OMNILEX : Une Base de Données sur le Lexique du Français Contemporain

Alain Desrochers

Résumé : OMNILEX est une base de données lexicales conçue pour la recherche en psycholinguistique. Cette base de données assure présentement deux fonctions principales : a) la sélection de stimuli expérimentaux par l'application simultanée de filtres et b) l'analyse quantitative des propriétés du lexique du français. Nous faisons un retour sur le concept de base de données lexicale et ses applications en recherche. Puis, nous décrivons le contenu et l'interface graphique du premier prototype d'OMNILEX accessible par l'Internet : www.omnilex.uottawa.ca. Enfin, nous discutons quelques pistes d'expansion de cette base de données.

Mots-clés : Psycholinguistique, base de données lexicales, sélection des stimuli expérimentaux, lexique du français, propriétés lexicales
Keywords: Psycholinguistics, Lexical database, Selection of experimental stimuli, French lexicon, Lexical properties

1. Introduction

Le but de cet article est de présenter les principales caractéristiques d'OMNILEX, une base de données sur le lexique du français contemporain. Nous faisons d'abord un retour sur le concept de base de données lexicales et nous en présentons quelques applications. Nous décrivons ensuite le contenu actuel d'OMNILEX en mettant, tour à tour, l'accent sur ses entrées lexicales constitutives, leur classement et leur mode de saisie. Puis, nous résumons les caractéristiques de l'interface d'interrogation de la base de données. Enfin, nous évoquons quelques pistes que nous nous proposons de poursuivre dans l'expansion de ce projet.

2. Le Concept de Base de Données Lexicales

Une base de données lexicales fournit une description structurée des entrées lexicales d'une langue. Le niveau d'abstraction de ces entrées peut varier selon les objectifs poursuivis par les concepteurs de la base de données. Le choix le plus typique consiste à utiliser le lexème comme forme de citation orthographique principale. On peut définir le lexème comme la forme de citation non fléchie d'un mot. Chaque forme de citation peut alors être associée à des variantes de forme (par énumération ou par l'indication d'une classe flexionnelle) ou à des variantes d'unités forme-sens. Comme le soulignent Sáenz et Vaquero (2005), la

conception des bases de données lexicales évoluent progressivement vers une standardisation, mais aucun code strict ne guide présentement les pratiques courantes.

Il reste que les bases de données sont ordinairement organisées en tables. Une table est un fichier constitué d'enregistrements (des ensembles d'informations organisés en rangées) et de champs (des catégories de données organisées en colonnes). Les unités lexicales qui servent de point d'ancrage dans une table peuvent également être projetées dans d'autres tables, elles-mêmes constituées d'enregistrements et de champs. Pour illustrer, considérons l'exemple du lexème « bout » dans une Table 1. et sa projection dans une Table 2. qui distingue ses différents sens.

Table 1.

Graphie	**Catégorie Grammaticale**	**Genre**
Bout	**Nom**	**masc**
Boutade	Nom	fém
bout-dehors	Nom	masc
boute-en-train	Nom	masc
Boutefas	Nom	masc
Boutefeu	Nom	masc

Table 2.

SENS
Partie terminale d'un objet
Limite d'un espace
Fin d'une durée
Partie de quelque chose
Ce qui est petit, incomplet
Cordage

Nous évoquions, plus haut, l'idée qu'une base de données lexicales est constituée d'enregistrements qui décrivent les entrées lexicales d'une langue. Rappelons que cette description peut porter autant sur les formes que sur les sens des citations. Les modalités descriptives ne sont limitées que par des considérations conceptuelles ou technologiques. Un enregistrement dans une table peut très bien réunir des informations catégorielles (p.ex., la catégorie grammaticale ou sémantique), quantitatives (p.ex., le nombre de phonèmes, la fréquence d'occurrence), relationnelles (p.ex., sur les synonymes ou les

homographes), figuratives (p.ex., un dessin d'objet ou une photographie) ou auditives (p.ex., la prononciation du mot). Le contenu d'une base de données lexicales est généralement dicté par les applications auxquelles on la destine.

3. L'utilité d'une Base de Données Lexicales

L'application la plus courante d'une base de données lexicales pour la recherche en psycholinguistique est la sélection des stimuli expérimentaux. Les logiciels de gestion de base de données (p.ex., Microsoft Access) permettent aux utilisateurs d'élaborer une requête en activant des filtres de sélection. Ces filtres servent à spécifier les critères d'inclusion ou d'exclusion sur des variables particulières. Par exemple, on pourra appliquer ces critères à la forme orthographique ou phonologique des mots recherchés, à leur catégorie grammaticale ou à leur fréquence d'occurrence, etc., selon les besoins de la recherche. Le lancement d'une requête résulte typiquement en une liste sélective d'entrées lexicales et de ses caractéristiques. La base OMNILEX a d'abord été conçue pour faciliter cette application, mais il ne s'agit pas d'une caractéristique distinctive. D'autres bases lexicales peuvent également être utilisées aux mêmes fins : BRULEX (Content, Mousty, & Radeau, 1990), NOVLEX (Lambert & Chesnet, 2001), LEXIQUE (New, Pallier, Ferrand, & Matos, 2001), VOCOLEX (Dufour, Peereman, Pallier, & Radeau, 2002) et MANULEX (Lété, Sprenger-Charolles, & Colé, 2004). Ce qui distingue OMNILEX des autres bases, c'est son interface graphique explicitement axée sur la sélection des stimuli expérimentaux, l'étendue de ses entrées lexicales et de ses champs de données constitutifs ainsi que ses données normatives établies auprès d'échantillons de répondants canadiens d'expression française.

Une deuxième fonction d'une base de données lexicales est de permettre l'analyse de la structure interne des mots et celle des relations entre les mots. Une base de plus de plusieurs milliers de mots fournit un matériel idéal pour extraire les patrons de régularité dans la structure des mots (p.ex., les patrons syllabiques ou morphologiques, les procédés de formation lexicale). Par ailleurs, plusieurs variables relationnelles ne peuvent être calculées que si on dispose d'un

échantillon de mots considérable. C'est le cas, par exemple, de l'indice N de voisinage orthographique proposé par Coltheart, Davelaar, Jonasson et Besner (1977). Les voisins orthographiques d'un mot comprennent ceux de même longueur qui se différencient de lui par une seule lettre (p.ex., noir – soir – voir; pour un traitement détaillé, voir Mathey, 2001). Le recours à une riche base de données lexicales est essentiel pour l'établissement des indices de similitude de formes (p.ex., les voisins orthographiques ou phonologiques, les homographes) ou de sens (p.ex., les synonymes).

Si on adjoint à une base de données lexicales des définitions, on peut alors lui faire jouer un rôle central dans l'apprentissage des langues assisté par ordinateur. Selva, Verlinde et Binon (2003), par exemple, ont élaboré le Dictionnaire d'Apprentissage du Français Langue Étrangère ou Seconde (DAFLES) qu'ils ont ensuite exploité pour générer des exercices lexicaux en contexte et pour fournit aux apprenants du français une rétroaction sur la justesse de leurs réponses. Si, au lieu d'adjoindre une définition aux entrées lexicales, on différencie leurs composantes sémantiques, la base de données lexicales peut alors constituer l'ensemble des noeuds terminaux d'un réseau sémantique interrogeable. Dutoit et ses collaborateurs (Dutoit & Nugues, 2002; Dutoit, Nugues, & Torcy, 2003) ont élaboré une base de données, appelée le Dictionnaire Intégral, dont les entrées lexicales sont organisées en graphes hiérarchiques de concepts. Les arcs entre les concepts permettent de spécifier divers types de relations sémantiques entre les mots (p.ex., la synonymie, l'hypernymie, l'hyponymie). Cette application permet, notamment, aux utilisateurs de soumettre une définition et de recevoir en rétroaction les termes qui lui correspondent le mieux dans le lexique (p.ex., une personne qui vend des fleurs → fleuriste, floriculteur, bouquetière, maraîcher). Dans toutes ces applications, la base de données lexicales fournit les matériaux de base pour établir des relations entre des unités de forme ou de sens.

4. Le Contenu d'OMNILEX 1

Le contenu du premier prototype d'OMNILEX peut être décrit de diverses manières. Si on axe cette description sur la répartition des entrées lexicales par catégorie grammaticale, nous obtenons une première vue sur sa composition.

Catégorie grammaticale	Nombre d'entrées lexicales
Nom	48,570
Adjectif	27,289
Verbe	13,845
Adverbe	1,878
TOTAL	96,031

Il est également possible de décrire le contenu d'OMNILEX en mettant l'accent sur la fonction des classes de données inscrites dans les enregistrements. Par exemple, les *données de base* (p.ex., l'orthographe, la transcription phonétique, la catégorie grammaticale) servent principalement à différencier les unités lexicales les unes des autres. Les *données structurelles* visent à caractériser soit la structure interne des mots (p.ex., la structure phonologique or morphologique) et les rapports de forme entre les mots (p.ex., la similitude orthographique ou phonologique). Les *données distributionnelles* se rapportent aux caractéristiques statistiques des unités langagières dans la langue (p.ex., la fréquence d'occurrence des mots). Enfin, les *données sémantiques* visent à caractériser la valeur de symbole des formes lexicales (p.ex., la catégorie sémantique, la typicité catégorielle, la valeur d'imagerie).

Le schème de classement optimal des entrées d'une base de données dépend du cadre théorique qui en guide la conception ou de la fonction à laquelle elle est destinée. Peu importe le cadre théorique auquel on se rattache ou la fonction que l'on cherche à assurer, les modalités de saisie des données pertinentes seront nécessairement variables. Au stade initial de l'élaboration d'une base de données, certaines données devront être saisies manuellement (p.ex., l'orthographe, la catégorie grammaticale). Il en ira de même pour les informations qui n'existent dans aucune autre base de données similaires (p.ex., la fréquence subjective, l'âge d'acquisition, la valeur d'imagerie, l'indice de typicité). Une fois la saisie des

données primitives complétée, il devient alors possible d'élaborer des algorithmes pour produire de nouvelles données et ainsi enrichir la base. Par exemple, si on prend la forme orthographique ou phonologique comme input, il est possible de calculer des indices de longueur (p.ex., en lettres, en phonèmes, en syllabes) ou de similitude (p.ex., le voisinage, le point d'unicité). Les analyses de corpus peuvent également contribuer à enrichir une base de données en fournissant des indicateurs supplémentaires (p.ex., la distance sémantique ou les co-occurrences entre les mots). Dans l'élaboration d'OMNILEX 1, nous avons réuni des informations provenant de toutes les classes et en exploitant toutes les modalités de saisie (sauf l'analyse de corpus) citées plus haut.

La présente version d'OMNILEX comprend les champs de données suivants :

- L'orthographe de l'entrée lexicale
- Sa transcription phonétique d'après les dictionnaires usuels
- Sa catégorie grammaticale
- Son genre grammatical, s'il y a lieu
- Son nombre grammatical
- Sa longueur en lettres
- Sa longueur en phonèmes
- Sa fréquence d'occurrence dans la langue écrite dans la 2^{e} moitié du 20^{e} siècle d'après le Dictionnaire des fréquences du Trésor de la langue française (Imbs, 1971)
- Sa fréquence subjective sur une échelle de Likert en 7 points d'après les données normatives de Desrochers et Bergeron (2000)
- Sa valeur d'imagerie sur une échelle de Likert en 7 points d'après les données normatives de Desrochers et Bergeron (2000)
- Le nombre de ses voisins orthographiques
- Le nombre de ses voisins phonologiques

Examinons maintenant comment l'interface graphique d'OMNILEX permet d'exploiter ces informations et de construire des listes.

5. L' interface Graphique d'OMNILEX 1

L'interface graphique d'OMNILEX 1 a été largement influencée par celle que Coltheart (1981) a conçu pour la MRC Psycholinguistic Database. Elle fournit aux utilisateurs un ensemble de fonctions simples et centrées sur la sélection des entrées lexicales. L'organisation de cette interface conduit l'utilisateur à répondre

à quatre questions : a) quels champs de données désirez-vous retenir dans votre liste de mots? b) quels filtres désirez-vous activer dans votre recherche? c) sur quelles variables désirez-vous trier les mots dans votre liste? et d) dans quel ordre désirez-vous appliquer vos clés de tri? Une fois que l'utilisateur a fixé ses choix, il peut lancer sa requête par un simple clic et le résultat apparaît à l'écran. Il a alors l'option d'imprimer la liste de mots qu'il a obtenue ou de la sauvegarder dans un fichier (p.ex., Microsoft Excel ou Word).

Ce premier prototype d'OMNILEX est accessible sur l'Internet à l'adresse suivante : www.omnilex.uottawa.ca. S'il est facile de retracer les étapes qui ont conduit à l'état actuel de la base de données, il est plus difficile de prédire jusqu'où nous la mènerons.

6. L'Expansion d'OMNILEX

Dans l'immédiat, nous nous proposons de poursuivre le travail déjà amorcé et d'enrichir OMNILEX sur le plan du contenu et de la versatilité pour la sélection des stimuli expérimentaux à des fins de recherche. La prochaine étape vise à élargir la collection des variables lexicales en y ajoutant des données sur la structure syllabique et morphologique, à augmenter à plus de 6,000 le nombre de mots pour lesquels des données normatives sur la fréquence subjective et la valeur d'imagerie sont disponibles et à ajouter la représentation graphique des mots présentement en cours de rectification orthographique.

A plus long terme, nous explorons la possibilité d'enrichir la base de données de quatre façons. Premièrement, nous souhaiterions lui ajouter des données sur la fréquence d'occurrence à l'écrit et à l'oral, de préférence basées sur des corpus d'origine canadienne française, en prenant en compte les considérations méthodologiques évoquées par Lété et al. (2004) et Zeno, Ivens, Millard et Duvvuri (1995). Deuxièmement, nous nous proposons d'étayer la caractérisation sémantique des entrées lexicales de la base. Troisièmement, il nous serait utile de doter OMNILEX de divers outils de traitement automatique du langage tels un transcripteur phonétique et un syllabeur. Enfin, à l'instar des concepteurs du English Lexicon Project, David Balota et ses collaborateurs, nous pourrions

ajouter à OMNILEX des données sur la justesse et la latence du traitement lexical en situation de décision lexicale et de lecture orale. Ces ajouts auraient pour conséquence d'augmenter l'utilité de la base de données pour la sélection des stimuli expérimentaux à des fins de recherche et de modélisation computationnelle.

7. Conclusion

Le projet OMNILEX est né d'une nécessité pratique, celle d'élaborer un outil efficace et versatile pour sélectionner des stimuli expérimentaux conformes à des critères stricts. Au fil du temps, cette base s'est enrichie de plusieurs milliers d'entrées lexicales, que nous avons cherché à décrire de manière de plus en plus détaillée faisant ressortir les caractéristiques qui leur sont spécifiques et celles qu'elles partagent avec d'autres entrées. Plusieurs facteurs pourront jouer un rôle déterminant dans l'évolution de ce projet tels les besoins de la recherche en psycholinguistique, les applications parallèles et les innovations technologiques.

8. Remerciements

Ce projet a été rendu possible grâce aux subsides de recherche reçus du Programme Éduc-Action du Ministère de l'éducation de l'Ontario, du Conseil de recherche en sciences humaines du Canada, du Conseil de recherche en sciences naturelles et en génie du Canada et de la Faculté des sciences sociales de l'Université d'Ottawa. Notre reconnaissance va également à Hubert Séguin et à Marie-Hélène Côté pour leurs conseils sur des questions linguistiques, à Alain Côté pour sa contribution à la programmation de l'interface graphique et aux membres du Laboratoire de psychologie cognitive de l'Université d'Ottawa qui, collectivement et au fil des ans, ont conduit OMNILEX à son état d'achèvement actuel.

Bibliographie

Coltheart, Max. (1981) The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology* 33A: 497-505.

Coltheart, Max, Eileen Davelaar, Jon Torfi Jonasson and Derek Besner (1977) Access to the internal lexicon. In Stan Dornic (Ed.), *Attention and performance* VI : 535-555. Hillsdale, NJ: Erlbaum.

Content, Alain, Philippe Mousty and Monique Radeau (1990) BRULEX: une base de données lexicales informatisées pour le français écrit et parlé. *L'Année Psychologique* 90 : 551-566.

Desrochers, Alain and Mylène Bergeron (2000) Valeurs de fréquence subjectiveet d'imagerie pour un échantillon de 1,016 substantifs de la langue française. *Revue canadienne de psychologie expérimentale* 54 : 274-325.

Dufour, Sophie, Ronard Peereman, Christophe Pallier and Monique Radeau (2002) VOCOLEX: une base de données lexicales sur les similarities phonologiques entre les mots français. *L'année Psychologique* 102:725-746. Dutoit, Dominique and Pierre Nugues (2002) A lexical database and an algorithm to find words from definitions. Actes de *European Conference on Artificial Intelligence* : 450-454. Lyon, France.

Dutoit, Dominique, Pierre Nugues and Patrick de Torcy (2003) *The Integral Dictionary : A lexical network based on computational semantics*. Communication présentée à l'International Conference on Computational Science and its Applications, Calgary, Canada.

Imbs, Paul (1971) *Études statistiques sur le vocabulaire français - Dictionnaire des fréquences I: Table alphabétique*. Paris: Didier.

Lambert, Eric and David Chesnet, (2001) Novlex : une base de données lexicales pour les élèves de primaire. *L'Année Psychologique* 101 : 277-288.

Lété, Bernard, Liliane Sprenger-Charolles and Pascale Colé (2004) MANULEX : A grade-level lexical database from French elementary school readers. *Behavioral Research Methods, Instruments, and Comptuters* 36 :156-166.

Mathey, Stéphanie (2001) L'influence du voisinage orthographique lors de la reconnaissance des mots écrits. *Revue canadienne de psychologie expérimentale* 55 : 1-23.

New, Boris, Christophe Pallier, Ludovic Ferrand and Rafael Matos (2001) Une base de données lexicales du français contemporain sur internet : LEXIQUE. *L'Année Psychologique* 101: 447-462.

Sáenz, Fernando and Antonio Vaquero (2005) Knowledge representation issues and implementation of lexical data bases. In Jesus Cardeñosa, Alexander Gelbukh, & Edmundo Tovar (Eds.), Universal network language: Advance in theory and application. *Research on Computer Science* 12: 430-442.

Selva, Thierry, Serge Verlinde and Jean Binon (2003) Vers une deuxième génération de dictionnaires électroniques. In Michael Zock & John Carroll (Eds.), Les dictionnaires électroniques. *Traitement automatique des langues* 44.2 :177-197.

Zeno, Susan M., Stephen H. Ivens, Robert T. Millard, and Raj Duvvuri (1995) *The educator's word frequency guide*. Brewster, NY : Touchstone Applied Science Associates.