

Présentation, démonstration et mise en pratique des bases de données lexicales (Lexique3, etc)

Boris New

**Laboratoire Cognition et Comportement -Université Paris Descartes
71, avenue Edouard Vaillant
92100 Boulogne Billancourt**

1- La base *Lexique 3*

Les bases de données lexicales sont extrêmement utiles car elles permettent d'étudier ou de contrôler l'influence de multiples facteurs tels que la fréquence chez l'adulte à l'écrit et à l'oral, la catégorie grammaticale, le nombre de lettres, de phonèmes, de syllabes, le nombre de voisins orthographiques, etc. En effet tous ces facteurs sont connus pour avoir une influence sur les temps de lecture (la lecture), la reconnaissance auditive et la production de mots.

Brulex (Content, Mousty et Radeau, 1990) fut la première base de données disponible pour les psycholinguistes. Si *Brulex* fut extrêmement utilisée, elle 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 calculées uniquement à partir 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.

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.

2- Open Lexique

La base de données Lexique est interrogeable en ligne à travers trois outils que nous vous invitons à découvrir (<http://www.lexique.org>).

Ces outils permettent non seulement d'interroger Lexique mais aussi d'interroger d'autres bases de données telles que:

- Manulex (Lété, Sprenger-Charolles, et Lété, 2004) afin d'avoir la fréquence des mots selon des manuels scolaires répartis en trois niveaux
- L'âge d'acquisition (Ferrand et Alario, 1998; Ferrand, Grainger et New, 2003) estimé par des adultes (plusieurs études montrent que ces données sont fortement corrélées avec l'âge d'acquisition objectif)

- La fréquence subjective (fréquence obtenues grâce à l'estimation donnée par des adultes) et le degré d'imageabilité (Desrochers et Bergeron, 2000)
- D'autres bases de données seront disponibles dans le futur...

C'est l'utilisation, dans le cadre de vos besoins personnels, de ces outils et de ces bases de données que nous proposons de faire découvrir dans cette école thématique.

Références

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.

Desrochers, A., Bergeron, M. (2000). Subjective frequencies and imagery values of a sample of 1,916 nouns in French. *Canadian Journal of Experimental Psychology*, 54, 274-325

Ferrand, L., & Alario, X. (1998). Normes d'associations verbales pour 366 noms d'objets concrets. *L'Année Psychologique*, 98, 689-739.

Ferrand L., Grainger J., New B. (2003) Normes d'âge d'acquisition pour 400 mots monosyllabiques. *L'Année Psychologique*, 104, 445-468.

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.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A New French Lexical Database. *Behavior Research Methods, Instruments, & Computers*, 36, 516-524.

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