

Manuel de Lexique 3

Document version 3.1

Boris New¹, Christophe Pallier²

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

²Unité 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

1.1 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
 - o déterminer dans quelle base se trouve l'information que vous cherchez (le plus souvent c'est la base *Lexique3*)
 - o comprendre comment cette base est structurée (quel sont le ou les champs dont vous avez besoin)
 - o 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.

Comment citer Lexique ?

Lexique étant souvent mis à jour, faites attention à bien citer la version de Lexique que vous avez utilisé. (p.ex. Lexique 3.01 ou Lexique 2.50, etc.)

- New B., Pallier C., Ferrand L., Matos R. (2001) Une base de données lexicales du français contemporain sur internet: LEXIQUE, *L'Année Psychologique*, 101, 447-462. <http://www.lexique.org>
- New, B., Pallier, C., Brysbaert, M., Ferrand, L. (2004) Lexique 2 : A New French Lexical Database. *Behavior Research Methods, Instruments, & Computers*, 36 (3), 516-524.

Historique de cette documentation

- 3.1 Définition des nouveaux champs de Lexique 3.6 et 3.7.
- 3.03 Nouvelle définition des fréquences de sous-titres
Nouvelles définitions pour les champs sur le nombre de définitions
- 3.02 Rajout de "Comment citer Lexique?"
Avertissement concernant la fréquence des sous-titres
- 3.01 Rajout d'explications à propos du code phonétique utilisé
Suppression de la partie sur les fréquences
- 3.00 Introduction de o ouvert/fermé; suppression de o d'origine étrangère
Changement de caractère pour le schwa
Annexe décrivant l'historique du champs "Phonologie"
Définition du nouveau champs cgramlem
- 3.00b3 Nombre de mots des corpora recalculé
Rajout de la catégorie grammaticale LIA (liaison euphonique)
- 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

Introduction rapide pour le nouveau venu	2
1 ETAT DE L'ART DES BASES DE DONNÉES LEXICALES EN FRANÇAIS	6
2 CONSTITUTION DES CORPUS	8
2.1 Le corpus de textes (Frantext)	8
2.2 Le corpus de films (ou corpus de sous-titres)	8
3 ETIQUETAGE GRAMMATICAL DU CORPUS	8
4 ESTIMATION DE LA FIABILITÉ DES FRÉQUENCES	9
5 AVANTAGES DU CORPUS DE SOUS-TITRES	9
6 ORGANISATION DE LA BASE LEXIQUE 3	9
6.1 Organisation de la table <i>Lexique3</i>	10
6.2 Organisation de la table <i>lex3.lemmes.txt</i>	19
7 LES AUTRES BASES	19
8 LES OUTILS	20
8.1 Les outils "en ligne"	20
8.1.1 La recherche de fréquence dans les corpus	20
8.1.2 La recherche par mots	20
8.1.3 La recherche par propriété	21
8.2 <i>Open Lexique</i>	23
8.3 Les outils "hors ligne" : Undows	24
9 DISPONIBILITÉ ET SITE WEB	25
10 LICENCE	25
11 CONCLUSION	25
Bibliographie	27
Annexe A: Open Lexique - Noms des champs	29
Annexe B: Historique de l'obtention des codes phonologiques de Lexique	30

TABLE DES TABLEAUX

Tableau 1 Présentation d'un extrait de <i>Lexique3.txt</i>	13
Tableau 2 Codes phonémiques.....	14
Tableau 3: Codes des catégories grammaticales.....	15
Tableau 4: Nombre et exemples de lemmes selon leur fréquence (corpus de sous-titres).....	15
Tableau 5: Informations complémentaires sur les verbes.....	16
Tableau 6: Nombre de mots dans Lexique 3 en fonction du nombre de syllabes et du nombre de lettres.....	17
Tableau 7 Présentation des opérateurs utilisés dans recherches simples.....	20
Tableau 8 Présentation des opérateurs utilisés dans les expressions régulières.....	21

TABLE DES FIGURES

Figure 1 Exemple de requête de type "Recherche par Mots".....	20
Figure 2 Exemple de requête effectuée sur la base Lexique3.....	21
Figure 3 Résultats obtenus suite à la requête présentée dans la Figure 2.....	22
Figure 4 Exemple de recherche utilisant les possibilités d' <i>Open Lexique</i>	22
Figure 5 Exemples de requêtes effectués "hors ligne".....	23

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

2 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. *MHATLex* (Pérennou et Calmès, 2000) est une base payante qui contient 81 000 lemmes et 854 000 formes fléchies ainsi leurs représentations phonologiques, des informations morpho-syntaxiques et fréquentielles.

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

3 Constitution des corpus

3.1 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 14,7 millions d'items.

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

Ce corpus

Attention ce sont **deux** corpus de sous-titres **différents** qui ont été utilisés pour Lexique avant la version 3.40 et celui utilisé pour la version 3.40 et les versions ultérieures.

Tous les détails sur la constitution de ce corpus sont dans l'article ci-joint :

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (in press). [The use of film subtitles to estimate word frequencies](#). *Applied Psycholinguistics*.

Nous avons téléchargé les sous-titres de 9474 films ou saisons de séries représentant en tout 50 millions de mots.

Ces films ont été classés en 4 catégories :

1. sous-titres de films français : 1.9 millions de mots (e.g., *Camille Claudel*, *C'est arrivé pr`es de chez vous*),
2. sous-titres de films anglo-saxons : 26.5 millions de mots (e.g., *Arizona Dream*, *Schindler's List*),
3. sous-titres de films et séries anglo-saxonnes: 19.5 millions de mots (e.g., *Friends*, *Ally Mc Beal*), and
4. sous-titres de films européens non anglo-saxons 2.5 million words (e.g., *Cria Cuervos*, *Good Bye Lenin!*).

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").

Nous avons calculé la fréquence des mots pour chacun de ces sous-corpus. Une fois ceci effectué, nous avons calculé la fréquence moyenne de chacun de ces mots à travers les 4 corpus. De cette façon, nos fréquences étaient basées sur le corpus le plus large possible et nous évitions une trop forte influence des films et séries américaines très fortement représentés.

L'avantage de ce corpus est de se baser davantage sur le langage oral que le corpus de livres. En effet les sous-titres de films sont composés quasiment uniquement de dialogues très proches de ce qu'entendent les gens à la télévision. Ces fréquences ont montré dans plusieurs analyses un avantage, assez souvent non négligeables, en terme d'explication de la variance des temps de réaction de tâche de décision lexicale par rapport aux fréquences de livres.

3.3 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 *Franctext*; le second repose sur un corpus de sous-titres de films.

5 Avantages du corpus de sous-titres

Un premier 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*, *internauts*, *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;	nd:pre:1s;ind:pre:3s;	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 nbphons	17 cvcv	18 p cvcv	19 voisorth	20 voisphon	21 puorth	22 puphon	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); **freqlemlivres**: 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); **freqlivres**: 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 orthographique inversée; **phonrenv**: forme phonologique inversée; **orthosyll**: forme orthographique syllabée

-Mot (ortho): La graphie est la forme orthographique du mot (p. ex. *chienne*) Attention, les mots correspondent seulement aux mots qui sont apparus au moins une fois dans notre corpus (16 + 50 millions de mots). Il peut ainsi y avoir des lemmes de certains mots apparus dans le corpus qui ne sont pas listés comme entrées indépendante car il n'y sont pas apparus en tant que tels (seul le mot dérivé était dans le corpus). Lexique 2 comprenait 129 000 entrées tandis que Lexique 3 en comprenait 135 000 et Lexique 3.5 142 000.

-Phonie (phon): Représentation phonologique du mot. Les codes phonémiques utilisés sont présentés dans le Tableau 2. L'historique complet de la génération de ce champs est présenté dans l'Annexe B: Historique de l'obtention des codes phonologiques de Lexique.

Tableau 2 Codes phonémiques

Voyelles			Consonnes		
Codes Lexique	Exemples	Sons nommés	Codes Lexique	Exemples	Sons nommés
a	bat, plat	A	p	père, soupe	p (occlusive)
i	lit, émis	I	b	bon, robe	b (occlusive)
y	lu	U	t	terre, vite	t (occlusive)
u	roue	Ou	d	dans, aide	d (occlusive)
o	peau, mot	o (fermé)	k	carré, laque	k (occlusive)
O	éloge, fort	o (ouvert)	g	gare, bague	g (occlusive)
e	été	e-fermé	f	feu, neuf	f (fricative)
E	paire, treize	e-ouvert	v	vous, rêve	v (fricative)
°	abordera	schwa élidable	s	sale, dessous	s (fricative)
2	deux	e-fermé	z	zéro, maison	z (fricative)
9	œuf, peur	e-ouvert	S	chat, tâche	ch (fricative)
5	cinq, linge	in (voy. Nasale)	Z	gilet, mijoter	ge (fricative)
1	un, parfum	un (voy. nasale)	m	main, femme	m (cons. nasale)
@	ange	an (voy. nasale)	n	nous, tonne	n (cons. nasale)
§	on, savon	on (voy. nasale)	N	agneau, vigne	gn (c. nasale palat.)
3	parvenu	schwa non élidable	l	lent, sol	l (liquide)
Semi-Voyelles			R	rue, venir	R
j	yeux, paille	y (semi-voyelle)	x	jota	jota (emprunt espagn.)
8	huit, lui	ui (semi-voyelle)	G	camping	ng (emprunt angl.)
w	oui, nouer	w (semi-voyelle)			

- **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 démonstratif
ADJ:ind	Adjectif indéfini
ADJ:int	Adjectif interrogatif
ADJ:num	Adjectif numérique
ADJ:pos	Adjectif possessif
ADV	Adverbe
ART:def	Article défini
ART:inf	Article indéfini
AUX	Auxiliaire
CON	Conjonction
LIA	Liaison euphonique (l')
NOM	Nom commun
ONO	Onomatopée
PRE	Préposition
PRO:dem	Pronom démonstratif
PRO:ind	Pronom indéfini
PRO:int	Pronom interrogatif
PRO:per	Pronom personnel
PRO:pos	Pronom possessif
PRO:rel	Pronom relatif
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** (*freqlemfilm2*) : Elle correspond à la somme des fréquences des formes fléchies de chaque lemme fournie par notre sélection de films. Ex: freq (*arbre*) = freq ("*arbre*") + freq ("*arbres*")

Pour davantage de détails sur cette fréquence : **New, B.**, Brysbaert, M., Veronis, J., & Pallier, C. (in press). [The use of film subtitles to estimate word frequencies](#). *Applied Psycholinguistics*.

Attention cette fréquence a changé à partir de Lexique 3.40. (Elle est alors basée sur un plus gros corpus et un moyennage sur les sous-corpus)

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

Limite inférieure	Limite supérieure	Nombre de lemmes	Noms	Adjectifs	Verbes	Adverbes
1 000	30 000	117			aller, faire, voir	
100	1 000	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	1 165	secteur, sable, nuage	malin, joyeux, curieux	creuser, exciter	parfaitement, désormais, lentement
10	20	1 137	atmosphère, bouquin, individu	classique, féminin, fidèle	boucher, désigner, étrangler	soudain, clairement, volontiers
1	10	8 800	pupitre, éther	déconcertant, morose	vexer, assouvir, exporter	fièrement, bêtement
0	1	30 730	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 14,7 millions d'occurrences).

- **Fréquence par million selon le corpus de films (*freqfilm2*)** : Elle correspond à la fréquence par million d'occurrences du mot selon notre corpus de sous-titres. 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*). Attention, cette fréquence a changé à partir de Lexique 3.40.

- **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. (14,7 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 (lemmes et formes fléchies) dans Lexique 3 en fonction du nombre de syllabes et du nombre de lettres

F.fléchies		Nombre de syllabes											
		1	2	3	4	5	6	7	8	9	10	13	Tot
N o m b r e d e l e t t r e s	1	46	0	1	0	0	0	0	0	0	0	0	47
	2	122	5	4	0	0	0	0	0	0	0	0	131
	3	630	77	11	2	0	0	0	0	0	0	0	720
	4	1592	1038	9	3	0	0	0	0	0	0	0	2642
	5	2596	4024	222	0	3	0	0	0	0	0	0	6845
	6	2423	8639	1875	5	0	1	0	0	0	0	0	12943
	7	1319	11478	5905	136	1	7	0	0	0	0	0	18846
	8	450	10207	11022	788	3	0	0	0	0	0	0	22470
	9	77	6300	13334	2174	49	2	0	2	0	0	0	21938
	10	8	3050	11511	3693	211	5	1	0	0	1	0	18480
	11	1	1009	7572	4574	463	13	3	0	2	1	0	13638
	12	1	328	3747	4038	728	37	1	0	0	0	0	8880
	13	0	97	1487	2680	848	81	5	1	0	0	0	5199
	14	0	31	484	1406	706	123	7	0	0	0	0	2757
	15	0	6	137	617	468	143	16	0	0	0	0	1387
	16	0	4	39	241	244	121	13	0	0	0	0	662
	17	0	1	21	59	106	76	21	2	0	0	0	286
	18	0	0	6	29	52	46	9	1	0	0	1	144
	19	0	0	4	10	23	31	11	6	0	0	0	85
	20	0	0	0	5	8	7	7	4	0	0	0	31
	21	0	0	0	8	4	6	4	2	0	0	0	24
	22	0	0	0	1	4	4	1	2	0	0	0	12
	23	0	0	0	0	1	2	0	3	0	0	0	6
	24	0	0	0	0	0	0	2	0	1	0	0	3
	25	0	0	0	0	1	0	0	1	0	0	0	2
	Tot		9266	46296	57394	20473	3928	711	108	32	12	12	14

Lemmes		Nombre de syllabes									
N o m b r e d e l e t t r e s		1	2	3	4	5	6	7	8	9	Tot
	1	41	0	1	0	0	0	0	0	0	42
	2	103	5	4	0	0	0	0	0	0	112
	3	486	57	8	1	0	0	0	0	0	552
	4	967	604	8	2	0	0	0	0	0	1581
	5	1211	1957	98	0	0	0	0	0	0	3266
	6	659	3541	695	3	0	1	0	0	0	4899
	7	177	3906	1913	60	0	0	0	0	0	6056
	8	29	2535	3420	336	0	0	0	0	0	6320
	9	2	1152	3802	796	22	0	0	1	0	5775
	10	1	456	2863	1359	109	0	0	0	0	4788
	11	0	174	1447	1591	248	6	0	0	0	3466
	12	0	66	752	1278	393	26	0	0	0	2515
	13	0	34	317	764	435	68	4	0	0	1622
	14	0	9	124	435	324	94	6	0	0	992
	15	0	2	48	225	210	109	15	0	0	609
	16	0	2	11	75	111	88	11	0	0	298
	17	0	0	1	36	58	57	16	1	0	169
	18	0	0	1	20	30	34	6	1	0	92
	19	0	0	1	9	16	22	10	4	0	62
	20	0	0	0	0	3	5	4	3	0	15
	21	0	0	0	0	9	4	1	2	0	16
	22	0	0	0	0	2	0	3	2	0	7
	23	0	0	0	0	0	0	0	2	0	2
	24	0	0	0	0	0	0	1	0	1	2
	25	0	0	0	0	0	1	0	1	0	2
	Tot	3676	14500	15514	6990	1970	515	77	17	1	43260

- **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 ccvccv.

- **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 quelques différences avec l'algorithme de syllabation utilisée sur les formes phonologiques (p.ex. les schwas finaux sont comptés comme des voyelles). Il fait notamment des erreurs sur les mots composés et quand deux voyelles se suivent (*pays, yaourt, voyant, truand*).

- **Catégories grammaticales possibles de la forme orthographique** (*cgramortho*): Champs qui indique les différentes catégories grammaticales possibles pour une représentation orthographique donnée. P.ex. *danse* sera catégorisée comme "NOM,VER" c'est-à-dire un mot qui peut être soit un nom, soit un verbe.

-Pourcentage de personnes connaissant la définition du lemme (*deflem*). Ce champs indique le pourcentage de personnes ayant dit qu'il connaissait le lemme du mot sur le site « [Combien de mots connaissez-vous ?](#) ». Cet indicateur peut donc constituer une sorte de fréquence subjective objective en quelques sortes ;-).

-Nombre de personnes ayant répondu pour la définition de ce lemme (*defobs*). Ce champs indique le nombre de personnes à partir duquel nous avons construit le pourcentage du champs précédent (*deflem*).

-Distance de Levenshtein orthographique (*old20*) Récemment Yarkoni, Balota et Yap (2008) ont remis essayé d'améliorer l'indice de voisinage donné par le nombre de voisins. Pour cela, pour un mot donné, ils ont calculé sa distance de Levenshtein¹ en fonction de tous les autres mots anglais. Enfin, ils ont calculé la distance moyenne des 20 mots les plus proches. Dans une série de régressions multiples ils trouvent que cette distance de Levenshtein moyenne explique davantage de variance que le nombre de voisins orthographiques. Nous avons ici calculé la distance de Levenshtein des 20 mots les plus proches en ayant calculé préalablement la distance de Levenshtein de chaque mot avec les 125 653 entrées orthographiquement différentes de la base *Lexique 3.6*.

-Distance de Levenshtein phonologique (*pld20*) Ce champs est l'équivalent de la distance de Levenshtein orthographique appliqué à la représentation phonologique. (calculé en fonction des 71 358 formes phonologiques différentes de *Lexique 3.6*).

-Morphologie Dérivationnelle (*morphoder*) Ce champs donne la décomposition en morphèmes dérivationnels d'un mot donné. Ainsi *plumage* est décomposé en *plume-age*. Ce champs est le résultat du programme Dérif (Namer, 2003; <http://www.cnrtl.fr/outils/DeriF/>). Attention pour la version actuelle de ce programme de nombreux suffixes et préfixes étant encore non traités ou traités partiellement). Par exemple, *abandonner* n'est pas segmenté comme *abandon-er* mais comme un monomorphémique (*abandonner*). Nous sommes donc vivement intéressés par toute contribution concernant ce champs.

-Nombre de morphèmes (dérivationnels) (*nbmorph*) C'est le nombre de morphèmes dérivationnels directement calculé à partir du champs précédent.

Le Tableau 7 présente en résumé les principales caractéristiques de tous les champs sous formes numériques.

Tableau 7: Minimum, 1^{er} quartile, médiane, moyenne, 3^{ème} quartile et maximum pour les champs numériques

Champs	Min	1er Qu	Médiane	Moyenne	3eme Qu	Max
Fréquence du lemme (films)	0	0.12	1.02	64.83	7.71	33959.88
Fréquence du lemmes (livres)	0	0.54	2.64	48.37	13.31	38943.65
Fréquence des films	0	0	0.06	6.67	0.42	26198.37
Fréquence des livres	0	0.07	0.2	6.74	1.01	38943.65

¹ Distance de Levenshtein: Nombre de transformations (ajout, suppression, remplacement) pour arriver d'une chaîne de caractères à une autre. P.ex. la distance de Levenshtein entre *jupe* et *juge* est de 1 et de 2 entre *jupe* et *juger*.

Nombre d'homographes	1	1	1	1.27	1	6
Nombre d'homophones	1	2	2	3.45	4	24
Nombre de lettres	1	7	9	8.86	10	25
Nombre de phonèmes	1	5	6	6.44	8	20
Nombre de voisins orthographiques	0	0	1	1.43	2	26
Nombre de voisins phonologiques	0	0	2	3.93	5	38
Point d'unicité orthographique	0	0	0	2.38	5	21
Point d'unicité phonologique	0	0	4	3.55	6	17
Nombre de syllabes	1	2	3	2.76	3	9

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

L'équivalent de la base *Lemmes.txt* pour Lexique 2 peut être généré 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
- [Prénoms](#) : 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"

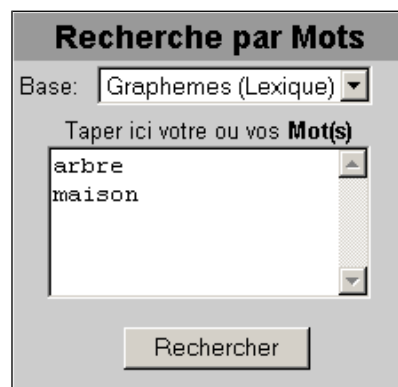
8.1.1 La recherche de fréquence dans les corpus

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 Error: Reference source not found présente un exemple d'un tel type de requête.

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



Recherche par Mots

Base: Graphemes (Lexique) ▼

Taper ici votre ou vos **Mot(s)**

arbre
maison

Rechercher

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 8 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 Error: Reference source not found. Un exemple de recherche complexe utilisant les expressions régulières est la recherche `^[^aeiouyââçèéêôîû]*[aeiouyââçèéêôîû][^aeiouyââçèéêôîû]*$` 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 Error: Reference source not found. 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 9 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	[^aeiouéèîê]	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.

☐ Utiliser la **Recherche Simple** [Aide]
 ☒ Utiliser les **Expressions Régulières** [Aide]

graphemes

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

Trier par le champs graphemes.frantfreqparm Ordre Croissant

Afficher les champs:

graphemes.graph	graphemes.phon	graphemes.cgram
graphemes.phon	Non Spécifié	Non Spécifié

Rechercher Afficher 500 résultats par page

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 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 Error: Reference source not found présente les résultats obtenus suite à la requête présentée dans la Error: Reference source not found.

Figure 3 Résultats obtenus suite à la requête présentée dans la Error: Reference source not found

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 *Error: Reference source not found* 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*.

Utiliser la Recherche Simple [Aide]
 Utiliser les Expressions Régulières [Aide]

graphemes

graphemes.nbsyll = 2
 graphemes.graph =
 graphemes.graph =

brulex

brulex.nhomg = 3
 brulex.graph =
 brulex.graph =

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 *Error: Reference source not found* présente des exemples de requêtes effectuées avec ces outils.

Figure 5 Exemples de requêtes effectués "hors ligne"

```

MS Marquer - bash
8 x 12
abducteur
abduction
c:/Lexique/Bases+Scripts>gawk 'FS="\t";if ($1 ~ /^abc/){print $1}' Graphemes>
abcisses
abcps
<<FS="\t";if ($8 > 17000){print $1,$8}>' Graphemes.txt
0_graph 7_frantfreqparm
de 37524.35
et 18621.71
la 23889.00
le 17901.87
c:/Lexique/Bases+Scripts>gawk 'FS="\t";if ($8 > 10000){print $1,$8}>' Graphem>
0_graph 7_frantfreqparm
d' 12502.19
de 37524.35
des 12299.45
en 10644.13
et 18621.71
il 12021.52
la 23889.00
le 17901.87
les 16011.00
un 11468.61
ó 16994.68
c:/Lexique/Bases+Scripts>

```

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 Nijmegen, *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 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 de combler ce retard il restait encore un certains nombres de caractéristiques absentes pour la langue française (fréquences orales, fréquences des homographes). *Lexique 3* permet non seulement de rattraper ce retard mais apporte aussi un certain nombre de nouveautés par rapport aux bases de données existantes.

D'autre part, les estimations de la fréquence d'usage à l'oral sont basées 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). VoCalex: A lexical database on phonological similarity between French words. *L'Année Psychologique*, 102, 725-746.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113, 256-281.

Keller, E., & Zellner, B. (1998). Motivations for the prosodic predictive chain. *Proceedings of ESCA Symposium on Speech Synthesis*, 76, 137-141.

Lambert, E., & Chesnet, D. (2001). NOVLEX: Une base de données lexicales pour les élèves de primaire. *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.

Namer F. (2003). Automatiser l'analyse morpho-sémantique non affixale: le système DériF. *Cahiers de Grammaire*, Nabil Hathout, Michel Roché et Nicole Serna (éds). Toulouse: ERSS, pp.31-48.

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.

Peereman, R., & Dufour, S. (2003). Un correctif aux notations phonétiques de la base de données LEXIQUE [A corrective to the phonetic notations of the LEXIQUE database]. *L'Année Psychologique*, 103, 103-108.

Pythoud, C. (1996). Problèmes de la correction automatique de l'orthographe lexicale du Français à travers une étude de cas: Le correcteur orthographique ispell et le dictionnaire Français-IREQ [Automatic spell-checking problems: The ispell program and the 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.

Yarkoni, T, Balota, D.A., & Yap, M.J. (2008). Moving Beyond Coltheart's N: A New Measure of Orthographic Similarity. *Psychonomic Bulletin & Review*, 15, 971-979.

Annexe A: Open Lexique - 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](#)

Fréquence subjective et valeur d'imagerie (Desrochers & Bergeron): [Ce document](#)

Annexe B: Historique de l'obtention des codes phonologiques de Lexique

Le problème de l'obtention des codes phonologiques de Lexique provient du fait que Lexique a toujours compris un grand nombre de formes fléchies. Nous n'avons donc pas pu utiliser de dictionnaire comme c'était le cas pour Brulex.

Lexique 1

Pour *Lexique 1*, nous avons dérivé la forme phonologique de nos entrées grâce au logiciel [LAIPPTS 1.13](#) (Keller & Zellner, 1998). 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 %. Cependant, ce logiciel (*LAIPPTS*) était un logiciel prévu pour générer de la parole à partir de textes continus et non de mots isolés (cadre dans lequel nous l'avons utilisé).

Lexique 2

Pour *Lexique 2*, Peereman et Dufour (sous presse) ont examiné, les codes phonémiques de *Lexique 1* 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*.

Lexique 3

Pour *Lexique 3*, les représentations phonologiques ont été obtenues à partir de *Lexique 2* 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

Lexique 3.2

Pour *Lexique 3.2*, Christian Lachaud a effectué un grand nombre de corrections. Il a tout d'abord réintroduit après un gros travail de vérification manuelle la différence entre o ouvert (noté o) et o fermé (noté O). Il a supprimé la présence du phonème *h* qui indiquait la possibilité d'une liaison ou pas. En outre, il a corrigé environ 1400 entrées (1% du Lexique) pour des problèmes diverses:

Pourquoi ces codes phonétiques ?

Nous avons commencé par utiliser les codes du premier logiciel de text2speech que nous avons employé. (LAIPPTS). Ensuite nous avons légèrement changé ces codes pour le rendre plus pratique (remplacement du schwa * par ° pour que les recherches soient plus faciles)

Nous n'utilisons pas API car nos fichiers sont des fichiers textes (nous passerons à l'unicode, une fois que celui-ci sera parfaitement supporté par toutes les applications), et pas X-Sampa car nous voulons que chaque phonème soit représenté par un seul caractère. Nous voulons aussi que tous ces caractères soient facilement utilisables par des anglo-saxons. Normalement, notre code phonétique est relativement proche de X-Sampa. Voici les changements de X-Sampa vers Lexique :

X-Sampa	Lexique
e~	5
a~	@
o~	§
9~	1
A	a
N	G
J	N
H	8
@	°