Notes

Le document suivant reprend l'article décrivant Brulex qui a été publié dans la revue L'Année Psychologique, 1990, *90*, 551-566.
Quelques modifications ont été apportées lors de la création de la version pour Apple Macintosh:

1. Les codes alphabétiques et phonétiques sont légèrement différents dans les deux versions. Un tableau reprend l'ensemble des codes utilisés dans les deux versions. Les codes phonétiques ont été choisis pour être lisibles avec la police standard GENEVA.

2. La version MAC est fournie uniquement sous la forme d'un fichier texte ASCII standard. Les utilisateurs auront donc le loisir de l'importer directement dans le logiciel de traitement de bases de données de leur choix. L'organisation des champs est conforme à la description donnée au tableau I de l'article, et les séparateurs sont des marques de tabulation.

3. Les adresses Email signalées dans l'article n'existent plus: elles sont remplacées par

acontent@ulb.ac.be
pmousty@ulb.ac.be
moradeau@ulb.ac.be

4. Les utilisateurs sont priés de renvoyer la fiche d'engagement complétée et signée à l'adresse suivante;

BRULEX
Laboratoire de Psychologie Expérimentale CP 191
Avenue F.D. Roosevelt, 50
B-1050 Bruxelles
Belgique

BRULEX
Une base de données lexicales informatisée pour le français écrit et parlé

Alain Content, Philippe Mousty et Monique Radeau
Université Libre de Bruxelles

Résumé

Cet article présente un outil développé pour la recherche en psycholinguistique. Brulex donne, pour environ 36.000 mots de la langue française, l'orthographe, la prononciation, la classe grammaticale, le genre, le nombre et la fréquence d'usage. Il contient également d'autres informations utiles à la sélection de matériel expérimental (notamment, point d'unicité, comptage des voisins lexicaux, patrons phonologiques, fréquence moyenne des digrammes).

BRULEX
Une base de données lexicales informatisée pour le français écrit et parlé[1]

Alain Content, Philippe Mousty et Monique Radeau[2]
Université Libre de Bruxelles

October 11, 2000

L'apparition de micro-ordinateurs plus puissants permet actuellement de réaliser des applications qui nécessitaient auparavant l'utilisation de systèmes centraux multi-utilisateurs. Les micro-ordinateurs s'avèrent avantageux parce qu'ils sont à la fois plus souples et plus simples d'accès pour l'utilisateur. La rapidité de traitement, les capacités de stockage sur mémoire externe, et les logiciels spécifiques disponibles permettent notamment d'envisager la réalisation et la gestion de bases de données de grande taille.

Cette solution présente plusieurs particularités intéressantes. Les applications restent extensibles de manière souple et à peu de frais, ce qui permet d'envisager le développement de manière progressive. Elles offrent des possibilités de consultation et d'exploitation dans des délais raisonnables, y compris dans un mode interactif d'utilisation. Elles peuvent s'appuyer sur des logiciels répandus, qui fournissent d'emblée à l'utilisateur un nombre important de primitives spécifiquement conçues pour ce type

---

[1] Ce travail a été réalisé grâce à l'aide du Ministère Belge de la Politique Scientifique, Services de Programmation de la Politique scientifique (Action de Recherche Concertée "Processus cognitifs dans la lecture, 1984-1990); du Programme National d'Impulsion à la Recherche fondamentale en Intelligence Artificielle (Projet "Lexical Processes", 1987-1991), ainsi que du Fonds National Belge de la Recherche Fondamentale Collective (Conventions 2.4505.80 et 2.4532.88). La responsabilité scientifique et technique incombe aux auteurs.

Nous remercions très chaleureusement toutes les personnes qui ont contribué, à un stade ou à un autre, au développement de Brulex, à savoir, Dominique Brodkom, Monique Declercq, Anne Deneubourg, Agnès De Wier, Karin d'Hoore et Claire Genevrois.

[2] Laboratoire de Psychologie Expérimentale, Avenue Adolphe Buyl, 117, B-1050 Bruxelles. EMail: R07208@BBRBFU01.BITNET

d'application, ce qui facilite le travail de consultation, de tri, de sélection ou d'édition. Enfin, le coût relativement faible et la grande diffusion des micro-ordinateurs et de leurs logiciels ont pour corollaire un certain degré de standardisation. Il en résulte une meilleure portabilité des applications.

La préparation de matériel d'expérience en psycholinguistique implique souvent de prendre en considération ou de contrôler nombre de propriétés susceptibles d'affecter les performances. La tâche est particulièrement difficile en français, dans la mesure où diverses normes et mesures ne sont pas disponibles, soit parce qu'elles n'ont jamais été calculées, soit parce qu'elles n'ont jamais été publiées.

La base de données que nous avons constituée, BRULEX, vise à combler certaines de ces lacunes tout en tirant parti de l'utilisation d'un support informatique. Nous espérons fournir ainsi un outil efficient pour la recherche et l'expérimentation en psycholinguistique, en neuropsychologie du langage, voire même en linguistique descriptive.

Brulex contient 35.746 entrées lexicales. Pour chaque entrée, différentes informations sont disponibles. Une partie de ces informations, qu'on appellera informations de base, résulte de l'introduction manuelle de données provenant de dictionnaires et de travaux antérieurs. Les autres, les informations générées, sont des variables qui ont été calculées automatiquement à partir des informations de base et des propriétés statistiques du corpus.

Corpus

La base de données a été créée en 1986 en reprenant la majeure partie des entrées du dictionnaire *Micro-Robert* (Robert, 1986). Ce corpus a été choisi parce qu'il semblait bien correspondre aux besoins de la recherche psycholinguistique. En effet, selon les auteurs, la majorité des mots repris appartient à la langue courante et la langue parlée contemporaine est bien représentée. En outre, il inclut un certain nombre de termes scientifiques considérés comme indispensables ainsi que des mots "littéraires ou archaïques nécessaires à la lecture des classiques" (p. IX). Le dictionnaire contient environ 30.000 mots. De l'avis des auteurs, cette

nomenclature représenterait un vocabulaire riche, trois fois supérieur au vocabulaire habituel de l'adulte moyen.

Toutes les entrées du dictionnaire, comportant 20 caractères ou moins, ont été enregistrées à l'exception des noms propres et des affixes. Les verbes sont repris à l'infinitif uniquement. Toutes les formes des articles et des pronoms sont mentionnées. Des entrées distinctes ont été créées pour les homographes variant par la classe grammaticale (ex. BIEN, substantif, vs BIEN, adverbe). Les formes féminines des noms et des adjectifs ont été ajoutées et donnent lieu à des entrées séparées dans Brulex. Notons que seules les formes du pluriel dont la prononciation diffère de celle du singulier sont mentionnées par le *Micro-Robert* (ex. CHEVAL - CHEVAUX).

Le tableau 1 présente de façon synoptique la structure de Brulex, tandis que le tableau 2 fournit des éléments descriptifs sur la composition du corpus. Un extrait de Brulex pour des mots de 4 lettres est fourni en annexe.

Tableaux 1 et 2 environ ici

Bien que les logiciels de traitement de bases de données permettent aisément de réaliser un tri des entrées selon l'ordre désiré par l'utilisateur, nous avons choisi de présenter le répertoire dans l'ordre communément utilisé par les dictionnaires. Les entrées sont dont triées par ordre alphabétique, sans tenir compte des signes diacritiques dans le classement. Les entrées homographes sont classées en fonction de la classe grammaticale, selon l'ordre de mention dans les tableaux 2 et 5 du présent article.

Informations de base

1. GRAPH: Identité orthographique.

Dans la mesure où le code ASCII enrichi utilisé par les micro-ordinateurs le permet, nous avons employé des représentations orthographiques identiques à l'orthographe standard, à une exception près: étant donné que le è provoque des erreurs avec certaines versions de DBASE III+, il a été codé ε (code ASCII 238).

2. PHONS: Transcription phonologique segmentale.

La représentation phonologique a été définie sur base d'une compilation des informations disponibles dans le *Micro-Robert* et le

Petit Robert (Robert, 1987), parce que le *Micro-Robert* ne fournit pas les spécifications phonologiques pour certains dérivés.

A ce niveau également, nous avons pris le parti d'utiliser au maximum les possibilités offertes par le jeu de caractères étendu disponible. Ainsi, tous les segments phonétiques ont pu être codés par un symbole unique, qui est en outre presque toujours évocateur de la valeur phonétique correspondante. Le tableau 3 reprend la liste des codes informatiques utilisés et leur transcription phonétique dans la notation traditionnelle (Warnant, 1987).

Tableau 3 environ ici

Une des difficultés rencontrées dans l'encodage de la forme phonologique des mots concerne les variations de prononciation selon le dialecte, l'usage, ou le contexte. La stratégie adoptée a consisté à sélectionner dans Brulex la forme considérée comme la plus fréquente en présentation isolée, et à marquer les entrées pour lesquelles des variantes phonologiques étaient identifiées. Les variantes sont contenues dans une base de données annexe qui pourra être fournie séparément sur demande.

Trois types de variantes ont été distingués. Une source importante de variation dans la prononciation est liée au [ə] caduc. Pour ces formes, la représentation adoptée inclut toujours le [ə]. Pour les variantes créées par la prononciation de consonnes géminées (ex. SYLLABE prononcé [si.lab] ou [sil.lab]), la représentation adoptée spécifie la prononciation sans redoublement, plus courante (Warnant, 1987). Dans tous les autres cas, le choix a été fait par référence à Warnant (1987). Trois champs (SCHWA, GEMIN, IVARP) décrits ci-dessous permettent d'identifier, de sélectionner ou d'éliminer certaines classes de mots définies en fonction des variations de leur prononciation. Au total, 4303 entrées comportent au moins une forme de variation dans la prononciation.

3. SCHWA: indicateur de [ə] caduc.

Ce champ prend la valeur 1 si le mot inclut un ou plusieurs [ə] caducs (exemples: PETIT, prononcé [pəti] ou [pti] ; BARBE, prononcé [baRbə] ou [baRb]) et 0 ailleurs. 3324 mots présentent ce type de variation.

4. GEMIN: indicateur de redoublement optionnel de consonnes.

Ce champ prend la valeur 1 si le mot inclut une ou plusieurs consonnes pouvant être redoublées, et 0 ailleurs. 471 mots présentent ce type de variation.

5. IVARP: indicateur d'autres variantes phonologiques.

Ce champ prend la valeur 1 si, en dehors des cas SCHWA et GEMIN, le mot accepte plus d'une prononciation (exemples: ANANAS, [anana] vs [ananas], ICEBERG, [isbɛ̲Rg] ou [ajsbɛ̲Rg]), et 0 ailleurs. 680 mots présentent ce type de variation.

6. GENRE: genre grammatical du mot.

Le genre est marqué m (masculin) ou f (féminin) pour les substantifs, adjectifs, articles et pronoms; rien n'est indiqué pour les autres classes grammaticales. Les substantifs et adjectifs admettant les deux genres (ex. SECRETAIRE) ont été marqués h. L'information a été reprise dans le *Micro-Robert*. Le tableau 4 donne la répartition en genre des noms et des adjectifs.

Tableau 4 environ ici

7. NMBRE: indicateur du pluriel.

Les formes plurielles (N=663) sont marquées p; rien n'est indiqué pour le singulier.

8. FRFRM: fréquence d'usage des formes.

Ce champ reprend la fréquence relative associée aux formes orthographiques, c'est-à-dire, aux séquences de caractères, sans distinction de classe syntaxique ni de signification. Cette information permet notamment le calcul de fréquences textuelles de chaînes de caractères ou de phonèmes (cf. Content et Radeau, 1988).

La fréquence introduite est reprise des tables publiées par le Centre de recherche pour un Trésor de la Langue Française (Imbs, 1971). Elle représente le nombre d'occurrences d'une chaîne de caractères rapporté à un total de 100 millions, pour un échantillonnage de textes de la seconde moitié du XXème siècle. Le corpus (23,5 millions de mots) est constitué de textes littéraires (romans, essais, recueils de poèmes, oeuvres dramatiques) publiés

entre 1919 et 1964. Le code -1 a été attribué aux mots qui n'apparaissent pas dans TLF.

Les formes féminines et les formes plurielles, qui ne constituent généralement pas des entrées séparées dans TLF, n'ont donc pas de fréquence formelle (-1). Pour les homographes syntaxiques (ex.: DEJEUNER, verbe, vs. le DEJEUNER, nom), la fréquence formelle a été assignée de manière arbitraire à la première occurrence de la chaîne de caractères dans Brulex, les autres entrées prenant la valeur -1. Pour certains homographes, TLF fournit des valeurs de fréquence distinctes. Dans ce cas, la fréquence formelle a été calculée en sommant les fréquences de toutes les entrées homographiques.

9.FRLEX : fréquence lexicale.

Par opposition à FRFRM, FRLEX vise à fournir une information sur la fréquence d'usage associée à chaque entrée lexicale. Comme pour FRFRM, l'information est la fréquence relative tirée de TLF (2ème moitié du XXème siècle). La valeur -1 est affectée aux mots absents de TLF.

Les mots homographes différant par la classe grammaticale reçoivent leur fréquence propre lorsque l'information est disponible dans TLF (table alphabétique et table de répartition des homographes). Dans les autres cas, FRLEX est indéterminé et vaut -2 pour la série de formes homographiques. Rappelons que les homographes sémantiques ne sont pas distingués. Les formes du féminin et du pluriel ont reçu la même fréquence que la forme masculin singulier du même mot, sauf quand des significations différentes y sont associées (exemples: BARBU, BARBUE; ACCUSE, ACCUSEE). Dans ce dernier cas, le code -3 a été indiqué dans le champ FRLEX.

10. CGRAM: classe grammaticale.

La classe grammaticale a été notée d'après *Micro-Robert*. Les catégories et les codes utilisés sont présentés dans le tableau 5.

Tableau 5 environ ici

11. NVARS: nombre de variantes sémantiques.

Ce champ vaut 0 sauf pour les mots polysémiques (N= 1232) appartenant à une même catégorie grammaticale (ex. BAIE, "fruit

charnu" vs. BAIE, "petit golfe", vs. BAIE, "ouverture pratiquée dans un mur"). Ont été comptabilisées comme variantes sémantiques les définitions d'un mot apparaissant comme des paragraphes distincts et numérotés dans le *Micro-Robert*.

12. VIMAG: valence d'imagerie.

Cette information reprend la valence d'imagerie selon les normes publiées par Hogenraad et Orianne (1981). Ce champ concerne 1086 mots.

Informations générées

13. NGRAPH: nombre de lettres.

14. NPHONS: nombre de phonèmes.

15. NSYLL: nombre de syllabes.

Des critères intuitifs ont été utilisés pour la syllabation. Pour rappel, dans les cas de [ə] caduc, la représentation phonologique inclut le [ə] et la syllabation effectuée en a tenu compte (exemples: CIERGE [sjɛR.zə] et PETIT [pə.ti] ont deux syllabes).

16. IGRAPH: orthographe inversée.

Ce champ permet de réaliser plus facilement des tris ou des sélections basées sur la finale orthographique.

17. IPHONS: phonologie segmentale inversée.

Ce champ permet de trier ou rechercher des mots sur base de la finale (ex. choix de rimes).

18. GRAPHM: orthographe en majuscules sans diacritiques.

19. PUGRAPH: point d'unicité orthographique.

Le point d'unicité orthographique d'un mot correspond au nombre de lettres, compté à partir de la gauche, qui est nécessaire pour identifier le mot sans ambiguïté dans Brulex. Le calcul a été effectué sur la forme orthographique de base (GRAPH) incluant les marques diacritiques. Les homographes d'un mot n'interviennent pas dans le calcul de son point d'unicité.

20. PUPHONS: point d'unicité phonologique.

Le point d'unicité phonologique d'un mot correspond au nombre de phonèmes, compté à partir de la gauche, qui est nécessaire pour identifier le mot sans ambiguïté dans le corpus constitué par Brulex. Les homophones d'un mot n'interviennent pas dans le calcul de son point d'unicité phonologique.

21. NBHOM: nombre de variantes grammaticales.

Ce champ reprend le nombre d'entrées qui sont à la fois homographes et homophones. Dans la mesure où les mots polysémiques ne donnent pas lieu à des entrées distinctes, il s'agit toujours de variantes grammaticales (exemple: BOUCHER, nom vs. verbe). Le nombre de mots présentant ce type de variantes est de 6773, le cas le plus fréquent étant les couples nom-adjectif.

22. NBHOMG: nombre d'homographes.

Le champ NBHOMG reprend pour chaque mot, le nombre d'entrées qui ont une forme orthographique identique, qu'elles soient homophones ou non. Dans le corpus considéré, quelques formes homographiques (N=19) ont des prononciations distinctes (exemple: SUPPORTER, nom vs. verbe). Ces homographes non-homophones peuvent être identifiés par comparaison entre la valeur pour ce champ et pour le précédent.

23. NBHOMP: nombre d'homophones.

Ce champ indique, pour chaque mot, le nombre d'entrées pour lesquelles la forme phonologique est identique. Les cas d'homophonie incluent donc à la fois les variantes grammaticales et les homophones non homographes (exemple: CENT, SANG, SANS), qui peuvent également être isolés par différence.

24. NCOUNT: nombre de voisins orthographiques.

Suivant Coltheart, Davelaar, Jonasson et Besner (1977), nous appelons voisins orthographiques d'un mot les entrées de même longueur dans Brulex qui diffèrent par une lettre, toutes les lettres communes étant aux mêmes positions (ex.: BAC a comme voisins LAC, SAC, BEC, BIC, BOB, BAH, BAI, BAL, BAN, BAS et BAT). Le calcul a été effectué sur la forme orthographique de base (GRAPH) incluant les marques diacritiques.

25. CFRLEX: classe logarithmique de fréquence.

Cette valeur est calculée comme la partie entière de 1000*LOG10 (FRLEX+1) lorsque la fréquence lexicale (FRLEX) est supérieure ou égale à 0. Dans les cas où la valeur de FRLEX est négative (codes -1, -2 et -3), celle-ci a simplement été recopiée.

26. PHONS1: patron phonologique majeur.

Ce champ spécifie la structure du mot, en termes de la séquence de consonnes (notées C), voyelles (V) et semi-voyelles (Y) qui le composent. Cette donnée n'est toutefois définie que pour les 10 premiers phonèmes des mots.

27. PHONS2: patron phonologique détaillé.

Ce champ spécifie la structure des 10 premiers phonèmes du mot, en termes de la séquence de consonnes fricatives voisées (Z), non-voisées (S), occlusives voisées (B), non-voisées (P), nasales (N), liquides (L), voyelles (V) et semi-voyelles (Y) qui le composent.

28. MODIGR: fréquence moyenne des digrammes.

Les fréquences des digrammes et trigrammes en position initiale, interne et finale, ont été calculées précédemment sur base de Brulex (Content et Radeau, 1988). Les fréquences présentées ici sont les moyennes des logarithmes décimaux des fréquences textuelles (établies sur la base d'un comptage pondéré par les fréquences d'usage des mots) de tous les digrammes constituant le mot. La transformation logarithmique a été introduite en raison des distributions très asymétriques des fréquences des n-grammes (cf. Content et Radeau, 1988). Cette mesure donne une indication du degré de régularité séquentielle de la forme orthographique.

29. RENUM: numéro d'enregistrement.

Chaque fiche porte un numéro d'enregistrement qui permet de l'identifier de manière univoque.

Extensions

Tel qu'il est présenté, Brulex est susceptible de répondre à un grand nombre de demandes. Cependant, l'ajout d'autres variables pertinentes peut être envisagé (notamment dans le domaine de la morphologie et de la sémantique). La flexibilité des logiciels permet facilement d'étendre la base de données en introduisant de

nouvelles informations, ou en augmentant le corpus. Il est évident que tous les utilisateurs seraient intéressés à connaître l'existence de tels développements. Pour répondre à ce besoin, nous suggérons que les utilisateurs nous fassent part de leurs projets et de leurs réalisations en cette matière. Nous pourrions ainsi diffuser régulièrement ces informations à tous les utilisateurs de manière à favoriser les échanges.

Pour constituer Brulex, un grand nombre d'informations ont été introduites manuellement par différents opérateurs. Bien qu'un effort important de vérification et de correction ait été fait, nous ne pouvons garantir que la base de données ne contient aucune erreur de transcription. Pour autant que les utilisateurs nous fassent part des inexactitudes qu'ils rencontreraient, nous envisageons de faire circuler périodiquement des listes correctives.

Par ailleurs, un programme (Content, 1989) a été écrit spécifiquement pour faciliter l'exploitation de bases de données lexicales telles que Brulex. Une de ses fonctions est le calcul d'informations générées. Il permet également de créer de nouvelles bases de données à partir d'une base de référence. Les sous-ensembles peuvent être obtenus par la sélection des entrées (ex. les mots d'une longueur donnée), par la sélection des informations (ex. constitution d'un répertoire des mots contenant uniquement l'orthographe, la fréquence, le nombre de voisins,...), ou par la combinaison de ces deux modes de sélection. Enfin, il permet de calculer des informations pour des formes qui ne figureraient pas dans la base de référence, que ce soient des mots absents du corpus de référence ou des logatomes (ex. nombre de voisins orthographiques, mesures de la régularité orthographique de pseudomots).

Exploitation

Brulex a été créé sur micro-ordinateur IBM-compatible, à l'aide du logiciel Clipper (Nantucket, 1987). Il peut donc être exploité notamment avec DBASE III+ (Ashton-Tate, 1986), sur n'importe quel micro-ordinateur du même type, équipé d'un disque dur de taille suffisante. Chaque enregistrement occupe 175 octets, et le fichier entier occupe environ 6.3 Moctets. Il est loisible aux utilisateurs de sélectionner une partie du corpus selon leurs

besoins. Les lecteurs désireux de se procurer une copie de la base de données sont priés de prendre contact avec les auteurs pour des informations complémentaires sur les modalités d'obtention.

Références

Ashton-Tate (1986). dBAse III Plus 1.1 Culver City, Californie: Ashton-Tate.

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

Content, A. (1989). SyLex: un outil d'exploitation de bases de données lexicales. Rapport interne du Laboratoire de Psychologie Expérimentale, Septembre 1989.

Content, A. et Radeau, M. (1988) Données statistiques sur la structure orthographique du français. Cahiers de Psychologie Cognitive, Septembre 1988, Numéro spécial.

Hogenraad, R. & Orianne, E. (1981) Valence d'imagerie de 1.130 noms de la langue française parlée. Psychologica Belgica, 21, 21-30.

Imbs, P. (1971). Etudes statistiques sur le vocabulaire français. Dictionnaire des fréquences. Vocabulaire littéraire des XIXe et XXe siècles. Centre de recherche pour un Trésor de la langue française (C.N.R.S.), Nancy. Paris: Librairie Marcel Didier.

Nantucket (1987). Clipper dBase III Compiler. Los Angeles, California: Nantucket Corporation.

Robert, P. (1986). Micro-Robert, dictionnaire du français primordial. Nouvelle édition revue et mise à jour.Paris: Dictionnaires le Robert.

Robert, P. (1987). Le petit Robert. Dictionnaire alphabétique et analogique de la langue française. Paris: Dictionnaires le Robert.

Warnant, L. (1987) Dictionnaire de la prononciation française dans sa norme actuelle. Paris et Gembloux: Duculot.

Tableau 1

Structure de la base de données

Nom du champ	Type	Nombre d'octets	Fonction
1. RENUM	N	5	Numéro d'enregistrement
2. GRAPH	C	20	Forme orthographique
3. PHONS	C	20	Forme phonologique
4. CGRAM	C	2	Classe grammaticale
5. GENRE	C	1	Genre grammatical
6. NMBRE	C	1	Nombre
7. FRFRM	N	8	Fréquence formelle
8. FRLEX	N	8	Fréquence lexicale
9. CFRLEX	N	3	Classe de fréquence (Log*1000)
10. IVARP	N	1	Indicateur de variante phonologique
11. SCHWA	N	1	Indicateur de [ ] caduc
12. GEMIN	N	1	Indicateur de consonnes géminées
13. NVARS	N	1	Compteur de variantes sémantiques
14. VIMAG	N	1	Valence d'imagerie
15. NGRAPH	N	2	Nombre de caractères
16. NPHONS	N	2	Nombre de phonèmes
17. NSYLL	N	2	Nombre de syllabes
18. IGRAPH	C	20	Forme orthographique inversée
19. IPHONS	C	20	Forme phonologique inversée
20. GRAPHM	C	20	Orthographe sans diacritiques
21. PUGRAPH	N	2	Point d'unicité orthographique
22. PUPHONS	N	2	Point d'unicité phonologique
23. NBHOM	N	2	Nombre d'homographes homophones
24. NBHOMG	N	2	Nombre d'homographes
25. NBHOMP	N	2	Nombre d'homophones
26. NCOUNT	N	2	Nombre de voisins orthographiques
27. PHONS1	C	10	Patron phonologique (V,C)
28. PHONS2	C	10	Patron phonologique détaillé
29. MODIGR	N	4	Fréquence moyenne des digrammes

Tableau 2

Description du Corpus

Nombre de lettres	Nombre	Pourcentage
1	30	0,08
2	82	0,23
3	338	0,95
4	1065	2,98
5	2435	6,81
6	3891	10,89
7	5030	14,07
8	5555	15,54
9	5155	14,42
10	4231	11,84
11	3110	8,70
12	2079	5,82
13	1300	3,64
14	736	2,06
15-19	709	1,98

Nombre de Phonèmes	Nombre	Pourcentage
1	32	0,09
2	434	1,21
3	1479	4,14
4	3417	9,56
5	6107	17,08
6	6383	17,86
7	6123	17,13
8	4476	12,52
9	3161	8,84
10	1946	5,44
11	1123	3,14
12	600	1,68
13-18	465	1,31

Tableau 2 (Suite)

Nombre de Syllabes	Nombre	Pourcentage
1	2396	6,70
2	11.713	32,77
3	13.258	37,09
4	6157	17,22
5	1807	5,06
6	352	0,98
6	57	0,16
7	6	0,02

Catégorie Grammaticale	Nombre	Pourcentage
Substantifs	19.384	54,23
Adjectifs	10.431	29,18
Verbes	4.334	12,15
Adverbes	1150	3,22
Articles	10	0,03
Pronoms	80	0,22
Prépositions	55	0,15
Conjonctions	22	0,06
Interjections	103	0,29
Locutions	174	0,49
Participes Présents	3	0,01

Classes de Fréquence*	Nombre	Pourcentage
<0**	9399	26,29
0	553	1,55
<10	1630	4,56
<100	8616	24,10
<1000	10818	30,26
<10.000	4065	11,37
<100.000	613	1,71
<1.000.000	44	0,12
<10.000.000	8	0,02

* En nombre d'occurrences pour 100.000.000

** Les classes négatives correspondent aux valeurs codées -1 (mots absents dans TLF), -2 (entrées ambiguës), et -3 (féminins et pluriels distincts du masculin singulier).

Tableau 3

Codes phonétiques

Symbole usuel	Code informatique	Exemples*
i	i	idée, ami
e	é	ému, ôté
	_	perdu, modèle
a	a	alarme, patte
	A	bâton, pâte
	o	obstacle, corps
o	O	auditeur, beau
u	u	coupable, loup
y	y	punir, élu
	E	creuser, deux
	e	malheureux, peur
	^	petit, fortement
	ê	peinture, matin
	â	vantardise, temps
	ô	rondeur, bon
	û	lundi, brun
j	ï	piétiner, briller
w	ü	oui, fouine
	ÿ	huile, nuire
p	p	patte, repas, cap
t	t	tête, ôter, net
k	k	carte, écaille, bec
b	b	bête, habile, robe
d	d	dire, rondeur, chaude
g	g	gauche, égal, bague
f	f	feu, affiche, chef
s	s	soeur, assez, passe
	/	chanter, machine, poche
v	v	vent, inventer, rêve
z	z	zéro, raison, rose
	j	jardin, manger, piège
l	l	long, élire, bal
R	R	rond, chariot, sentir
m	m	madame, aimer, pomme
n	n	nous, punir, bonne
	N	agneau, règne
	£	jumping, Sterling
'	'	hollandais, haricot (non-liaison)
x	x	Bach, esp. Hijo

* D'après Warnant (1987).

Tableau 4

Répartition des substantifs et adjectifs selon le genre

	Masculin	Féminin	Masculin et Féminin
Substantifs	10.226	8.327	831
Adjectifs	4.144	3.748	2.539
Total	14.370	12.075	3.370

Tableau 5

Code	Catégorie grammaticale
NO	substantif
AJ	adjectif
VB	verbe
AV	adverbe
AR	article
PN	pronom
PR	préposition
CO	conjonction
IN	interjection
LO	locution ou partie de locution
PP	participe présent

BRULEX : MODALITES D'UTILISATION

BRULEX est une base de données lexicales créée par le Laboratoire de Psychologie Expérimentale de l'Université libre de Bruxelles. Elle fournit l'orthographe, la prononciation, la classe grammaticale, le genre, le nombre, la fréquence d'usage, ainsi qu'une série d'autres informations qui peuvent être utiles dans les recherches en psychologie du langage, pour environ 36.000 mots français.

Ces informations ont été collationnées sur support informatique à l'usage des chercheurs en sciences du langage et en particulier en psycholinguistique, pour faciliter des opérations de sélection, de consultation et d'analyse statistique.

Une présentation détaillée de la base de données a été publiée sous la référence suivante :

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

Les utilisateurs de BRULEX s'engagent:

- à utiliser BRULEX exclusivement à des fins de recherche scientifique, et uniquement dans le cadre de l'équipe de recherche ayant acquis le produit.

- à ne pas publier, reproduire, communiquer au public ou distribuer BRULEX, en tout ou partie, sans l'accord des auteurs.

- à mentionner la référence de l'article de présentation de BRULEX dans toute publication scientifique pour laquelle la base de données aurait été utilisée.

Laboratoire:

.....
.....

Adresse:
.....
.....

Nom:.....
Fonction : .....

Date:.....
Signature : .....

Installation notes

The distribution includes the following files

Brulex Folder
• Read me first (the present file)
• Brulex.str (a Foxpro file containing the Brulex database structure)
• Brulex.txt (a text file containing the Brulex data)

1. Please note the the alphabetical and phonetic codes are slightly different in the PC version (as well as in the published paper) and in the current Mac version. There is a conversion table below. Phonetic codes have been selected to be readable in the Geneva font and approximate usual symbols.

2. The Brulex data are provided as a text file (standard ASCII), to leave the opportunity to users to import it in any database software of their choice. As we have primarily used Foxbase/Foxpro, a Foxpro database structure file is provided, conforming to the preceding description of the fields.

You can create the Brulex data base with FoxPro using the following commands:

```
CREATE "<YHDN>:Brulex Folder:brulex.dbf" from "<YHDN>:Brulex
Folder:brulex.str"
APPEND FROM "<YHDN>:Brulex Folder:brulex.txt" TYPE DELIMITED WITH TAB
```

in which the <YHDN> specification is replaced by Your Hard Disk Name, assuming that the Brulex Folder has been created at the root level of the disk.

3. Our email addresses have changed, the address mentioned in the published paper is no more valid:
acontent@ulb.ac.be
pmousty@ulb.ac.be
moradeau@ulb.ac.be

4. Users are kindly requested to send back the completed form to

BRULEX
Laboratoire de Psychologie Expérimentale CP 191
Avenue F.D. Roosevelt, 50
B-1050 Bruxelles
Belgique

and to mention a reference to Brulex in any publication resulting from the use of the database.

Phonetic codes conversion table

Code Mac	Code PC	Exemples*
i	i	idée, ami
é	é	ému, ôté
è	ɛ	perdu, modèle
a	a	alarme, patte
A	A	bâton, pâte
o	o	obstacle, corps
O	O	auditeur, beau
u	u	coupable, loup
y	y	punir, élu
ø	E	creuser, deux
œ	e	malheureux, peur
e	^	petit, fortement
ê	ê	peinture, matin
â	â	vantardise, temps
ô	ô	rondeur, bon
û	û	lundi, brun
ï	ï	piétiner, briller
ü	ü	oui, fouine
ÿ	ÿ	huile, nuire
p	p	patte, repas, cap
t	t	tête, ôter, net
k	k	carte, écaille, bec
b	b	bête, habile, robe
d	d	dire, rondeur, chaude
g	g	gauche, égal, bague
f	f	feu, affiche, chef
s	s	soeur, assez, passe
ʃ	/	chanter, machine, poche
v	v	vent, inventer, rêve
z	z	zéro, raison, rose
j	j	jardin, manger, piège
l	l	long, élire, bal
R	R	rond, chariot, sentir
m	m	madame, aimer, pomme
n	n	nous, punir, bonne
ñ	N	agneau, règne
£	£	jumping, Sterling
'	'	hollandais, haricot (non-liaison)
x	x	Bach, esp. Hijo

* D'après Warnant (1987).

For the orthographic codes, the only difference with the PC version is that the e-grave (è), which had been replaced with ɛ in the PC version because of a bug in DBASE, is now coded with the standard character.