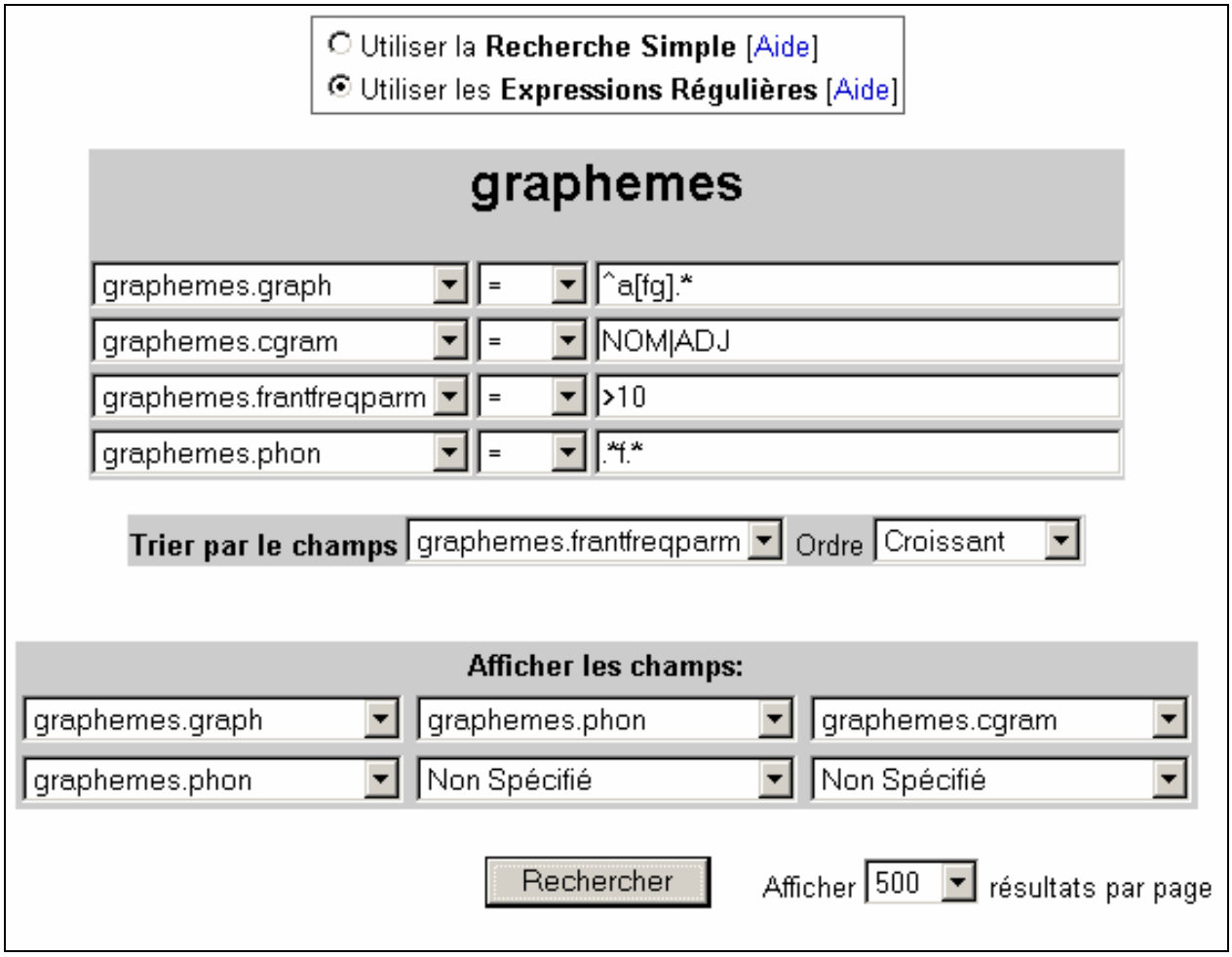
2) soit une recherche utilisant à la fois les opérateurs disponibles en recherche simple et les expressions régulières. Les expressions régulières permettent d'effectuer des recherches très complexes de chaînes de caractères. Tous les opérateurs disponibles dans la recherche par "Expressions Régulières" sont présentés dans le Tableau 15. Un exemple de recherche complexe utilisant les expressions régulières est la recherche *^[^aeiouyàâçèéêôîû]*[aeiouyàâçèéêôîû][^aeiouyàâçèéêôîû]*$* qui permet de rechercher tous les mots ne contenant qu'une seule voyelle.

Ensuite il sélectionne les champs sur lesquels il effectue sa recherche puis tape l'expression recherchée. L'utilisateur peut aussi choisir les colonnes qu'il désire afficher et sur quelle colonne il désire qu'un tri soit effectué. Une requête est présentée dans la Figure 4. Cette requête utilise les expressions régulières et demande tous les mots commençant par la lettre *a* suivie d'un *f* ou d'un *g*, qui soient *nom* ou *adjectif*, dont la fréquence est supérieure à 10 occurrences par million et dont la représentation phonémique comprend la fricative */f/*. Cette requête demande en outre que les résultats soient triés selon leur fréquence par ordre croissant et de n'afficher que 4 colonnes (le mot, sa représentation phonémique, sa catégorie grammaticale et sa fréquence).

Tableau 15 Présentation des opérateurs utilisés dans les expressions régulières

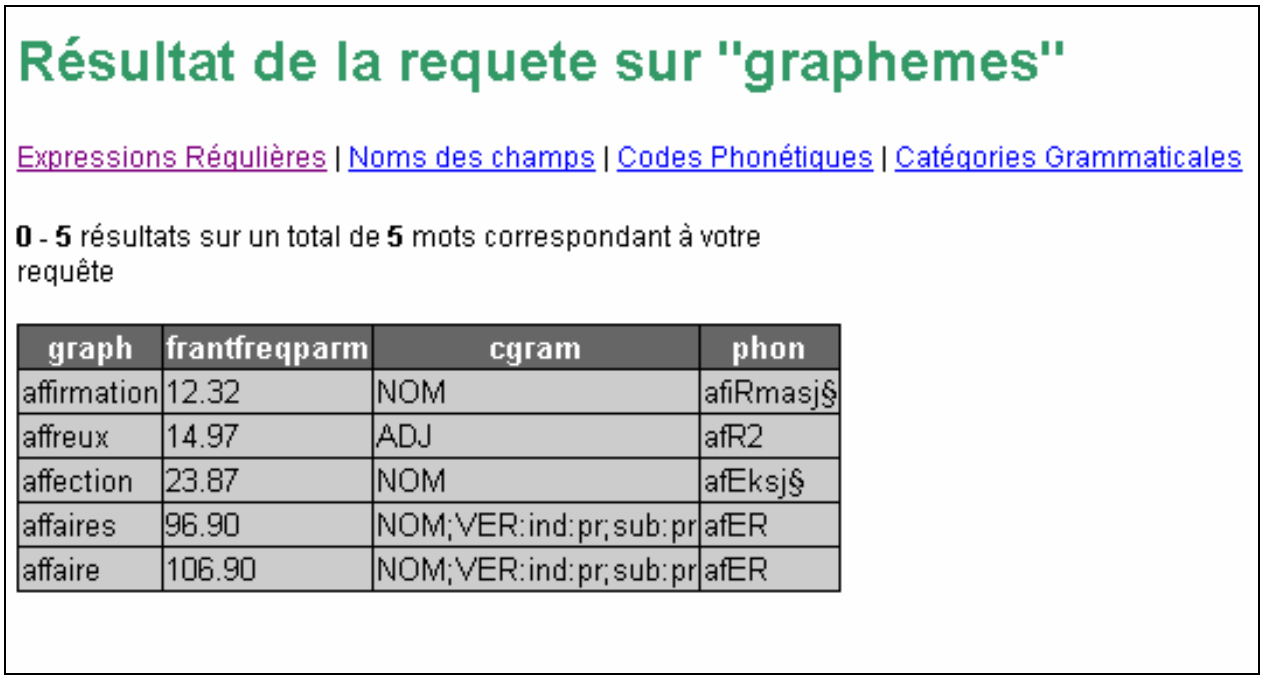
Symbole	Signification	Exemple	Résultat
^	Début de chaîne	^a	arbre, arbuste
$	Fin de chaîne	e$	tente, mare
.	Tout caractère	^a..e$	arme, acte
[xyz]	Les caractères x, y ou z	a[bc]	raccroché, abruti
[x-z]	La tranche de caractères de x à z	a[l-n]	amener, alourdi, anneau
[^xyz]	Tous les caractères sauf xyz	[^aeiyouéèïê]	Toutes les consonnes
*	Désigne le caractère qui précède répété un nombre quelconque de fois, y compris zéro	m*	emmener, amender, entasser
+	Désigne le caractère qui précède répété au moins une fois	m+	emmener, amender
?	Désigne le caractère qui précède répété au plus une fois	m?	amender, entasser
|	ou	(buv|parl)ant	buvant, parlant
{n}	désigne le caractère qui précède exactement n fois	nn{2}	patronne mais pas patron

Figure 4 Exemple de requête effectuée sur la base Graphemes.



Le nombre de résultats et les entrées correspondant à la requête sont alors affichés dans un tableau que l'utilisateur pourra copier et coller dans un tableur par exemple, afin de les retravailler. Pour de ne pas rendre les recherches trop lourdes pour le serveur, nous avons limité celles-ci à 2 000. Si la requête de l'utilisateur dépasse les 500 résultats, celui-ci pourra naviguer 2 000 par 2 000. La Figure 5 présente les résultats obtenus suite à la requête présentée dans la Figure 4.

Figure 5 Résultats obtenus suite à la requête présentée dans la Figure 4



Résultat de la requete sur "graphemes"

Expressions Régulières | Noms des champs | Codes Phonétiques | Catégories Grammaticales

0 - **5** résultats sur un total de **5** mots correspondant à votre requête

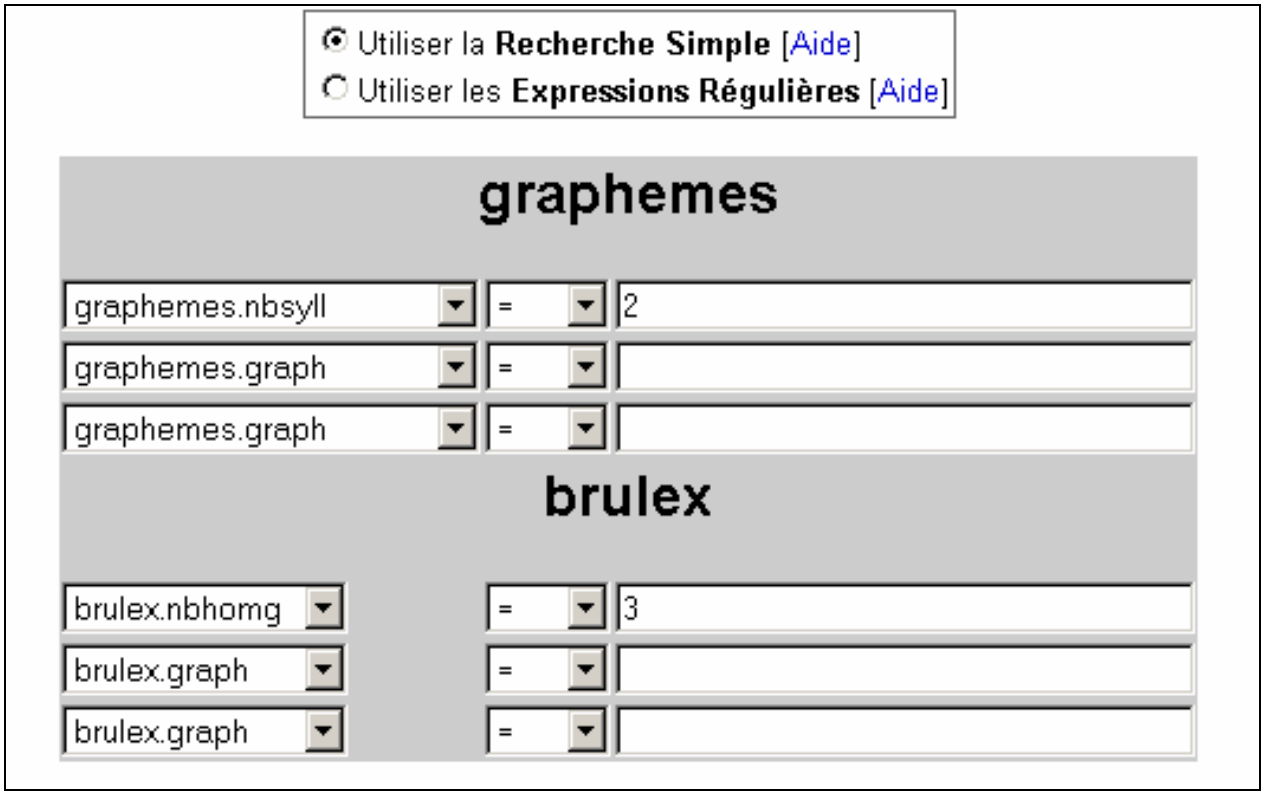
graph	frantfreqparm	cgram	phon
affirmation	12.32	NOM	afiRmasj§
affreux	14.97	ADJ	afR2
affection	23.87	NOM	afEksj§
affaires	96.90	NOM;VER:ind:pr;sub:pr	afER
affaire	106.90	NOM;VER:ind:pr;sub:pr	afER

De plus, deux pages html présentent beaucoup d'exemples d'utilisation à la fois de la recherche simple et de la recherche par expressions régulières.

8.2 Open Lexique

Un des problèmes de toute base de données est le souhait d'avoir la base la plus riche possible. Or, le fait de rajouter de nouveaux champs pose certains problèmes : la taille de la base de données devient de plus en plus importante et la base devient de ce fait de plus en plus lente à télécharger, interroger et corriger. Afin de résoudre ce problème nous avons développé *Open Lexique* : il s'agit d'un moteur de recherche permettant d'interroger plusieurs bases de données simultanément. Cet outil nous permet donc d'ajouter des bases de données et des informations aux entrées lexicales de *Lexique* sans pour autant alourdir notre base. Cela rend aussi *Lexique* facilement extensible. La Figure 6 présente un exemple de requête utilisant *Open Lexique* où nous demandons tous les mots de 2 syllabes selon *Graphemes* qui ont 3 homographes selon *Brulex*.

Figure 6 Exemple de recherche utilisant les possibilités d'*Open Lexique*. Nous demandons ici tous les mots de 2 syllabes selon Graphemes qui ont 3 homographes selon *Brulex*.



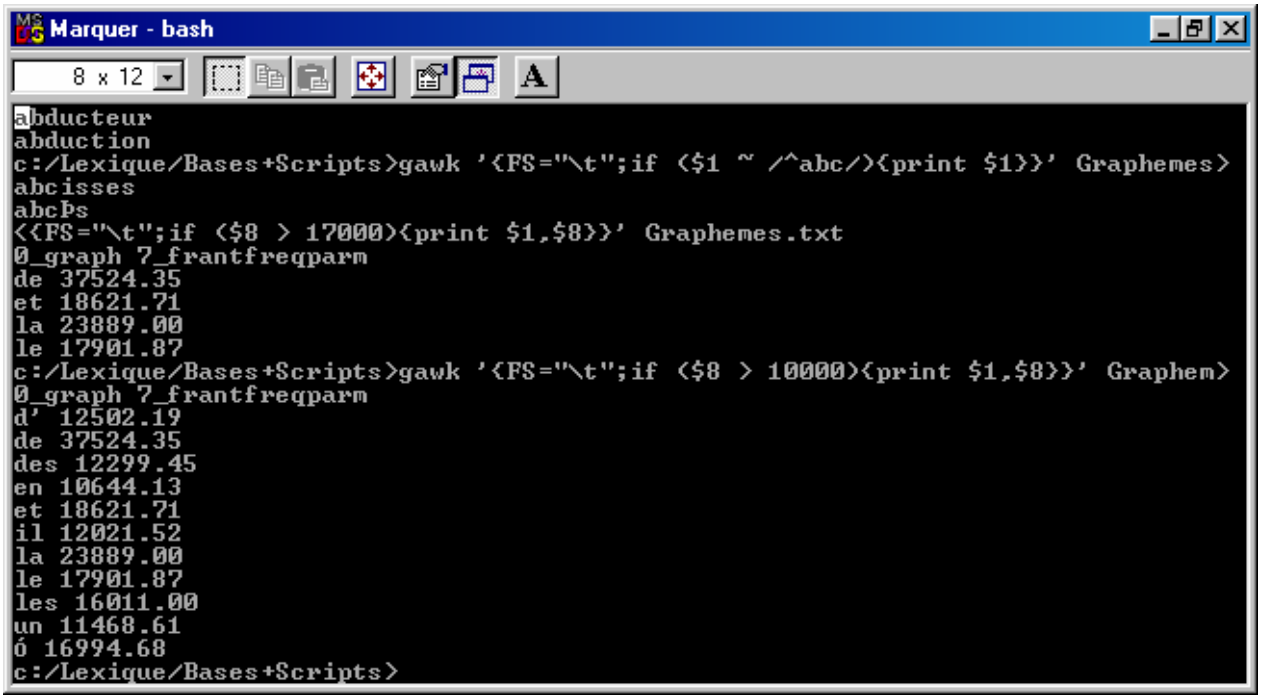
Pour l'instant, les bases interrogeables en plus des bases *Graphemes*, *Lemmes* et *Surface* constituant *Lexique* sont les bases d'Alario et Ferrand (1999), *Brulex* (Content et al., 1990) et la base sur l'âge d'acquisition de Ferrand, Grainger et New (sous presse). *Open Lexique* permet donc aux utilisateurs de *Lexique* d'accéder, pour certains items, à l'âge d'acquisition, le nombre de voisins orthographiques et phonologiques, le nombre d'homographes et d'homophones, le nombre d'homonymes sémantiques, la valence d'imagerie, etc.

8.3 Les outils "hors ligne" : Undows

Compte tenu des différentes limites imposées par les moteurs "en ligne", nous avons mis à disposition tout un ensemble d'outils permettant d'effectuer des recherches beaucoup plus puissantes que celles "en ligne".

Ainsi, nous avons regroupé dans une application facilement utilisable dénommée Undows (http://www.lexique.org/undows) des outils libres tels que *gawk*, *perl*, *bash*, et les *textutils*. Nous avons choisi d'utiliser les outils *awk* et *perl* car ce sont des langages de programmation spécialisés dans le traitement de données de type "texte". Ces langages permettent d'effectuer facilement des requêtes simples de types "sélection de données" ou des programmes beaucoup plus complexes. En démarrant cette application, l'utilisateur a accès à plusieurs exemples de recherches courantes à effectuer sur *Lexique* telles qu'une recherche sur tous les mots ayant la catégorie grammaticale *NOM*, tous les mots commençant par *b*, tous les mots finissant par *t*, ou tous les mots compris dans une certaine gamme de fréquence. La Figure 7 présente des exemples de requêtes effectuées avec ces outils.

Figure 7 Exemples de requêtes effectué "hors ligne"



Des exemples de scripts *awk* ou *perl* sont aussi inclus qui permettent de faire des opérations plus complexes telles que l'écriture des mots de la base à l'envers, le calcul des points d'unicité, l'algorithme de syllabation utilisé pour la constitution des formes syllabées de *Lexique*, le calcul des voisins (orthographiques ou phonologiques) et de leurs fréquences, etc.

De plus nous mettons à disposition de nombreuses documentations avec les outils "hors ligne". Cet ensemble de documentation comprend toutes les documentations officielles des outils disponibles ainsi que deux documentations que nous avons rédigées. Nous avons notamment écrit une rubrique "Foire Aux Questions" essayant de répondre aux principales questions des utilisateurs concernant l'utilisation de *Undows* avec *Lexique* ainsi qu'une documentation expliquant comment utiliser le langage *awk* afin d'interroger *Lexique.*

8.4 Évolutivité

Depuis la première version de Lexique rendu publique le 19 octobre 2000, la communauté d'utilisateurs de *Lexique* n'a cessé de grandir. Aujourd'hui, notre site accueille, chaque mois, 1500 visiteurs en moyenne. Depuis cette première version, la base Lexique en elle-même, le site et les outils permettant de l'interroger ont été mis à jour et enrichis régulièrement. L'amélioration la plus importante qui a été faite, concerne les représentations phonologiques. En effet celles-ci ont été corrigées de manière exhaustive par Peereman et Dufour (sous presse).

Nous travaillons d'ors et déjà sur la troisième version de Lexique qui devrait voir notamment apparaître de nouveaux champs tels que le nombre d'homographes, d'homophones, ainsi que le nombre de voisins phonologiques et orthographiques. Nous avons aussi commencé à développer de nouveaux outils permettant aux utilisateurs d'interroger Lexique sans être connectés à internet par le biais d'une interface simplifiée. Enfin nous envisageons aussi de mettre à disposition *Lexique* sous la forme de CD-ROM.

9 Conclusion

Pour conclure, les points forts de cette nouvelle base de données lexicales du français sont les suivants.

- Les fréquences de *Lexique* sont basées sur un corpus de 31 millions de mots issus de textes récents publiés entre 1950 et 2000.

- Elle est la première base disponible gratuitement du français à inclure, les formes fléchies des mots (formes verbales conjuguées, formes plurielles et féminines des noms et adjectifs).

- Afin de réduire les problèmes de représentativité de corpus, deux estimations de la fréquence sont fournies : l'une fondée sur le corpus original de *Frantext*, et l'autre sur les pages web françaises indexées par le moteur de recherche *FastSearch*. Afin que les fréquences de *Lexique* restent actualisées, ces deux estimateurs pourront facilement être mis à jour dans quelques années.

- Elle est organisée autour de deux tables qui ont pour clés principales, soit les formes orthographiques, soit les lemmes. Elle fournit de nombreuses informations fréquentielles concernant les lettres, les bigrammes, les trigrammes, les phonèmes et les syllabes.

- La création du récent projet *Open Lexique* permet d'interroger plusieurs bases simultanément à *Lexique* dont les bases *Brulex*, la base d'Alario et Ferrand (1999) et la base de Ferrand et al. (sous presse).

- Elle est gratuite, libre d'accès, téléchargeable, et des outils sont fournis pour l'interroger. Elle est actualisée très régulièrement.

Annexe A: Noms des champs

A quoi correspond les différents champs de telle ou telle base (comment les informations ont-elles été obtenues) ?

400 images (Alario & Ferrand) : Article d'Alario et Ferrand

400AoA (Ferrand, Grainger & New) : Article de Ferrand, Grainger & New

Anagrammes (Lexique) : Page Web de la base Anagramme

Brulex (Content, Mousty & Radeau) : Documentation Brulex

Graphemes (Lexique) : Ce document

Lemmes (Lexique) : Ce document

Manulex Lemmas (Lété, Sprenger-Charolles, & Colé) : Page Web Manulex

Manulex Wordforms (Lété, Sprenger-Charolles, Colé) : Page Web Manulex

Prénoms (Mike Campbell) : Page Web de Prénoms

Surface (Lexique) : Ce document

Voisins (Lexique) : Page Web de Voisins