Le lexique mental de l'enfant de 6 à 11 ans.

Présentation des bases lexicales

MANULEX et MANULEX-INFRA.

Pr. LÉTÉ Bernard*
Université de Lyon 2
Institut de Psychologie
Laboratoire d'Étude des Mécanismes Cognitifs (EA 3082)

Résumé:

Pour étudier le lexique mental, il est nécessaire de développer des bases de données lexicales pour estimer la nature des unités lexicales représentées dans la mémoire lexicale des individus. Le but de notre intervention sera de présenter les bases lexicales françaises MANULEX et MANULEX-INFRA qui quantifient le lexique écrit adressé à l'enfant de 6 à 11 ans dans les manuels de lecture. A l'aide de ces bases, nous tenterons également d'estimer la taille du vocabulaire (i.e., la taille du lexique d'usage) d'un enfant de l'école primaire et de décrire son évolution.

MANULEX (Lété, Sprenger-Charolles et Colé, 2004 ; voir également Lété, 2003, 2004) est une base de données lexicales qui fournit les fréquences d'occurrences de mots calculées à partir d'un corpus de 54 manuels scolaires (1,9 millions de mots). Les listes de fréquences de mots sont fournies pour trois 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 CEI (7 ans) où se construit le lexique orthographique par automatisation progressive de la reconnaissance du mot écrit et le cycle 3 (CE2-CM2, 8-11 ans) où se consolide et s'enrichit le stock lexical par exposition répétée à l'écrit. Un quatrième niveau est constitué par le regroupement de l'ensemble des manuels du CP au CM2. MANULEX est constituée de deux lexiques pour les 4 niveaux considérés : un lexique de 48 900 formes orthographiques différentes rencontrées dans les manuels (chantons, chanteraient, bateau, bateaux, ...) et un lexique de 23 900 lemmes pour lesquels la

fréquence correspond à la somme des fréquences des formes orthographiques (chanter = chantons + chanteraient, bateau = bateau + bateaux, ...). MANULEX-INFRA (Peereman, Lété, Sprenger-Charolles, 2007) décrit le lexique adressé à l'enfant à l'école primaire à un niveau infra-lexicale (associations graphèmes-phonèmes,syllabes, bigrammes, trigrammes, ...). Nous avons en particulier développé une métrique de la consistance des relations grapho-phonologiques et phono-graphiques à chaque niveau d'âgepour tous les mots du lexique des formes orthographiques de MANLILEX. La consistance dechaque mot est mesurée sur une échelle de 1 à 100 à chaque niveau car celle-ci peut varier en fonction du lexique adressé à l'enfant à une étape particulière de son apprentissage.

Mots clés: Lexique mental- MANULEX– MANULEX-INFRA - Fréquence.

*Auteur correspondant : Bernard LETE

المعجم الذهني للطّفل من ستّة (6) إلى إحدى عشر (11) سنة: عرض لبنكي المعطيات مانولاكسومانولاكسأنفرا

ملخص:

من أجل دراسة المعجم الذهني، لا بد من إعداد بنوك للمعطيات المعجمية، وذلك بغية تقدير طبيعية الوحدات المعجمية الممثلة على المستوى الذاكرة المعجمية للأشخاص. يتمثل الهدف من مداخلتنا هذه في عرض بنكين معجميين فرنسيين: مانولاكس، وكانولاكس أنقرا، اللذان يعدّان المعجم المكتوب الموجه للأطفال من 6إلى11سنة في مناهج القراءة. وبالاعتماد على هذين البنكين، سنحاول أيضا تقدير حجم المفردات (أي حجم المعجم المستعمل) لطفل في المدرسة الابتدائية وكيفية تطوره.

يعد مانولاكس بنكا للمعطيات المعجمية، والذي يمنح التواترات اللفظية للكلمات المحسوسة انطلاقا من 54 دليلا مدرسيا(1.9مليون كلمة).

تحدد قوائم تواترات الكلمات حسب ثلاثة مستويات لمهارة القراءة:

- السنة الأولى ابتدائي (6سنوات) أين يتطور المعجم من خلال التوسط الفونولوجي.
- السنة الثانية(7سنوات) أين يتطور المعجم الإملائي من خلال آلية تدريجية للتعرف على الكلمة.
- الطور 3(11-8سنة):أين تتلاحم ويُثرى المخزن المعجمي من خلال العرض المتكرر للكلمات المكتوبة.
- ويأتي الطور (4) تجميعاً لمجموع الدلائل، من السنة الأولى إلى السنة السادسة.

يتكون مانولاكس من معجمين للمستويات الأربع: معجم يضم48900لمختلف الأشكال الإملائية الموجودة في المناهج، على سبيل المثال(chantons, chanterons bateau, bateaux…) ومعجم متكون من:23900والتي تناسب درجة تواترها مجموع تواترات الأشكال الإملائية

(...= bateau, bateaux chanter= chantons, chanterons, bateau)

يصف مانولاكس انفرا (Peereman, Lété, Sprenger-Charolles, 2007) المعجم الموجه للطفل في المدرسة الابتدائية الخاص بالمستوى دون المعجمي (ربط الحرف الشفاهي بالحرف المكتوب، المقاطع، الحروف الثنائية، الحروف الثلاثية...).

صممنا أداة قياس التوافق بين العلاقات الكتابية والفونولوجية والعلاقات الفونولوجية-الكتابية لكل فئة عمرية لجميع كلمات المعجم للأشكال الإملائية لمانولاكس. يقاس توافق كل كلمة من خلال سلم يتدرج بين 1-100 لكل مستوى. وذلك لأن التوافق قد تغير تبعا للمعجم الموجه للطفل في مرحلة معينة من التعلم.
الكلمات المفتاحية:المعجم الذهني- مانولاكس- مانولاكسأنفرا- تواتر الكلمات.

The mental lexicon of the child from 6 to 11 years old: presentation of the lexical bases MANULEX and MANULEX-INFRA

Abstract:

To study the mental lexicon, it is necessary to develop lexical databases to estimate the nature of the lexical units represented in the lexical memory of individuals.The aim of our intervention will be to present the French lexical bases MANULEX and MANULEX-INFRA which quantify the written lexicon addressed to children from 6 to 11 years old in reading manuals.Using these databases, we will also try to estimate the size of the vocabulary (i.e., the size of the lexicon of use) of a child in elementary school and to describe its evolution.

MANULEX (Lété, Sprenger-Charolles, &Colé, 2004; see also Lété, 2003, 2004) is a lexical database that provides word occurrence frequencies calculated from a corpus of 54 textbooks (1.9 million words). The word frequency lists are provided for three levels of reading expertise: first grade (6 years old), where the child's lexicon is built on the basis of phonological mediation; second grade (7 years old), where the orthographic lexicon is built through the gradual automation of written word recognition; and third grade (8-11 years old), where the lexical stock is consolidated and enriched through repeated exposure to the written word. A fourth level is made up of all the textbooks from CP to CM2. MANULEX consists of two lexicons for the 4 levels considered: a lexicon of 48,900 different orthographic forms encountered in the textbooks (chantons, chanteraient, bateau, bateaux, ...) and a lexicon of 23,900 lemmas for which the frequency corresponds to the sum of the frequencies of the orthographic forms (chanter = chantons + chanteraient, bateau = bateau + bateaux, ...). MANULEX-INFRA (Peereman, Lété, Sprenger-Charolles, 2007) describes the lexicon addressed to children in elementary school at an infra-lexical level (grapheme-phoneme associations, syllables, bigrams, trigrams, ...). In particular, we have developed a metric of the consistency of grapho-phonological and phono-graphical relations at each age level for all words in the MANLILEX lexicon of orthographic forms. The consistency of each word is measured on a scale of 1 to 100 at each level because it can vary according to the lexicon addressed to the child at a particular stage of learning.

Keywords: Mental Lexicon- MANULEX - MANULEX-INFRA - Frequency .

Introduction

MANULEX[1] (Lété, Sprenger-Charolles, &Colé, 2004*)* et MANULEX-INFRA (Peereman, Lété, &Sprenger-Charolles, 2007) sont des bases de données lexicales extraites d'un corpus de 54 manuels scolaires de lecture. Librement accessible sur le Web[2], elles sont à la disposition des chercheurs qui travaillent notamment sur l'acquisition de la lecture.

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. Ehrlich, Brameraud du Boucheron et Florin (1978) notent en effet une importante accélération du développement des connaissances lexicales entre le CE1 et le CM2 (7 ans et 10 ½ 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.

MANULEX : Une estimation du vocabulaire des enfants de 6 à 11 ans

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. L'ensemble des manuels a également été regroupé 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. Pour chaque niveau, deux lexiques ont été constitués : un lexique des formes orthographiques de 48 886 entrées (formes non fléchies) et un lexique des lemmes de 23 812 entrées (formes fléchies équivalentes aux entrées d'un dictionnaire).

---

[1] Pour *Lexique* des *Manuels.*

[2] eMANULEX : http://www.manulex.org/fr/home.html

Pour chaque mot, MANULEX fournit quatre indices : *F* : la fréquence brute du mot dans le corpus ; *D* : l'indice de dispersion du mot parmi les manuels ; *U* : la fréquence par million estimée à partir de *D* ; *IFC* : un indice de fréquence courant calculé à partir de *U* par transformation logarithmique. La fréquence par million *U* est dérivée de *F* avec un ajustement avec *D.U* approche la valeur réelle de *F* dans un corpus théorique de taille infinie. C'est la fréquence d'*u*sage du mot dans les manuels qui traduit mieux son importance que la seule fréquence *F*. Par exemple, les fréquences brutes respectives de *"papa"* et *"point"* sont de 1 567 et 1 601 par million. Le problème est que "papa" est rencontré dans tous les manuels alors que *"point"*est rencontré dans très peu de manuels mais avec une forte occurrence dans chacun de ces manuels. Pondérées par *D* (0.79 et 0.24), leur fréquence d'usage par million (*U)*est 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). 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 dus 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.

Pour faire une estimation du stock lexical chez l'enfant de l'école primaire, 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 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 Anglin (1993). L'auteur considère que le stock

lexical d'un enfant de CP est de 10 400 mots[3], 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 Anglin aux 59 000 entrées du *Larousse* (2004) sur les noms communs. Dire qu'en enfant de CM2 connaît près de 40 000 mots revient à considérer qu'il connaît près des 2/3 des entrées d'un dictionnaire (et même plus si on considère la classification de Anglin, cf. note 3) ce qui est loin d'être le cas.

L'erreur de ces estimations est de partir d'un lexique de référence représentant l'ensemble des mots de la langue. 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.

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 le nombre de mots observés dansun manuel (voir Lété, 2003, pour des analyses plus poussées). C'est une bonne estimation 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.Au CP, un manuel a en moyenne 13 300 mots, 2 500 formes orthographiques différentes et 1 900 lemmes (une forme = 1.3 lemme). Un manuel de CE1 a en moyenne 1 000 lemmes de plus. Au cycle 3, la progression est de 2 000 lemmes. Avec 4 900 lemmes considérés comme acquis, on arrive à une estimation près de 5 fois inférieure au *lexique* des manuels qui représente les mots qui peuvent être rencontrés par les enfants (au cycle 3, le nombre de lemmes rencontrés dans les manuels scolaires est de 22 500). Au cycle 3, un manuel a en moyenne autant de formes orthographiques et de lemmes qu'un livre de la littérature comme *Les Lettres de mon moulin* d'Alphonse Daudet.

Même si un enfant lit beaucoup plus, son stock de vocabulaire ne s'accroîtra pas forcément beaucoup. Pourquoi ? Prenons l'exemple du livre de Daudet : si l'enfant lit son manuel de lecture du cycle 3 plus *Les Lettres de mon*

[3] En réalité, ce sont des familles morphologiques de mots. Selon sa classification, *chant* et *chanter* constituent un même mot de base ce qui réduit le nombre d'entrées par rapport à notre comptage sur les lemmes.

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 : 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[4].

Au CP, les 1 000 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 1 000 mots (lemmes). Si l'enfant lit 11 pages de plus (11%), il va rencontrer 2 000 lemmes de plus mais ceux-ci seront beaucoup moins répétés (une seule fois vraisemblablement). Ceci dit, même si l'enfant connaît bien 1 000 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 5 000 lemmes 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 cinq 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.

Avec 5 000 lemmes à connaître, l'enseignement du vocabulaire n'est donc pas une tâche insurmontable à l'école. Ceci dit, connaître 5 000 lemmes ne suffit pas à être lecteur. Les écrits adultes comportent approximativement près de 73 000 lemmes (d'après la base adulte LEXIQUE, New, Pallier, Ferrand, & Matos, 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.

---

[4]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 dote chacun d'eux d'un rang traduit par un nombre *r*, le produit *f* * *r* tend à être constant.

MANULEX-INFRA et l'incertitude des associations graphèmes-phonèmes dans l'orthographe du français

Une façon objective de rendre compte de la complexité orthographique d'une langue consiste à calculer le degré d'incertitude associé à l'écriture d'une association phonème-graphème ou à la lecture d'une association graphème-phonème. Ce niveau d'incertitude, plus communément appelé "consistance", permet de prédire la difficulté à écrire ou lire un mot.

La consistance phonologie vers orthographe (ci-après consistance PO), impliquée par exemple dans une tâche de copie de mots sous dictée, réfère à la variabilité des graphèmes qui peuvent être assignés à un phonème donné. Par exemple, la consistance PO d'un mot baisse si un de ses phonèmes peut être associé à plusieurs graphèmes (par exemple /C/ en français est orthographié différemment dans les mots *"saint"*, *"pin"*, et *"rein")* alors qu'elle augmente si un de ses phonèmes peut être associé à un seul graphème (par exemple /U/ est toujours orthographié *"ou"* en français comme dans les mots *"fou"*, *"cou"* et *"bijou"*). De façon opérationnelle, la consistance est calculée comme la proportion de mots pour lesquels le phonème est associé à un graphème particulier relativement au nombre total de mots dans lesquels le phonème apparaît quelle que soit sa façon d'être orthographiée. La valeur résultante s'étend de 0 (consistance minimale) à 1 (consistance maximale). (Multipliée par 100 dans MANULEX-INFRA.) La consistance peut aussi être estimée dans la direction orthographe vers phonologie (ci-après consistance OP), direction qui est celle de la lecture à voix haute. La consistance OP réfère alors à la variabilité des phonèmes qui peuvent être assignés à un graphème donné. Les consistances OP et PO peuvent varier indépendamment l'une de l'autre. Par exemple, l'anglais et le français ont des consistances PO sensiblement équivalentes alors que les voyelles françaises sont plus consistantes dans le sens OP (Peereman& Content, 1999).

La description des associations graphèmes-phonèmes a été effectuée à partir de 37 phonèmes (plus un muet) et 125 graphèmes. Au total, 290 associations graphèmes-phonèmes ont été recensées dans les mots de MANULEX (45 080 mots retenus dans MANULEX-INFRA).Pour calculer l'indice de consistance, un mot comme *main* se voit d'abord affecté de la séquence de ses deux associations graphème↔phonème ([m]-/m/ et [ain]-/C/). Puis les fréquences de chaque association, précédemment calculées sur l'ensemble des mots de la base, sont associées au début, au milieu et à la fin du mot considéré. Les fréquences des 290 associations sont équivalentes quel que soit le sens considéré (OP et PO). La consistance PO (dans le sens de l'écriture) est obtenue en divisant la fréquence de l'association /C/-[ain] par la fréquence

totale de toutes les variantes graphémiques du phonème /C/ (incluant donc [in], [ein], [aim], …). La valeur obtenue (entre 0 et 1) est multipliée par 100. Pour notre exemple, sur 100 apparitions du phonème /C/ en fin de mot, il s'écrit [ain] dans 18% des cas et sur 100 apparitions du phonème /m/ en début de mot, il s'écrit toujours [m]. Le mot *main* se verra donc affecté d'un indice de consistance PO de 100 en début de mot et de 18 en fin de mot. Son indice de consistance total est obtenu en faisant la moyenne des indices de début et de fin, soit 59. Pour ce qui est de la consistance OP (lecture à voix haute), les indices sont de 100 tant en début ([m] se lit toujours /m/ en début de mot) qu'en fin de mot ([ain] se lit toujours /C/ en fin de mot). Ainsi le mot *main* a une consistance OP totale de 100. Selon ces mesures, le mot *main* est deux fois plus difficile à écrire qu'à lire. Pour un mot de trois associations et plus, un indice de consistance de milieu de mot est obtenu en faisant la moyenne de tous les indices de consistance du milieu du mot.

Les indices de consistance ont été calculés pour le début du mot (consistance de la première association graphème-phonème), la fin du mot (consistance de la dernière association graphème-phonème) et le milieu du mot (consistance moyenne de toutes les associations graphème-phonème intra-mot). Enfin, un indice de consistance pour la totalité du mot correspond à la moyenne de toutes les consistances. Dans le sens OP (lecture à voix haute), la consistance moyenne est de 94 pour le début du mot, 78 pour le milieu et 85 pour la fin. L'orthographe du français est donc consistante dans le sens de la lecture. Il n'en va pas de même dans le sens PO (écriture sous dictée) pour laquelle les consistances moyennes sont de 83 et 74 pour le début et le milieu du mot (semblable à la consistance OP) mais de seulement 44 pour la fin du mot. L'orthographe du français est inconsistante dans le sens de l'écriture, mais seulement pour la fin du mot à cause de toutes les marques morphosyntaxiques (féminin, pluriels, conjugaison des verbes).

Lété, Peereman et Fayol (2007) ont mené la première étude en langue française cherchant à expliquer l'évolution des connaissances orthographiques d'enfants de CP au CM2 à l'aide des mesures de consistance de MANULEX-INFRA. Via une méthode d'analyse en régressions multiples, plusieurs prédicteurs de la réussite en orthographe ont été liés à un large ensemble de mots (3430 mots) mono et polysyllabiques (1 à 6 syllabes) reflétant le système d'écriture à partir duquel un enfant apprend à lire et à écrire. Il y avait quatre grandes classes de prédicteurs : la longueur du mot, la fréquence de la forme orthographique, le voisinage phonographique[5] et la complexité orthographique indexée par la consistance de début, milieu et fin de mot. L'étude constitue

[5]Un voisin phonographique est un mot qui se distingue d'un autre par une lettre et un phonème.

ainsi une exploration complète de la production orthographique chez des enfants de 6 à 11 ans.

On a cherché à distinguer et à étudier l'évolution de deux types de procédures pour écrire un mot : le recours aux connaissances infra-lexicales que l'on a indexé par les mesures de consistance de MANULEX-INFRA (pour le début, le milieu et la fin du mot) et la récupération directe de la représentation orthographique du mot que l'on a indexé par les mesures de fréquence lexicale (forme orthographique) de MANULEX. Une base de données récente sur des productions écrites en dictée de mots (EOLE, Pothier & Pothier, 2003) nous a permis de disposer de mesures comportementales sur lesquelles on pouvait mener les analyses de régressions[6].

Le raisonnement était que de meilleures performances obtenues pour des mots fréquents par rapport à des mots rares traduiraient l'utilisation de la procédure lexicale de récupération de la forme orthographique du mot en mémoire. D'un autre côté, de meilleures performances sur les mots consistants par rapport aux inconsistants traduiraient l'utilisation de la procédure infra-lexicale. Selon les modèles en stades (Frith, 1985), la contribution significative d'une procédure à un niveau donné devrait s'accompagner d'une diminution de la contribution de l'autre procédure, car le passage d'une procédure à l'autre se fait par abandon de l'une au profit de l'autre. Dans une conception plus interactive des processus cognitifs, la contribution significative d'une procédure à un niveau donné devrait s'accompagner d'un maintien de la contribution de l'autre procédure, car il n'y a pas passage d'une procédure à l'autre mais automatisation accrue des procédures qui deviennent de plus en plus rapidement et correctement mobilisées dans la tâche de production orthographique.

Les résultats ont infirmé la conception en stades du développement en confirmant l'existence des deux procédures à chaque niveau d'acquisition : dès le CP, la procédure lexicale est active pour les mots monosyllabiques et la procédure infra-lexicale est active pour les polysyllabiques. Autrement dit, c'est la longueur du mot qui détermine la procédure à partir de laquelle l'enfant produit l'orthographe du mot. Les résultats ont confirmé également qu'une difficulté importante dans l'acquisition de l'orthographe est imputable à la consistance des relations phonème vers graphème. La difficulté liée à la complexité orthographique (i.e., l'effet de consistance) ne disparaît pas en

[6]EOLE liste, pour près de 12 000 mots, le pourcentage de graphies correctes obtenu du CP au CM2 dans une tâche de copie de mots sous dictée. Chaque mot a été orthographié par 40 enfants à chaque niveau (42 000 enfants au total). L'ensemble des mots étudiés a été ramené à 3 430 dans notre étude.

fonction de l'expertise, elle reste même à un niveau sensiblement équivalent à travers les niveaux et ne diminue que pour les mots monosyllabiques au CE2.

Conclusion

Les analyses distributionnelles des unités linguistiques dans des corpus d'écrits, telles que celles ici rapportées avec MANULEX et MANULEX-INFRA permettent de mieux définir les régularités statistiques, présentes dans l'input langagier, susceptibles d'être exploitées lors de l'acquisition du langage. Un effet avéré de ces variables statistiques (comme la fréquence lexique ou la consistance) permet de considérer que l'enfant apprenti-lecteur exploite ces régularités dans la compréhension et la production du langage. Cette sensibilité de l'enfant aux propriétés statistiques de son environnement langagier –ce qu'on dénomme l'apprentissage probabiliste du langage (Chater & Manning, 2006)– pourrait dès lors être invoquée comme le mécanisme essentiel à la construction de son lexique mental dans lequel des unités probables sont renforcées en mémoire, les autres plus ambigües ou plus rares étant progressivement oubliées.

Cette hypothèse n'est pourtant pas totalement exempte de circularité. Si certains indices distributionnels semblent exploitables *a priori*, l'extraction de ces régularités ne peut s'opérer qu'à partir d'un lexique déjà structuré, au moins partiellement. Il faut donc prendre en compte la capacité de généralisation du système de l'apprenant par extraction de règles, que celles-ci émergent de sa base linguistique ou de leur enseignement explicite à l'école. Grâce aux règles formelles du langage, l'enseignement pourrait en effet "mettre de l'ordre" dans le contenu des apprentissages probabilistes en organisant l'espace de stockage des informations, mais également, et surtout, en identifiant et en éliminant les erreurs durant l'apprentissage. En fait, un des problèmes des apprentissages probabilistes est qu'ils encodent les informations fréquentielles sans isoler ces erreurs, les apprenants se familiarisant aussi bien avec les associations erronées qu'avec les associations correctes.

Comme le note Ellis (2002), "*linguistic regularities emerge as central tendencies in the conspiracy of the database of memories of utterances*" (p.166). Dans ce cadre, tout effetd'une variable statistique (fréquenceouconsistance) "*describe the tuning of the language system through use*"[7] (p.173).

[7] *Les régularités linguistiques apparaissent comme des tendances centrales dans l'organisation de la mémoire lexicale*. Dans ce cadre, tout effet d'une variable statistique *rend compte du réglage du système langagier par son usage*.

Références

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

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences, 10*, 287-291.

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

Ellis, N. C. (2002). Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143-188.

Frith, U. (1985). Beneath the surface of developmental dyslexia. In Karalyn E. Patterson, John C. Marshall & Max Coltheart (Eds.), *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading* (pp. 301-330). London: Erlbaum.

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

Lété, B., Peereman, R., & Fayol, M. (2008). Phoneme-to-grapheme consistency and word-frequency effects on spelling among first- to fifth-grade French children: A regression-based study. *Journal of Memory and Language, 58,* 952-977.

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*, 166-176.

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

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.

Peereman, R., & Content, A. (1999). LexOP. A Lexical database with Orthography-Phonology statistics for French monosyllabic words. *Behavior Research Methods, Instruments, and Computers*, *31*, 376-379.

Peereman, R., Lété, B., &Sprenger-Charolles, L. (2007). MANULEX-INFRA: Distributional characteristics of grapheme-phoneme mappings, infra-lexical and

lexical units in child-directed written material.*BehaviorResearchMethods, 39*, 593-603.

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