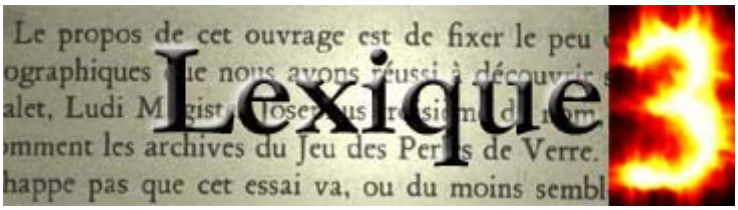


Manuel de Lexique 3

Document version 3.00 beta 2 **!!! En cours d'élaboration !!!**

Boris New[1], Christophe Pallier[2]

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

[2]Unité de neuroimagerie cognitive INSERM U562
Service Hospitalier Frédéric Joliot, CEA
F91401 Orsay FRANCE

E-mail :boris.new AT psycho.univ-paris5.fr

Remerciements: Nous remercions Agnès Bontemps-New qui a eu l'idée de constituer un corpus à base de dialogues films. Nous remercions le projet Technolangue qui nous a permis de financer une partie de Lexique 3. Nous tenons aussi à remercier l'ATILF, Jacques Dendien, Jean-Marie Pierrel, Claude de Loupy, et Jean Veronis pour leur précieuse aide.

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

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 *Lexique3*)
    - 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.

Historique de cette documentation

3.00b2 Refonte de la conclusion et du début de l'état de l'art
3.00b1 Mise à jour afin de rendre compte des nouveautés de Lexique 3

TABLE DES MATIERES

TABLE DES TABLEAUX

TABLE DES FIGURES

Ce manuel explique pourquoi et comment utiliser la base de données *Lexique 3*. Si *Lexique 1 et 2* avaient apporté quelques avantages importants par rapport aux bases de données existant à l'époque (présence des formes fléchies, actualisation, différents indices de fréquence), il y avait encore des améliorations possibles. En effet, les fréquences étaient basées sur de la langue écrites exclusivement (et pas de fréquences orales), il n'était pas possible d'obtenir les fréquences de cooccurrences de mots (ou fréquences d'expressions), les mots composés n'étaient pas présentés, et nous n'avions pas accès aux fréquences des différentes formes grammaticales d'un même mot (p.ex. fréquence de *danse* utilisé comme nom ou utilisé comme verbe). Ce sont tous ces avantages que cette nouvelle version de *Lexique* apporte.

En résumé voici les principales nouveautés de *Lexique 3* par rapport à *Lexique 2*:

- Nouvelles fréquences écrites et orales (basées sur des sous-titres de films)
- Nouvelles entrées de mots récents ou populaires (ex: *internet*, mail, *télécharger*)
- Fréquences des films plus réalistes
- Fréquences des homonymes et homographes (la "*danse*" vs je "*danse*")
- Fréquence des syntagmes de n'importe quelle longueur (ex: *la verte prairie*)
- Formes orthographiques syllabées
- Nouvelles formes phonologiques (15 000)
- Présence des mots composés (ex: *garde-chasse*)
- Fréquences des chiffres et des nombres

1 Etat de l'art des bases de données lexicales en français

La première base de données lexicales informatisée mis à disposition des psycholinguistes fut *Brulex* (Content, Mousty et Radeau, 1990) . *Brulex* regroupait les 35 746 entrées lexicales du *Petit Robert* et leurs fréquences selon le *TLF* (Imbs, 1971). Ces fréquences étaient estimées sur un corpus de textes littéraires datant de 1919 à 1964 et comprenant 26 millions de mots.

Brulex connaissait cependant trois limitations importantes. La première 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, les lettres, les bigrammes, ou les phonèmes. La seconde était que les fréquences étaient basées uniquement sur des textes relativement anciens de surcroît (le plus récent datant de 1964). La troisième de ces

limitations était l'absence de mise à jour. Les auteurs avaient clairement indiqués que leur base ne serait pas mise à jour par de nouvelles champs ou des corrections aux données déjà existantes.

Manulex ou NOVLEX sont deux bases de données plus récente (Lambert et Chesnet, 2001), qui fournissent les formes fléchies et leurs fréquences. En revanche, elles se fondent sur des corpus de manuels scolaires pour les enfants (*Novlex*: CE2, 417000 mots; *Manulex*: CP-CM2, 1,9 millions de mots).

Morphalou (Romary, Salmon-Alt et Francopoulo, 2004) est une base encore plus récentes comprenant 539 413 formes ainsi que des informations morpho-syntaxiques (catégorie grammaticale, genre, nombre et lemme). Cependant *Morphalou* ne contient ni les mots composés (garde-chasse, pomme de terre), ni les fréquences de ces entrées. *Vocolex* est encore une autre base de données qui fournit un ensemble d'indicateurs statistiques sur les similarités entre mots de la langue française.

Afin d'avoir une base de données comprenant les formes fléchies, ainsi que des estimations de fréquences plus actuelles, nous avons créé la base de données *Lexique 1* puis *Lexique 2*. Les fréquences de *Lexique 1 & 2* furent constituées à partir d'une sélection de textes publiés après 1950 du corpus de textes *Frantext*. *Lexique 2* comprenait ainsi 130 000 formes fléchies ainsi que leur fréquence. Si *Lexique 2* apportait un certain nombre d'innovations comparativement aux bases de données existantes, il subsistait encore quelques limitations. Ainsi, les mots composés n'étaient pas présents dans la base. Un autre défaut provenait du fait que n'ayant pas eu accès aux textes, nous n'avions pas la fréquence des formes homographes telles que danse (dans sa forme nominale (*la danse*) et dans sa forme verbale (*je danse*)). Nous avons donc développé *Lexique 3* afin de lever ces limitations.

Pour avoir la fréquence des formes homographes, il nous fallait avoir accès à d'importants corpus de textes. Nous avons donc demandé aux auteurs de *Frantext*, l'autorisation d'utiliser la partie la plus récente de leur corpus. Cependant, *Frantext* est un corpus de textes littéraires (ex d'auteurs: Françoise Sagan, Michel Tournier, mais aussi Georges Perec ou Marguerite Duras). Il y a donc un style assez soutenu et le vocabulaire utilisé ne reflète peut-être pas toujours l'usage de la langue française.

Pour cette raison, nous avons recherché un deuxième corpus reflétant davantage l'usage de la langue. Nous avons d'abord pensé au corpus du journal "*Le Monde*" mais le style utilisé était encore une fois assez élaboré et, du coup, paraissait éloigné de l'usage courant de la langue française.

Ensuite, nous avons eu l'idée de télécharger un corpus de pages web. Pour autant le contenu textuel des pages web n'est pas utilisable directement en raison des menus, des mentions légales, etc. Il exige donc un important travail de prétraitement des données différent pour chaque site web téléchargé. Ce travail de prétraitement rendait donc difficile l'obtention d'un gros corpus.

En troisième tentative, nous avons essayé de scanner des livres ou des journaux populaires tels que des romans de gare ou des journaux télé. Là encore, la tâche s'est révélée ardue en raison de la mise en page relativement complexe des magazines. Se posait aussi le problème du temps de scannage des ouvrages afin d'obtenir un corpus conséquent.

Enfin, nous avons eu l'idée de travailler sur des dialogues de films et de séries et plus précisément sur les sous-titres. En effet les sous-titres de films et de séries présentent trois avantages non négligeables:

-ils existent déjà sous forme numérique de fichiers textes

-ils proviennent de films et de séries souvent américaines très populaires (ex: Ally McBeal, 24h) qui correspondent donc à ce qui peut être entendu en regardant la télévision.

-enfin, ils correspondent à des dialogues parlés et peuvent, de ce fait, servir à estimer l'usage de la langue parlée

2 Constitution des corpus

2.1 Le corpus de textes (Frantext)

L'Atilf nous a donné accès à 218 textes littéraires (romans) publiés entre 1950 et 2000 : cela représente un corpus de 14,8 millions d'items.

2.2 Le corpus de films (ou corpus de sous-titres)

Nous avons téléchargé les sous-titres de 2960 films ou saisons de séries. Ces films pouvaient être étrangers (américains, asiatiques, etc) ou français. Des exemples de films sont *Fight Club* ou *The Postman.*

Comme beaucoup de sous-titres avaient été obtenues par reconnaissance automatique de caractères, nous avons d'abord du effectuer un gros travail de sélection et de correction des fautes d'OCR. (p.ex. "i" remplacé par "l"). Un deuxième travail a été d'enlever les parties ne concernant pas le dialogue mais l'auteur des sous-titres.

3 Etiquetage grammatical du corpus

Afin d'étiqueter grammaticalement nos corpus, nous avons utilisé l'étiqueteur Cordial Analyseur. Pour l'instant, Cordial semble parmi les tout meilleurs catégoriseurs grammaticaux pour le français.

Nous avons obtenu une liste de 293 000 items distincts incluant les mots composés ainsi que leur fréquence Ces items comprenaient des symboles (dont la ponctuation), des abréviations, des mots étrangers et des noms propres. Pour "nettoyer" cette liste, nous avons employé *Aspell*, 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 157 920 items.

4 Estimation de la fiabilité des fréquences

La fréquence des mots est un facteur très important dans la reconnaissance des mots. Les mots utilisés couramment sont plus facilement et plus rapidement reconnus que les mots utilisés plus rarement. Beaucoup d'étude montrent que c'est le facteur expliquant le plus de variance dans la tâche de décision lexicale.

Cet effet n'existe pas uniquement entre les mots très fréquents et les mots très peu fréquent (comme entre *porte et osselet*) mais il joue aussi pour des différences plus subtiles (comme entre *danger* et *nuage*). C'est donc un facteur extrêmement important à contrôler dès lors que l'on veut mettre en évidence l'importance d'un autre facteur dans la reconnaissance de mots.

Gernsbacher (1984) a suggéré que les fréquences basées sur des corpus écrit (comme les fréquences de Brulex ou de Lexique 1 et 2) n'étaient pas de très bons estimateurs de la fréquence d'usage. Elle a notamment argumenté que ces fréquences écrites "classiques" ne prennent pas en compte la fréquence d'occurrence parlée. De plus ces fréquences reposent souvent sur des corpus anciens et non actualisés. Elle a ainsi montré que la familiarité pouvait être un meilleur prédicteur des temps de décision lexicale (notamment pour les mots de basse fréquence) que les fréquences utilisées à l'époque. Il ressort donc de ces études qu'il est crucial d'avoir les fréquences les plus actualisées et les plus proches de l'usage parlé possible.

Dans *Lexique 3*, nous proposons deux estimateurs des fréquences d'usage : le premier est fondé sur un sous-ensemble de de textes littéraires récents (romans) tirés du corpus *Frantext*; le second repose sur un corpus de sous-titres de films.

Etant donné l'importance de la fréquence dans les principaux paradigmes utilisés en psycholinguistique (décision lexicale, dénomination, mais aussi dénomination d'image) mais aussi en traitement automatique de la langue, nous avons voulu avoir une estimation de la fiabilité de ces différents indices de fréquences.

Pour cela, nous avons cherché dans la littérature, des annexes présentant des temps de réaction à une tâche de décision lexicale simple dans la littérature. Nous avons trouvé un article de Bonin et al. présentant trois expériences en décision lexicale simple. La tâche de décision lexicale se décompose ainsi: un sujet voit des mots ou des nonmots présentés à l'écran et il doit décider le plus rapidement possible si la chaîne de caractères présentée est un mot ou un nonmot. On enregistre alors le temps de réaction du sujet.

Nous avons alors effectué une corrélation entre les temps de réaction et les fréquences les logs des fréquences des lemmes[1] données par *Brulex*, *Lexique* 2, *Lexique 3 (films) et Lexique 3 (livres).* Pour Lexique 3 lorsqu'un mot pouvait appartenir à plusieurs catégories grammaticales (ex: *prise*), nous prenions la fréquence de sa forme nominale.

[1] Nous avons effectué nos comparaisons sur les fréquences des lemme car *Brulex* ne donne pas la fréquence de la forme. D'autre part, plusieurs expériences (New, Brysbaert, Segui, Ferrand, & Rastle, 2004; Baayen et al., 1997) montrent que la fréquence du lemme est plus importante que la fréquence de la forme sauf pour certaines catégories assez particulières d'items (p.ex. les mots ayant un pluriel de haute fréquence ou les mots de classe fermée).

4.1 Décision lexicale

4.1.1 Expérience 1

Dans leur première expérience, Bonin et al. présentent 36 mots et 36 nonmots sélectionnés à partir de la base d'Alario et Ferrand (1999). Trente sujets prennent part à l'expérience

Voici le tableau des corrélations entre les temps de réaction et les différents indices de fréquences:

Indice de fréquence	R^2
Fréquence de *Brulex*	0.19
Fréquence de *Lexique 2* (livres)	0.27
Fréquence de *Lexique 3* (films)	0.44
Fréquence de *Lexique 3* (livres)	0.26

Nous constatons que le pourcentage de variance expliquée est le plus faible pour Brulex et le plus fort pour nos fréquences basées sur le corpus de films.

4.1.2 Expérience 2

Dans leur deuxième expérience, Bonin et al. présentent 34 mots et 34 nonmots sélectionnés à partir de la base d'Alario et Ferrand (1999). Trente sujets prennent part à l'expérience

Voici le tableau des corrélations entre les temps de réaction et les différents indices de fréquences:

Indice de fréquence	R^2
Fréquence de *Brulex*	0.37
Fréquence de *Lexique 2* (livres)	0.41
Fréquence de *Lexique 3* (films)	0.48
Fréquence de *Lexique 3* (livres)	0.39

Nous constatons à nouveau que notre fréquence basée sur le corpus de films est l'indice qui explique le plus de variance tandis que la fréquence de *Brulex* est celle qui en explique le moins.

4.1.3 Expérience 3

Dans leur première expérience, Bonin et al. présentent 237 mots et 237 nonmots sélectionnés à partir de la base d'Alario et Ferrand (1999). Trente sujets prennent part à l'expérience. Seules les réponses correctes sont analysées. Trois mots ne sont pas pris en compte dans les résultats car leur taux d'erreurs est supérieur à 50%.

Voici le tableau des corrélations entre temps de réaction et les différents indices de fréquences:

Indice de fréquence	R^2
Fréquence de *Brulex*	0.37
Fréquence de *Lexique 2* (livres)	0.38
Fréquence de *Lexique 3* (films)	0.42
Fréquence de *Lexique 3* (livres)	0.40

A nouveau, pour cette troisième expérience comportant un bien plus grand nombre de mots, nous observons que la fréquence basée sur le corpus de sous-titres expliquent la plus grande part de la variance tandis que la fréquence de *Brulex* en explique le moins.

4.1.4 Conclusions

Dans trois expériences indépendantes, nous observons que la fréquence de *Brulex* est celle qui explique le moins de variance des temps de réaction tandis que la fréquence basée sur notre corpus de films est celle qui en explique le plus. Les fréquences de *Lexique 2* et de *Lexique 3* (livres) donnent des résultats intermédiaires relativement similaires. Ce dernier résultat n'est pas surprenant puisque ces deux fréquences sont dérivées du même corpus *Frantext*.

Comment pouvons-nous expliquer la supériorité de *Lexique 3* (films) par rapport aux autres indices de fréquences? Deux explications sont envisageables: la première serait que la fréquence des sous-titres est plus proche de la fréquence utilisée à l'oral au quotidien. Cela signifierait que même dans une tâche visuelle de décision lexicale, c'est la fréquence orale qui prévaut. (contrairement à l'idée couramment répandue selon laquelle pour les tâche de lecture ce serait la fréquence écrite qui prévaudrait). Cette idée de la prévalence des fréquences orales est compatible avec les théories accordant une grande importance à l'accès phonologique durant la lecture. La deuxième explication fait appel à la façon dont sont constitués les deux corpus. En effet, le corpus de films est construit à partir de 3000 sources différentes (films ou séries) alors que les corpus de livres sont constitués de beaucoup de moins de sources (pour *Lexique 2* environ 460 ouvrages, pour *Lexique 3* environ 220). Pour tester cette hypothèse, il faudrait constituer deux nouveau sous-corpus comprenant le même nombre de mots et le même nombre de sources et voir si ces nouvelles fréquences présentent les mêmes différences ou pas.

5 Autres avantages du corpus de sous-titres

Outre le fait qu'il semble que les fréquences du corpus de sous-titres soient plus représentatives des fréquences du lexique mental que les fréquences tirées du corpus de livres (voir paragraphe précédent), on peut aussi y voir d'autres avantages.

Un autre avantage de la présence de ce corpus de sous-titres est qu'il contient beaucoup de films très récents ce qui permet d'avoir du vocabulaire plus actuels. Il a permis d'ajouter un grand nombre d'entrées récentes qui avaient beaucoup moins de chance de se trouver dans des textes littéraires. C'est le cas de mots tels que *techno, téléchargement, internautes, internet*.

Un second avantage provient du fait que les fréquences de termes caractéristiques du langage parlé tels que *Salut, Bonjour, Au revoir, Oui,* ou *Non* sont au moins 5 fois plus fréquents dans le corpus de sous-titres que dans le corpus de livres.

Enfin, un dernier avantage provient du fait qu'il sera très facile de réactualiser ce corpus très régulièrement.

6 Organisation de la base Lexique 3

Lexique 3 est fournie sous forme de fichiers textes, les champs étant séparés par des tabulations. Cela permet de les importer facilement avec la plupart des logiciels.

6.1 Organisation de la table *Lexique3*

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

Tableau 1 Présentation d'un extrait de *Lexique3.txt*

1_ortho	2_phono	3_lemme	4_cgram	5_genre	6_nombre	7_freqlemfilms	8_freqlemlivres	9_freqfilms	10_freqlivres	11_infover	12_nbhomogr	13_nbhomoph	14_islem
dansant	d@s@	danser	VER			108.14	92.57	2.34	5.54	par:pas;	2	3	0
dansante	d@s@t	dansant	ADJ	f	s	1.65	6.89	0.48	1.76		1	2	0
dansantes	d@s@t	dansant	ADJ	f	p	1.65	6.89	0.21	1.96		1	2	0
dansants	d@s@	dansant	ADJ	m	p	1.65	6.89	0.37	0.61		1	3	0
danse	d@s	danse	NOM	f	s	41.06	35.14	38.62	29.19		2	8	1
danse	d@s	danser	VER			108.14	92.57	18.46	9.8	imp:pre:2s;	2	8	0
dansé	d@se	danser	VER	m	s	108.14	92.57	5.27	4.32	par:pas;	1	4	0
dansée	d@se	danser	VER	f	s	108.14	92.57	0.11	0.27	par:pas;	1	4	0
dansent	d@s	danser	VER			108.14	92.57	3.14	5.54	ind:pre:3p;	1	8	0

1_ortho	15_nblettres	16_nbphon	17_cvcv	18_p_cvcv	19_voisorth	20_voisphor	21_puort	22_puphor	23_syll	24_nbsyll	25_cv-cv	26_orthrenv	27_phonrenv	28_orthosyll
dansant	7	4	CVCCVCC	CVCV	3	14	5	4	d@-s@	2	CV-CV	tnasnad	@s@d	dan-sant
dansante	8	5	CVCCVCCV	CVCVC	1	3	0	0	d@-s@t	2	CV-CVC	etnasnad	t@s@d	dan-san-te
dansantes	9	5	CVCCVCCVC	CVCVC	0	3	0	0	d@-s@t	2	CV-CVC	setnasnad	t@s@d	dan-san-tes
dansants	8	4	CVCCVCCC	CVCV	1	14	0	4	d@-s@	2	CV-CV	stnasnad	@s@d	dan-sants
danse	5	3	CVCCV	CVC	6	18	5	3	d@s	1	CVC	esnad	s@d	dan-se
danse	5	3	CVCCV	CVC	6	18	5	3	d@s	1	CVC	esnad	s@d	dan-se
dansé	5	4	CVCCé	CVCV	4	54	0	4	d@-se	2	CV-CV	ésnad	es@d	dan-sé
dansée	6	4	CVCCéV	CVCV	2	54	0	4	d@-se	2	CV-CV	eésnad	es@d	dan-sée
dansent	7	3	CVCCVCC	CVC	2	18	0	3	d@s	1	CVC	tnesnad	s@d	dan-sent

Légende: **ortho**: le mot; **phon**: les formes phonologiques du mot; **lemme**: les lemmes de ce mot; **cgram**: les catégories grammaticales de ce mot; **genre**: le genre; **nombre**: le nombre; **freqlemfilms**: la fréquence du lemme selon le corpus de sous-titres (par million d'occurrences); **freqlivres**: la fréquence du lemme selon le corpus de livres (par million d'occurrences); **freqfilms**: la fréquence du mot selon le corpus de sous-titres (par million d'occurrences); **freqlemlivres**: la fréquence du mot selon le corpus de livres (par million d'occurrences);**infover**: modes, temps, et personnes possibles pour les verbes; **nbhomogr**: nombre d'homographes; **nbhomoph**: nombre d'homophones; **islem**: indique si c'est un lemme ou pas; **nblettres**: le nombre de lettres; **nbphons**: nombre de phonèmes; **cvcv**: la structure orthographique; **p-cvcv**: la structure phonologique; **voisorth**: nombre de voisins orthographiques; **voisphon**: nombre de voisins phonologiques; **puorth**: point d'unicité orthographique; **puphon**: point d'unicité phonologique; **syll**: forme phonologique syllabée; **nbsyll**: nombre de syllabes **; cv-cv** : structure phonologique syllabée; **orthrenv**: forme orthograhique inversée; **phonrenv**: forme phonologique inversée; **orthosyll**: forme orthographique syllabée

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

-Phonie (*phon*): Représentation phonologique du mot. Les codes phonémiques utilisés sont présentés dans le Tableau 2. La représentation phonologique a été obtenue à partir de *Lexique 2* (voir le manuel de *Lexique 2* pour davantage de détails) pour les entrées qui le permettaient. Pour les entrées ne le permettant pas, nous avons utilisé le logiciel *Multitel Elite* 2.0.1 (Pagel, Black et Lenzo, 1998; Black, Lenzo et Pagel, 1998). Comme pour tout logiciel de "text to speech" adapté à la parole continue et employant un système de règles, des erreurs ont pu être introduites, notamment sur les mots d'origine étrangère. Nous en avons d'ores et déjà corrigé un certain nombre mais il peut en rester. Si vous en trouvez, n'hésitez pas à en faire part sur le forum de Lexique.

Tableau 2 Codes phonémiques

Voyelles		
Codes Lexique	**Exemples**	**Sons nommés**
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)

Consonnes		
Codes Lexique	**Exemples**	**Sons nommés**
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

- 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.*

- Classe grammaticale (*cgram*) : Les différents codes utilisés pour représenter les catégories grammaticales sont présentés dans le Tableau 3.

Tableau 3: Codes des catégories grammaticales

Abréviations	Catégorie grammaticale
ADJ	Adjectif
ADJ:dem	Adjectif demonstratif
ADJ:ind	Adjectif indéfini
ADJ:num	Adjectif numérique
ADJ:pos	Adjectif possessif
ADV	Adverbe
ART:def	Article défini
ART:inf	Article indéfini
AUX	Auxiliaire
CON	Conjonction
NOM	Nom commun
ONO	Onomatopée
PRE	Préposition
PRO:dem	Pronom démonstratif
PRO:ind	Pronom indéfini
PRO:int	Pronom interrogatif
PRO:per	Pronom personnel
PRO:rel	Pronom relation
VER	Verbe

- Genre (*genre*) : Un mot peut être masculin (m) ou féminin (f).

- Nombre (*nombre*) : Un mot peut être singulier (s) ou pluriel (p)

- Fréquence du lemme par million selon le corpus de films (*freqlemfilm*) : Elle correspond à la somme des fréquences des formes fléchies de chaque lemme fournie par notre sélection de films, normalisée par une division par 14,8 (le corpus original comprenant 14,8 millions d'occurences). Ex: freq (*arbre*) = freq ("*arbre*") + freq ("*arbres*")

Tableau 4: Nombre et exemples de lemmes selon leur fréquence (corpus de sous-titres)

Limite inférieure	Limite supérieure	Nombre de lemmes	Noms	Adjectifs	Verbes	Classe fermée
1000	Infini	117			aller, faire, voir	la, le, un, dans,
100	1000	589	porte, voiture, café, police	désolé, grand, bon	fermer, couper, courir	beaucoup, même, souvent
50	100	490	coin, conseil, danger	calme, idiot, sympa	laver, traverser, regretter	doucement, ailleurs, pourtant
20	50	1165	secteur, sable, nuage	malin, joyeux, curieux	creuser, exciter	parfaitement, désormais, lentement
10	20	1137	atmosphère, bouquin, individu	classique, féminin, fidèle	boucher, désigner, étrangler	soudain, clairement, volontiers
1	10	8800	pupitre, éther	déconcertant, morose	vexer, assouvir, exporter	fièrement, bêtement
0	1	30730	filmographie, radiologue, osselet	équatorial, moutonnier	harponner, auréoler	hygiéniquement

- Fréquence du lemme par million selon le corpus de livres (*freqlemlivre*) : Elle correspond à la somme des fréquences des formes fléchies de chaque lemme fournie par notre sélection de livres de *Frantext*, normalisée par une division par 14,8 (le corpus original comprenant 14,8 millions d'occurences).

- Fréquence par million selon le corpus de films (*freqfilm*) : Elle correspond à la fréquence par million d'occurrences du mot selon notre corpus de sous-titres. (18,8 millions de mots). Contrairement à Lexique 2, *danse* aura deux entrées et deux fréquences, une pour sa forme nominale (p.ex. *la danse*) et une pour sa forme verbale (*je danse*).

- Fréquence par million selon le corpus de livres (*freqlivre*) : Elle correspond à la fréquence par million d'occurrences du mot selon notre corpus de livres. (14,8 millions de mots).

- Informations verbales (*infover*): Ce sont les informations de mode, de temps, et de personne que sont susceptibles de prendre les formes verbales

Tableau 5: Informations complémentaires sur les verbes

Mode	
ind	indicatif
cnd	conditionnel
sub	subjonctif
par	participe
inf	infinitif
imp	impératif

Personne	
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

Temps	
pre	présent
fut	futur
imp	imparfait
pas	passé

- Nombre d'homographes (*nbhomogr*): Nombre d'entrées ayant la même forme orthographique mais pouvant différer de par leur catégorie grammaticale ou de par leur lemme.

- Nombre d'homophones (*nbhomoph*): Nombre d'entrées ayant la même forme phonologique.

- Nombre de lettres (*nblettres*)

Tableau 6: Nombre de mots dans Lexique 3 en fonction du nombre de syllabes et du nombre de lettres

	Nombre de syllabes							
Nombre de lettres	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	44	3	1	0	0	0	0	0
2	122	8	4	0	0	0	0	0
3	631	77	12	2	0	0	0	0
4	1606	1042	9	4	0	0	0	0
5	2639	4030	222	0	3	0	0	0
6	2455	8644	1876	5	0	1	0	0
7	1323	11487	5909	137	1	8	0	0
8	450	10223	11021	788	3	0	0	0
9	77	6307	13336	2174	49	2	0	2
10	8	3056	11510	3694	212	6	1	0
11	1	1010	7572	4574	463	13	3	0
12	1	328	3748	4037	728	37	1	0
13	0	97	1486	2680	848	81	5	2
14	1	31	483	1406	706	123	7	0
15	0	6	137	617	468	143	16	0
16	0	4	39	241	244	121	13	0
17	0	1	21	59	106	76	21	2
18	0	0	6	29	52	46	9	1
19	0	0	4	10	23	31	11	6
20	0	0	0	5	8	7	7	4
21	0	0	0	8	4	6	4	2
22	0	0	0	1	4	4	1	2
23	0	0	0	0	1	2	0	3
24	0	0	0	0	0	0	2	0
25	0	0	0	0	1	0	0	1

- 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.

- Nombre de voisins orthographiques (*voisorth*) : Le nombre de voisins orthographiques calculés à partir toutes les entrées de la base. 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 de toutes les entrées phonologiques de la base *Lexique3*.

-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 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 lemmes, le point d'unicité orthographique est de 0.[avant la version 2.60 les voisins n'étaient pas calculés sur les lemmes mais sur toutes les entrées de *Lexique3s*]

- 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, le point d'unicité phonologique est 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*

- 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.

-Représentation orthographique syllabée (*orthosyll*): Champs encore expérimental donnant la représentation orthographique syllabée (Ex *mai-son*). L'algorithme utilisé montre quelque différences avec l'algorithme de syllabation utilisée sur les formes phonologiques (p.ex. les schwas finaux sont comptés comme des voyelles).

6.2 Organisation de la table *lex3.lemmes.txt*

L'équivalent de la base *Lemmes.txt* pour Lexique 2 est générable quand on télécharge Lexique 3. Il suffit de double-cliquer sur le fichier *Lemmes.bat* et cela générera la base lex3.lemmes.txt. *Lex3.lemmes.txt* est organisée de la même façon que la base *Lemmes* de Lexique 2. Vous pouvez donc avoir sa description dans le manuel de Lexique 2.

7 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

8 Les Outils

Afin de rendre *Lexique* disponible au plus grand public, nous avons mis à disposition plusieurs outils gratuits permettant de l'interroger. Il existe trois moteurs de recherche "en ligne" facilement utilisables: un moteur permettant de connaître la fréquence de n'importe quelle chaîne de caractères dans l'un des deux corpus (corpus de sous-titres ou corpus de livres),un moteur permettant de faire des requêtes à partir d'une simple liste de mots, et enfin un moteur permettant de connaître tous les mots partageant certaines propriétés.

8.1 Les outils "en ligne"

8.1.1 La recherche de fréquence dans les corpus

C'est un nouvel outil de recherche disponible avec Lexique 3 qui permet de connaître la fréquence de n'importe quelle chaîne de caractères. Il est ainsi possible de savoir dans combien de fois apparaissent *sel* et *poivre* dans la même phrase. Il est aussi possible de savoir dans combien de phrases apparaît le syntagme "*pomme d'Adam*" ou encore "*la pomme d'Adam*". Cela permet d'effectuer tout un pan de nouvelles recherches qui n'étaient pas possible auparavant telles que des recherches concernant les relations associatives ou sémantiques entre les termes, ou encore des recherches sur les expressions idiomatiques (*broyer du noir*, *monts et merveilles*).

8.1.2 La recherche par mots

Ce moteur 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 **Erreur ! Source du renvoi introuvable.** présente un exemple d'un tel type de requête.

Figure 1 Exemple de requête de type "Recherche par Mots"

8.1.3 La recherche par propriété

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 7 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

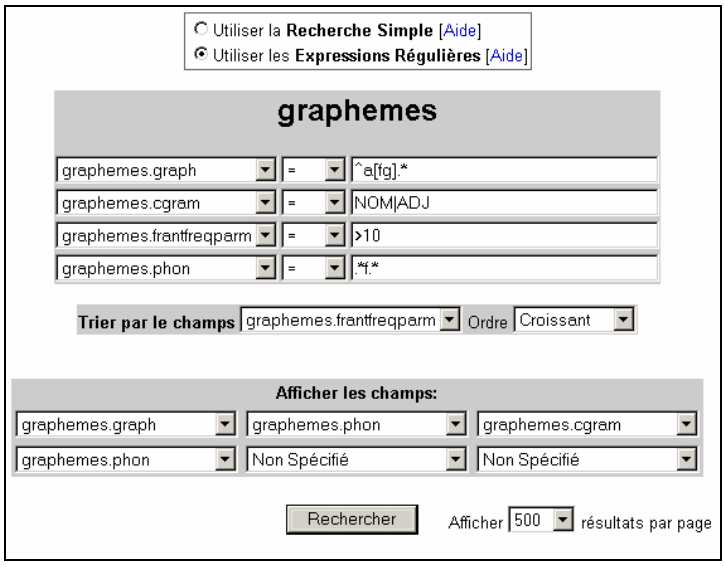
2) soit une recherche utilisant à la fois les opérateurs disponibles en recherche simple et les expressions régulières. Les expressions régulières permettent d'effectuer des recherches très complexes de chaînes de caractères. Tous les opérateurs disponibles dans la recherche par "Expressions Régulières" sont présentés dans le **Erreur ! Source du renvoi introuvable.**. 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 **Erreur ! Source du renvoi introuvable.**. 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 8 Présentation des opérateurs utilisés dans les expressions régulières

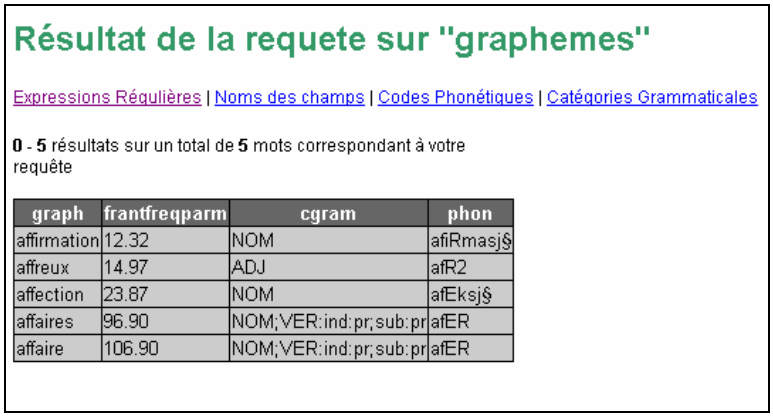
Symbole	Signification	Exemple	Résultat
^	Début de chaîne	^a	arbre, arbuste
$	Fin de chaîne	e$	tente, mare
.	Tout caractère	^a..e$	arme, acte
[xyz]	Les caractères x, y ou z	a[bc]	raccroché, abruti
[x-z]	La tranche de caractères de x à z	a[l-n]	amener, alourdi, anneau
[^xyz]	Tous les caractères sauf xyz	[^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 2 Exemple de requête effectuée sur la base Lexique3.



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 **Erreur ! Source du renvoi introuvable.** présente les résultats obtenus suite à la requête présentée dans la **Erreur ! Source du renvoi introuvable.**.

Figure 3 Résultats obtenus suite à la requête présentée dans la Erreur ! Source du renvoi introuvable.



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

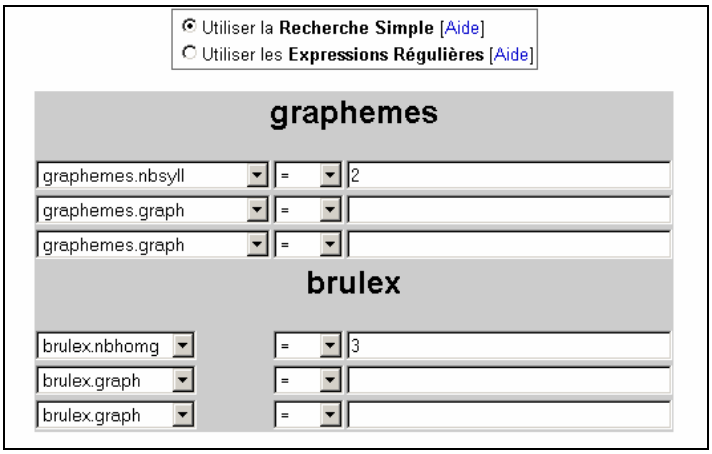
De plus, deux pages html présentent beaucoup d'exemples d'utilisation à la fois de la recherche simple et de la recherche par expressions régulières.

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 **Erreur ! Source du renvoi introuvable.** présente un exemple de requête utilisant *Open Lexique* où nous demandons tous les mots de 2 syllabes selon *Lexique3* qui ont 3 homographes selon *Brulex*.

Figure 4 Exemple de recherche utilisant les possibilités d'*Open Lexique*.

Nous demandons ici tous les mots de 2 syllabes selon Lexique3 qui ont 3 homographes selon *Brulex*.



Pour l'instant, les bases interrogeables en plus des bases de *Lexiqu3* et de *Lexique 2* sont les bases *Manulex* (Lété et al., 2004), la base 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://undows.lexique.org/) 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 **Erreur ! Source du renvoi introuvable.** présente des exemples de requêtes effectuées avec ces outils.

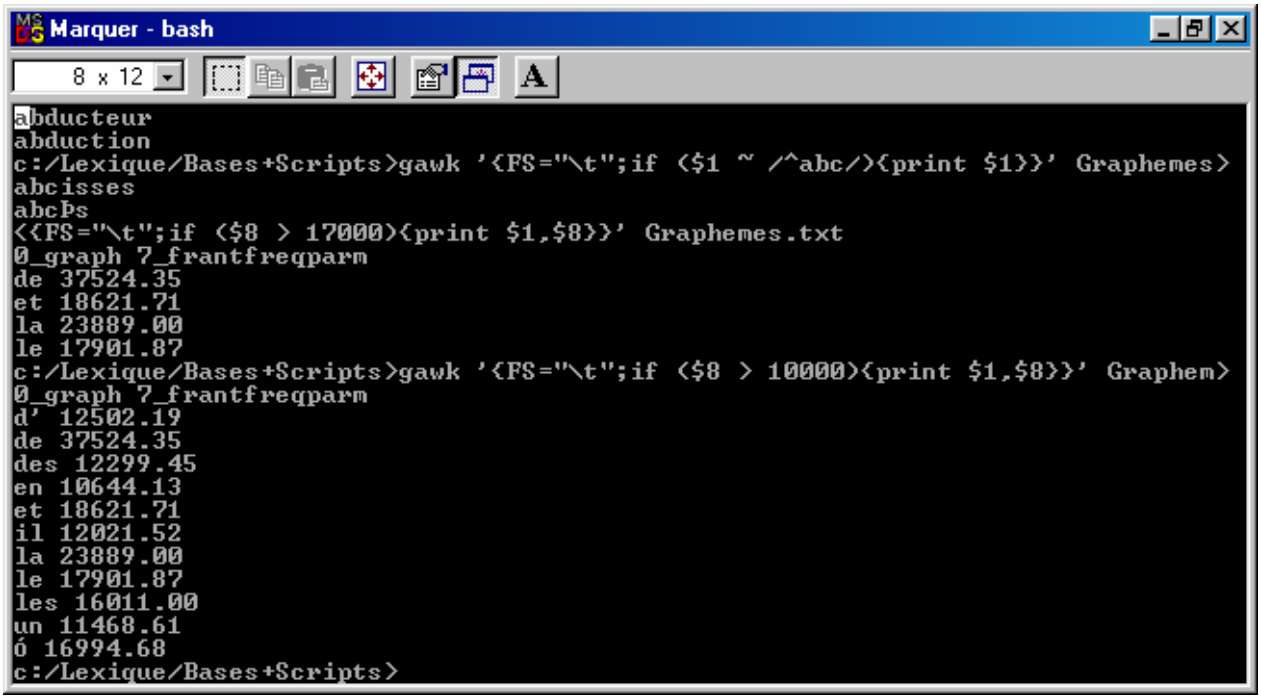


Figure 5 Exemples de requêtes effectué "hors ligne"

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.*

9 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*. 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, 3000 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. Nous avons aussi développé de nouveaux outils permettant aux utilisateurs d'interroger Lexique sans être connectés à internet.

10 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.

11 Conclusion

Depuis plus d'une dizaine d'années, les psycholinguistes travaillant sur l'anglais, l'allemand ou le hollandais disposaient sous réserve d'une modeste contribution de *Celex*, une base de données donnant les fréquences des formes ambiguës grammaticalement, des formes fléchies, et des fréquences des mots à l'écrit et à l'oral. Si *Brulex*

puis *Lexique 1 & 2* ont permis progressivement aux chercheurs désireux de travailler sur le français de pouvoir travailler sur de nouveaux sujets de recherche, *Lexique 3* permet enfin de rattraper ce retard par rapport à nos voisins.

Lexique 3 permet non seulement de rattraper ce retard mais elle dispose aussi de certains avantages: ces estimations de la fréquence d'usage à l'oral sont basés sur un corpus plus important que ceux disponibles jusqu'alors. (19 millions de mots vs 5 millions de mots pour *Celex* anglais).

Lexique 3 dispose aussi d'un nouvel outil permettant de chercher la fréquence de cooccurrence de n'importe quelle suite de mots. A notre connaissance, c'est la première fois qu'un outil de ce type est disponible pour des corpus aussi larges.

Enfin, la façon dont le corpus estimant l'usage de la langue oral permettra d'étendre et de mettre à jour ces fréquences très facilement. En effet la langue était quelque chose de vivant, il est très important de ne pas disposer de fréquences figées mais au contraire d'avoir des fréquences qui suivent l'évolution de cette langue.

Bibliographie

Alario F-X., Ferrand L., Laganaro M., New B., Frauenfelder U., & Segui J. (2004) Predictors of Picture Naming Speed. *Behavior Research Methods, Instruments, & Computers*,36 (1), 140-155.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, 37, 94–117.

Black, A.W. and Lenzo, K. and Pagel, V. (1998). Issues in building general Letter to Sound Rules. *Proceedings of 3rd ESCA/COCSADA Workshop on Speech Synthesis*, 77-81.

Bonin, P., Chalard, M., Méot, A., & Fayol, M. (2001). Age-of-acquisition and word frequency in the lexical decision task: Further evidence from the French language. *Current Psychology of Cognition*, 20, 401-443.

Bonin, P., Barry, C., Méot, A., & Chalard, M. (2004). The influence of age of acquisition in word reading and other tasks: A never ending story? Journal *of Memory and Language*, 50, 456-476.

Coltheart, M., Davelaar, E., Jonasson, J.T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance* (Vol. 6, pp. 535-555). New York : Academic Press.

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.

Dufour, S., Peereman, R., Pallier, C, Radeau, M. (2002). VoColex: A lexical database on phonological similarity between French words. *L'Année Psychologique*, 102, 725-746.

Lambert, E., & Chesnet, D. (2001). NOVLEX: Une base de données lexicales pour les élèves de primaire. *L'Année Psychologique*, 101, 277-288. [Available: http://www2.mshs.univ-poitiers.fr/novlex/]

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 (Edit) et G. Humphreys (Edit), *Basic processes in reading: Visual word recognition*, Hillsdale, NJ, (Lawrence Erlbaum Associates), 148-197.

Morrison C., Ellis A. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, (1), 116-133.

New, B., Brysbaert, M., Segui, Ferrand, L., Rastle, K. (2004) The Processing of singular and plural nouns in French and English. *Journal of Memory and Language*, 51, 568–585.

Pagel, V. and Black, A.W. and Lenzo, K. (1998). Letter-to-Sound Rules for Accented Lexicon Compression. *Proceedings of ICSLP'98*, 252-255.

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. (1996). Le Grand Robert Electronique, Havas Interactive.

Romary L., Salmon-Alt S., Francopoulo G. (2004). Standards going concrete : from LMF to Morphalou. *Workshop on Electronic Dictionaries,* Coling Geneva, Switzerland.

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 2) : Ce document

Lemmes (Lexique 2) : 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 2) : Ce document

Voisins (Lexique) : Page Web de Voisins