Les normes de fréquences lexicales et infra-lexicales chez l'enfant.

Bernard Lété
INRP-Lyon 2
EMC/DDL (CNRS 5596)
5 avenue Pierre Mendès France
69676 - BRON cedex

1- Les normes de fréquences lexicales : la base *Manulex*

Manulex (pour *Lexique* des *Manuels)* a été développée pour mettre à la disposition des chercheurs un outil actualisé rendant compte du lexique écrit adressé à l'enfant. En effet, en France, les chercheurs sur le développement du langage se servent de bases extraites de corpus d'écrits adressés à l'adulte comme *Brulex* (Content et al., 1990) ou *Lexique* (New et al., 2001, 2004). Le problème est que ces bases ne donnent pas une indication précise sur les mots que l'enfant est susceptible de rencontrer en lisant. Or, les études actuelles sur le développement du langage (en particulier les modèles connexionnistes) accordent une place prépondérante à ce qui est appelé "*l'exposition à l'écrit*". Selon ces modèles, on peut rendre compte du développement du système lexical d'un enfant par sa capacité à extraire implicitement des régularités statistiques de son environnement langagier. Avoir des mesures de fréquence de mots basées sur un corpus d'écrits adressés à l'enfant, c'est avoir en quelque sorte une mesure indirecte de ces régularités.

Manulex (Lété, 2003, 2004 ; Lété, Sprenger-Charolles et Colé, 2004) a été construite à partir de 54 manuels scolaires de lecture. Quatre sous-corpus ont été définis : CP, CE1, CE2-CM2, et CP-CM2 (1,9 millions de mots au total). Chaque corpus a été traité par *Cordial Analyseur*© pour effectuer un étiquetage morpho-syntaxique des mots. Nous avons travaillé ensuite à la construction des fréquences de deux lexiques : un lexique des formes orthographiques (48 886 entrées) et un lexique des lemmes (23 812 entrées). Pour calculer les fréquences, nous nous sommes appuyés sur "*The American Heritage Word Frequency Book"* de Carroll et al. (1971) qui est la référence, en langue Anglaise, pour les normes de fréquences chez l'enfant. Pour les quatre niveaux et les deux lexiques, nous avons calculé 3 indices : *F* : la fréquence brute relevée dans le corpus en question ; *D* : un indice de dispersion du mot dans les manuels ; et *U* : la fréquence par million pondérée par *D* . Le fait de pondérer la fréquence par *D* permet d'obtenir une mesure plus fiable et donc d'approcher la fréquence "réelle" dans un corpus de taille infinie.

Manulex est librement téléchargeable à l'adresse http://unpc.univ-lyon2.fr/~lete/manulex/index.htm et interrogeable sur le site de *Lexique* à l'adresse http://www.lexique.org/moteur/

2- Les normes de fréquences infra-lexicales : la base *Manulex-infra*

Manulex-infra (Peereman, Lété, Sprenger-Charolles, soumis) fournit plusieurs normes statistiques décrivant le système d'écriture du français auquel est confronté l'enfant du CP au CM2 dans ses manuels scolaires. Nous avons en particulier développé une métrique de la consistance des relations grapho-phonologiques et phono-graphiques à chaque niveau d'âge pour tous les mots du lexique des formes orthographiques de *Manulex*. La consistance de chaque mot est mesurée sur une échelle de 1 à 100 à chaque niveau car celle-ci peut varier en fonction du lexique adressé à l'enfant à une étape particulière de son apprentissage. En plus de la consistance, les fréquences positionnelles des bigrammes, trigrammes et syllabes sont fournies.

Pour calculer l'indice de consistance et les fréquences des associations, un mot comme *main* se voit d'abord affecté de sa séquence d'associations graphèmes ↔ phonèmes : [m-m.ain-5]. Puis les fréquences de chaque association précédemment calculées sur l'ensemble des mots de la base sont associées au début, au milieu (moyenne des fréquences) et à la fin du mot considéré. Pour *main* , l'association [ain-5] en fin de mot a une fréquence de 2199 par million au CP. Pour calculer les indices positionnels de consistance, toutes les associations possibles sont d'abord référencées dans les deux sens possibles (graphème → phonème et phonème → graphème). On calcule ensuite la probabilité d'apparition d'une association particulière rapportée à l'ensemble des cas possibles (multipliée par 100). Pour *main*, l'association finale phonème → graphème [5-ain] a un indice de 17.64. Cela signifie que, sur 100 apparitions de la relation [5-ain] en fin de mot, le phonème /5/ s'écrit "ain" dans 17.64% des cas (en fin de mot et dans le corpus CP, il s'écrit "in" dans 37.57% des cas, "en" dans 41.32% des cas, …). L'association phonème → graphème [m-m] de début de mot a un indice de 100 ce qui veut dire que, sur 100 apparitions de la relation [m-m] en début de mot, le phonème /m/ s'écrit toujours "m". Le mot *main* se verra affecter d'un indice moyen de consistance phonème → graphème de 58.82. Considérons maintenant la consistance dans le sens graphème → phonème (lecture à voix haute). Les indices sont de 100 tant en début de mot ("m" se lit toujours /m/ en début de mot) qu'en fin de mot ("ain" se lit toujours /5/ en fin de mot). Le mot main a ainsi un indice moyen de consistance graphème → phonème de 100. Le mot *main* est donc deux fois plus difficile à écrire qu'à prononcer.

La discriminabilité de chaque mot à chaque niveau est également évaluée grâce au calcul du voisinage orthographique, des homophones, des homographes et du point d'unicité orthographique. *Manulex-infra* n'a pas d'équivalent dans les autres langues.

Manulex-infra est librement téléchargeable à l'adresse http://leadserv.u-bourgogne.fr/ (lien à préciser).

3- Un exemple d'utilisation des normes de fréquence pour estimer les répertoires lexicaux des enfants de 6 à 11 ans[1]

Un vocabulaire de base a été extrait de *Manulex* en sélectionnant toutes les entrées communes aux trois niveaux (Lété, 2003 ; 2004). De plus, nous avons pris un critère de dispersion supérieur à .25 signifiant que chaque mot a été trouvé dans 3/4 des manuels à chaque niveau. Le vocabulaire extrait comporte 3 215 lemmes qui couvrent près de 95% des formes orthographiques relevées au CP, 91% de celles du CE1 et 83% de celles du cycle 3. Autrement dit, un enfant de CE1 connaissant ces mots est capable de lire près de 90% de l'écrit de son niveau.

Nous avons ensuite comparé trois sources de données pour estimer le répertoire lexical des enfants :

a) Ehrlich, Bramaud du Boucheron et Florin (1978) : les auteurs ont proposé à des élèves de CE1 au CM2 une tâche de jugement de mots (450 par élève) sur une échelle en 5 points (*je connais très bien, …, je ne connais pas*). La mesure du répertoire correspond au nombre de mots jugés "*connus*" à "*très bien connus*" rapportés au nombre total de mots de l'échantillon (13 500). Il s'agit donc d'une estimation.

b) *Échelle d'Orthographe Lexicale* (EOLE) de Pothier et Pothier (2003) : les auteurs ont fait orthographier 11 694 mots à des élèves de CP au CM2 (50 mots par élève). Le nombre de mots correctement orthographiés par 75% des élèves sert de mesure du répertoire.

c) *Manulex* : Lété et al. (2003) : nous avons estimé le nombre moyen de mots susceptibles d'être rencontrés par les élèves dans leur manuel respectif aux trois niveaux considérés. Le nombre moyen de lemmes en réception de l'écrit fournit la mesure du répertoire lexical en réception d'écrit.

[1] Cette partie synthétise la 3ème partie de Lété (2004) (cf. document PDF fourni).

Les données indiquent qu'en réception d'écrit, les enfants sont confrontés en moyenne dans leur manuel à un stock de 2 000 mots au CP, 3 000 au CE1 et 5 000 au cycle 3. Au niveau orthographique, les 3/4 des enfants orthographient correctement 300 mots au CP, 1 000 au CE1, 2 000 au CE2, 3 500 au CM1 et 5 000 au CM2. Les données en jugement de connaissance surestiment certainement le stock lexical puisqu'elles donnent près de 5 500 mots connus dès le CE1 et près de 9 500 au CM2. Cela est donc supérieur à ce qui peut être rencontré dans un manuel de lecture à chaque niveau.

En résumé, la taille du vocabulaire d'un enfant à la fin du cycle 3 (11 ans) est de l'ordre de 5 000 mots (lemmes). L'enseignement du vocabulaire n'est donc pas une tâche insurmontable à l'école. Ceci dit, connaître 5 000 mots ne suffit pas à être lecteur. Les écrits adultes comportent approximativement près de 60 000 lemmes (cf. *Lexique*, New et al., 2001, 2004). Pour que l'enfant enrichisse sa base de vocabulaire, il doit donc lire énormément étant donné que la probabilité de rencontrer ces mots est faible : ce sont principalement des mots rares pour lesquels plusieurs "rencontres" sont nécessaires afin de garder une trace en mémoire lexicale.

Références

Carroll, J. B., Davies, P., & Richman, B. (Eds.) (1971). *The American Heritage Word-Frequency Book*. Boston, MA: Houghton Mifflin.

Ehrlich, S., Bramaud du Boucheron, G., & Florin, A. (1978). *Le développement des connaissances lexicales à l'école primaire*. Paris : PUF.

Lété, B. (2003). Building the mental lexicon by exposure to print: A corpus-based analysis of French reading books. In P. Bonin (Ed.), *Mental lexicon. "Some words to talk about words"* (pp. 187-214). Hauppauge, NY : Nova Science Publisher.

Lété, B. (2004). MANULEX : Le lexique des manuels scolaires de lecture. Implications pour l'estimation du vocabulaire des enfants de 6 à 11 ans. In E. Calaque & J. David (Eds.), *Didactique du lexique : Contextes, démarches, supports* (pp. 241-257). Bruxelles : De Boeck.

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.

Peereman, R., Lété, B., & Sprenger-*Charolles*, L. (soumis). Manulex-Infra: Grade-level statistics upon grapheme-phoneme associations from child-directed written material.. *Behavior Research Methods.*

Pothier, B, & Pothier, P. (2003). *EOLE : Échelle d'acquisition en orthographe lexicale (du CP au CM2)*. Paris : Retz.