# ESTIMATION DU VOCABULAIRE ENCYCLOPÉDIQUE SCOLAIRE MÉMORISÉ À L'ÉCOLE ÉLÉMENTAIRE

**Moïse Déro, Fabien Fenouillet**

**Groupe d'études de psychologie** | « Bulletin de psychologie »

2014/2 Numéro 530 | pages 127 à 141

ISSN 0007-4403

# Estimation du vocabulaire encyclopédique scolaire mémorisé à l'école élémentaire

Déro Moïse*
Fenouillet Fabien**

## INTRODUCTION

De nombreuses études ont mis en évidence que les connaissances en vocabulaire étaient l'un des meilleurs prédicteurs d'une aptitude verbale générale, et même des tests composites d'intelligence (Anderson, Freebody, 1979 ; Sternberg, Powell, 1983 ; Jenkins, Stein, Wysocki, 1984). Cependant, ces études sont très variées, certaines faisant un recensement, ce qui est déjà fort complexe, comme Nagy et Anderson (1984) ou Lété, Sprenger-Charolles et Colé (2004). D'autres recherches s'intéressent plutôt aux mécanismes d'acquisition, notamment chez le jeune enfant (Hepburn, 2010 ; Marulis, Neuman, 2010), mais aussi pour des vocabulaires spécialisés, comme celui des mathématiques (Brown, 2008). D'autres, enfin, s'intéressent aux relations avec d'autres performances scolaires ou cognitives, la lecture notamment ou la compréhension d'histoires (Verhoeven, van Leeuwe, Vermeer, 2011 ; Lee, 2011). Cette importance du vocabulaire s'observe également dans la corrélation avec les performances scolaires chez des élèves au collège, où le vocabulaire encyclopédique (histoire, mathématiques...) est mieux corrélé que la mémoire de travail ou les tests de raisonnement (Lieury, 1996 ; Van Acker, Vrignaud, Lieury, 1997 ; Lieury, Lorant, Le Cam, Rocher, 2013).

### Les inventaires de vocabulaire

Combien de mots un enfant connaît-il ? La question paraît simple, mais la diversité des réponses scientifiques fournies depuis plus d'un siècle démontre bien la complexité de cette interrogation. Si on examine les recherches les plus anciennes, quand Terman (1916) estime à 3 600 mots le vocabulaire moyen d'un enfant de langue anglaise de 8 ans, Smith (1941) l'évalue à 44 000 mots. Si, pour Seashore et Eckerson (1940) et Hartman (1941), les estimations du lexique des étudiants anglophones sont très vastes (respectivement 155 000 et 215 000 mots), celle de Kirkpatrick (1891) n'en inclut que 19 000.

Beaucoup de recherches ont déjà noté cette grande variabilité des estimations de vocabulaire (Anderson, Freebody, 1979 ; Anglin, 1993 ; Graves, 1986 ; Lorge, Chall, 1963 ; Miller, 1978, 1991 ; Nagy, Anderson, 1984 ; Nagy, Anderson, Herman, 1987 ; Nagy, Herman, 1987 ; Seashore, Eckerson, 1940). Trois principaux facteurs expliquent ces variations. Le premier porte sur le choix de la source d'échantillon (d'après des dictionnaires, des lettres, des textes, des transcriptions de productions orales...) et sa méthodologie d'extraction. Le second facteur est celui de la définition du « mot », que prend en compte telle ou telle recherche (la question de la réduction des occurrences de formes rencontrées et leur classification). Enfin, la manière dont les auteurs estiment qu'un mot est connu (évaluation de la connaissance, par définition, par production, par QCM...), concourt également à expliquer les variations observées entre les études.

### Définition adoptée pour entrées lexicales

Dans ce travail de recherche, nous utilisons le terme d'entrées lexicales, dont la définition est fondée sur celle utilisée par Lieury (1991) dans ses recensements. Par entrée lexicale est désignée toute forme graphique repérée dans le texte, lemmatisée – c'est-à-dire réduite à sa forme canonique qu'est son lemme [1] – qui admet une signification bien marquée. Premièrement, cela veut dire que des lemmes, comme « article » ont deux entrées sémantiques distinctes : le sens de « déterminant » en grammaire française (dont 22 occurrences seront rencontrées dans notre inventaire des manuels

* Laboratoire Centre de recherche en éducation et formation (Cref, EA1589), Université Paris Ouest Nanterre.

** Laboratoire cognitions humaine et artificielle (CHArt, EA 4004), Université Paris Ouest Nanterre.

Correspondance : Moïse Déro, ESPE Lille Nord de France, site Arras, 7bis rue Raoul François, BP 30927, 62022 Arras Cedex.

<moise.dero@espe-lnf.fr>

1. Pour les mots à formes variables, réduction au masculin singulier des formes du genre, du nombre pour les substantifs et les adjectifs, réduction à l'infinitif pour les différents temps des verbes.

scolaires) et le sens de « journal », qui n'est rencontré qu'une fois dans notre inventaire. Deuxièmement, il n'y a pas de familles de mots comme l'entendent Nagy et Anderson (1984). Les locutions ont également été comptées comme entrées (exemples : « Afrique du Nord », « robe de chambre »).

### Inventaires de langue française et évaluations du vocabulaire chez les élèves

En matière d'estimation du vocabulaire en France, il existe peu d'études précises et de grande ampleur avant les travaux d'Ehrlich, Bramaud du Boucheron et Florin (1978). Dans ces travaux, les auteurs utilisent, comme source de vocabulaire, un extrait du *Dictionnaire français contemporain* de Dubois et coll. (1971), en retenant des « mots primaires » (mots de base) représentatifs des familles sémantiques du dictionnaire. Cet extrait passe par différents processus d'évaluation auprès d'adultes, pour aboutir à un échantillon de 13 500 mots de référence, dont un extrait de 2 700 mots est soumis à des élèves, lors d'épreuves écrites. À l'issue des résultats de l'épreuve de jugements, des estimations des mots « moyennement connus » en rapport aux 13 500 mots de référence sont calculées. Elles indiquent 3 026 en CE1 (22,4 %), 3 913 en CE2 (29 %), 5 193 en CM1 (38,5 %) et 6 143 (45,5 %) en CM2. L'analyse sémantique des définitions de l'étude indique un développement sensible des connaissances des élèves par rapport aux objets, événements et actions en lien aux mots. Le nombre d'élèves parvenant à définir les mots croît de manière très importante.

Plus récemment, Lété, Sprenger-Charolles et Colé (2004) ont fait un recensement lexical très approfondi, à partir de 54 manuels scolaires de lecture, édités entre 1964 et 1995, et en usage entre 1995 et 1998. Cette base de données lexicales comprend 23 000 lemmes et 48 900 formes orthographiques exactes et fournit, de manière détaillée, plusieurs normes de l'exposition à l'écrit (fréquences, catégories syntaxiques...), pour des élèves de CP (âgés de 6 ans environ) au CM2 (environ 10 ans). Cet inventaire, examinant strictement le vocabulaire du corpus étudié selon des indicateurs psycholinguistiques, les auteurs n'ont pas étudié les connaissances qu'en ont les élèves, ni les relations avec une performance scolaire.

### Le vocabulaire « encyclopédique » du collège français de la 6e à la 3e, de Lieury (1991-1996)

Pour qualifier les connaissances qui se réfèrent à des matières spécialisées (histoire, biologie, etc.), Lieury a proposé le terme de « mémoire encyclopédique » (Lieury, Van Acker, Durand, 1995a, 1995b). En effet, le vocabulaire a, semble-t-il, été utilisé d'une façon empirique, sans référence théorique ou comme une épreuve représentant un facteur « verbal » de l'intelligence, sans référence à la mémoire. D'un autre côté, les tests de vocabulaire présentent, généralement, des mots du vocabulaire courant. À l'inverse, dans les connaissances et le système scolaire, apparaissent des disciplines d'une grande spécificité, histoire, géographie, mathématiques, biologie, géologie, physique, littérature, langues... Certes, ces savoirs spécifiques reposent probablement, en grande partie, sur la mémoire lexicale et la mémoire sémantique. Mais, par leur spécificité – lien avec les nombres (mathématiques, physique et chimie, histoire, géographie), avec des visages (histoire), des cartes spatiales (géographie) – ils reposent, peut-être, sur des mécanismes psychologiques et neurologiques en partie différents. Ainsi, de nombreux mots sont des noms propres (Ramsès, Charlemagne, Lac Victoria, Manhattan) ou des concepts qui ont des sens différents de leur équivalent dans la vie courante, comme le mot « disque » qui rappelle le support musical, alors que le disque solaire en Histoire ou le disque au sens mathématique, ont des significations différentes.

En étudiant les savoirs scolaires et des savoirs non scolaires (cinéma, musique, sport, etc.), Postal et Lieury (1998) ont considéré la mémoire encyclopédique dans sa globalité. Or, ils remarquent que la mesure des connaissances encyclopédiques extra-scolaires ne peut, à elle seule, prédire la réussite scolaire d'un élève. Ils constatent que, tout au plus, la quantité de savoir littéraire est le plus en rapport avec la moyenne générale des élèves. L'étude de la mémoire encyclopédique scolaire reste donc essentielle.

Pour estimer l'étendue du lexique de connaissances que rencontrent les élèves, durant leur scolarité, une étude longitudinale a été menée (Lieury, Van Acker, Clévédé, Durand, 1992a, 1992b ; Lieury, Van Acker, Durand, 1995a, 1995b ; Lieury, 1996, 2012), en suivant le parcours scolaire et l'acquisition de mots nouveaux de 200 élèves de huit classes du collège français (de la 6e à la 3e). Ayant échoué à recenser les cours des professeurs sur la base de notes fragmentaires (Lieury, 2012), l'inventaire s'est fondé sur les manuels scolaires utilisés dans le collège. Ce travail préliminaire a permis d'aboutir à la mise au point de questionnaires à choix multiples (QCM), qui ont permis d'examiner, chez 200 élèves, l'acquisition des mots techniques des différentes disciplines. Pour comptabiliser le vocabulaire encyclopédique, c'est-à-dire les mots techniques, en plus du vocabulaire courant, les études de Lieury n'ont retenu que les mots supplémentaires à l'inventaire basique de Dottrens et Massarenti (édition de 1963) et l'estimation de juges. Ces juges étaient des étudiants en

3e année de licence, chargés de repérer, par la lecture, dans les manuels scolaires du collège (3 à 4 étudiants par manuel), les mots techniques, donc hors vocabulaire courant, provenant de la liste du Dottrens et Massarenti. Ces inventaires révèlent environ 6 000 mots (en plus du vocabulaire courant) dans les manuels de 6e, jusqu'à près de 24 000 dans les manuels de 3e. À partir de ces inventaires par matière (histoire, biologie, etc.), ont été construits des QCM avec des échantillons au hasard de cent mots par matière et remplis par les élèves en fin d'année, afin d'estimer le stock de mots mémorisés en mémoire à long terme. Les résultats mettent en évidence que la mesure de la mémoire encyclopédique, exprimée par le score « réussites – erreurs » (R-E) des questionnaires à choix multiple, est fortement corrélée avec la moyenne générale annuelle des élèves. En classe de 6e, la corrélation est de. 60, pour un effectif de 190 élèves. Elle est de. 72 en 5e (n = 138), puis. 59 dans les classes de 4e (n = 147) et, enfin, de. 61 en classe de 3e (n = 174). Ce même score R-E explique 41 % de la variance des résultats au brevet des collèges (r =. 64 ; n = 174). On constate que ces corrélations sont plus importantes que celle obtenue entre le test de raisonnement du D70 (.50) avec la moyenne générale en 5e (Lieury, 1997). Pourtant, ce score du D70 est légèrement supérieur aux corrélations relevées par N'GuyenXuan (1969) et Aubret (1987) pour d'autres tests de raisonnement, sur des effectifs plus nombreux. La mémoire à long terme des connaissances scolaires (mémoire encyclopédique) est un meilleur prédicteur des résultats scolaires que les tests de raisonnement classiques.

## ÉTUDE 1. INVENTAIRE DU VOCABULAIRE À L'ÉCOLE PRIMAIRE

Cependant, les études de Lieury et de ses collaborateurs ont principalement utilisé des juges pour séparer le vocabulaire technique du collège (« Ramsès », « hypoténuse »...) du vocabulaire courant. Cette procédure subjective, de recours à des juges, sous-estime le vocabulaire technique, comme l'indique la réévaluation de Déro (1998), qui a utilisé une procédure informatique (numérisation des textes et traitements logiciels d'analyse lexicale, etc.) sur la base des travaux de Lieury. Ainsi, Déro aboutit à une estimation encore plus grande, des connaissances encyclopédiques abordées au collège français, que celles de Lieury (Lieury et coll., 1995a, 1995b). Il était nécessaire de faire une estimation plus objective de ce vocabulaire acquis à la fin de l'école élémentaire (CM2). À cette fin, notre étude a pour objectif de réaliser un inventaire du vocabulaire scolaire français, rencontré par les élèves, de 6 à 10 ans, du CP au CM2.

Dans la ligne des travaux de Lieury, nous nous sommes intéressés, aussi, aux connaissances à l'école élémentaire. L'étude 1 propose un nouvel inventaire du vocabulaire de l'école élémentaire (CP à CM2, élèves de 6 ans à 10 ans environ), sur la base de 10 manuels scolaires, édités entre 1977 et 1989 et en usage en 1996 dans deux écoles de la région Bretagne. Dans le cadre de la conception de la mémoire encyclopédique, l'inventaire n'est pas restreint au vocabulaire courant, mais utilise un manuel d'histoire-géographie, un de mathématiques et un de sciences, des niveaux CM1 et CM2. Ce choix est différent de Lété et coll. (2004), qui exploitent uniquement des manuels de lecture. Enfin, sur la base de cet inventaire, nous étudierons le vocabulaire mémorisé en fin d'année (étude 2 et étude 3).

### Méthode

Les traitements détaillés ci-dessous relatent la méthodologie de travail utilisé par Déro en 1996 (non publié).

#### *Critères de mesures de l'inventaire informatisé*

Le but de l'étude était de prendre en compte la sémantique du mot et non seulement l'unité lexicale. Ici, le terme d'entrée sémantique recouvre toute forme graphique repérée dans le texte, lemmatisée, puis réduite, qui admet une signification bien marquée. D'une part, cela signifie que, pour des mots polysémiques, deux entrées distinctes sont créées et non une seule. D'autre part, il n'y pas de famille de mots comme l'entendent Nagy et Anderson (1984) : les adjectifs sont considérés comme des entrées principales séparées. Les locutions ont également été comptées comme entrées uniques et non en autant d'entrées distinctes.

#### *Choix des manuels scolaires*

Les dix livres inventoriés sont ceux utilisés en classe dans les écoles qui nous ont reçus. Pour le CP, un ouvrage, composé de trois livrets progressifs, a été utilisé. En CE1 et CE2, cinq livres ont été examinés : trois livres de contes et deux livres de français. En CM1 et CM2, quatre ouvrages ont servi de référence : un en français, un en mathématiques, un en histoire et géographie, un en sciences. Le lecteur trouvera en tableau 1, le titre de ces dix livres.

#### *Procédure du comptage informatisé*

Les dix manuels ont été numérisés, puis traités par un logiciel de reconnaissance des caractères. Après vérifications, informatisée et manuelle, des textes par rapport aux ouvrages papier, le corpus a

été reformaté selon les règles du logiciel : élimination des attributs de textes, colonnage... Durant cette étape, les formes graphiques brutes ont été cotées, de façon à correspondre à la définition d'entrée lexicale retenue.

Seules, les majuscules initiales des noms propres ont été conservées. Les mots pleins (substantifs et adjectifs) et les mots fonctionnels (articles, pronoms, prépositions) ont eu, principalement en début de phrases, leur majuscule initiale convertie en minuscule. Ce travail s'est fait essentiellement manuellement pour une question de fiabilité des changements de casse. L'automatisation de ce traitement n'est pas sans risque, si l'on veut correctement respecter et distinguer les mots. Par exemple, le « a » est la forme la plus fréquente des conjugaisons du verbe avoir. S'il est écrit « a » ou « A », certains logiciels compteront ces graphies comme deux formes graphiques distinctes. Par ailleurs, « A » risque d'être confondu avec la préposition « à » mise en majuscule non accentuée, comme c'est le cas la plupart du temps en début de phrase. Cette confusion prend de l'importance si le chercheur s'intéresse aux formes pôles du verbe « avoir », par exemple.

De même, les césures inutiles des mots, les numérotations de pages et les caractères non alphanumériques ont été éliminés. Les lettres avec un signe diacritique (exemple « à », « ï »), des initiales de mots communs ont été converties en minuscules accentuées. Les abréviations utiles ont été remplacées par leurs mots non abrégés. La syntaxe des nombres a, également, été modifiée, même si, par la suite, les dates et les valeurs n'apparaissent pas dans les inventaires terminaux. Pour spécifier une locution et éviter que le traitement logiciel ne sépare ses constituants, nous avons employé le caractère non-délimiteur « _ », en lieu et place du caractère délimiteur qu'est le blanc « » (exemple, « lapin de garenne » devient « lapin_de_garenne »).

### Résultats

Les résultats (tableau 1) indiquent un total de lemmes de 11 095 mots sur l'ensemble du primaire, sur les 290 059 mots rencontrés dans les manuels. Avec les formes lexicales différentes, conjugaisons différentes notamment, c'est un total de 19 549 unités lexicales que les élèves ont théoriquement à mémoriser. Par rapport à Lieury, l'inventaire a été complété par plusieurs champs, comme la nature des mots, leur désambiguïsation quant à la polysémie et leurs occurrences dans la totalité des ouvrages.

| | | Niveau scolaire | Entrées lexicales (lemmes) | Entrées lexicales (formes brutes) | Nbre tot. de mots (occurrences) |
|---|---|---|---|---|---|
| Livres | Au fil des mots | CP | 1 718 | 2 195 | 12 741 |
| | Français au CE1 | CE1 | 2 447 | 3 673 | 25 556 |
| | La ruche aux livres | CE1 | 2 887 | 4 593 | 30 168 |
| | Les belles histoires | CE | 1 993 | 3 041 | 18 665 |
| | Le livre des bêtes | CE | 1 604 | 2 465 | 17 193 |
| | La ronde des mots | CE | 2 522 | 3 617 | 22 104 |
| | Français au CM1 | CM1 | 5 361 | 9 499 | 72 453 |
| | Mathématiques | CM1 | 1 374 | 2 070 | 20 515 |
| | Histoire | CM | 4 029 | 6 371 | 31 210 |
| | Livre de sciences | CM | 3 683 | 5 881 | 39 454 |
| | Sous-total par niveau scolaire | CP (grade 1) | 1 718 | 2 195 | 12 741 |
| | | CE (grades 2-3) | 5 464 | 9 303 | 113 686 |
| | | CM (grades 4-5) | 9 095 | 16 274 | 163 632 |
| | Total | CP-CM (grades 1-5) | 11 095 | 19 549 | 290 059 |

Tableau 1. Inventaires des manuels de l'école élémentaire utilisés.

Le total d'entrées lexicales est de 11 095 en ne comptant que les mots différents de chaque niveau scolaire : 1 718 dans les livres de CP, 5 464 pour les livres de CE, et 9 095 pour les livres de CM.

Mais nos résultats sont bien inférieurs, de moitié, à l'inventaire de Lété et coll. (2004), puisque ces auteurs trouvent 23 000 lemmes, ce que nous identifions comme des entrées sémantiques. Cette différence vient de leur recensement très approfondi, puisqu'il porte sur 54 manuels et livres de lecture, ce qui aboutit, dans leur comptage, à 1 925 584 mots, contre 290 059 mots différents dans notre étude. Cependant, on peut penser que les élèves d'une école donnée ne sont pas confrontés à une telle quantité de mots. L'objectif de Lété et coll. était de faire un recensement de la langue, tandis que le nôtre est de rechercher quelles sont les capacités de mémorisation des élèves. De même, Lambert et Chesnet (2001) ont constitué une base de données lexicales, afin de préparer des listes de mots pour des expériences ou avec des objectifs pédagogiques. L'inventaire a été fait uniquement pour un niveau donné (CE2 : élèves de 8-9 ans), mais sur la base d'un très grand nombre de livres (38), à la fois scolaires et extra-scolaires des années précédentes. Notons que, par rapport à notre objectif de vocabulaire encyclopédique, où les noms propres sont importants, ces auteurs ont retiré les noms propres et termes géographiques. À l'inverse, les formes orthographiques différentes sont considérées comme distinctes ; il s'agit donc là d'un vocabulaire « lexical », alors que nous avons considéré les unités sémantiques. Les auteurs ont élaboré un logiciel original, qui permet de compter le nombre de mots nouveaux lors de la numérisation d'un nouveau livre. Ils notent, ainsi, qu'au bout d'un moment, le lexique se stabilise et que l'enregistrement des cinq nouveaux livres n'apporte que 7 % de mots nouveaux. Au total, 417 000 mots ont été analysés, desquels sont extraites 20 600 entrées différentes et 9 600 racines lexicales (ce que nous appelons « entrées sémantiques »). Pour les entrées lexicales différentes, leur décompte de 20 600 mots n'est pas très différent de notre inventaire (19 549). En revanche, Lété et coll. trouvent beaucoup plus d'entrées sémantiques, 9 600 (contre 5 464 dans notre étude, pour le CE2), ce qui s'explique par leur nombre plus important de livres sélectionnés et par le recours à des livres extra-scolaires. Cependant, ce décompte reste inférieur aux 11 095 de notre inventaire pour la totalité des livres pour le primaire (jusqu'au CM2). Le fait que le nombre de formes lexicales brutes (avec l'orthographe différenciée) soit équivalente, dans l'étude de Lambert et Chesnet et dans la nôtre, montre, vraisemblablement, que ce sont les mêmes mots outils et les mêmes verbes conjugués, qui restent les plus utilisés dans les textes (exemple : verbe « être » et « avoir »). Ce sont les entrées sémantiques qui se développent le plus, quantitativement et qualitativement, par rapport aux lemmes, au fur et à mesure de la progression dans les niveaux scolaires des élèves, comme l'avaient montré Ehrlich et coll. (1978).

## ÉTUDE 2. ESTIMATION DU VOCABULAIRE ACQUIS CHEZ LES ÉLÈVES DE L'ÉCOLE PRIMAIRE

L'objectif de notre étude est d'estimer le nombre de mots acquis par les élèves en fin d'année, sur les cinq années de l'école primaire. Nous avons repris le principe de la méthode de Lieury et coll. (1995, 2013), d'évaluer les connaissances scolaires à partir de questionnaires sur les mots. Le but est d'estimer le nombre de mots acquis, à la fois, sémantiquement et lexicalement. L'échantillonnage des mots est fait au hasard, de façon à estimer le nombre de mots acquis, en reportant la proportion d'items réussis sur le total de l'inventaire.

### Méthode

*Construction des questionnaires*

Pour sélectionner les items du questionnaire à choix multiple (QCM), nous avons effectué une extraction aléatoire par inventaire à chaque niveau scolaire. Quand le tirage aléatoire portait sur un terme polysémique, la bonne réponse était construite dans le contexte de la matière spécifique d'où le mot était extrait (français, histoire, sciences, mathématiques).

La consigne donnée aux élèves est de repérer le mot le plus proche du sens de celui qui leur est proposé. Les distracteurs lexicaux ont été construits sur la base d'une confusion possible sur le plan orthographique ou phonétique, soit des voisins orthographiques, soit des voisins phonétiques. Toutefois, tous ces voisins n'ont pas été conçus systématiquement pour ne pas donner à l'épreuve un caractère artificiel. Quant aux distracteurs sémantiques, qui ont été introduits dans les modalités de réponses, ils peuvent être des propositions de sens contraire, d'une catégorie sémantique différente mais proche, ou encore sans aucune parenté avec l'item. Pour certains mots, il n'a pu être possible de proposer, systématiquement, un distracteur lexical et un distracteur sémantique : dans ce cas, deux distracteurs sémantiques ont été utilisés.

Répondant à une demande formulée par les enseignants des écoles, le nombre de QCM passés par un élève était différent, en fonction de son niveau scolaire, en prévision de la durée de passation totale, prévue sur une quinzaine de jours. C'est pourquoi, les classes de CP ont eu un échantillon de 100 QCM sur les 1 728 mots du CP (1/17e), les élèves de CE ont eu 200 QCM sur les 5 464 mots du CE (1/27e) et les plus grands (CM1/CM2) ont passé 400 QCM sur les 9 095 mots du CM (1/22e).

Deux écoles, situées en Bretagne, ont participé à cette étude, Bréhan et Chantepie [2].

| | |
|---|---|
| **seau** | |
| a. où l'on met de l'eau | *(Bonne réponse)* |
| b. idiot | *(Distracteur lexical [sot])* |
| c. brouette | *(Distracteur sémantique)* |
| d. je ne sais pas | |
| **australopithèque** | |
| a. ancêtre de l'éléphant | *(Distracteur sémantique)* |
| b. homme préhistorique | *(Bonne réponse)* |
| c. dinosaure | *(Distracteur sémantique)* |
| d. je ne sais pas | |

Tableau 2. Exemples tirés de la version du QCM de 400 mots.

### Déroulement de l'expérience

C'est une étude de type papier-crayon, sous la forme d'un cahier de questions, qui a été menée. Sur une période de 15 jours et pendant leurs cours, les enseignants remettaient aux élèves leurs livrets de QCM. À leurs rythmes, les enseignants ont administré l'expérience en passation collective, en fonction de l'aménagement de leurs programmes de cours. Les enseignants n'ont donné d'indications sur les réponses que lorsque qu'ils avaient collecté les cahiers. Généralement, ils corrigeaient le cahier de questions en commun, en repérant distracteurs et bonnes réponses et en les expliquant. Afin de maximiser les possibilités d'apprentissage des élèves, les QCM furent administrés aux élèves dans la seconde quinzaine de juin.

### Résultats

Des tests de Student sur échantillons indépendants ont été calculés pour vérifier s'il existait des différences significatives entre les classes de ces écoles pour chaque niveau. Devant l'absence de différences significatives entre les deux écoles, pour chaque niveau, quant aux scores aux QCM (bonnes réponses noté R pour « réussites », le total des erreurs (distracteurs lexicaux et sémantiques) ou la modalité « je ne sais pas »), nous avons rassemblé les résultats des deux écoles pour chaque niveau. Au total, le nombre d'élèves, par niveau scolaire, était le suivant : 47 élèves pour le CP, 43 élèves pour le CE1, 45 élèves pour le CE2, 48 pour le CM1 et 51 pour le CM2.

*Développement du vocabulaire au cours des niveaux scolaires du primaire*

Notre objectif principal était d'estimer la quantité de vocabulaire stocké en fin d'année, en mémoire à long terme, chez nos élèves. C'est pour cette raison que les mots faisant l'objet de QCM ont été sélectionnés au hasard.

Au CP, un élève obtient, en moyenne, 60,53 réussites sur les 100 QCM ($\sigma = 15,56$) pour 20,85 erreurs ($\sigma = 12,85$) et 18,62 réponses où il ignore la signification des mots ($\sigma = 14,63$). Sur les modalités erreurs et « je ne sais pas », les coefficients de variation (respectivement 0,61 et 0,78) sont relativement élevés par rapport à celui de la Réussite (0,25), témoignant d'une large dispersion des notes autour de ces moyennes. Si on compare les deux types d'erreurs en pourcentage, introduits dans les QCM (distracteurs lexicaux *versus* distracteurs sémantiques), il n'apparaît pas de différence statistiquement significative en CP, quant au type de distracteurs choisis par les élèves ($t(46) = 0,60$ ; n.s.). Les élèves commettent donc, en moyenne, autant d'erreurs lexicales que de sens, en CP, aux QCM : 10,64 ($\sigma = 6,94$) distracteurs lexicaux contre 10,24 ($\sigma = 6,72$) distracteurs sémantiques.

Pour les classes de CE1, répondant à 200 QCM, on observe une moyenne de quelque 127,63 items, auxquels les élèves ont correctement répondu ($\sigma = 18,23$). Les élèves se trompent pour 27,51 items ($\sigma = 13,82$) et choisissent la modalité « je ne sais pas » dans 44,86 cas sur 200 ($\sigma = 20,49$). Si on relativise, en comparant les pourcentages de ces scores, en CE1, les élèves réussissent 63,81 % des items, pour 13,75 % d'erreurs et 22,43 % de réponses « je ne sais pas ». Par rapport aux items du niveau CP, le nombre d'erreurs est en diminution, alors que les bonnes réponses sont en augmentation (de plus de 3 points). En nous intéressant aux deux types d'erreurs, introduites dans les QCM, on constate que la différence entre les moyennes des erreurs formelles et des erreurs sémantiques n'est pas significative ($t(42) = -0,83$, n.s.). Des erreurs formelles sont rencontrées pour 13,45 items contre 14,04 pour les erreurs sémantiques, soit, respectivement, 6,73 % et 7,02 % des items.

En CE2, les élèves ont une réussite moyenne, au QCM, de 155,69 ($\sigma = 12,84$). Ils se distinguent au niveau des erreurs, les élèves bréhannais commettant, en moyenne, plus d'erreurs que leurs camarades de Chantepie, respectivement 35,95 ($\sigma = 13,11$) contre 27,33 ($\sigma = 11,73$). La modalité « je ne sais pas » indique que les élèves de Bréhan répondent plus souvent aux items bonnes réponses

2. Tous nos remerciements renouvelés aux directeurs d'école d'alors, MM. Jean-François Rolland et Michel Gauthier, et mesdames et messieurs les enseignants nous ayant ouvert leurs classes.

et pièges que les élèves de Chantepie. Rapporté aux pourcentages et tous les élèves confondus, le nombre de bonnes réponses représente 77,84 % des réponses, les erreurs comptant pour 15,68 % et les réponses marquant l'ignorance de la solution 6,48 % des items. Là encore, on note une progression par rapport au CE1. Toutefois, on remarque une augmentation du nombre d'erreurs pour les classes de CE2, par rapport aux classes de CE1, sur les mêmes QCM du cours élémentaire.

Au CM1, sur les 400 items auxquels ont été soumis les 48 élèves des deux écoles, le score de bonnes réponses est de 286,02 ($\sigma = 40,96$) pour 46,58 erreurs ($\sigma = 27,46$) et 67,40 réponses « je ne sais pas » ($\sigma = 42,06$). Il est facile d'observer, une fois encore, la grande variabilité des réponses de modalité « je ne sais pas », qui offre une grande dispersion autour de la moyenne. Ramenées en pourcentage, les bonnes réponses représentent 71,51 % des réponses, contre 11,65 % pour les erreurs. Au niveau des types d'erreurs commises, les écoles ne diffèrent pas sensiblement l'une de l'autre, ni pour les pièges lexicaux ($F(1,46) = 0,042$, ns) ni pour les pièges sémantiques ($F(1,46) = 0,489$, n.s.). Cependant, les deux types d'erreurs diffèrent l'une de l'autre, un t de Student, sur mesures appareillées, faisant apparaître des différences significatives entre les deux moyennes ($t(47) = 4,94$, $p < .0001$). Les erreurs lexicales sont, ici, plus importantes que les erreurs sémantiques. Toutefois, restons prudent : en effet, alors que, pour le CP et le CE, le rapport pièges formels/pièges sémantiques était relativement constant, de l'ordre de 92 % en CP et de 96 % en CE pour les pièges formels, il y a beaucoup moins de pièges lexicaux pour le CM (seulement 193 pièges formels pour 607 pièges sémantiques pour les 400 items, soit 48,25 %). La construction des items a, sans doute, été moins rigoureuse sur ces 400 items. Aussi, la pondération de ces items, par rapport aux autres, se trouve plus importante. Si écarts il y a, ces écarts sont accrus. La statistique nous renseigne, pour partie, de cette pondération particulière.

Au CM2, les 51 élèves obtiennent un score de 313,29 bonnes réponses (noté R pour réussite) pour les 400 items ($\sigma = 28,08$). Les erreurs rencontrées (notées E), lexicales et sémantiques, sont en moyenne de 35,55 ($\sigma = 17,12$). En pourcentage, on obtient une réussite, aux QCM, de l'ordre de 78,32 % pour 8,89 % d'erreurs.

Une analyse de la variance, introduisant en facteur le niveau scolaire et en variable dépendante l'estimation du vocabulaire, montre que l'acquisition du vocabulaire progresse effectivement d'année en année ($F(4,229) = 822,17$ ; $p < .001$).

| | Niveaux scolaires | | | | |
|---|---|---|---|---|---|
| | CP grade 1<br>6-7 ans | CE1 grade 2<br>7-8 ans | CE2 grade 3<br>8-9 ans | CM1 grade 4<br>9-10 ans | CM2 grade 5<br>10-11 ans |
| Moyenne | 1 039,94 | 3 486,79 | 4 253,42 | 6 503,40 | 7 123,53 |
| Ecart-type | 267,25 | 497,94 | 350,68 | 931,26 | 638,52 |
| Minimum | 188,98 | 2 513,44 | 3 688,80 | 4 843,09 | 5 343,31 |
| Maximum | 1 511,84 | 4 699,04 | 4 917,60 | 8 867,63 | 8 140,03 |

Tableau 3. Estimation du vocabulaire acquis en fin d'année par les élèves.

Mais ce ne sont que des estimations moyennes et le tableau 3 nous montre les différences énormes entre les maximum et minimum à chaque niveau scolaire. Par exemple, si l'élève ayant le meilleur score en CP a un stock de mots estimés à 1 511 mots (sur un inventaire de 1 718), l'élève ayant la moins bonne performance n'a acquis que 188 mots, ce qui ne représente que 12 % de l'inventaire des manuels. Ces écarts sont particulièrement frappants, lorsqu'ils sont exprimés sous forme de courbe. On voit, par exemple, que les élèves de chaque niveau ayant la meilleure estimation ne sont pas loin de l'inventaire lui-même des manuels. En revanche, l'écart est immense entre ces élèves et ceux qui, au contraire, ont l'estimation la plus basse (figure 1).

Cet écart est aussi impressionnant dans les autres niveaux et, en CM2, les écarts entre minimum et maximum sont d'environ 3 000 mots. Ce résultat rejoint les études de Lieury au collège, qui avait mentionné cette divergence énorme. Ce sont ces écarts qui produisent les corrélations si fortes entre la connaissance du vocabulaire et les performances scolaires, dont la lecture, comme nous allons le voir.

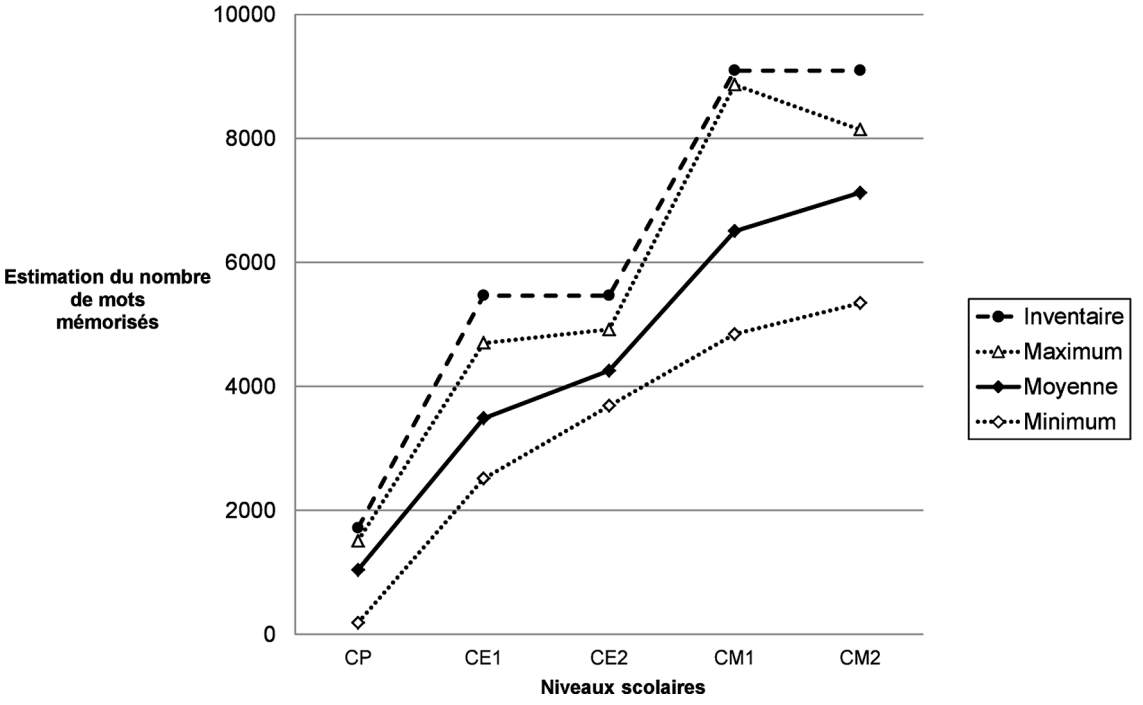


**Figure 1.** Estimation du vocabulaire mémorisé au primaire en fonction du niveau scolaire, par rapport à l'inventaire des manuels des mêmes niveaux (Note : Les courbes Maximum et Minimum représentent les élèves ayant la meilleure ou la moins bonne estimation par niveau scolaire).

*Relation entre le vocabulaire et les évaluations scolaires*

Parallèlement à la correction des QCM, nous avons demandé, aux enseignants, de nous fournir une note subjective de lecture et une note subjective d'appréciation générale, ce qui a été fait, bien que le système de notation ne soit plus en usage dans le primaire. À la demande des enseignants, il n'a pas été utilisé d'épreuves standardisées pour la lecture. Initialement, leurs notes correspondaient à des appréciations allant de « A+ », « C- », que nous avons converties en notes scolaires, d'après les indications des enseignants. Voici les résultats détaillés pour le CM2, et nous verrons, ensuite, des résultats synthétiques pour tous les niveaux scolaires (voir résultats synthétiques pour les niveaux scolaires).

Pour analyser ces corrélations et les suivantes, nous reprendrons les valeurs définies par Cohen (1988, 1992), pour qualifier une corrélation de faible, modérée ou forte. Une corrélation sera considérée comme faible, si elle est (en valeur absolue) inférieure à .20, modérée si elle est comprise entre .20 et .40, forte si elle est supérieure à .40 (Corroyer, Rouanet, 1994). Ces valeurs-repères peuvent être utilisées pour analyser les corrélations observées.

Tout d'abord, notons, pour les évaluations scolaires (tableau 4), que la corrélation entre lecture et appréciation générale est forte (.94), ce qui confirme la très grande importance de la lecture dans les premières années, comme l'ont montré de nombreux auteurs.

Quant aux relations avec le vocabulaire encyclopédique (réussites-erreurs), la corrélation la plus forte est avec la lecture (.71), mais très importante également (.60), avec l'appréciation générale. Notons que, du fait des pratiques françaises à l'école élémentaire (absence de notes pour évaluer les élèves), l'appréciation générale est un indicateur assez subjectif dans nos études, car elle ne se fonde pas, comme au collège, sur la moyenne générale des matières. La corrélation forte et continue sur les cinq années étudiées entre appréciation de lecture et appréciation générale suggère que cette dernière renvoie aux objectifs institutionnels de la maîtrise de la lecture. Mais, au cours moyen, il semble que cette appréciation générale s'étoffe d'autres critères pour déterminer les qualités d'un élève du primaire.

Les corrélations entre appréciation générale, lecture et erreur sont également très fortes. Cela laisse penser qu'à ces étapes de construction du lexique, la lecture est fortement gênée par des mots, dont l'unité lexicale et le sens ne sont pas établis. Ainsi, la lecture est corrélé à -.79 avec les erreurs, ce qui est considérable. Notre distinction entre sémantique et lexicale (l'aspect formel du mot) est particulièrement intéressante et montre que les erreurs sémantiques sont plus fortement corrélées

(-.79), que les erreurs lexicales (-.63). Ceci confirme que la compréhension des mots est plus décisive. Cependant, les erreurs lexicales, confusion avec un mot proche, entravent aussi la lecture, ce que notent couramment les enseignants lorsqu'ils relèvent des coquilles dans les productions de leurs élèves. Ce type de résultat montre, sur le plan pédagogique, tout l'intérêt de tels questionnaires, qui permettent, non seulement, d'évaluer le niveau d'un élève, mais aussi de faire apparaître des difficultés insoupçonnées de l'adulte, notamment, des enseignants.

| | Appréciation générale | Lecture | Score R – E | Réussites R | Erreurs E | Erreurs lexicales |
|---|---|---|---|---|---|---|
| Lecture | .94** | | | | | |
| Score R – E | .60** | .71** | | | | |
| Réussite R | .40* | .49** | .89** | | | |
| Erreurs E | -.70** | -.79** | -.68** | -.29* | | |
| Erreurs lexicales | -.56** | -.63** | -.68** | -.36** | .88** | |
| Erreurs sémantiques | -.68** | -.76** | -.62 | -.22 | .96** | .72** |

Tableau 4. Corrélations entre plusieurs indicateurs aux QCM par niveau du CM2 (n = 51).(* significatif à p < .05 ; ** significatif à p < .0).

*Résultats synthétiques pour les niveaux scolaires*

Les corrélations observées entre l'appréciation générale, l'appréciation en lecture et le score réussites-erreurs sont, en général, fortes, quel que soit le niveau scolaire (tableau 5). Ces corrélations vont de .49 à .76, avec une majorité de corrélations égales ou supérieures à .60. La performance dans la lecture et à l'école en général est liée, d'une façon très importante, à la mémoire du vocabulaire. Pour vérifier que le vocabulaire encyclopédique est en mesure de prédire l'appréciation générale de l'élève, indépendamment du niveau en lecture, nous avons réalisé une régression standard, en incluant l'appréciation en lecture comme deuxième paramètre. Le résultat au QCM de mémoire encyclopédique est bien en mesure de prédire significativement l'appréciation scolaire ($\beta = 0{,}27$, $p < .01$), tout comme l'appréciation en lecture ($\beta = 0{,}68$, $p < .001$).

Quelques comparaisons, en fonction du niveau scolaire, ont été réalisées (tableau 6). Les distracteurs lexicaux et sémantiques n'étant pas présents de façon équivalente, il n'est pas possible d'en comparer l'évolution différentielle. Mais on peut voir, à travers ces erreurs, l'impact de l'acquisition de ces aspects au niveau de la lecture. Ainsi, d'après les corrélations avec le type d'erreur et la performance en lecture, il semble que l'aspect lexical ait une plus grande importance chez les jeunes élèves, que le sémantique (CP à CE1), tandis que le sémantique prend plus d'importance en CM2. Mais nos résultats ont été établis sur quelques dizaines d'élèves et demandent à être confirmés.

| Niveaux | Appréciation générale et Réussite-Erreurs | Lecture et Réussite-Erreurs |
|---|---|---|
| CP | .73** | .64** |
| CE1 | .49** | .76** |
| CE2 | .72** | .63** |
| CM1 | .53** | .49** |
| CM2 | .60** | .71** |

Tableau 5. Corrélations entre l'appréciation générale et la lecture et le score de Réussites-Erreurs au questionnaire évaluant la connaissance du vocabulaire (** significatif à p < .01).

| Niveaux | Lecture et erreurs lexicales | Lecture et erreurs sémantiques |
|---|---|---|
| CP | -.40** | -.33** |
| CE1 | -.50** | -.30** |
| CE2 | -.58** | -.40** |
| CM1 | -.52** | -.54** |
| CM2 | -.63** | -.76** |

Tableau 6. Corrélations entre la lecture et le type (lexical ou sémantique) d'erreurs en fonction du niveau scolaire (** significatif à p < .01).

## ÉTUDE 3 : ESTIMATION DU VOCABULAIRE MÉMORISÉ SUR L'ENSEMBLE DES INVENTAIRES DES NIVEAUX SCOLAIRES

Dans l'étude précédente (étude n° 2), l'estimation du vocabulaire mémorisé en fin d'année était faite par niveau scolaire, sur la base de l'inventaire spécifique à ce niveau, c'est-à-dire sur les seuls livres de classe utilisés à ce niveau. Par exemple, le QCM de vocabulaire était construit à partir d'un échantillonnage aléatoire de cent mots pour le CP, à partir, uniquement, de l'inventaire des livres de CP et ainsi de suite pour les autres niveaux jusqu'au CM2. Or, il est vraisemblable que les enfants d'un niveau connaissent d'autres mots, soit par le contexte linguistique dans la famille, la télévision ou d'autres histoires racontées par leurs enseignants. C'est pourquoi, dans l'étude 3, nous avons voulu estimer le vocabulaire mémorisé en fin d'année, mais sur la base de l'inventaire de tous les manuels que nous avons recensés, soit les 11 095 mots issus des dix livres utilisés par les écoles.

### Méthode

*Construction des questionnaires*

Bien que la base de mots, pour l'échantillon des questionnaires, soit plus large (11 095 mots), nous avons continué à construire des questionnaires moins longs pour les élèves les plus jeunes, à la demande des enseignants : la classe de CP n'a eu qu'un échantillon de 100 mots sur les 11 095 du primaire (1/111e), les élèves de CE1/CE2 n'ont eu que 200 mots (1/55e) et les plus grands (CM1/CM2) ont passé 400 mots (1/27e). Pour réaliser cette étude, 400 mots de l'inventaire du primaire ont donc été tirés au hasard. Ils ont fait l'objet de la construction de nouveaux QCM : bonnes réponses, erreurs, et modalité « je ne sais pas ». Mais, par économie, quand les items étaient identiques à ceux des échantillons de l'étude n°2 précédente, nous reprenions les QCM déjà réalisés. Le questionnaire de mémoire encyclopédique est celui utilisé dans l'étude 2, pour le niveau CM2. L'item (par exemple, « méridien de Greenwich ») est suivi de trois réponses, la bonne réponse et deux distracteurs, sémantique ou lexical. Une réponse « je ne sais pas » est présentée, pour que l'élève ne soit pas obligé de commettre une erreur. S'il coche une erreur, c'est une confusion dans ses connaissances. Il y a cent items, extraits au hasard de l'inventaire des manuels du primaire (étude 1). Les élèves ont une heure pour répondre, ce qui est largement suffisant. Le tableau 7 présente un extrait de ces items.

| méridien de Greenwich | submersible | décagramme |
|---|---|---|
| 1 ▫ en mathématiques | 1 ▫ avion | 1 ▫ mille grammes |
| 2 ▫ en géographie | 2 ▫ sous-marin | 2 ▫ dix grammes |
| 3 ▫ en musique | 3 ▫ voiture très rapide | 3 ▫ cent grammes |
| 4 ▫ je ne sais pas | 4 ▫ je ne sais pas | 4 ▫ je ne sais pas |
| **branchies** | **Zeus** | **Apollo** |
| 1 ▫ arbre | 1 ▫ homme politique | 1 ▫ fusée vers la Lune |
| 2 ▫ chez les poissons | 2 ▫ dieu grec | 2 ▫ dieu grec |
| 3 ▫ cœur | 3 ▫ ville d'Asie | 3 ▫ cosmonaute |
| 4 ▫ je ne sais pas | 4 ▫ je ne sais pas | 4 ▫ je ne sais pas |

Tableau 7. extraits du Questionnaire de mémoire encyclopédique pour le CM2 (Note : la bonne réponse est soulignée).

*Déroulement de l'expérience*

Comme dans l'étude n° 2, le QCM a été administré sous la forme papier-crayon, avec un livret de questions. Les enseignants remettaient aux élèves leur livret de QCM, administré en passation collective, durant les enseignements. La consigne de la page des exemples du cahier rappelait, à l'élève, qu'un système de points était utilisé, afin qu'il ne prenne pas l'expérience pour un jeu ou ne réponde au hasard. Il devait entourer la réponse la plus proche de l'item. Afin de maximiser les possibilités d'apprentissage des élèves, les QCM sont administrés aux élèves dans la seconde quinzaine de juin.

*Participants*

Les participants sont des élèves des 5 niveaux du primaire, de la même école de Chantepie, située en Bretagne, élèves n'ayant pas participé à l'étude 2 : 23 élèves pour le CP ; 27 élèves pour le CE1, 22 pour le CE2, 24 pour le CM1 et enfin 21 pour le CM2.

Nous avons inclus une classe de 6e du collège (42 élèves), pour vérifier que ce vocabulaire encyclopédique du primaire était intégralement connu, après un an au collège.

**Résultats**

*Développement du vocabulaire mémorisé, par niveau scolaire, en référence à l'inventaire total du primaire (11 095 mots)*

La figure 2 résume la progression des élèves dans leur connaissance sémantique et lexicale des mots de notre échantillon de vocabulaire du primaire. La progression est rapide entre le CP et le CE2, comme Ehrlich et ses collègues l'avaient remarqué. Les élèves vont presque doubler leurs connaissances de CP (4 500 mots) au CE2 (8 300), c'est-à-dire en deux ans, puis la progression ralentit pour atteindre environ 9 000 mots à la fin du primaire.

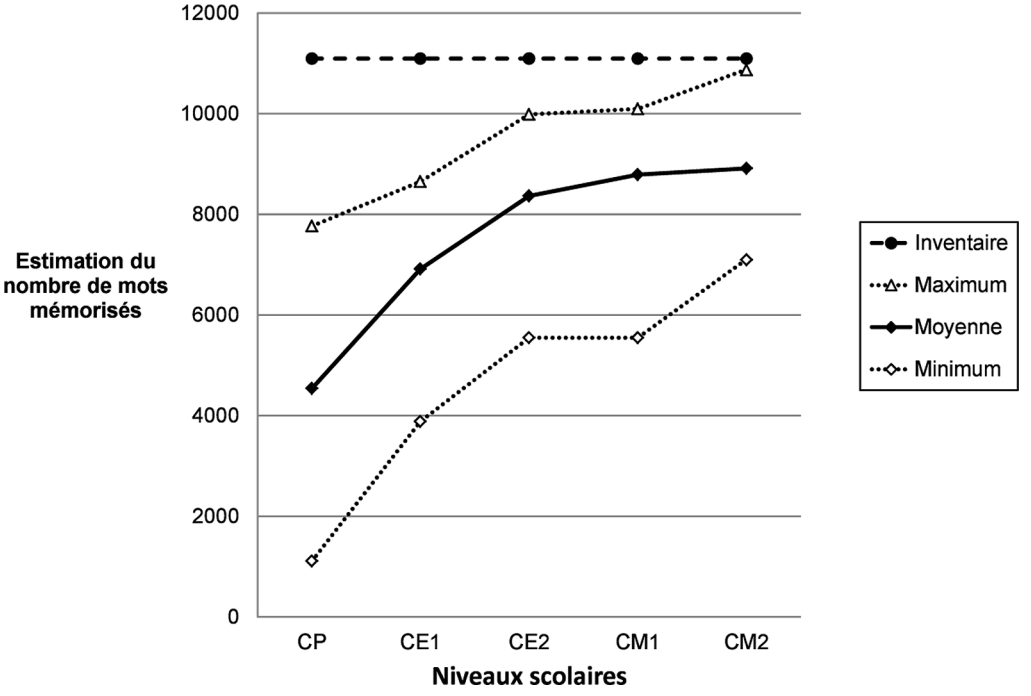


**Figure 2.** Estimation du vocabulaire mémorisé à la fin de chaque niveau scolaire, du CP au CM2 en fonction de l'inventaire total des manuels (11 095 mots) (Note : Les courbes Maximum et Minimum représentent les élèves ayant la meilleure ou la moins bonne estimation par niveau scolaire).

À nouveau (figure 2), les écarts entre le minimum et le maximum des évaluations sont considérables, notamment en CP, puisqu'il y a un écart de 1 100 mots à 7 600 mots, c'est-à-dire sept fois plus. Les élèves, dont l'estimation est maximum, auraient, en mémoire, plus de 10 000 mots, dès le CE2, pour l'âge de 9 ans.

| Niveaux | Estimation par niveau (étude 2) | Estimation sur le total du primaire (étude 3) |
|---|---|---|
| CP | 1 039 | 4 544 |
| CE1 | 3 486 | 6 911 |
| CE2 | 4 253 | 8 366 |
| CM1 | 6 503 | 8 788 |
| CM2 | 7 123 | 8 912 |

Tableau 8. Estimation du vocabulaire mémorisé selon l'inventaire de référence.

Sur le plan méthodologique, il est essentiel de comparer nos deux études (tableau 8) : l'étude 2, échantillonnant les mots sur la base des seuls livres d'un niveau scolaire et l'étude 3, dont l'échantillonnage se fait sur la totalité (11 095 mots). En fait, se restreindre aux livres d'un seul niveau conduit à une grande sous-estimation des connaissances des élèves, en particulier dans les premiers niveaux. Ainsi, pour les élèves de CP, on aboutit, en référence à l'inventaire total, à une estimation de quatre fois plus de mots connus par les élèves. Les enfants d'un niveau ont donc un vocabulaire bien au-delà des livres qui leurs sont présentés, soit par leur environnement familial, leurs camarades, soit parce que les enseignants eux-mêmes ne se limitent pas aux livres usuels. Même si cet écart diminue avec l'augmentation des niveaux scolaires, il reste, tout de même, une sous-estimation de 1 800 mots en CM2, soit une sous-estimation de 25 %. Enfin, il est intéressant de noter que notre estimation d'environ 9 000 mots mémorisés est tout à fait comparable à celle de la seule étude française, qui avait fait une estimation (voir introduction) des

mots connus, d'après des échantillons de mots issus du dictionnaire (Ehrlich, Bramaud du Boucheron, Florin, 1978 ; Florin, 1999).

Nous n'avons pas inclus, dans le graphique, le nombre de mots connus par les élèves de 6e de collège, puisque l'inventaire ne couvrait pas les manuels de 6e, qui contiennent plusieurs milliers de mots supplémentaires (Lieury, 2012, 1re éd. 1991 ; Déro, 1998). Notre but était simplement de vérifier qu'il n'y avait pas d'oubli de ce vocabulaire du primaire. Ce n'est pas le cas, puisque l'estimation, pour les élèves de 6e, est de 9 190 mots en moyenne. En revanche, ce qui nous intéresse est de vérifier si ce vocabulaire de base est encore prédictif des résultats en 6e de collège (voir plus bas, *Caractère prédictif avec les résultats scolaires de 6e de collège*).

*Relations avec les résultats scolaires*

La quantité de vocabulaire encyclopédique, c'est-à-dire englobant du vocabulaire des matières spécialisées, histoire, sciences, mathématiques, présente de bonnes corrélations, parfois très élevées, avec l'appréciation générale des enseignants et la maîtrise de la lecture. Mis à part les corrélations pour le CE1, qui ne sont pas très bonnes, sans que nous puissions l'expliquer, les corrélations entre le QCM de vocabulaire et l'appréciation générale vont de .62 (CP) à .82 (CE2). Les corrélations avec la lecture sont de .76 (CP) et .84 (CE2). Rappelons que l'appréciation de la lecture n'est faite que jusqu'au CE2, les enseignants de CM1 et CM2 n'ayant pas fourni cette évaluation (tableau 9).

| Niveaux | Appréciation générale et Réussite-Erreurs | Lecture et Réussite-Erreurs |
|---|---|---|
| CP | .78** | .76** |
| CE1 | .49** | .29 |
| CE2 | .82** | .84** |
| CM1 | .69** | – |
| CM2 | .62** | – |

Tableau 9. Corrélations entre l'appréciation générale et la lecture et le score de Réussites-Erreurs au questionnaire évaluant la connaissance du vocabulaire (** significatif à $p < .01$ ; Note : il n'y avait d'évaluation de la lecture en CM1 et CM2).

*Caractère prédictif avec les résultats scolaires de 6e de collège*

Le vocabulaire encyclopédique présente de bonnes corrélations (.56) avec la moyenne générale de 6e en fin d'année, soit un an plus tard que la passation du QCM de mémoire encyclopédique, à la fin du primaire (tableau 10). Les corrélations sont également bonnes avec les matières spécialisées, de la corrélation la moins forte avec les mathématiques (.46), jusqu'à la corrélation la plus forte avec l'histoire (.75). Ceci justifie notre objectif de s'intéresser aux connaissances encyclopédiques et pas seulement au vocabulaire du français usuel.

| Matières de la 6e (grade 6) | QCM Mémoire encyclopédique en fin de Primaire (R-E) | Moyenne générale 6e |
|---|---|---|
| Français | .58** | .91** |
| Biologie | .56** | .74** |
| Histoire | .75** | .85** |
| Mathématiques | .46** | .73** |
| Technique | .60** | .85** |
| Moyenne générale 6e | .56** | 1 |

Tableau 10. Corrélations entre la mémoire encyclopédique à la fin du Primaire et les résultats un an plus tard en 6e de collège ($n = 42$ ; ** significatif à $p < .01$).

## DISCUSSION GÉNÉRALE

Combien de mots un enfant connaît-il ? C'est cet objectif qui a conduit à la construction de nos trois études. L'étude 1 consistait à effectuer l'inventaire des mots rencontrés à l'école primaire élémentaire en France, du CP au CM2. À la différence de l'inventaire de Lété et coll. (2004), nous nous sommes limités aux manuels utilisés par les enseignants de nos études (une dizaine de manuels) comme l'échantillon de mots les plus probablement étudiés par les élèves. Le deuxième critère qui a guidé notre inventaire était de poursuivre les recherches sur la mémoire encyclopédique de Lieury, en prenant en considération les mots techniques des connaissances, par exemple, en Histoire, Sciences et Mathématiques, et pas seulement le vocabulaire français usuel. Les résultats de l'inventaire informatisé (Étude 1) indiquent 290 059 occurrences, dont 19 549 entrées lexicales (formes brutes), dont 11 095 entrées sémantiques (ou lemmes).

Le vocabulaire s'acquérant également hors contexte scolaire, il serait intéressant de poursuivre les travaux étudiant les savoirs scolaires et non scolaires, considérés comme connaissances encyclopédiques, à la manière des recherches de Postal et Lieury (1998) ou comme, tout récemment, l'étude de Le Cam, Rocher, Lorant et Lieury

(2013), pour des activités extrascolaires avec les médias numériques.

L'étude 2 avait, pour objectif, l'estimation du vocabulaire acquis en fin de chaque niveau scolaire. Des QCM ont donc été construits, mais seulement sur la base de l'inventaire de mots trouvés par niveau : 1 718 pour le CP ; 5 464 pour les niveaux CE1 et CE2 ; et 9 095 pour les niveaux CM1 et CM2. L'estimation, sur la base des résultats aux QCM, correspond à 1 039 mots acquis en fin de CP, jusqu'à 7 123 mots en fin de CM2. Le vocabulaire mémorisé en fin d'année est fortement variable selon les élèves. Ainsi, à la fin du CP, l'estimation la plus faible est de 188 mots mémorisés à la fin de l'année, tandis que l'estimation la plus élevée est de 1 511 mots. Cet écart reste aussi fort tout au long des niveaux scolaires, et l'estimation la plus faible au CM2 est de 5 343 mots, contre 8 140 mots, pour l'estimation la plus élevée. Ce sont ces écarts énormes qui produisent les corrélations fortes, entre le vocabulaire acquis et les performances scolaires. Ainsi, les corrélations entre vocabulaire et lecture sont de .49 au minimum, jusqu'à .72 ; les corrélations entre vocabulaire et appréciation générale sont de .49 à .76. Ces résultats tendent à confirmer de nombreuses études antérieures, insistant sur le rôle du vocabulaire. Les chercheurs conviennent, en effet, que l'exposition à l'écrit est un facteur essentiel dans l'apprentissage – principalement contextuel – du vocabulaire (par exemple, Cunningham, Stanovich, 1991, 2001 ; Anderson, Freebody, 1979 ; Just, Carpenter, 1987 ; Nagy, Anderson, 1984). Cunningham et Stanovich (1991, 2001) ont, en effet, montré que l'exposition à l'écrit agit sur le niveau de lecture, même quand les effets de l'âge, du quotient intellectuel ou d'autres facteurs linguistiques sont annulés statistiquement.

Étant donné ces écarts gigantesques entre les élèves, il nous est venu l'idée que les élèves connaissent vraisemblablement d'autres mots, soit par le contexte linguistique dans la famille, la télévision ou d'autres histoires racontées par leurs enseignants. C'est pourquoi, dans l'étude 3, nous avons voulu estimer le vocabulaire mémorisé en fin d'année, mais sur la base de l'inventaire de tous les manuels que nous avons recensés, soit les 11 095 mots issus des dix livres utilisés par les écoles. La méthode était la même pour la construction des QCM, c'est-à-dire un échantillonnage au hasard des mots, de façon à pouvoir estimer le nombre de mots mémorisés en fin d'année. Effectivement, les résultats sont très différents et montrent qu'évaluer la connaissance en vocabulaire uniquement *via* le vocabulaire du niveau scolaire en cours, peut montrer une limite en sous-estimant les connaissances réelles des élèves. Ainsi, alors que l'estimation, pour le CP, était de 1 039 mots, l'estimation, sur la base des quelques 11 000 mots de la totalité des manuels, est de 4 500, c'est-à-dire quatre fois plus. L'écart se réduit, toutefois, avec le développement de l'élève et si l'estimation était de 7 123 mots pour le CM2, il est de près de 9 000 mots (8 912), sur la base de tous les manuels. Il est intéressant de noter que notre estimation d'environ 9 000 mots mémorisés est tout à fait comparable à celle de la seule étude française qui avait fait une estimation (Ehrlich, Bramaud du Boucheron, Florin, 1978 ; Florin, 1993, 1999). D'un autre côté, les corrélations avec le score au QCM et les évaluations scolaires sont aussi élevées que dans l'étude 2 : de .76 à .86 de .62 à .82 avec l'appréciation générale ; à l'exception d'une classe (.29 avec la lecture et .49 avec l'appréciation générale). Toutefois, il serait judicieux de prendre en compte les évaluations nationales en CM2 et de palier 2, pour répliquer ces observations, afin d'avoir d'autres échelles d'évaluations plus standardisées que les appréciations générales et de lecture dont nous disposions. Enfin, sur le plan appliqué et institutionnel en France, notre étude montre que les manuels de l'école primaire ne constituent pas une surcharge, puisque les élèves ont acquis 9 000 mots sur près de 11 000, ce qui semble tout à fait raisonnable, loin des 2 500 mots acquis sur 6 000 mots des manuels en 6e de collège (Lieury, 2012).

Une particularité de notre étude était de distinguer les aspects sémantiques et lexicaux du vocabulaire. Ainsi, d'après les corrélations avec le type d'erreur et la performance en lecture, il semble que l'aspect lexical ait une plus grande importance, chez les jeunes élèves, que le sémantique (CP à CE1), tandis que le sémantique prend plus d'importance en CM2. Cependant, comme nous n'avons pas construit systématiquement les distracteurs en nombre équivalent, ce résultat demande à être confirmé.

Mais l'originalité principale de notre étude était de s'intéresser au vocabulaire encyclopédique, c'est-à-dire, non pas seulement à la langue usuelle (français), mais au vocabulaire technique des connaissances en histoire, sciences et mathématiques. Dans cette voie, nous avons fait passer le QCM de vocabulaire encyclopédique du primaire à des élèves de 6e de collège, où débute la spécialisation des matières avec des cours d'histoire, biologie, mathématiques, etc. Les corrélations entre le QCM et la moyenne annuelle des notes, en fin de 6e, sont élevées, de .56 avec les mathématiques à .75 avec l'histoire. Ce résultat montre qu'il est intéressant de porter attention au vocabulaire encyclopédique, c'est-à-dire à la variété des connaissances. En présentant, en début d'année scolaire, un tel test de vocabulaire, peut-être pourrions-nous, en fin d'année, le mettre en relation avec les évaluations scolaires et déterminer, ainsi, ses qualités de validité prédictive.

## RÉFÉRENCES


ANDERSON (Richard C.), FREEBODY (Peter).– *Vocabulary knowledge – Technical report n° 136*, Center for the study of reading, Champaign Ill., University of Illinois, 1979.

ANGLIN (Jeremy M).– Vocabulary development : A morphological analysis, *Monographs of the Society for research in child development*, *LVIII*, 10, 238, 1993.

AUBRET (Françoise).– Pronostic de la scolarité en second cycle secondaire : validité prédictive de certains tests collectifs et d'appréciations scolaires en classe de 3e, *L'orientation scolaire et professionnelle*, *XVI*, 2, 1987, p. 151-158.

BROWN (George D.).– Mathematics vocabulary instruction for current non-proficient students with and without IEPs : A study of three methods of instruction, *Dissertation abstracts international section A*, *LXIX*, 1-A, 2008, p. 176-176.

COHEN (Jacob).– *Statistical power analysis for the behavioral sciences*, Hillsdale, New Jersey, Lawrence Erlbaum, 2e éd., 1988.

COHEN (Jacob).– A power primer, *Quantitative methods in psychology*, *112*, 1, 1992, p. 155-159.

CORROYER (Denis), ROUANET (Henry).– Sur l'importance des effets et ses indicateurs dans l'analyse statistique des données, *L'année psychologique*, *94*, 4, 1994, p. 607-624.

CUNNINGHAM (Anne E.), STANOVICH (Keith E.).– Tracking the unique effects of print exposure in children : Associations with vocabulary, general knowledge and spelling, *Journal of educational knowledge*, *LXXXIII*, 2, 1991, p. 264-274.

CUNNINGHAM (Anne E.), STANOVICH (Keith E.).– What reading does for the mind, *Journal of direct instruction*, *1*, 2, 2001, p. 137-149.

DÉRO (Moïse).– Corrections méthodologiques aux inventaires du vocabulaire du collège français, *Psychologie et psychométrie*, *XIX*, 3, 1998, p. 27-49.

DOTTRENS (Robert), MASSARENTI (Dino).– *Vocabulaire fondamental du français : contribution à un enseignement rationnel de l'orthographe d'usage*, Neuchâtel, Delachaux & Niestlé, 3e éd., 1963.

DUBOIS (Jean), LAGANE (René), NIOBEY (Georges), CASALIS (Didier), CASALIS (Jean), MESCHONNIC (Henri).– *Dictionnaire du français contemporain*, Paris, Larousse, 1971.

EHRLICH (Stéphane), BRAMAUD DU BOUCHERON (Geneviève), FLORIN (Agnès).– *Le développement des connaissances lexicales à l'école primaire*, Paris, Presses universitaires de France, 1978.

FLORIN (Agnès).– Les connaissances lexicales des enfants d'école primaire. Pour une didactique des activités lexicales à l'école, *Repères, recherches en didactique du français langue maternelle*, VIII, 1993, p. 93-112.

FLORIN (Agnès).– *Le développement du langage*, Paris, Dunod, 1999.

GRAVES (Mickael F.).– Vocabulary learning and instruction, *Review of research in education*, *XIII*, 1986, p. 49-89.

HARTMANN (George W).– A critique of the common method of estimating vocabulary size, together with some data on the absolute word knowledge of educated adults, *Journal of educational psychology*, *XXXII*, 5, 1941, p. 351-358.

HEPBURN (Emma).– Vocabulary acquisition in young children : The role of the story, *Journal of early childhood literacy*, *X*, 2, 2010, p. 159-182.

JENKINS (Joseph R.), STEIN (Marcy L.), WYSOCKI (Katherine).– Learning vocabulary through reading, *American educational research journal*, *XXI*, 4, 1984, p. 767-787.

JUST (Marcel A.), CARPENTER (Patricia A.).– Vocabulary acquisition, dans Just (M. A.), Carpenter (P. A.), *The psychology of reading and language comprehension*, Newton, MA, Allyn & Bacon Inc, 1987, p. 103-128.

KIRKPATRICK (E. A.).– The number of words in an ordinary vocabulary, *Science*, *XVIII*, 1891, p. 107-108.

LAMBERT (Éric), CHESNET (David).– NOVLEX : une base de données lexicales pour les élèves de primaire, *L'année psychologique*, *CI*, 2, 2001, p. 277-288.

LE CAM (Marion), ROCHER (Thierry), LORANT (Sonia), LIEURY (Alain).– Les enfants du numérique : activités extrascolaires et caractéristiques chez 30 000 élèves de 6e du collège français, *Bulletin de psychologie, 66*, 1, 2013, p. 37-60.

LEE (Joanne).– Size matters : Early vocabulary as a predictor of language and literacy competence, *Applied psycholinguistics*, *XXXII*, 1, 2011, p. 69-92.

LÉTÉ (Bernard), SPRENGER-CHAROLLES (Liliane), COLÉ (Pascale).– MANULEX : A grade-level lexical database from French elementary school readers, *Behavior research methods, instruments, & computers*, *XXXVI*, 1, 2004, p. 156-166.

LIEURY (Alain).– Et si le meilleur prédicteur de la réussite scolaire était la biologie !, *Cahiers de Beaulieu*, 13, 1991, p. 38-51.

LIEURY (Alain).– Mémoire encyclopédique et devenir scolaire : étude longitudinale d'une cohorte sur les quatre années du collège français, *Psychologie et psychométrie*, *XVII*, 3, 1996, p. 33-44.

LIEURY (Alain).– *Mémoire et réussite scolaire*, Paris, Dunod, 4e éd., 2012.

LIEURY (Alain), LORANT (Sonia).– Encyclopedic memory : Long-term memory capacity for knowledge vocabulary in middle school, *International journal of educational psychology*, *II*, 1, 2013, p. 56-80.

LIEURY (Alain), LORANT (Sonia), LE CAM (Marion), ROCHER (Thierry).– Évaluation de la mémoire

encyclopédique chez 30 000 élèves de 6e du collège français, *Bulletin de psychologie*, *66*, 1, 2013, p. 9-21.

Lieury (Alain), Van Acker (Philippe), Clévédé (Marielle), Durand (Paul).– Les facteurs de la réussite scolaire : raisonnement ou mémoire sémantique ?, 2e année d'une étude longitudinale en cycle secondaire (5e), *Psychologie et psychométrie*, *XIII*, 1, 1992a, p. 33-46.

Lieury (Alain), Van Acker (Philippe), Clévédé (Marielle), Durand (Paul).– Mémoire des connaissances et réussite en cinquième, *Le langage et l'homme*, *XXVII*, 1, 1992b, p. 61-76.

Lieury (Alain), Van Acker (Philippe), Durand (Paul).– Mémoire encyclopédique et réussite en quatrième de collège, *Psychologie et psychométrie*, *XVI*, 1, 1995a, p. 25-48.

Lieury (Alain), Van Acker (Philippe), Durand (Paul).– Mémoire encyclopédique et réussite en 3e et au brevet des collèges, *Psychologie et psychométrie*, *XVI*, 3, 1995b, p. 35-59.

Lorge (Irvin), Chall (Jeanne).– Estimating the size of vocabularies of children and adults : An analysis of methodological issues, *The journal of experimental education*, *XXXII*, 1963, p. 147-157.

Marulis (Loren M.), Neuman (Susan B.).– The effects of vocabulary intervention on young children's word learning : A meta-analysis, *Review of educational research*, *LXXX*, 3, 2010, p. 300-335.

Miller (George A.).– The acquisition of word meaning, *Child development*, *XLIX*, 1978, p. 999-1004.

Miller (George A.).– *The science of words*, New York, Scientific American library, 1991.

N'GuyenXuan (Anh).– Étude par le modèle factoriel d'une hypothèse sur les processus de développement : recherche expérimentale sur quelques aptitudes intellectuelles chez des élèves du premier cycle de l'enseignement secondaire, Thèse de doctorat, Paris, Laboratoire de psychologie différentielle, INETOP, 1969.

Nagy (William E), Anderson (Richard C.), Herman (Patricia A.).– Learning word meanings from context during normal reading, *American educational research journal*, *XXIV*, 1987, p. 237-270.

Nagy (William E), Anderson (Richard C.).– How many words are there in printed school English, *Reading research quarterly*, *XIX*, 3, 1984, p. 304-330.

Nagy (William E), Herman (Patricia A.).– Breadth and depth of vocabulary knowledge : Implications for acquisition and instruction, dans McKeown (M. G.), Curtis (M. E.), *The nature of vocabulary acquisition*, Hillsdale, New Jersey, Lawrence Erlbaum, 1987, p. 19-35.

Postal (Virginie), Lieury (Alain).– Organisation de la mémoire encyclopédique : étendue et spécificité, implication dans la réussite scolaire, *Revue européenne de psychologie appliquée*, *XLVIII*, 2, 1998, p. 113-126.

Seashore (Robert H.), Eckerson (Lois D.).– The measurement of individual differences in general English vocabularies, *The journal of educational psychology*, *XXXI*, 1940, p. 14-37.

Smith (Mary K.).– Measurement of the size of general English vocabulary through the elementary grades and high school, *Genetic psychology monographs*, *XXIV*, 1941, p. 311-345.

Sternberg (Robert J.), Powell (Janet S.).– Comprehending verbal comprehension, *American psychologist*, *XXXVIII*, 1983, p. 878-893.

Terman (Lewis M.).– *The measurement of intelligence*, Boston, Houghton Mifflin, 1916.

Van Acker (Philippe), Vrignaud (Pierre), Lieury (Alain).– Mémoire de travail, mémoire encyclopédique et performance scolaire en troisième, *L'orientation scolaire et professionnelle*, *XXVI*, 4, 1997, p. 571-596.

Verhoeven (Ludo), van Leeuwe (Jan), Vermeer (Anne).– Vocabulary growth and reading development across the elementary school years, *Scientific studies of reading*, *XV*, 1, 2011, p. 8-25.

## Annexe 1. Liste des manuels scolaires étudiés

Blanc (J.-P.), Bramand (P.), Degeilh (J.), Faye (P.), Gély (J.), Grégoire (I.), Vargas (A.).– *Physique – biologie – technologie – informatique – CM*, Paris, Hachette écoles, 1986.

Drouet (Jean-Pierre), Martinez (Yves).– *Histoire : CM1 CM2,* Paris, Magnard, 1985.

Dupré (Jean-Paul).– *La balle aux mots : langue française CE1*, Paris, Nathan, 1986.

Dupré (Jean-Paul.), Olive (Martin), Schmitt (R.).– *La balle aux mots : langue française CM1*, Paris, Nathan, 1992.

Eiller (Robert), Ravenel (Simone), Ravenel (Roger).– *Math et Calcul CM1*, Paris, Hachette éducation, 1987.

Houblain (L.), Vincent (R.).– *Les belles histoires de Daniel et Valérie : premier livre de lecture courante,* Paris, Nathan, 1973.

Lemoine (J.), Lore (Y.), Massane (Jacques), Thomas (Jean-Louis).– *La ruche aux livres : CE1*, Paris, Hachette, 1989.

Mareuil (André), Goupil (Marie).– *Le livre des bêtes*, Paris, Librairie Istra, 1969.

Pernice (Vincent).– *La ronde des mots, deuxième livre de lecture courante*, Paris, Nathan, 1991.

Touyarot (Charles), Rollant (Charles), Giribone (Claude).– *Au fil des mots*, Paris, Nathan, 1977.