17 MANULEX : une base de données du lexique écrit adressé aux élèves

Bernard LÉTÉ
INRP-Université de Lyon 2
Laboratoire EMC/DDL (CNRS UMR 5596)

A ■ Introduction

Manulex[1] (Lété, Sprenger-Charolles & Colé, 2004) est une base de données lexicales qui fournit les fréquences d'occurrences de 23 900 lemmes et 48 900 formes orthographiques extraits d'un corpus de 54 manuels scolaires de lecture. Librement accessible sur le Web[2], la base est à la disposition des chercheurs qui travaillent notamment sur l'acquisition de la lecture. Ils y trouveront des normes fiables et actualisées de l'exposition à l'écrit des enfants à trois niveaux d'âge : CP, CE1 et cycle 3. Pour l'enseignement du vocabulaire, les enseignants du primaire pourront extraire de ces tables de fréquence des listes de mots adaptées aux enfants de cycle 2 et 3.

Les modèles actuels sur l'acquisition de la lecture et notamment les modèles connexionnistes qui simulent, sous forme de réseaux de neurones, l'apprentissage de l'écrit (cf. pour une discussion Lété, 2003) accordent une place prépondérante à ce qui est appelé l'exposition à l'écrit. Selon ces modèles (par exemple Plaut *et al.*, 1996), 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

1. Pour Lexique des Manuels.
2. Version française téléchargeable à http://unpc.univ-lyon2.fr/~lete/manulex/index.html
Version anglaise téléchargeable à http://www.lpl.univ-aix.fr/lpl/ressources/manulex/

de fréquence de mots basées sur un corpus d'écrits adressé à l'enfant, c'est avoir en quelque sorte une mesure indirecte de ces régularités.

Chez l'adulte, on peut considérer que le stock lexical n'évolue plus de façon significative à partir de 20-25 ans. Cette stabilité permet aux données quantitatives obtenues à partir de corpus d'écrits adressés à l'adulte d'être relativement fiables pour étudier le lexique. Il n'en va pas de même pour l'enfant pour lequel le stock de vocabulaire se constitue dès le plus jeune âge (9 à 15 mois en moyenne) et subit une nette accélération à partir de 2 ans, mais surtout à partir de l'apprentissage de la lecture. S. Ehrlich, G. Brameraud du Boucheron & A. Florin (1978) notent en effet une importante accélération du développement des connaissances lexicales entre le CE1 et le CM2 (7 ans et 10 ans) : le nombre de mots appris chaque année est supérieur de 50 % environ à ce qu'il était en moyenne entre 1 an et 7 ans. Il était donc nécessaire de disposer d'estimations du répertoire lexical de l'enfant en réception du langage à chaque niveau d'âge, et ceci au moins durant l'école primaire. De plus, il convenait de combler un vide dans les outils langagiers de langue française mis à la disposition des chercheurs : en effet, contrairement à ce qu'il existe pour la langue anglaise, il existe peu de bases de données lexicales utilisables pour les recherches en psycholinguistique.

Chez l'adulte, on compte essentiellement Brulex[3] (Content *et al.*, 1990) et Lexique[4] (New *et al.*, 2001) qui est désormais la référence en langue française pour les recherches chez l'adulte. Afin de constituer la base initiale de mots, B. New *et al.* ont travaillé sur un corpus de 31 millions de mots extraits de textes de la base Frantext[5] publiés entre 1950 et 2000. La base possède 130 000 entrées sur les formes orthographiques, 73 000 entrées sur les lemmes et 17 attributs linguistiques (nombre de lettres, nombre de phonèmes, nombre de syllabes, etc.).

En ce qui concerne les études chez l'enfant, les chercheurs en sont réduits à utiliser les vocabulaires de base (Catach & Jejcic, Groupe HESO, 1984 ; Gougenheim *et al.*, 1964 ; Ters *et al.*,1969) dont le but est de déterminer les termes fondamentaux d'une langue qui doivent être appris selon une progression définie. Ils ne permettent cependant pas d'appréhender le répertoire lexical de l'enfant et la façon dont se développe l'acquisition du vocabulaire aux différents âges. Le manque est plus criant en ce qui concerne les fréquences de mots pour mesurer l'exposition à l'écrit : jusqu'à la mise à disposition de Novlex[6] (Lambert & Chesnet, 2001), il n'existait pas de tables de fréquence de mots, en langue française, relatives à l'écrit adressé à l'enfant. E. Lambert & D. Chesnet ont sélectionnés 19 manuels scolaires de CE2 et 19

3. Brulex est téléchargeable sur le site de Manulex.
4. Téléchargeable à http://www.lexique.org/moteur/
5. Frantext est une base de données textuelles développée par l'INALF-Nancy, devenu aujourd'hui l'ATILF. Elle regroupe à l'heure actuelle 3600 textes représentatifs du français des XIX[e] et XX[e] siècles et renferme 183 M de mots. Site officiel : http://atilf.atilf.fr/frantext.htm
6. Téléchargeable à http://www2.mshs.univ-poitiers.fr/novlex/

livres de littérature de jeunesse (417 000 mots au total). La base donne la fréquence d'occurrence pour 100 millions de mots de 20 600 formes orthographiques et 9300 lemmes.

Manulex fournit les fréquences d'occurrences des mots à 3 niveaux d'expertise de la lecture : le CP (6 ans) où se construit le lexique de l'enfant sur la base de la médiation phonologique, le CE1 (7 ans) où se construit le lexique orthographique par automatisation progressive de la reconnaissance du mot écrit et le cycle 3 (8-11 ans) où se consolide et s'enrichit le stock lexical par exposition répétée à l'écrit.

Pour calculer les fréquences des mots, nous nous sommes appuyés sur le travail de J.B. Carroll *et al.* (1971). Leur *American Heritage Word Frequency Book* est largement utilisé dans les recherches en langue anglaise. Leur base contient 86 800 mots extraits d'un corpus de 6000 titres de livres adressés aux enfants américains de 7 à 15 ans. Pour chaque titre, ils ont extrait un passage de 500 mots (3 millions de mots au total) et ont calculé quatre indices pour rendre compte des fréquences d'occurrences. Ces indices sont repris dans Manulex.

Nous détaillerons ici la construction de la base et les calculs des fréquences basés sur J.B. Carroll *et al.* (1971). Sur la base du lexique obtenu dans les manuels de lecture, nous nous interrogerons ensuite sur le stock lexical d'un enfant de 6 à 11 ans à mesure que s'accroît son exposition avec l'écrit.

B ■ Le corpus de manuels scolaires de Manulex

1 Échantillonnage du corpus

Nous nous sommes appuyés sur les chiffres des ventes de l'année 1994 pour constituer un échantillon de manuels représentatifs de ceux utilisés dans les écoles primaires françaises. Nous nous sommes donnés un critère proche de 75 % des ventes cumulées à chaque niveau considéré. Nous avons ainsi constitué un échantillon de 54 manuels de lecture dont le détail est donné au tableau 1. L'ensemble est composé de 45 manuels d'apprentissage et de 9 manuels de français (les titres des manuels sont donnés en annexe 1). Ainsi, au CE1, l'échantillon des 10 manuels de lecture représente 72 % des ventes totales de manuels de lecture de ce niveau en 1994 ; les 3 manuels de français retenus représentant 87 % de ces ventes[7].

7. Nous n'avons pas retenu de livres de littérature de jeunesse parce que nous ne disposions pas de données précises quant à leur impact auprès des enfants de 6 à 11 ans. Ceci dit, chacun sait que les manuels de lecture font un usage très important d'extraits tirés de livres de jeunesse. Autrement dit, le lexique des manuels reflète largement celui adressé à l'enfant dans ces livres.

Tableau 1 : Corpus des manuels de Manulex

	Manuels d'apprentissage		Manuels de français	
	Nombre	Pourcentage cumulé des ventes 1994	Nombre	Pourcentage cumulé des ventes 1994
CP	13	73 %	-	-
CE1	10	72 %	3	87 %
CE2	9	75 %	2	84 %
CM1	7	73 %	2	76 %
CM2	6	79 %	2	81 %
Total	**45**		**9**	

1.1 Le traitement automatique du corpus

Après numérisation des 8800 pages des manuels, une reconnaissance optique des caractères a été effectuée. Lors de la vérification des textes numérisés produits par le logiciel, nous avons uniquement ôté les entêtes de page et nous avons laissé l'ensemble des mots que l'enfant était susceptible de lire (entêtes de chapitre, notes de bas de page, tables des matières, légendes des tableaux et des figures, etc.).

La construction de tables de fréquence suppose que l'on dispose à la fois des fréquences sur les formes orthographiques des mots rencontrés dans le corpus – les flexions des mots –, mais également des fréquences sur les lemmes, le tout étant associé à la catégorie morphosyntaxique du mot. En effet, une entrée dans un dictionnaire de langue est généralement normalisée à sa forme la plus simple. Pour les verbes, par exemple, on retrouve l'infinitif plutôt que la forme conjuguée. Un adjectif sera présenté dans sa forme la moins marquée morphologiquement, généralement le masculin singulier. Pour désigner ces entrées, on emploie souvent le terme de *lemme*. Certaines entrées peuvent être multiples pour tenir compte de graphies identiques correspondant à des racines et à des catégories grammaticales différentes (par exemple, grand peut être un nom, un adjectif ou un adverbe).

L'ampleur de ce travail de lemmatisation a été considérablement réduit grâce à l'utilisation d'un logiciel de traitement automatique du langage disponible dans le commerce : Cordial Analyseur (version 8.13) de Synapse Développement. Cordial est un étiqueteur morphosyntaxique qui reconnaît la fonction syntaxique d'un mot en contexte et lui attribue une valeur grammaticale unique à l'aide de 130 catégories. Nous avons ramené le nombre de ces catégories à 11 dans la version diffusée de la base : noms, adjectifs, adver-

bes, verbes, noms propres, déterminants, interjections, pronoms, prépositions, conjonctions et abréviations.

Les listes des mots à chaque niveau ont ensuite été construites sur la base des fichiers étiquetés obtenus. Les fréquences brutes des lemmes et des formes orthographiques ont été calculées et ont été associées à leur catégorie morphosyntaxique respective. De plus, pour chaque entrée, on a également relevé le nombre de manuels dans lesquels apparaissait chaque mot de façon à pouvoir calculer un indice de dispersion du mot dans chaque sous-corpus (cf. infra).

2 La base lexicale Manulex

2.1 Les lexiques de Manulex

À partir de l'étiquetage des textes des manuels, deux lexiques ont été constitués : un lexique des formes orthographiques (48 886 entrées au total) et un lexique des lemmes (23 812 entrées au total). Les lexiques fournissent les fréquences aux trois niveaux définis précédemment caractérisant un niveau de maîtrise de l'écrit : CP, CE1 et CYCLE 3[8]. Nous avons également regroupé l'ensemble des manuels dans un quatrième niveau (TOTAL) afin que le chercheur travaillant sur le lexique au delà de 11 ans (au collège par exemple) puisse disposer de données intermédiaires entre l'écrit adressé à l'enfant et l'écrit adressé à l'adulte.

Les données du tableau 2 indiquent le nombre d'entrées respectives des deux lexiques pour chaque niveau (les données sont arrondies à la centaine supérieure).

Tableau 2 : Nombre de mots et nombre d'entrées du lexique des formes orthographiques et du lexique des lemmes à chaque niveau

	CP	CE1	CYCLE 3	TOTAL
Mots	172 400	351 100	1 386 600	1 910 000
Lexique des formes orthographiques	11 400	19 100	45 600	48 900
Lexique des lemmes	6 800	10 500	22 500	23 900

Le corpus total représente près de 2 millions de mots. On note une progression de 3700 nouveaux lemmes du CP au CE1 et de 12 000 nouveaux lemmes du CE1 au cycle 3. À titre de comparaison, le nombre de lemmes de

8. Respectivement 6-7, 7-8 et 8-11 ans.

notre corpus total représente un tiers du nombre des lemmes de la base Lexique qui ont été extraits d'un corpus 15 fois supérieur au nôtre en nombre de mots.

2.2 Le calcul des fréquences

Pour établir les listes des fréquences des mots, nous avons utilisé les indices de J.B. Carroll *et al.* (1971) que nous avons calculé à chaque niveau considéré (CP – CE1 – CYCLE 3 – TOTAL) et pour chaque entrée de la base du lexique des lemmes et du lexique des formes orthographiques. Les indices sont au nombre de quatre (les formules exactes des calculs sont données en annexe 2) :

1. F : la fréquence brute du mot dans le corpus.
2. D : l'indice de dispersion du mot parmi les manuels.
3. U : la fréquence par million estimée à partir de D.
4. IFC : un indice de fréquence courant calculé à partir de U par transformation logarithmique.

Pourquoi calculer un indice de dispersion ? Comme nous l'avons précisé plus haut, Manulex veut proposer des normes statistiques fiables sur l'exposition à l'écrit des enfants. Les valeurs obtenues des fréquences de mots renvoient en quelque sorte à la qualité (ou à la force) des représentations cognitives des mots en mémoire lexicale chez un enfant. Or les fréquences lexicales ne sont jamais indépendantes du corpus à partir duquel on calcule ces valeurs. Même si on peut estimer que notre corpus est suffisamment représentatif du matériel écrit susceptible d'être rencontré par un enfant de 6 à 11 ans, tout ajout ou retrait d'un manuel modifierait les fréquences obtenues. De plus, un mot peut être sur-représenté dans un manuel et ne pas apparaîte dans les autres manuels. Par exemple, un manuel de CE2 propose à chaque fin de chapitre des mots croisés. Les mots « horizontalement » et « verticalement » sont dans ce manuel sur-représentés par rapport à la fréquence qu'on pourrait trouver, en moyenne, dans les autres manuels. La valeur de F (fréquence brute) ne reflètera donc pas la probabilité de rencontrer ce mot dans un corpus de textes de niveau CP mais la probabilité de rencontrer ce mot dans le manuel en question. Pour prendre un exemple plus précis, au CP, les mots *papa* et *point* ont respectivement une fréquence brute de 270 et 276. Les fréquences sont donc équivalentes et donnent une fréquence par million de 1.567 et 1.601 (en tenant compte du nombre de mot du corpus de CP). Cependant alors que *papa* a été trouvé 21 fois dans chaque manuel de CP, *point* a été trouvé 242 fois dans un seul manuel. Dès lors, la probabilité de rencontrer le mot *point* pour un élève de CP est beaucoup plus faible que celle de rencontrer *papa*.

Il faut donc pondérer les fréquences par un indice de dispersion (D) qui va permettre de réduire ces erreurs d'échantillonnage. Sa valeur est comprise

entre 0 (le mot a été trouvé dans un seul manuel) et 1 (le mot a été trouvé dans tous les manuels avec la même proportion). La fréquence par million U est dérivée de F avec un ajustement avec D. U va approcher la valeur réelle de F dans un corpus théorique de taille infinie. C'est la fréquence d'usage du mot dans les manuels qui traduit mieux son importance que la seule fréquence F. Pour l'exemple qui nous intéresse ici, les fréquences brutes respectives de papa et point (270 vs 276) qui auraient donné respectivement une fréquence par million de 1.567 et 1.601, seront pondérées par D (0.79 et 0.24) et traduites en une fréquence par million (U) qui sera respectivement de 1 270 et 507, cette dernière étant donc deux fois moins élevée que celle qui aurait été obtenue sans cette pondération (1.601).

C. Muller (1992) montre que, pour les mots à très haute fréquence, D modifie très peu le classement primitif, alors qu'il reclasse fortement les mots à fréquence moyenne ou faible. Les écarts des premiers ne sont dûs en général qu'à des raisons stylistiques, alors que pour les derniers, les écarts sont dus, dans la terminologie de Muller, à des raisons stylistiques et thématiques. Pour nous, c'était un élément important à considérer, car pour donner des normes fiables d'exposition à l'écrit, il ne faut pas identifier un mot comme fréquent alors qu'il est en réalité rare et contexte dépendant.

Le tableau 3 donne la distribution des valeurs de U et les compare à celles de Lexique[9]. Le chercheur ou l'enseignant peuvent utiliser ces valeurs pour déterminer les limites des mots fréquents et rares. On observe d'abord que la moyenne s'échelonne de 135 occurrences par million au CP, à 94 au CE1 et 47 au cycle 3. Un corpus de textes adressés à l'adulte donne une fréquence par million de l'ordre de 20 en moyenne. Ensuite, les valeurs des médianes (centile 50), largement en dessous des valeurs moyennes, indiquent que beaucoup de mots ont une fréquence d'occurrence faible. Ceci est vrai à partir du cycle 3 où 50 % des mots sont en dessous de 1 occurrence par million, cette valeur rejoignant celle observée dans un corpus d'écrits adressés à l'adulte. Enfin, les différences des valeurs de U entre nos trois niveaux de base (CP, CE1 et CYCLE 3) nous indiquent qu'il est important de différencier ces 3 lexiques. En effet, les valeurs observées au CP montrent que le lexique est fortement répété dans les manuels, car 25 % des mots ont une fréquence d'occurrence supérieure à 43 par million. C'est déjà moins vrai au CE1 où 25 % des mots a une fréquence d'occurrence supérieure à 19 par million. Au cycle 3, le centile 75 approche celui observé dans Lexique.

9. Lexique donne simplement les valeurs F par million.

Tableau 3 : Distributions de U (fréquences pondérées par million) pour Manulex et comparaison avec Lexique

	MANULEX				LEXIQUE
	CP	CE1	CYCLE 3	TOTAL	
Moyenne	135	94	47	45	20
Centile 25	1	0	0	0	0
Centile 50 (médiane)	8	3	1	1	1
Centile 75	43	19	8	6	3

3 Une estimation du vocabulaire des enfants de 6 à 11 ans

Nous distinguerons ici le lexique des manuels, c'est-à-dire les mots qu'un enfant de 6 à 11 ans « risque » de rencontrer dans son parcours scolaire, du vocabulaire de l'enfant, c'est-à-dire son stock lexical en réception ou en production de l'écrit. Ce stock peut justement être estimé à partir d'une analyse du lexique des manuels. Pour nous psycholinguistes, c'est un peu une démarche inverse à celle d'un linguiste – un lexicographe par exemple – qui part des fréquences des mots d'un texte – le vocabulaire du texte – pour remonter au lexique d'une langue en estimant les probabilités d'emploi des mots rencontrés. Ici, en descendant d'un lexique de situation – le lexique de notre corpus de manuels – à un lexique individuel hors situation – le vocabulaire –, nous descendons du niveau du discours à la mise en pratique de la langue par un individu, véritable champ d'étude de l'apprentissage du lexique.

Les estimations de la taille du vocabulaire des enfants varient considérablement d'un auteur à l'autre et les chiffres avancés semblent souvent aberrants. Une étude continuellement citée dans la littérature scientifique en langue anglaise est celle de J.M. Anglin (1993). L'auteur considère que le stock lexical d'un enfant de CP est de 10 400 mots[10], de 19 500 chez un enfant de CE2 et de 40 000 chez un enfant de CM2. Même si la langue est différente, on peut par exemple rapporter l'estimation de J.M. Anglin au nombre des entrées de la dernière édition du dictionnaire Larousse (2004) : le dictionnaire contient 59 000 entrées sur les noms communs et 28 000 sur les noms propres. On a vu également que la base Lexique construite à partir d'un corpus de 32 millions de mots possède 73 000 lemmes différents.

10. En réalité, ce sont des familles morphologiques de mots : selon sa classification, chant et chanter constituent un même mot de base.

L'erreur de ces estimations est de partir d'un lexique de référence représentant l'ensemble des mots de la langue. Par exemple, J.M. Anglin extrapole ses données à partir des 460 000 entrées du Webster. Ces estimations ne peuvent refléter le stock lexical puisqu'il est impossible qu'un enfant puisse être confronté à l'ensemble des mots de sa langue et, même si cela était possible, il faudrait plusieurs lectures pour qu'il puisse garder une trace des occurrences de mots en mémoire lexicale. Pour estimer la taille du vocabulaire d'un enfant ou d'un adulte, il faut, à notre avis, raisonner sur l'exposition à l'écrit et se demander ce à quoi a pu être confronté l'adulte et l'enfant lors de ses lectures. C'est lors de cette confrontation à l'écrit que se construit le stock lexical de l'enfant et que se consolide et s'enrichit celui de l'adulte. C'est pourquoi il faut faire la différence entre le lexique – la liste des mots d'une langue – et le vocabulaire – la mise en pratique du lexique d'une langue par un individu.

Le tableau 2 indiquait que le lexique des manuels scolaires de lecture était de 6800 lemmes au CP, de 10 500 au CE1 et de 22 500 au cycle 3. Nous notions également que ces valeurs étaient étroitement dépendantes du corpus considéré et que si nous avions ajouté des manuels au CP, nous aurions pu nous approcher du lexique de CE1. Pour estimer le nombre de mots auquel un enfant est exposé de 6 à 11 ans, et ainsi quantifier son vocabulaire, on peut raisonner sur un manuel. C'est une bonne estimation chez un enfant, car un enfant lit peu et lorsqu'il lit un peu plus que ses pairs, il lit des livres avec un lexique de mots fréquents : comme on le verra, il accroît peu son stock de vocabulaire, car il est exposé de nombreuses fois aux mêmes mots.

Le tableau 4 donne une estimation du vocabulaire en réception de l'écrit du CP au cycle 3 basée sur l'analyse moyenne des lemmes extraits des manuels. Une comparaison a été effectuée avec un livre de la littérature (*Les Lettres de mon moulin* d'Alphonse Daudet).

Tableau 4 : Estimation du vocabulaire en réception de l'écrit d'un enfant du CP au cycle 3. Les écarts-types sont entre parenthèses

	MANULEX			DAUDET
	CP	CE1	CYCLE 3	
Total des mots	13 300 (4 200)	27 100 (10 200)	49 600 (20 100)	49 000
Formes orthographiques	2 500 (700)	4 300 (1 300)	7 600 (2 400)	8 000
Lemmes	1 900 (500)	2 900 (800)	4 900 (1 300)	5 500
Ratio Formes/Lemmes	1.3	1.5	1.6	1.5
Progression des lemmes	.	+ 1 000	+ 2 000	

Au CP, un manuel a en moyenne 13 300 mots, 2500 formes orthographiques différentes et 1900 lemmes (une forme = 1.3 lemme). Un manuel de CE1 a en moyenne 1000 lemmes de plus. Au cycle 3, la progression est de 2000 lemmes. Avec 4900 lemmes considérés comme acquis, on arrive à une estimation 5 fois inférieure au lexique des manuels qui représente les mots qui peuvent être rencontrés par les enfants. Au cycle 3, un manuel a en moyenne autant de formes orthographiques et de lemmes que le livre de A. Daudet pris en exemple.

Même si un enfant lit beaucoup plus, son stock de vocabulaire ne s'accroîtra pas forcément beaucoup. Pourquoi ? Prenons l'exemple du livre de A. Daudet : si l'enfant lit son manuel de lecture du cycle 3 plus *Les Lettres de mon moulin* dans l'année, il lira beaucoup de mots communs aux deux livres ce qui renforcera ses connaissances lexicales pour ces mots. Par exemple, le nombre de mots du livre de Daudet qui n'ont pas été trouvés dans les manuels du cycle 3 sont au nombre de 884 (15 % des lemmes du livre) et, parmi eux, il y a 215 noms propres (4 % des lemmes du livre). Autrement dit, si un enfant lit le livre de Daudet après avoir acquis le vocabulaire de son livre de cycle 3, il rencontrera de nouveaux lemmes mais, s'il veut consolider leurs traces en mémoire lexicale, il devra les rencontrer à nouveau plusieurs fois (au moins 2 à 3 fois). Mais la probabilité de rencontrer ces mêmes mots dans un autre livre est faible. Ceci est dû à un phénomène bien connu des lexicographes et qui est illustré au tableau 5 : un petit nombre de mots occupent une forte proportion des occurrences d'un texte ; un grand nombre, au contraire, ont une très faible occurrence et sont rencontrés une seule fois dans le corpus considéré[11].

Tableau 5 : Taux de couverture des mots selon les rangs ordonnés des lemmes aux trois niveaux de Manulex

Rangs des lemmes	CP	CE1	Cycle 3
10	30 %	29 %	28 %
100	57 %	58 %	56 %
500	76 %	77 %	72 %
1 000	85 %	84 %	80 %
2 000	93 %	91 %	87 %
3 000	96 %	94 %	91 %

11. G.K. Zipf (1932) a traduit cette observation sous forme d'une loi : *si* on classe les mots d'un texte par fréquences décroissantes *f* et que l'on doie chacun d'eux d'un rang traduit por un nombre *r*, le produit *f* * *r* tend à être constant.

Au CP, les 1000 premiers lemmes représentent à eux seuls près de 85 % des mots. Autrement dit, sur 100 pages d'un manuel, on a sur 85 pages une répétition de seulement 1000 mots (lemmes). Si l'enfant lit 11 pages de plus (11 %), il va rencontrer 2000 lemmes de plus mais ceux-ci seront beaucoup moins répétés (une seule fois vraisemblablement). Ceci dit, même si l'enfant connaît bien 1000 mots, il risque de buter sur 15 mots sur 100 rencontrés (plus d'un mot sur 10). Un mot sur 10, cela représente presque un mot par phrase. Ces mots peu fréquents sont spécifiques à un contexte phrastique et sont indispensables pour particulariser sémantiquement un énoncé.

En résumé, un enfant à la fin du cycle 3 possède, en réception de l'écrit, un stock de mots de base de l'ordre de 5000 unités pour lesquels on peut estimer qu'ils sont suffisamment consolidés en mémoire pour être bien compris et, pour la plus grande partie d'entre eux, produits à l'oral et à l'écrit. C'est quatre fois moins de ce qu'il est susceptible de rencontrer dans un corpus d'écrit de son niveau scolaire et 15 fois moins de ce qu'il est susceptible de rencontrer plus tard dans des écrits adressés à l'adulte.

Afin de compléter cette estimation, nous allons comparer ces données en réception de l'écrit à deux autres études. L'une (Ehrlich *et al.*, 1978) estime le répertoire lexical d'enfants du CE1 au CM2 à partir de jugements de connaissance ; l'autre, plus récente (Pothier & Pothier, 2003), fournit une échelle actualisée d'acquisition en orthographe lexicale (*EOLE*) et donne ainsi une estimation en production du langage.

S. Ehrlich *et al.* ont proposé à des élèves du CE1 au CM2 une tâche de jugement de mots (450 mots par élève) sur une échelle en 5 points. 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). B. Pothier & P. Pothier ont fait orthographier 11 700 mots à 48 900 élèves du CP au CM2 (50 mots par élève). Le nombre de mots correctement orthographiés par 75 % des élèves sert de mesure du répertoire. Le tableau 6 fait la synthèse des données : a) le lexique en réception tel que nous l'avons présenté ici dans la première partie ; b) le vocabulaire en réception de l'écrit estimé en début de cette partie ; c) le vocabulaire relatif aux mots jugés très bien connus dans l'étude d'Ehrlich *et al.* ; d) le vocabulaire correctement orthographié par 75 % des enfants dans l'étude de Pothier et Pothier.

Le recoupement de ces données indique un stock lexical beaucoup moins élevé que certaines estimations souvent rapportées dans la littérature scientifique, souvent de l'ordre de plusieurs dizaines de milliers de mots (Anglin, 1993 ; Nagy & Anderson, 1984). En réception de l'écrit, les enfants sont confrontés en moyenne à 1900 mots au CP, 2900 au CE1 et 4900 au cycle 3. Ces valeurs s'éloignent du lexique théorique auquel peut être confronté un enfant (22 500 lemmes au cycle 3). Les données en jugement de connaissance sur les mots très bien connus (pour lesquels on peut considérer que les juge-

Au CP, les 1000 premiers lemmes représentent à eux seuls près de 85 % des mots. Autrement dit, sur 100 pages d'un manuel, on a sur 85 pages une répétition de seulement 1000 mots (lemmes). Si l'enfant lit 11 pages de plus (11 %), il va rencontrer 2000 lemmes de plus mais ceux-ci seront beaucoup moins répétés (une seule fois vraisemblablement). Ceci dit, même si l'enfant connaît bien 1000 mots, il risque de buter sur 15 mots sur 100 rencontrés (plus d'un mot sur 10). Un mot sur 10, cela représente presque un mot par phrase. Ces mots peu fréquents sont spécifiques à un contexte phrastique et sont indispensables pour particulariser sémantiquement un énoncé.

En résumé, un enfant à la fin du cycle 3 possède, en réception de l'écrit, un stock de mots de base de l'ordre de 5000 unités pour lesquels on peut estimer qu'ils sont suffisamment consolidés en mémoire pour être bien compris et, pour la plus grande partie d'entre eux, produits à l'oral et à l'écrit. C'est quatre fois moins de ce qu'il est susceptible de rencontrer dans un corpus d'écrit de son niveau scolaire et 15 fois moins de ce qu'il est susceptible de rencontrer plus tard dans des écrits adressés à l'adulte.

Afin de compléter cette estimation, nous allons comparer ces données en réception de l'écrit à deux autres études. L'une (Ehrlich *et al.*, 1978) estime le répertoire lexical d'enfants du CE1 au CM2 à partir de jugements de connaissance ; l'autre, plus récente (Pothier & Pothier, 2003), fournit une échelle actualisée d'acquisition en orthographe lexicale (*EOLE*) et donne ainsi une estimation en production du langage.

S. Ehrlich *et al.* ont proposé à des élèves du CE1 au CM2 une tâche de jugement de mots (450 mots par élève) sur une échelle en 5 points. 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). B. Pothier & P. Pothier ont fait orthographier 11 700 mots à 48 900 élèves du CP au CM2 (50 mots par élève). Le nombre de mots correctement orthographiés par 75 % des élèves sert de mesure du répertoire. Le tableau 6 fait la synthèse des données : a) le lexique en réception tel que nous l'avons présenté ici dans la première partie ; b) le vocabulaire en réception de l'écrit estimé en début de cette partie ; c) le vocabulaire relatif aux mots jugés très bien connus dans l'étude d'Ehrlich *et al.* ; d) le vocabulaire correctement orthographié par 75 % des enfants dans l'étude de Pothier et Pothier.

Le recoupement de ces données indique un stock lexical beaucoup moins élevé que certaines estimations souvent rapportées dans la littérature scientifique, souvent de l'ordre de plusieurs dizaines de milliers de mots (Anglin, 1993 ; Nagy & Anderson, 1984). En réception de l'écrit, les enfants sont confrontés en moyenne à 1900 mots au CP, 2900 au CE1 et 4900 au cycle 3. Ces valeurs s'éloignent du lexique théorique auquel peut être confronté un enfant (22 500 lemmes au cycle 3). Les données en jugement de connaissance sur les mots très bien connus (pour lesquels on peut considérer que les juge-

ments sont les plus fiables) sont du même ordre de grandeur : on arrive à 3300 mots à la fin du cycle 3. Les données rapportées par S. Ehrlich *et al.* sur les mots moyennement connus sont de l'ordre de 6000 au CM2. Cette estimation rend compte, à notre avis, des capacités de généralisation de l'enfant pour comprendre des mots nouveaux, mais n'est pas indicatrice du stock lexical de l'enfant, c'est-à-dire des mots pour lesquels il existe une fluence d'usage dans les activités de compréhension et de production du langage. Au niveau orthographique, les trois quarts des enfants orthographient correctement 300 mots au CP, 1000 au CE1, 2000 au CE2, 3500 au CM1 et 5000 au CM2. La production des graphies correctes équivaut alors au répertoire lexical estimé dans les deux autres études.

Tableau 6 : Synthèse des données analysées. Les chiffres renvoient au nombre de lemmes

	CP	CE1	CE2	CM1	CM2
a) lexique en réception (Manulex)	6 800	10 500	22 500	22 500	22 500
b) vocabulaire estimé en réception de l'écrit	1 900	2 900	4 900	4 900	4 900
c) S. Ehrlich *et al.* (1978) : vocabulaire estimé par des jugements de connaissance (mots très bien connus)	–	2 400	2 900	3 000	3 300
c) B. Pothier & P. Pothier (2003) : vocabulaire estimé en production d'orthographe lexicale (graphies correctes par 75 % des enfants)	400	1 100	2 200	3 700	5 100

C ■ Conclusion

Les données ici rapportées font apparaître que la taille du vocabulaire d'un enfant à la fin du cycle 3 (11 ans) est de l'ordre de 5000 mots (lemmes). L'enseignement du vocabulaire n'est donc pas une tâche insurmontable à l'école. Ceci dit, connaître 5000 mots ne suffit pas à être lecteur. Les écrits adultes comportent approximativement près de 73 000 lemmes (cf. Lexique, New *et al.*, 2001). 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 expositions sont nécessaires afin de garder une trace en mémoire lexicale.

Cependant, le simple comptage des mots ne suffit pas à rendre compte de la richesse lexicale d'un individu. Lorsqu'il apprend à lire, l'environnement linguistique de l'enfant est finalement bien pauvre. Mais rien ne l'empêchera de comprendre la phrase donnée en exemple par N. Chomsky (1957)

« d'incolores idées vertes dorment furieusement »[12] pour peu que son lexique contiennent les six lemmes qui la composent. Ce constat est toujours d'actualité et questionne aujourd'hui les modèles connexionnistes où l'apprentissage est fait de confrontations répétées avec des événements linguistiques. On peut comprendre cette phrase, la discuter, l'interpréter de différentes façons bien que nous ne l'ayons jamais rencontrée. Nous devons donc repenser notre conception de l'apprentissage du lexique en intégrant l'extraordinaire capacité de généralisation du cerveau humain et sa singulière propension à créer du sens par tous les moyens.

D ■ Références bibliographiques

Anglin J.M. (1993). Vocabulary development : A morphological analysis. Monographs of the Society for Research. *Child Development*, serial n° 238, vol. 58-10.

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

Catach N., Jejcic F. & HESO groupe (1984). *Les Listes orthographiques de base du français (LOB). Les mots les plus fréquents et leurs formes fléchies les plus fréquentes*. Paris, Nathan.

Chomsky N. (1957). *Syntactic structures*. Haag, Mouton.

Content A., Mousty P. & Radeau M. (1990). Brulex : Une base de données lexicales informatisée pour le français écrit et parlé. *L'année psychologique* 90, pp. 551-566.

Dubois F. & Buyse R. (1940). Échelle Dubois-Buyse. *Bulletin de la Société Alfred Binet* 405, pp. 19-52.

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

Gougenheim G., Michéa R., Rivenc P. & Sauvageot A. (1964). *L'élaboration du français fondamental (1er degré)*. Paris, Didier.

Lambert E. & Chesnet D. (2001). NOVLEX : Une base de données lexicales pour les élèves de primaire. *L'année psychologique* 101, pp. 277-288.

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

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, pp. 156-166.

Muller C. (1992). *Principes et méthodes de statistique lexicale*. Paris, Champion.

Nagy W. E. & Anderson R. C. (1984). How many words are there in printed school English ? *Reading Research Quarterly* 19, pp. 304–330.

12. Cette phrase (en anglais : *Colorless green ideas sleep furiously*) est morphologiquement correcte mais est sémantiquement inacceptable (si on fait abstraction de sa poésie). Chomsky l'a donnée en exemple dans sa démonstration sur l'autonomie de la syntaxe pour illustrer la non-pertinence des considérations sémantiques pour poser un jugement de grammaticalité.

New B., Pallier C., Ferrand L. & Matos R. (2001). Une base de données lexicales du français contemporain sur Internet : LEXIQUE. *L'Année psychologique* 101, pp. 447-462.

Plaut D.C., McClelland J.L., Seidenberg M.S. & Patterson K. (1996). Understanding normal and impaired word reading : Computational principles in quasi-regular domains. *Psychological Review* 103, pp. 56-115.

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

Ters F., Mayer G. & Reichenbach D. (1969). *L'Échelle Dubois-Buyse d'orthographe usuelle française*. Neuchâtel, Messeiller.

Zipf G.K. (1932). *Selective studies and the principle of relative frequency in language.* Cambridge, MA, Harvard University Press.

Annexe 1 : Titres des manuels scolaires constituant le corpus de Manulex

CP

Titre	Éditeur	Année
Au fil des mots (3 livrets)	Nathan	96
Bien lire à l'école (premiers entraînements)	Nathan	96
Bigoudi et compagnie (3 livrets)	Nathan	95
C'est à lire	Hachette	96
Daniel et Valérie (2 livrets)	Nathan	96
Gafi le fantôme (2 livrets)	Nathan	96
Je lis seul, Tu lis seule (fiches de lecture)	Nathan	96
La Ruche aux livres	Hachette	97
Lecture à croquer	Magnard	96
Lecture en fête (2 livrets)	Hachette	96
Lire au CP (2 livrets)	Nathan	96
Paginaire	Hachette	95
Ratus et ses amis	Hatier	95

CE1

Titre	Éditeur	Année
a.r.t.h.u.r	Nathan	96
C'est à lire	Hachette	95
Éclats de lire	Magnard	95
Gafi Livre	Nathan	98
Je lis seul, Tu lis seule (fiches de lecture)	Nathan	97
La lecture silencieuse (fiches de lecture)	Nathan	96
La Ruche aux livres	Hachette	97
La Semaine de français	Nathan	96
Langue française (NBM)	Nathan	96
Le français au CE1	Hachette	96
Les 7 clés pour lire et pour écrire	Hatier	96
Paginaire	Hachette	95
Ratus découvre les livres	Hatier	96

CE2

Titre	Éditeur	Année
À la croisée des mots	Istra	96
a.r.t.h.u.r	Nathan	96
Bien lire à l'école (niveau 1)	Nathan	96
C'est à lire	Hachette	96
Éclats de lire	Magnard	95
Ixel sait lire	Hachette	96
Je lis seul, Tu lis seule (fiches de lecture)	Nathan	96
La lecture silencieuse (2 livres)	Nathan	96
La Ruche aux livres	Hachette	96
La Semaine de français	Nathan	96
Langue française (NBM)	Nathan	96
Les 7 clés pour lire et pour écrire	Hatier	95

CM1

Titre	Éditeur	Année
a.r.t.h.u.r	Nathan	96
Bien lire à l'école (niveau 2)	Nathan	96
C'est à lire	Hachette	94
Éclats de lire	Magnard	95
Je lis seul, Tu lis seule (fichier autocorrectif))	Nathan	96
La lecture silencieuse (premier livre)	Nathan	96
La Ruche aux livres	Hachette	96
La Semaine de français	Nathan	95
Langue française (NBM)	Nathan	96
Les 7 clés pour lire et pour écrire	Hatier	95

CM2

Titre	Éditeur	Année
a.r.t.h.u.r	Nathan	96
Bien lire à l'école	Nathan	96
C'est à lire	Hachette	96
Éclats de lire	Magnard	95
Je lis seul, Tu lis seule (fichier autocorrectif))	Nathan	96
La Lecture silencieuse (2 livres)	Nathan	96
La Semaine de français	Nathan	96
Langue française (NBM)	Nathan	96
Livre miroir	Magnard	91
Les 7 clés pour lire et pour écrire	Hatier	95

Annexe 2 : Formules de calcul des 4 indices des tables de fréquence de Manulex

Fréquence (F) : le nombre de fois où le mot apparaît dans le corpus.
Dispersion (D) : l'indice est basé sur la dispersion des occurrences à travers les manuels. Sa valeur s'échelonne de 0.00 to 1.00. D est égal à 0 quand toutes les occurrences du mot apparaissent dans un seul manuel. Il est égal à 1 si les occurrences sont distribuées en exacte proportion à travers les manuels. Les valeurs entre ces extrêmes indiquent le degré de dispersion du mot.
La formule pour calculer D est la suivante :

$$D = [\log (\sum p_i) - [(\sum p_i \log p_i) / \sum p_i]] / \log (n)$$

où :
n est le nombre de manuels à chaque niveau considéré (*n* = 13 au CP, 13 au CE1, 28 au cycle 3, et 54 pour l'ensemble du corpus, du CP au CM2)
i est le numéro du manuel (*i* = 1, 2, ..., *n*)
p_i est la fréquence (*F*) du mot dans le $i^{ème}$ manuel, avec $p_i \log p_i = 0$ si $p_i = 0$.

Fréquence estimée par million de mots (*U*) : la fréquence est dérivée de *F* avec un ajustement avec *D*. Quand *D* égale 1, *U* est calculé simplement comme la fréquence par million de mots. Mais quand *D* est inférieur à 1, la valeur de *U* est diminuée. On considère que *U* est une meilleure estimation de la fréquence réelle d'un mot qui serait trouvée dans un corpus de taille infinie. Cela permet des comparaisons directes des valeurs obtenues dans les quatre sous-corpus.
L'ajustement est effectué à l'aide de la formule suivante :

$$U = (1{,}000{,}000 / N) [FD + (1 - D) * f_{min}]$$

où :
N est le nombre total de mots dans le corpus (172 348 au CP ; 351 024 au CE1 ; 1 386 546 au cycle 3 ; et 1 909 918 du CP au CM2)
F est la fréquence du mot dans le corpus
D est l'indice de dispersion
f_{min} est 1/*N* fois la somme des produits de f_i et s_i, où f_i est la fréquence dans le manuel *i* et s_i est le nombre de mots de ce manuel.

L'Indice Standard de Fréquence (*ISF*) : il est dérivé directement de *U*. C'est un moyen simple d'indiquer les fréquences d'occurrences. Avec le logarithme, on considère qu'une distance, par exemple, de 10 occurrences entre deux mots de basse fréquence est plus importante qu'une différence de 10 entre deux mots de haute fréquence. La constante multiplicative « 10 » et la constante additive « 4 » permet d'avoir des références faciles à appréhender : un *ISF* de 90 signifie que le mot survient tous les 10 mots, un *ISF* de 80 signifie que le mot survient tous les 100 mots, un ISF de 70 signifie que le mot survient tous les 1000, ..., un *ISF* de 40 signifie que le mot survient une fois par million.
La formule de calcul est la suivante :

$$ISF = 10 * [\log_{10} (U) + 4]$$