

QUELLES SONT LES RELATIONS ENTRE LECTURE ET ÉCRITURE POUR L'APPRENTISSAGE DU LEXIQUE ?



Bernard LÉTÉ

Professeur à l'Université Lyon 2

L'apprentissage de l'écrit implique à la fois la maîtrise de la lecture et de l'orthographe. Dans les modèles intégrateurs, on considère que ces deux compétences se développent en interaction. Par exemple, la *self-teaching hypothesis* de Share (1995) considère que chaque identification réussie d'un nouveau mot par l'intermédiaire du décodage fournit une occasion d'acquérir les représentations orthographiques des mots sur lesquelles reposent la lecture et l'orthographe. L'exposition à l'écrit (les lectures de l'enfant) a un rôle important dans la production orthographique car celle-ci repose en grande partie sur l'exploitation des connaissances acquises en réception de l'écrit. Les calculs des fréquences d'occurrences textuelles¹ des unités linguistiques dans des corpus d'écrits permettent de mieux définir quelles sont les régularités statistiques traitées lors de la lecture, susceptibles d'être exploitées lors de la production de l'orthographe. Ces fréquences d'occurrences textuelles peuvent être utilisées pour quantifier la consistance orthographique d'un mot c'est-à-dire le degré d'incertitude associé à l'écriture d'une association phonème-graphème (tâche d'écriture sous dictée par exemple) ou à la lecture d'une association graphème-phonème (tâche de lecture à voix haute). Un effet avéré de ces variables statistiques permet de considérer que l'enfant exploite ces régularités dans la compréhension et la production du langage. Cette sensibilité de l'enfant aux propriétés statistiques de son environnement langagier pourrait dès lors être invoquée comme mécanisme essentiel à la construction de sa base de connaissances linguistiques dans laquelle des unités probables sont renforcées en mémoire, les autres plus ambiguës ou plus rares étant progressivement oubliées.

Le but de cette contribution est de **caractériser le matériau linguistique rencontré par les enfants dans leur manuel de lecture² (i.e., l'input linguistique) pour tenter de caractériser et expliquer la nature de leur production orthographique (i.e., l'output linguistique).**

¹ Ou *fréquences des tokens* c'est-à-dire le nombre de fois (en fréquence par million) qu'une unité est rencontrée dans un corpus de textes. Ces fréquences des tokens sont à distinguer des *fréquences du type*, c'est-à-dire le nombre de fois qu'une unité est rencontrée dans la liste des mots du corpus (chaque mot a donc une fréquence de 1). Les fréquences des tokens sont les seules mesures appropriées pour caractériser un apprentissage puisqu'elles reposent sur l'expérience réelle d'un individu avec les unités linguistiques de sa langue.

² Voir Lété (2018), pour une description plus complète.

I. Apprendre à orthographier les mots en production écrite

Les modèles en stades sont souvent utilisés pour rendre compte de l'apprentissage de l'orthographe des mots en production écrite (*cf.* pour une synthèse, Fayol 2017 et Fayol et Jaffré, 2014). L'apprentissage y est vu comme une progression à travers des étapes distinctes dans lesquelles l'introduction d'une stratégie plus avancée aide l'enfant à progresser vers l'étape suivante. Les changements surviennent sur une courte période ou plus graduellement selon les modèles. L'enfant découvre une nouvelle stratégie et l'applique quand il la considère comme pertinente, ou utilise de façon de plus en plus experte une stratégie déjà existante. D'autres modèles s'inscrivent dans le cadre du modèle de la double voie utilisé dans l'étude de la lecture : lors d'une écriture sous dictée par exemple, la conversion phonème-graphème repose sur une voie infra-lexicale pour laquelle on suppose que chaque graphème est produit à partir de sous unités phonologiques (phonèmes, syllabes) et une voie lexicale pour laquelle on suppose que la forme orthographique du mot est directement récupérée en mémoire. À l'instar de la conception en lecture, la voie lexicale est considérée comme plus rapide que la voie infra-lexicale car elle est régit par des processus automatiques alors que la voie infra-lexicale est ralentie par des processus analytiques mobilisant plus de ressources attentionnelles.

Le fait que les enfants utilisent de multiples stratégies orthographiques peut s'accorder avec l'hypothèse que, de façon générale, le traitement du langage englobe une exploitation complexe, de nature probabiliste, de l'input linguistique. De nombreuses études ont en effet montré que même de très légères variations dans les propriétés statistiques des inputs linguistiques ont des conséquences sur les performances. Par exemple, dès l'âge de huit mois, l'enfant découvre les frontières des mots dans le flux de parole en utilisant les fréquences de co-occurrences des syllabes³ (Saffran *et al.*, 1996). Aussi, il serait surprenant que les enfants plus âgés ne manifestent pas une sensibilité aux propriétés statistiques des événements linguistiques lorsqu'ils apprennent à lire et à écrire. Par exemple, les travaux de Pacton et collaborateurs en français (*e.g.*, Pacton, Fayol et Perruchet, 2002) montrent que des enfants de CP possèdent un savoir implicite sur les régularités orthographiques plus précocement que ne le prédiraient les modèles d'acquisition en stades. Ils savent que certaines consonnes ne peuvent pas être doublées, que des consonnes doublées peuvent apparaître à certaines positions mais pas à d'autres et que les voyelles, par contre, ne peuvent pas être doublées dans un mot français. Le fait que ces connaissances soient acquises sans un enseignement explicite suggère que **les enfants sont capables d'extraire automatiquement certaines régularités statistiques des propriétés de l'écrit au fur et à mesure de leur exposition à l'écrit.**

³ Les fréquences de co-occurrences des syllabes entre les mots sont inférieures à celles à l'intérieur des mots. Elles donnent donc un indice pour segmenter les unités lexicales dans un flux continu de parole.

II. Les manuels de lecture et l'apprentissage statistique de l'écrit

L'apprentissage statistique de l'écrit est un type particulier d'apprentissage implicite, non enseigné, pour lequel on considère que l'apprenant s'ajuste aux régularités statistiques des événements linguistiques auxquels il est confronté grâce à des processus associatifs élémentaires. Ces propriétés statistiques des événements linguistiques renvoient, entre autres, à : (1) leur répétition : dans le cas de la lecture, la fréquence d'exposition à un mot ; (2) son incertitude : la consistance orthographique d'une association graphème-phonème (papa / oignon) ; (3) sa discriminabilité : le voisinage orthographique ou phonologique d'un mot (foire / poire) ; (4) aux co-occurrences conditionnelles de certaines unités : par exemple, quand on considère les fréquences textuelles, le phonème /o/ s'écrit plus fréquemment "au" après un "p" que "o" alors qu'il s'écrit plus fréquemment "o" en fin de mot après "m".

Nous pouvons nous appuyer sur deux bases de données, librement accessibles sur le Web⁴, pour disposer de mesures objectives des fréquences d'occurrences des unités lexicales et infra-lexicales dans l'écrit adressé à l'enfant à l'école primaire dans les manuels de lecture : *Manulex* (Lété, Sprenger-Charolles et Colé, 2004) fournit les fréquences des unités lexicales pour les cinq niveaux de l'école primaire ; *Manulex-infra* (Peereman, Lété, et Sprenger-Charolles, 2007) donne une mesure graduée de la consistance orthographique des mots dans le sens orthographe vers phonologie (tâche de lecture à voix haute par exemple) et phonologie vers orthographe (tâche d'écriture sous dictée par exemple).

1. L'écrit adressé à l'enfant à l'école primaire : un vocabulaire proche de 5000 lemmes au CM2

Il convient de faire la différence entre la notion de lexique d'une langue qui est la liste des mots d'une langue telle qu'elle peut être fournie par *Manulex* par exemple, et la notion de vocabulaire qui correspond à la mise en pratique du lexique d'une langue par un individu, par la lecture et la production écrite. Pour estimer la taille du vocabulaire d'un enfant, on peut essayer d'estimer la quantité de mots à laquelle un enfant est exposé durant ses lectures. C'est lors de cette confrontation à l'écrit que se construit son vocabulaire. Dans *Manulex*, le lexique des manuels scolaires de lecture est de 6 800 lemmes au CP, de 10 500 au CE1 et de 22 500 au cycle 3.

Pour estimer le nombre de mots auquel un enfant est exposé de 6 à 11 ans, et ainsi quantifier son vocabulaire, on peut raisonner à partir de son manuel de lecture. Au CP, un manuel a en moyenne 13 300 mots, 2 500 formes orthographiques différentes correspondantes à 1 900 lemmes (un lemme = 1.3 forme orthographique). Un manuel de CE1 a en moyenne 1 000 lemmes de plus qu'en CP. Au cycle 3 (CE2-CM1-CM2), la progression est de 2000 lemmes. Avec 4 900 lemmes considérés comme acquis à la fin du CM2, on arrive à une estimation cinq fois inférieure au lexique des manuels, qui représente les mots qui peuvent être rencontrés par les enfants.

Un enfant de CM2 possède donc, en réception de l'écrit, un vocabulaire de l'ordre de 5000 lemmes (7500 formes orthographiques) pour lesquels on peut supposer 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

⁴ Site *eManulex*, version électronique des bases *Manulex* à <http://www.manulex.org/fr/home.html>

l'écrit. C'est quatre fois moins que ce qu'il est susceptible de rencontrer dans un corpus d'écrit de son niveau scolaire et 15 fois moins que ce qu'il est susceptible de rencontrer plus tard dans des écrits adressés à l'adulte. Cette taille de 5000 unités lexicales en français correspond très précisément à celle calculée en anglais par Nation (2001). De même, Laufer (1992) considère que, pour bien comprendre des textes en anglais langue seconde, il faut connaître environ 5000 mots anglais. Enfin, cette estimation est congruente avec deux autres études en français portant sur les connaissances lexicales des enfants de 6 à 11 ans. L'une (Ehrlich, Bramaud du Boucheron et Florin, 1978) estime le répertoire lexical d'enfants du CE1 au CM2 à partir de jugements de connaissance. Pour cela, ils ont proposé à des élèves du CE1 au CM2 une tâche de jugement de mots (450 mots par élève) sur une échelle en 5 points. La mesure du répertoire correspond au nombre de mots jugés *connus* à *très bien connus* rapportés au nombre total de mots de l'échantillon (13 500). Les données en jugement de connaissance sur les mots très bien connus (pour lesquels on peut considérer que les jugements sont les plus fiables) sont de l'ordre de 3 300 mots à la fin du cycle 3. L'autre étude est plus récente et concerne l'orthographe lexicale. Pothier et Pothier (2003) ont élaboré une échelle actualisée d'acquisition en orthographe lexicale (*Éole*). Ils ont fait orthographier 11 700 mots à 48 900 élèves du CP au CM2 (50 mots par élève). Le nombre de mots correctement orthographiés par 75 % des élèves sert de mesure du répertoire. **Les données montrent que trois quarts des enfants orthographient correctement 300 mots au CP, 1 000 au CE1, 2 000 au CE2, 3 500 au CM1 et 5 000 au CM2.** La production des graphies correctes équivaut alors au répertoire lexical estimé dans les deux autres études.

2. Toute la complexité orthographique du français est présente dans les manuels dès le CP

La complexité orthographique du français est d'emblée présente dans l'écrit adressé à l'enfant dans sa première année d'apprentissage : les 290 associations trouvées dans les 45 080 formes orthographiques de *Manulex-infra* (lexique du CP au CM2) sont toutes présentes dans les manuels de CP dans seulement 10 861 formes orthographiques, soit un quart seulement des formes orthographiques de *Manulex*. De plus, les valeurs de consistance ne varient pas d'un niveau à l'autre : quel que soit le segment considéré (début, milieu, fin et totalité du mot), les consistances moyennes restent identiques alors que la taille du lexique s'accroît. Ces données indiquent que **les enfants sont exposés à toutes les inconsistances orthographiques dès la première année de leur apprentissage de la lecture**, ce qui est pour eux une situation idéale pour l'acquisition de l'écrit. En effet, on sait que des réseaux de neurones entraînés seulement sur des mots consistants ont des difficultés à ajuster leurs connexions pour apprendre de nouvelles associations inconsistantes.

En définitive, l'enfant d'école primaire dispose, dès la fin de sa première année d'apprentissage, et pour peu qu'on lui enseigne correctement les règles de correspondance graphème-phonème, de l'ensemble du matériau linguistique utile pour lire et écrire des unités lexicales qu'il n'a jamais rencontrées. L'apprentissage du code écrit dépend donc fortement de ses capacités à généraliser ses connaissances à des unités lexicales dont la probabilité de rencontre est extrêmement faible. Il devra pour cela repérer correctement les unités graphémiques permettant de créer les associations orthographe ↔ phonologie pertinentes et les renforcer à chaque nouvelle rencontre des mots.

3. Les mots acquis au CP constituent un lexique orthographique de base

Près de 93 % des formes orthographiques du CP sont contenues dans le lexique du cycle 3 mais elles représentent seulement 24 % des formes orthographiques de ce lexique. Autrement dit, les trois-quarts des formes orthographiques du cycle 3 apparaissent seulement à partir de ce niveau. Quelles sont les caractéristiques de ce lexique précoce du CP et de ce lexique tardif au cycle 3 ? Les mots acquis tôt ont 7 caractères en moyenne et sont bisyllabiques. Ils se distinguent des mots acquis tardivement essentiellement par leur fréquence lexicale. Ils sont plutôt fréquents au CP (19 par million en moyenne) et deviennent encore plus fréquents au cycle 3 (car ils sont de plus en plus répétés). Les mots acquis tardivement sont rares et un peu plus longs que ceux du lexique précoce (9 caractères) et trisyllabiques. On observe enfin que la complexité orthographique n'est pas plus importante pour les mots du lexique tardif qui sont pourtant plus rares. Cet aspect laisse suspecter que la consistance orthographique d'un mot n'est pas liée à sa fréquence lexicale.

Le lexique des mots acquis tôt est donc en réalité un lexique de base qui continue à se retrouver plus tardivement au cycle 3 et certainement dans tous les écrits rencontrés à l'âge adulte. Les deux périodes d'acquisition considérées peuvent se réduire à deux corpus de taille différente où une petite proportion de mots fréquents est commune aux deux corpus et où une importante quantité de mots rares caractérise le corpus le plus grand (cycle 3). On peut considérer que ce "noyau" lexical de mots fréquents constitue le socle à partir duquel s'effectue l'apprentissage de l'écrit.

4. La complexité orthographique est indépendante de la fréquence

On trouve souvent dans la littérature l'idée selon laquelle la complexité orthographique (régularité ou consistance) est liée à la fréquence du mot : un mot fréquent serait plus irrégulier/inconsistant qu'un mot rare. Or, les données de *Manulex-infra* infirment cette conception. Mesurées sur l'ensemble des mots de *Manulex-infra*, les corrélations entre la fréquence des formes orthographiques et les mesures de consistance sont extrêmement faibles ($r < 0.06$). De façon générale, la corrélation est positive, la consistance s'élevant avec la fréquence du mot. On note cependant une corrélation négative ($r = -0.08$) dans le sens OP (orthographe-phonologie) pour la fin des mots.

La fréquence de la forme orthographique et la complexité orthographique mesurée par la consistance des relations orthographe ↔ phonologie sont donc deux mesures indépendantes qu'il est dès lors très intéressant de manipuler car elles peuvent permettre de caractériser deux types de processus dans le traitement du mot écrit : l'un, indexé sur la fréquence, traitant les mots fréquents par récupération directe du mot en mémoire ; l'autre, indexé par la consistance, traitant les mots rares par transcodage des relations graphème ↔ phonème.

III. Un exemple d'exploitation de corpus pour comprendre l'impact de la réception de l'écrit sur la production orthographique

En conformité avec le cadre de travail de l'apprentissage statistique, nous avons cherché à étudier si les distributions statistiques des unités lexicales et infra-lexicales décrites dans *Manulex* et *Manulex-infra* pouvaient rendre compte du niveau des productions orthographiques des enfants à l'école primaire (Lété, Peereman et Fayol, 2008 ; voir Lété, 2008, pour une synthèse en français). Pour cela, nous avons 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*. La base de données *Éole* déjà présentée (Pothier & Pothier, 2003) nous a permis de disposer de mesures comportementales de productions orthographiques de mots du CP au CM2. Via une méthode d'analyse en régressions multiples, plusieurs prédicteurs de la réussite en orthographe ont été liés à un large ensemble de mots mono et polysyllabiques 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 phonographique⁵ et la complexité orthographique indexée par la consistance de début, milieu et fin de mot. L'étude constitue ainsi une exploration complète de la production orthographique chez des enfants de 6 à 11 ans.

Notre 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, 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 confirmé 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 → graphème. La difficulté liée à la complexité orthographique (*i.e.*, l'effet de consistance) ne disparaît pas en 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. Ces résultats sont en accord avec ceux d'autres études qui observaient un effet de consistance à partir du CE2 jusqu'au

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

CM2. Cependant, contrairement à ces études, du fait qu'un effet de fréquence est observé dès le CP, notre travail confirme qu'un lexique orthographique se constitue très rapidement avec l'exposition à l'écrit.

Pour évaluer la prise en compte d'unités plus larges que le phonème dans la production orthographique, une information de nature plus contextuelle donc, neuf autres prédicteurs de consistance syllabique cette fois-ci ont également été pris en compte. Les résultats ont montré un pourcentage additionnel de variance expliquée de l'ordre de 13% au CE1 et de 17% au cycle 3. Au CP, la consistance syllabique n'ajoutait aucun effet significatif au modèle. Ce résultat suggère que les plus jeunes enfants gèrent leur production orthographique en prenant en compte des unités de taille réduite et échouent à prendre en compte des unités plus larges.

IV. Conclusion

À partir des fréquences textuelles des unités lexicales de *Manulex* et des unités infra-lexicales de *Manulex-infra*, il est intéressant de caractériser le matériau linguistique à partir duquel un enfant apprend à lire de façon à caractériser sa production orthographique. Sur la base de ces estimations, on peut considérer qu'il existe en mémoire lexicale une masse critique de près de 5 000 unités non fléchies et 6 500 unités fléchies. Ces unités ont été fortement rencontrées en réception de l'écrit et leur patron orthographique est très bien connu. Si le système lexical stocke une quantité réduite de possibilités, c'est parce qu'il peut se satisfaire d'une information limitée mais suffisante pour reconstruire les événements linguistiques rarement ou jamais rencontrés. L'information linguistique doit donc être stockée de façon optimale, par exemple sous forme d'instance pour les événements fréquents et sous forme de règles (de reconstruction) pour les événements rares.

L'exploitation des fréquences textuelles de *Manulex* permet de mieux comprendre la production orthographique (en écriture sous dictée) des élèves de primaire. Les enfants de CP semblent se fonder seulement sur des associations phonème → graphème pour orthographier les mots alors que les enfants plus âgés se servent de l'information contextuelle pour traduire la phonologie en orthographe. Ce résultat est compatible avec l'hypothèse que les enfants utilisent de plus en plus l'information contextuelle à mesure que leur expertise de l'écrit se développe. Par exposition répétée aux mots, les enfants apprennent graduellement que certaines graphies sont plus fréquentes que d'autres dans certains contextes. Le fait que les associations phonème → graphème continuent à jouer un rôle important chez les enfants plus âgés est un résultat attendu puisque l'information contextuelle ne permet pas de réduire toutes les ambiguïtés orthographiques. En outre, les connaissances sur les associations graphème ↔ phonème demeurent très influentes parce qu'elles sont dérivées à partir de beaucoup de mots et parce qu'elles sont généralement beaucoup plus fréquentes que des associations basées sur de plus grandes unités. Il est ainsi probable que les enfants utilisent des unités orthographiques de tailles différentes mais la façon dont elles sont gérées lors de la production on-line est toujours une question en suspens.

Éléments bibliographiques

- EHRlich, S., BRAMAUD DU BOUCHERON, G. & FLORIN, A. (1978). *Le développement des connaissances lexicales à l'école primaire*. Paris : PUF.
- FAYOL, M. (2013). *L'acquisition de l'écrit*. Paris : PUF (collection *Que sais-je ?*).
- FAYOL, M., & Jaffré, J.P. (2014). *L'orthographe*. Paris : PUF (collection *Que sais-je ?*)
- LETE, B. (2006). L'apprentissage implicite des régularités statistiques de la langue et l'acquisition des unités morphosyntaxiques. L'exemple des homophones-hétérographes. *Langue Française*, 151, 41-58.
- LETE, B. (2008). La consistance phonographique : une mesure statistique de la complexité orthographique. In C. BRISSAUD, J.-P. JAFFRE & J.-C. PELLAT, *Nouvelles recherches en orthographe* (pp. 85-99). Limoges: Lambert-Lucas.
- LETE, B. (2018). Apprendre le vocabulaire et l'orthographe dans les manuels de lecture. In L. Ferrand, B. LETE, & C. THEVENOT (Eds.), *Psychologie cognitive des apprentissages scolaires*. Paris: Dunod.
- LÉTÉ, B., PEEREMAN, R. & FAYOL, M. (2008). 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, 156-166.
- PACTON, S., FAYOL, M. & LÉTÉ, B. (2008). L'intégration des connaissances lexicales et infralexicales dans l'apprentissage du lexique orthographique. *ANAE*, 96-97, 47-52.
- PACTON, S., FAYOL, M. & PERRUCHET, P. (2002). The acquisition of untaught orthographic regularities in French. In L. VERHOEVEN, C. ERLBRO & P. REITSMA (Eds.), *Precursors of functional literacy* (pp.121-137). Dordrecht: Kluwer.
- 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. *Behavior Research Methods*, 39, 593-603.
- PERRUCHET, P. & PACTON, S. (2004). Qu'apportent à la pédagogie les travaux de laboratoire sur l'apprentissage implicite ? *L'Année Psychologique*, 104(1), 121-146.
- PERRUCHET, P. & PEEREMAN, R. (2005). Apprendre sa langue maternelle, une question de statistique ? *Pour La Science*, 327, 82-87.
- POTHIER, B. & POTHIER, P. (2003). *EOLE: Échelle d'acquisition en orthographe lexicale du CP au CM2*. Paris : Retz.