L'Année psychologique, 2001, *101*, 447-462

NOTE MÉTHODOLOGIQUE

*Laboratoire de psychologie expérimentale**
Université René-Descartes, Paris 5
CNRS UMR 8581[1]

*Laboratoire des sciences cognitives et psycholinguistique***
École des hautes études en sciences sociales (EHESS)
CNRS UMR 8554[2]

UNE BASE DE DONNÉES LEXICALES DU FRANÇAIS CONTEMPORAIN SUR INTERNET : LEXIQUE™

par Boris NEW*[3], Christophe PALLIER**,
Ludovic FERRAND* et Rafael MATOS*[4]

SUMMARY : A lexical database for contemporary french on internet : LEXIQUE

We present a new lexical database of French, named Lexique. Based on a corpus of texts written since 1950 which contained 31 million words, Lexique yields 130 000 entries including the inflected forms of verbs, nouns and adjectives. Each entry provides several kinds of information including frequency, gender, number, phonological form, graphemic and phonemic unicity points. Several tables give additional statistics such as the frequencies of various units : letters, bigrams, trigrams, phonemes and syllables. The database is available for free on the Internet.

Key words : word recognition, database, frequencies.

1. 71, avenue Édouard-Vaillant, 92774 Boulogne-Billancourt Cedex.
2. 54, boulevard Raspail, 75270 Paris Cedex 06.
3. E-mail : new@psycho.univ-paris5.fr.

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

Cet article décrit une base de données lexicales du français, dont les points forts sont les suivants :

— Elle est fondée sur des textes publiés entre 1950 et 2000 provenant du corpus Frantext de l'ATILF[1]. Ce corpus comprend 31 millions de mots.
— Elle inclut, entre autres, les formes fléchies des mots (formes verbales conjuguées, formes plurielles et féminines des noms et adjectifs).
— Deux estimations de fréquence sont fournies : l'une fondée sur le corpus original de Frantext, et l'autre sur les pages web françaises indexées par le moteur de recherche FastSearch[2].
— Elle est organisée autour de deux tables qui ont pour clés principales, soit les formes orthographiques soit les lemmes (un lemme est le mot choisi pour représenter toute une famille de formes apparentées. Par exemple : *manger* est le lemme de *mangea*, *mangeait*..., etc.).
— Elle fournit de nombreuses informations fréquentielles concernant les lettres, les bigrammes, les trigrammes, les phonèmes et les syllabes.
— Elle est gratuite, libre d'accès, téléchargeable, et des outils sont fournis pour l'interroger.
— Elle est actualisée et peut être mise à jour dans cinq ou dix ans.

Pendant longtemps, les psycholinguistes ont sélectionné manuellement le matériel verbal dans le Trésor de la langue française (Imbs, 1971). Leur travail a été grandement facilité quand Content, Mousty et Radeau (1990) ont mis à leur disposition BRULEX, une base de données informatisée regroupant les 35 746 entrées lexicales du *Petit Robert* et leurs fréquences selon le TLF. Ces fréquences étaient estimées sur un corpus de textes littéraires datant de 1919 à 1964 et comprenant 26 millions de mots. Une limitation notable de Brulex était l'absence des formes fléchies telles que les verbes conjugués ou certaines formes écrites plurielles ou féminines. Cela pose problème par exemple pour estimer des fréquences d'unités telles que les syllabes. Nov-

1. Laboratoire d'analyses et traitements informatiques du lexique français (cf. http://www.inalf.fr).
2. http://www.alltheweb.com.

lex, une base de données plus récente (Lambert et Chesnet, 2001) fournit les formes fléchies mais se fonde sur un corpus spécialisé de textes pour enfants de 417 000 mots. C'est pourquoi nous avons entrepris de construire une nouvelle base de données avec des estimations de fréquences plus complètes, plus actuelles, et comprenant les formes fléchies.

DESCRIPTION DU CORPUS ORIGINAL

Afin de constituer la base initiale de mots, nous avons sélectionné dans la base Frantext tous les textes publiés entre 1950 et 2000 : cela représentait un corpus de 31 millions d'items. Frantext est une base de données textuelles regroupant 3 200 textes représentatifs du français des XIX^{e} et XX^{e} siècles, développée par l'INALF-Nancy, devenu aujourd'hui l'ATILF et accessible à l'adresse : http://zeus.inalf.fr/frantext.htm. Ces textes étaient essentiellement des romans, mais comprenaient également quelques recueils de poésie, des essais et des traités scientifiques ou techniques. Nous avons obtenu une liste de 246 000 items distincts ainsi que leur fréquences[1]. Ces items comprenaient des symboles (dont la ponctuation), des abréviations, des mots étrangers et des noms propres. Pour nettoyer cette liste, nous avons employé le dictionnaire *Francais-Gutenberg 1.0*[2] (Pythoud, 1996), le logiciel Ispell et le dictionnaire *Le Grand Robert* (Robert, 1996). Le résultat de ce filtrage a produit une liste de 130 000 items ayant des formes orthographiques distinctes.

CALCUL DES FRÉQUENCES

La fréquence des mots joue un rôle fondamental dans la plupart des tâches psycholinguistiques (voir Monsell, 1991 pour une

1. Le logiciel d'interrogation ne traitait malheureusement pas correctement les noms composés : un mot comme « garde-manger » était identifié comme deux items distincts « garde » et « manger ».

2. http://www.unil.ch/ling/cp/frgut.htlm.

synthèse). De nombreuses études ont montré que les performances étaient meilleures pour les mots de haute fréquence que pour les mots de basse fréquence, que cela soit en termes de nombre d'erreurs ou de temps de réaction. Cependant, d'autres facteurs comme l'âge d'acquisition, ou la familiarité, généralement très corrélés avec la fréquence d'usage, interviennent (Morrison et Ellis, 1995 ; Connine *et al.*, 1990). Pour décorréler ces différents facteurs, il est primordial d'avoir de bonnes estimations de chacun d'entre eux.

Dans *Lexique*, nous proposons deux estimateurs des fréquences d'usage : le premier est fondé sur le corpus initial de Frantext, constitué de textes littéraires ; le second est fondé sur le nombre de pages web françaises contenant un mot donné. Ce deuxième estimateur, fondé sur quinze millions de pages web,

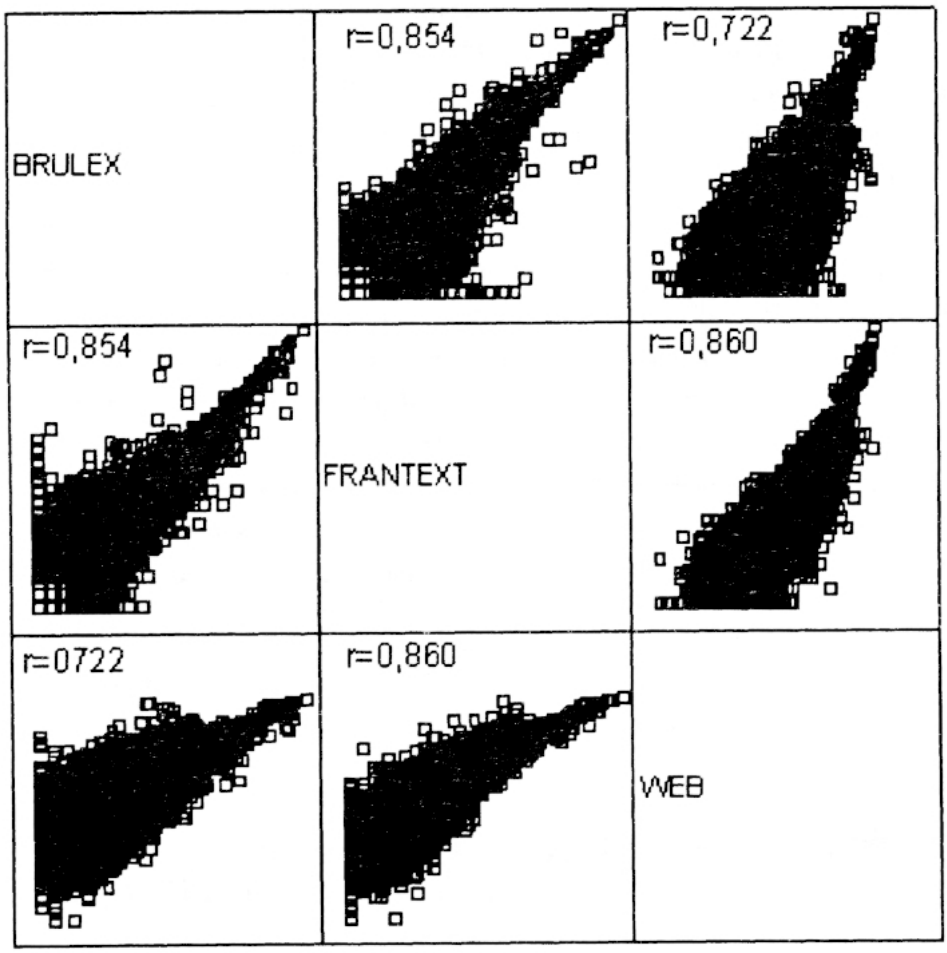

Fig. 1. — Matrices de corrélations entre les différentes bases de données Brulex, Frantext et le Web

Matrix of correlations between the different database Brulex, Frantext and the Web

nous a paru constituer une source d'information supplémentaire sur l'usage du français.

Plus précisément, nous avons soumis au moteur de recherche FastSearch (http: www.alltheweb.com), les 130 000 formes orthographiques obtenues à partir du corpus Frantext. L'interrogation était effectuée sur les 15 millions de pages françaises répertoriées, en mode SafeSearch pour éviter la sur-représentation des mots à connotation sexuelle. Pour chaque mot a été obtenu le nombre de pages dans lesquelles celui-ci apparaissait ; il ne s'agit donc pas exactement de la fréquence lexicale de la forme, mais néanmoins d'un estimateur de l'usage de ce mot. Par exemple, des mots tels que *publicité*, *entreprise* ou *télévision* se retrouvent avec des fréquences comparables à celles de mots tels que *champ*, *arbre* ou *chaise* selon FastSearch, mais avec des fréquences très divergentes selon Frantext. D'autres items tels que *kiwi* sont extrêmement rares selon Brulex ou Frantext alors que FastSearch les considère, de façon plus réaliste, comme « plutôt rares ». Pour comparer ces deux estimations de fréquence entre elles et par rapport aux fréquences du TLF, nous avons construit le diagramme de corrélation de la figure 1 à partir du logarithme des fréquences de 23 440 items selon le TLF, Frantext et FastSearch.

OBTENTION DES AUTRES DESCRIPTEURS

Pour obtenir la catégorie grammaticale, le genre, le nombre et le lemme des mots, nous avons utilisé conjointement le *Grand Robert*, et les deux lemmatiseurs : *Tree Tagger*[1] de Helmut Schmid et *Flemm*[2] *2.0* (Namer, soumis). En effet, aucune de ces sources seules permettait d'avoir une information suffisamment complète.

Dans une troisième étape, nous avons dérivé la forme phonologique de nos entrées grâce au logiciel *LAIPTTS 1.13*[3]. Ce logiciel utilise un noyau de 500 règles de conversion gra-

1. http://www.univ-nancy2.fr/pers/namer/.
2. http://www.ims.uni-stuttgart.de/projekte/corplex/DecisionTreeTagger.html.
3. http://www.unil.ch/imm/docs/LAIP/LAIPTTS.html.

phème-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 %.

ORGANISATION DE LA BASE

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

— *Graphèmes.txt* : une base organisée à partir des formes orthographiques.
— *Lemmes.txt* : une base organisée à partir des lemmes. Nous avons choisi la forme « infinitif » pour les verbes, et la forme « masculin singulier » pour les participes passés, adjectifs et noms.
— *Surface.txt* : un fichier qui résume les statistiques fréquentielles concernant les lettres, bigrammes, trigrammes, phonèmes et syllabes pour chaque mot.

Ces tables sont fournies sous forme de fichiers textes, les champs étant séparés par des tabulations. Cela permet de les importer facilement avec la plupart des logiciels. Deux dossiers supplémentaires, Surface et Outils, contiennent respectivement des informations fréquentielles détaillées à propos des lettres, bigrammes, trigrammes, phonèmes et syllabes, et des outils facilitant l'utilisation des tables.

ORGANISATION DE LA TABLE « GRAPHÈMES »

Voici les différents champs de cette table (tableau I).

— Graphie (graph) :

La graphie est la forme orthographique du mot (par ex. « chienne »).

TABLEAU I. — *Graphèmes.txt*
Sample of the file grapheme.txt

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

Légende. — **graph :** le mot ; **phon :** la forme phonologique du mot ; **cgram :** les catégories grammaticales de ce mot ; **genre :** le genre ; **nombre :** le nombre ; **lemme :** les lemmes de ce mot ; **freqfrant :** les fréquences de frantext par million d'occurences ; **freqweb :** les fréquences de fastsearch (web) par million de pages ; **nblettres :** le nombre de lettres ; **nbphons :** nombre de phonèmes ; **cvcv :** la structure orthographique ; **p-cvcv :** la structrure phonologique ; **puorth :** point d'unicité orthographique ; **puphon :** point d'unicité phonologique ; **syll :** forme phonologique syllabée ; **nbsyll :** nombre de syllabes ; **cv-cv :** structure phonologique syllabée.

— Phonie (phon) :

Les codes phonémiques utilisés sont présentés dans le tableau II.

TABLEAU II. — *Codes phonétiques*

Phonetic codes

Symbole	Exemples	Sons nommés
I	lit, émis	i-fermé
Y	lu	u-fermé
e	Eté	e-fermé
2	(deux) bleu	eu-fermé
E	Treize	e-ouvert
5	(cinq) cinq, linge	in (voy. nasale)
9	(neuf) neuf, oeuf	eu-fermé
1	(un) un, parfum	un (voy. nasale)
a	tabac	a-ouvert
A	il bat	a-fermé
@	ange	an (voy. nasale)
o	galop	o-fermé
O	éloge	o-ouvert
§	on, savon	on (voy. nasale)
u	roue	ou-fermé
*	premier	schwa d'expiration
%	alpes	schwa obligatoire (enlevé en fin de mots)
j	yeux, paille	y (semi-voyelle)
8	(huit) huit, lui	u (semi-voyelle)
w	oui, nouer	w (semi-voyelle)
p	père, soupe	p (occlusive)
b	bon, robe	b (occlusive)
m	main, femme	m (cons. nasale)
f	feu, neuf	f (fricative)
v	vous, rêve	v (fricative)
t	terre, vite	t (occlusive)
d	dans, aide	d (occlusive)
n	nous, tonne	n (cons. nasale)
N	agneau, vigne	gn (c. nasale palat.)
k	carre, laque	k (occlusive)
g	gare, bague	g (occlusive)
s	sale, dessous	s (fricative)
z	zero, maison	z (fricative)
S	chat, tâche	ch (fricative)
Z	gilet, mijoter	ge (fricative)
l	lent, sol	l (liquide)
R	rue, venir	r grassaye
r	rue, venir	r roule
h	hop!	h aspire
s	les haricots	arrêt glottique
x	jota	jota (emprunt espagn.)
G	camping	ng (emprunt angl.)
¤	abjureras	rr

— Classe grammaticale (cgram) :

Si une même entrée pouvait appartenir à plusieurs classes grammaticales différentes, celles-ci ont été séparées par un

point-virgule. Les différents codes utilisés pour représenter les catégories grammaticales sont présentés dans le tableau III.

TABLEAU III. — *Codes des catégories grammaticales*

Codes for syntactic categories

Abréviations	Signification
ABR	Abréviations
ADJ	Adjectif
ADV	Adverbe
CONJ	Conjonction
DET	Déterminant
INT	Interjection
NOM	Nom
NUM	Numéral
PRE	Préposition
PRO	Pronom
PRO:pers	Pronom personnel
PRO:poss	Pronom possessif
PRO:rela	Pronom relatif
SYM	Symbole
VER	Verbe
Ind	Indicatif
Cond	Conditionnel
Futu	Futur
Sub	Subjonctif
Infi	Infinitif
Imp	Impératif
Pr	Présent
Impf	Imparfait
Ps	Passé simple
Pper	Participe passé
Ppre	Participe présent

— Genre (genre) :

Il correspond au genre de l'item lexical :

m → masculin ;

f → féminin ;

é → épicène.

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

— Nombre (nombre) :

Les codes utilisés pour représenter le singulier, le pluriel, etc., sont indiqués dans le tableau IV.

— Lemme (lem) :

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

TABLEAU IV. — *Codes du champ nombre*

Codes for number

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

— Nombre de lettres (nbgraphs).

— 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 » sera représentée par ccvvccv.

— Structure de la forme phonologique (p-cvcv) :

C'est un découpage du mot en voyelles (V) et consonnes (C) selon sa représentation phonologique.

— Point d'unicité orthographique (pugraph) :

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.

— 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.

— Syllabation (syll) :

Les formes phonologiques ont été syllabées selon un algorithme décrit dans Pallier et New (en préparation).

— Nombre de syllabes (nbsyll).

— Structure phonologique syllabique (cv-cv) :

Elle décrit la structure phonologique du mot syllabé. Les consonnes sont notées C, les voyelles sont notées V et les semi-voyelles Y.

— Nombre aléatoire (rand) :

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

— Fréquence par million selon Frantext (frantfreqparm) :

Elle correspond à la fréquence fournie par Frantext, normalisée par une division par 31 (le corpus original comprenant 31 millions de tokens). La somme de ce champs ne fait pas un million en raison du premier filtrage effectué.

— Fréquence par million de pages selon FastSearch (fsfreqparm) :

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

ORGANISATION DE LA TABLE « LEMMES »

— Lemme (lem) :

Cette base est organisée selon ce champs qui est le lemme.

— Graphies (graph) :

Ce champs présente les graphies des formes fléchies associées à ce lemme. Ainsi pour le lemme « chien », les graphies sont « chien », « chienne », « chiens » et « chiennes ».

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

— Phonies (phon).
— Classes grammaticales (cgram).
— Genre (genre).
— Nombre (nombre).
— La fréquence cumulée du lemme selon Frantext (frantfreqcum) :

C'est la somme des fréquences des formes orthographiques (calculées ci-dessous).

— La fréquence des formes orthographiques selon Frantext (frantfreqgraph) :

TABLEAU V. — *Extraits de Lemmes.txt* — Examples from Lemmes.txt

lem	graph	phon	cgram	genre	nombre	freqfrantcum	freqfrantgraph	freqwebcum	freqwebgraph
danse	danse;danses	d@s;d@s	NOM;VER:imp:pr;ind:pr	f	s;1s;3s;2s	63.9	49.71;14.19	13148.23	10745.56;240
danser	danse	d@se	ADJ;NOM;VER:cond:pr	f;m	1s;3s;2s;2	116		18057	
danseur	danseur;danseurs;danseuse;	d@s9R;d@s9R;d	NOM	m;f	s;(p)	27.13	6.94;7.87;6.58;	3238.4	602.54;1440.
dansoter	dansota;dansotait;dansotter	d@sOta;d@sOtE;	VER:ind:impf;ind:ps;infi		3s	0.12	0.03;0.06;0.03	0.21	0;0.14;0.07
dansé	dansé;dansée;dansées;dans	d@se;d@se;d@s	ADJ;VER:pper	f;m	s;(p);p	4.06	3.16;0.35;0.10;	488.44	367.81;53.31;
dantesque	dantesque;dantesques	d@tEsk;d@tEsk	ADJ	é	(p)	0.25	0.19;0.06	83.99	55.69;28.30
danubien	danubien;danubienne;danubi	danybj5;danybjEn;	ADJ	f;m	(p)	0.39	0.10;0.13;0.10;	63.11	23.05;19.26;1
daphnie	daphnies	dafni	NOM	f	(p)	0.06	0.06	23.75	23.75
daphné	daphné	dafne	NOM	m	s	0.06	0.06	153.26	153.26
dard	dard;dards	daR;daR	NOM	m	s;(p)	2.03	1.35;0.68	393.45	304.35;89.10
dardant	dardant;dardantes	daRd@;daRd@t	ADJ;VER:ppre	m;f	3p;p;(p)	0.68	0.65;0.03	22.2	21.99;0.21
darder	darda;dardaient;dardait;darda	daRda;daRdE;da	ADJ;VER:imp:pr;ind:im	é;f;m	2s;1s;3s;3	2.5	0.10;0.06;0.32;	163.55	9.60;4.97;18.
dardillon	dardillon	daRdij§	NOM	m	s	0.03	0.03	0.56	0.56
dardé	dardé;dardée;dardées;dardés	daRde;daRde;daR	ADJ;VER:pper	é;f;m	s;(p);p	0.96	0.32;0.35;0.19;	26.05	17.02;3.64;1.
dargeot	dargif	daRZif	NOM	m	s	0.03	0.03	0.7	0.70
dariole	darioles	daRjOl	NOM	f	(p)	0.06	0.06	6.51	6.51
darique	darique;dariques	daRik;daRik	NOM	f	s;(p)	0.22	0.06;0.16	5.95	1.26;4.69
darne	darne	daRn	ADJ;NOM	é;f	s	0.06	0.06	95.89	95.89
daron	daron;daronne;daronnes;daro	daR§;daROn;daR	NOM	f;m	s;(p)	1.22	0.42;0.48;0.06;	15.27	12.75;2.03;0.
darse	darse;darses	daRs;daRs	NOM	f	s;(p)	0.58	0.32;0.26	58.42	45.53;12.89
dartre	dartres	daRtR	NOM	f	(p)	0.13	0.13	17.86	17.86

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

TABLEAU VI. — *Gros plan sur un verbe : « abaisser»* — Zoom on the verb « abaisser », to pull down

Abaisser	abaissa;abaissai;abaissaient;abaissait;abaissant;abaisse;abaissent;abaisser;abaissera;abaisserai;abaisseraient;abaisserait;abaisses;abaissez;abaissons;abaissât;abaissèrent;abaissé;abaissée;abaissées;abaissés						abEsa;abEsE;abEsE;abEsE;abEs@;abEs;abEs;abese;abEsRa;abEsRE;abEsRE;abEsRE;abEs;abEse;abEs§;abEsA;abEsER;abese;abese;abese;abese		
ADJ;NOM;VER:cond:pr;imp:pr;ind:futu;ind:impf;ind:pr;ind:ps;infi;pper;ppre;sub:impf;sub:pr	f;m	1s;2s;2p;1p;3s;3p;s;(p);p	658	45;2;8;40;74;167;42;138;3;3;4;6;2;3;1;1;7;66;24;4;18	45732	761;62;259;625;3560;7960;1730;16800;576;72;66;258;332;1190;120;13;143;6100;2820;855;1430			

Ce sont les fréquences des formes fléchies du lemmes. Ainsi le lemme « arbre » ayant deux formes fléchies « arbre » et « arbres », nous affichons 8004.64 ; 8448.17.

— La fréquence cumulée du lemme selon FastSearch (fsfreqcum).

— La fréquence des formes orthographiques selon FastSearch (fsfreqgraph).

ORGANISATION DE LA TABLE « SURFACE »

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

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

Par exemple la fréquence cumulée (token) des lettres de *perruche* correspondra à la moyenne des fréquences de *p* en première position, *e* en deuxième, etc.

Les fréquences pondérées sont toutes données en occurences par million.

Mot	*GrPond*	*GrPondEt*	*BigrPond*	*BigrrPondEt*	...
Perruche	1 494 397,25	1 614 786,74	99 154,57	116 978,78	

ORGANISATION DU DOSSIER SURFACE

Le dossier surface comprend des fichiers donnant des statistiques sur les lettres, bigrammes, trigrammes, phonèmes et syllabes calculées à partir de la table « Graphèmes ».

Il comprend 5 sous-dossiers correspondant chacun à une unité d'analyse : lettres, bigrammes, trigrammes, phonèmes et syllabes.

Chaque dossier est organisé de la même façon et comprend le même type de fichiers. Nous allons ici décrire le dossier concer-

nant les informations à propos des lettres mais l'organisation des autres dossiers est en tout point similaire à celui-ci.

FreqGr.txt

Exemple :

a	11 033 ; 1 743 071	18 307 ; 3 283 403	...

Cela signifie que « a » en première position est apparu 11 033 fois et qu'il a une fréquence pondérée de 1 743 071. Puis nous présentons ces statistiques pour la lettre « a » en deuxième position, etc.

GrMots.txt

Exemple :

a-b-a-i-s-s-a	11 033 ; 1 743 071	1 382 ; 83 367	...
e-t	6 832 ; 1 919 822	2 187 ; 994 764	

Il donne pour chaque mot les statistiques de chacun de ses lettres présentées dans le fichier FreqGr.txt.

GrMotsSomme.txt

Il donne pour chaque mot les moyennes pour l'ensemble de ses lettres.

Exemple :

e-t	4 509,50	1 457 293	2 322,50	462 529	2

1re col. : Mot ;

2e col. : Moyenne (4 509) de 6 832 + 1 382 (nombre de fois où e en 1re pos. + t en 2e pos.) ;

3e col. : Moyenne (1 457 293) de 1 919 822 + 994 764 (fréquence pondérée des lettres par pos.) ;

4e col. : Écart-type du nombre de fois ;

5e col. : Écart-type de la fréquence pondérée ;

6e col. : Nombre de lettres.

SommeFreqGr

Il donne pour chaque lettre les moyennes pour toutes ses positions

a	5 555,06	492 034	5 471,87	819 449	18

2e col. : Moyenne du nombre de fois toutes positions confondues ;

3e col. : Moyenne de fréquence pondérée toutes positions confondues ;

4e col. : Écart-type du nombre de fois toutes positions confondues ;

5e col. : Écart-type de fréquence pondérée toutes positions confondues.

CALCULS À PARTIR DE LA DERNIÈRE POSITION

Les dossiers lettres, bigrammes, trigrammes, phonèmes et syllabes contiennent tous un dossier DER (grder pour le dossier bigrammes par ex.) où se trouvent les mêmes fichiers mais avec un calcul commençant par la dernière unité utilisée. Ainsi freqgrder.txt présente la même information que freqgr.txt mis à part le fait que la première colonne correspond aux statistiques de la lettre en dernière position, la deuxième colonne à l'avant-dernière position, etc.

DISPONIBILITÉ

La base de données LEXIQUE peut être consultée et téléchargée sous forme compressée (zip) à partir du site *http://www.lexique.org*. Ce site contient diverses informations et outils. Un forum de discussion lexique-psycho@yahoogroups.com permet aux utilisateurs de poser des questions ou de proposer des améliorations. Étant donné que Frantext et FastSearch sont deux bases de données régulièrement actualisées, il sera facile de mettre *Lexique* à jour dans cinq ou dix ans.

RÉSUMÉ

Cet article présente une nouvelle base de données lexicales du français : Lexique. Fondée sur un corpus de textes écrits entre 1950 et 2000 contenant 31 millions de formes orthographiques, la base de données comprend 130 000 entrées incluant les formes fléchies (formes conjuguées des verbes, formes féminines ou plurielles des noms ou adjectifs). Chaque entrée fournit plusieurs informations dont la fréquence, le genre, le nombre, la forme phonologique canonique, les points d'unicité orthographiques et phonologiques. Des tables supplémentaires donnent les fréquences de diverses unités : lettres, bigrammes, trigrammes, phonèmes et syllabes. Cette base de données est accessible librement et téléchargeable par Internet.

Mots-clés : reconnaissance de mots, fréquence, base de données.

BIBLIOGRAPHIE

Connine C., Mullennix J., Shernoff E., Yelen J. — (1990) Word familiarity and frequency in visual and auditory word recognition, *Journal of Experimental Psychology : Learning, Memory and Cognition, 16* (6), 1084-1096.

Content A., Mousty P., Radeau M. — (1990) BRULEX : Une base de données lexicales informatisée pour le français écrit et parlé, *L'Année Psychologique*, 90, 551-566.

Imbs P. — (1971) *Études statistiques sur le vocabulaire français. Dictionnaire des fréquences, Vocabulaire littéraire des XIX^e^ et XX^e^ siècles*, Centre de la recherche pour un trésor de la langue française (CNRS), Nancy, Paris, Librairie Marcel-Didier.

Lambert E., Chesnet D. — (2001) Novlex : une base de données lexicales pour les élèves de primaire, *L'Année Psychologique*, 01, 277-288.

Monsell S. — (1991) The nature and locus of word frequency effects in reading, *in* D. Besner et G. Humphreys (Edit), *Basic processes in reading : Visual word recognition*, Hillsdale (NJ), Lawrence Erlbaum, 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.

Pallier C., New B. — (en préparation) *Un syllabaire de la langue française.*

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, *Mémoire de licence*, Université de Lausanne.

Robert P. — (1996) *Le Grand Robert électronique*, Havas Interactive.

(Accepté le 19 mars 2001.)