CORPUS

Corpus

19 | 2019
Corpus et pathologies du langage

ÉQOL : Une nouvelle base de données québécoise du lexique scolaire du primaire comportant une échelle d'acquisition de l'orthographe lexicale

ÉQOL : A new academic database of the Quebec primary school lexicon with an acquisition scale for lexical orthography

Brigitte Stanké, Marine Le Mené, Stefano Rezzonico, André Moreau, Christian Dumais, Julie Robidoux, Camille Dault et Phaedra Royle

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/3818
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Référence électronique
Brigitte Stanké, Marine Le Mené, Stefano Rezzonico, André Moreau, Christian Dumais, Julie Robidoux, Camille Dault et Phaedra Royle, « ÉQOL : Une nouvelle base de données québécoise du lexique scolaire du primaire comportant une échelle d'acquisition de l'orthographe lexicale », *Corpus* [En ligne], 19 | 2019, mis en ligne le 01 janvier 2019, consulté le 06 septembre 2019. URL : http://journals.openedition.org/corpus/3818

ÉQOL : Une nouvelle base de données québécoise du lexique scolaire du primaire comportant une échelle d'acquisition de l'orthographe lexicale

ÉQOL : A new academic database of the Quebec primary school lexicon with an acquisition scale for lexical orthography

Brigitte Stanké, Marine Le Mené, Stefano Rezzonico, André Moreau, Christian Dumais, Julie Robidoux, Camille Dault et Phaedra Royle

Introduction

1 L'élaboration de l'outil dont il est question dans cet article (la base de données ÉQOL) s'inscrit dans un projet québécois de plus grande envergure sur l'apprentissage de l'orthographe lexicale. Ce projet, financé par le FRQSC (Fonds de recherche *du* Québec - *Société et Culture*) et l'Institut universitaire en déficience intellectuelle et en trouble du spectre de l'autisme, s'est donné pour objectif principal de mieux cibler les différents facteurs en jeu dans l'apprentissage de l'orthographe des mots auxquels sont exposés les élèves à l'école primaire. Afin de répondre à cet objectif, il a d'abord été nécessaire de recenser le lexique scolaire qui constitue l'environnement écrit d'un élève de primaire scolarisé en français au Québec, et d'établir une échelle d'acquisition de ce lexique. Ces deux préalables sont décrits et discutés au sein du présent article.

1.1. Tour d'horizon des travaux sur l'apprentissage de l'orthographe lexicale

2 Les recherches sur l'apprentissage de l'orthographe sont relativement récentes comparées à celles consacrées à l'apprentissage de la lecture et sont également beaucoup moins nombreuses (Fayol & Jaffré 2016). Celles-ci ne sont d'ailleurs pas mentionnées au sein du rapport du National Reading Panel (2000). Ces études s'accordent sur le fait que l'apprentissage de l'orthographe s'avère être plus complexe que celui de la lecture et qu'il dépend de plusieurs facteurs, dont des facteurs environnementaux, cognitifs et linguistiques (Caravolas, Hulme & Snowling 2001).

3 Les recherches en psychologie, en neurosciences et en éducation ont étudié l'impact des facteurs environnementaux sur le développement et sur les difficultés d'apprentissage du langage écrit (Billard *et al.* 2008 ; Fluss *et al.* 2008 ; Noble & McCandliss 2005). Ces recherches montrent que la qualité de l'environnement familial et les activités précoces de lecture partagée et d'orthographes approchées contribuent à développer les capacités préalables au langage écrit (Ecalle & Magnan 2015 ; Morin & Montésinos-Gelet 2007).

4 Un autre courant d'études s'est focalisé sur la compréhension des processus cognitifs qui sous-tendent l'apprentissage du langage écrit et ses troubles. Un nombre important de recherches atteste du rôle important de la conscience phonologique sur les compétences en lecture, et plus encore sur l'orthographe (Gillon 2017, pour une synthèse), ainsi que de la conscience morphologique (Casalis & Colé 2018). Plus récemment, le rôle des capacités visuoattentionnelles a été mis en évidence, notamment dans l'apprentissage de l'orthographe conventionnelle des mots (orthographe lexicale) (Bosse 2005).

5 Les linguistes se sont quant à eux intéressés aux liens entre les caractéristiques linguistiques des différents systèmes d'écriture et l'acquisition de l'orthographe. Le système d'écriture dans laquelle se réalise l'apprentissage joue un rôle déterminant sur son acquisition. Plus particulièrement, la rapidité avec laquelle les élèves apprennent à lire et à orthographier dépend des caractéristiques linguistiques du système d'écriture de leur langue. Dans des langues dites transparentes ou consistantes, comme l'espagnol, les élèves apprennent plus rapidement à lire et à orthographier que dans des langues plus opaques (inconsistantes) comme le français et l'anglais (Caravolas 2004 ; Caravolas, Lervåg, Defior, Seidlová Málková & Hulme 2013 ; Seymour, Aro & Erskine 2003). Le degré de transparence d'un système d'écriture est souvent défini à partir du calcul de la consistance orthographique, qui renvoie à la stabilité des correspondances existant entre le code orthographique et le code phonologique (Fayol, Bonin & Collay 2008). La consistance orthographique n'est toutefois pas la seule variable linguistique ayant fait l'objet de recherches. Le voisinage orthographique et la fréquence lexicale jouent également un rôle déterminant, et ce, dès le début de l'apprentissage de la lecture et de l'orthographe (Lété, Peerman & Fayol 2008 ; Pacton, Fayol & Lété 2008). L'évaluation des facteurs linguistiques repose essentiellement sur les bases lexicales, établies à partir de corpus de mots, propres à chaque langue. Ces bases fournissent des estimations quantitatives sur la distribution des variables lexicales et infralexicales (unités phonologiques composant les mots : syllabes et phonèmes). Le choix des corpus de référence est donc déterminant pour établir ces estimations. Leur élaboration et leurs limites font l'objet des deux prochains points.

1.2. Le recours aux bases de données pour l'apprentissage de l'orthographe

6 Ces dernières années, plusieurs bases de données en français, conçues dans le cadre de travaux sur l'apprentissage de l'orthographe lexicale, ont été rendues accessibles aux chercheurs, enseignants ou orthophonistes œuvrant dans le domaine de l'apprentissage de l'écrit. Dans ce sens, il est important de citer Manulex-Infra (Lété, Sprenger-Charolles & Colé 2004), Manulex-Morpho (Peereman, Sprenger-Charolles & Messaoud-Galusi 2013), Lexique 3.80 (New 2006), Silex (Gingras & Sénéchal 2017), Omnilex (Desrochers 2006) ou encore Novlex (Lambert & Chesnet 2001). Bien que plusieurs de ces bases de données aient été élaborées dans le but commun d'évaluer le rôle de certains facteurs (lexicaux, sous-lexicaux ou cognitifs) sur l'apprentissage de l'orthographe lexicale, toutes n'ont pas été constituées en prenant en compte les mêmes mots ni les mêmes variables. Toutefois, le plus souvent, elles permettent de décrire leurs entrées lexicales sur la base de valeurs fréquentielles, grammaticales (catégorie, genre) ou phonologiques (transcription phonologique, nombre de syllabes, découpage syllabique, etc.).

7 Pour les construire, les auteurs de ces bases de données ont eu recours à des ressources textuelles de différentes natures. Certains d'entre eux ont par exemple sélectionné leurs entrées lexicales dans des dictionnaires. C'est le cas de la base Omnilex qui contient 102 000 entrées extraites entre autres du *Petit Robert*, du *Petit Larousse* et du *Trésor de la Langue Française*. D'autres, comme la base de données Lexique 3.80 ont pour corpus source des textes littéraires incluant ceux de la francophonie au sens large. Notons par ailleurs que cette même base de données s'appuie également sur un corpus de sous-titres de films. Parmi les bases de données précédemment citées, quatre d'entre elles se focalisent spécifiquement sur le lexique adressé aux élèves du primaire et sont donc issues essentiellement d'ouvrages de littérature jeunesse : Manulex-infra, Manulex-Morpho, Novlex et Silex.

1.3. Limites des bases de données existantes et création d'un outil dédié à l'apprentissage du lexique scolaire

8 Les ressources que nous venons de lister constituent un apport remarquable pour la recherche et l'enseignement, mais plusieurs raisons nous ont cependant fait privilégier l'élaboration d'un outil spécifiquement dédié à l'objet de notre recherche.

9 D'abord, les données écrites sur lesquelles reposent ces bases de données ne sont pas toujours représentatives du lexique adressé aux enfants en contexte scolaire. Pourtant, il semble justifié que les productions orthographiques des élèves soient évaluées à partir de mots auxquels ils sont davantage exposés et dont ils devront faire usage à l'école.

10 En outre, même lorsque ces ressources se basent sur un corpus spécifiquement adressé à des élèves de primaire, elles ne rendent pas compte des caractéristiques du lexique scolaire (c'est-à-dire spécifique à chacune des disciplines enseignées à l'école) rencontré par les élèves au cours de leur cursus. Or, de nombreuses recherches ont montré qu'une maîtrise insuffisante de ce lexique, qui diffère du lexique littéraire en ce qu'il serait selon

certains auteurs (*i.a.* Nagy & Townsend 2012) plus abstrait, plus soutenu et morphologiquement plus complexe, engendrerait des difficultés scolaires et, notamment, des difficultés de compréhension en lecture (Hirsch 2003 ; Perfetti & Stafura 2014 ; Schmitt, Jiang & Grabe 2011 ; Townsend, Filippini, Collins & Biancarosa 2012). Cette corrélation serait d'autant plus marquée pour les élèves allophones et ceux provenant de milieux socioéconomiques défavorisés (Lesaux, Kieffer, Kelley & Harris 2010).

11 Le lexique scolaire étant fortement dépendant des concepts enseignés dans chaque matière et à chacun des niveaux scolaires, les bases de données issues d'œuvres littéraires ne pourront pas en estimer de façon suffisamment pertinente la distribution et la fréquence. À titre d'exemple, les bases de données existantes attribueront une fréquence peu élevée aux items lexicaux liés aux mathématiques ou aux sciences, très peu représentés dans la littérature jeunesse, alors que ceux-ci sont en réalité très fréquemment rencontrés par les élèves en contexte scolaire.

12 Enfin, les bases de données existantes ne permettent pas de déterminer la façon dont se développe l'acquisition de l'orthographe lexicale pour, entre autres, guider son enseignement et remédier aux difficultés rencontrées par les apprenants. Pour ce faire, il est indispensable de disposer d'une échelle d'acquisition. À notre connaissance, seules deux normes objectives existent dans la littérature francophone : l'échelle Dubois-Buyse (Ters, Mayer & Reichenbach 1999) et l'échelle d'acquisition de l'orthographe lexicale (ÉOLE) (Pothier & Pothier 2003). L'apport de cette dernière est considérable puisqu'elle fournit le pourcentage de réussite orthographique de près de 12 000 mots relevés dans la presse nationale française et dictés à 48 900 élèves du primaire (entre le CP et le CM2). Toutefois, la question de la représentativité du lexique utilisé se pose à nouveau dans la mesure où cette échelle est construite sur la base d'articles de presse destinés aux adolescents et aux adultes, et non sur la base de matériel écrit scolaire dont devront faire usage les élèves du primaire.

13 À l'aune de ces constats, nous avons donc décidé d'élaborer une ressource lexicale *hybride*, rendant compte à la fois de la fréquence et des caractéristiques linguistiques du lexique scolaire (catégorie grammaticale, genre, nombre, nombre de lettres, nombre de lettres géminées, nombre de phonèmes, nombre de syllabes à l'oral et à l'écrit, consistance moyenne, fréquence cumulée, fréquence totale, taux de réussite orthographique) présent dans les manuels scolaires du primaire et, pour une partie des mots de ce lexique, de la capacité des élèves à les orthographier à chaque niveau scolaire.

2.1. Méthodologie de création de la base de données lexicale

14 Afin de recenser le plus fidèlement possible le lexique qui constitue l'environnement écrit d'un élève de primaire, nous avons collecté tous les mots figurant dans un ensemble de 12 manuels de mathématiques, 12 manuels de français et 4 manuels d'univers social[1] provenant de la maison d'édition *ERPI*. À ce corpus, nous avons également ajouté le contenu lexical de 24 livres de littérature pour les élèves de CP et de 90 courts extraits d'ouvrages de littérature jeunesse. Au total, plus de 14 800 mots (formes fléchies) différents ont été répertoriés. À chacune de ces entrées, nous avons associé des informations extraites des bases de données Lexique 3.80 et Manulex-infra. Parmi les renseignements fournis par ÉQOL figurent des indices grammaticaux (genre, catégorie

grammaticale), de longueur du mot (nombre de lettres, nombre de graphèmes ou nombre de syllabes à l'oral et à l'écrit) et de consistance (consistance moyenne entre phonèmes et graphèmes). De nouvelles mesures de fréquence ont par ailleurs été effectuées. Nous avons par exemple calculé la fréquence (pour mille) de l'item lexical sur le total des mots du corpus toutes années scolaires confondues, mais aussi année par année. Nous avons également calculé la fréquence cumulée de chacun des mots, c'est-à-dire la fréquence du mot dans le corpus de l'année scolaire ciblée ajoutée à la fréquence du mot dans les corpus des années précédentes.

2.2. Méthodologie de l'élaboration de la base de données menant à la construction de l'échelle d'acquisition de l'orthographe lexicale

15 Par la suite, 1 200 mots parmi les plus fréquents de ce répertoire ont été utilisés pour constituer la liste de mots à orthographier afin de créer l'échelle d'acquisition de l'orthographe lexicale. Soixante-six dictées ont été créées comportant chacune 20 phrases. Pour chacune de ces phrases, un seul mot cible a été dicté. Les phrases de chacune des dictées ont été rédigées en prenant soin d'intégrer les mots cibles au sein d'un environnement lexical auquel les élèves sont par ailleurs exposés dans les manuels scolaires.

16 Ont pris part à cette étude 4 733 élèves provenant de 46 écoles publiques et d'une école privée situées sur le territoire québécois. Afin de s'assurer que les élèves de l'étude ne présentent pas de difficultés d'apprentissage du langage écrit, une évaluation de leurs compétences en orthographe a été réalisée. Deux dictées tests ont ainsi été élaborées : une dictée adressée aux élèves de 1re et 2e années (CP et CE1) et une autre destinée aux élèves de 3e à la 6e année (CE2 à la 6e). La première dictée comportait 20 mots à orthographier : 5 non-mots, 6 mots consistants[2] et 9 mots inconsistants. La seconde comptait 40 mots : 8 non-mots, 8 mots irréguliers, 5 mots consistants et 19 mots inconsistants (voir l'Annexe 1 pour un exemple). Pour chacune des dictées-tests, l'orthographe phonologique[3] des mots et des non-mots ainsi que l'orthographe des mots ont été cotées. Les élèves ayant obtenu un score inférieur au 10e percentile à l'orthographe phonologique ou lexicale[4] ont été exclus de l'échantillon. La suite de l'étude repose donc sur les productions de 4 030 enfants (2 144 filles et 1 879 garçons ; 7 enfants n'ont pas indiqué leur sexe sur les fiches de renseignements).

17 L'évaluation des compétences orthographiques a été effectuée en trois temps de collecte distincts (mai et septembre 2015 et mai 2016). Sur une période de trois semaines consécutives, les élèves ont eu à réaliser une série de 10 dictées (dont la dictée-test) de 20 mots (voir l'Annexe 2 pour un exemple réalisé par un élève). 62 dictées de 20 mots ont donc été élaborées afin de couvrir les 1 200 mots visés. Chaque mot a été orthographié par un minimum de 50 élèves par niveau scolaire. Ainsi, des dictées comportant 20 mots cibles ont été proposées aux enfants. Certains enfants ont orthographié correctement les mots cibles, d'autres ont fait des erreurs. L'échelle d'acquisition indique le pourcentage de réussite orthographique pour chacun des mots cibles selon le niveau scolaire. Ce sont les enseignants des élèves qui étaient chargés de faire passer ces dictées en classe entière. Des consignes précises sur le déroulement de l'expérimentation avaient au préalable été remises aux enseignants afin de s'assurer du caractère homogène de l'expérimentation.

Bien que toutes les dictées aient été conçues comme une série de phrases complètes, les élèves ne devaient orthographier que le mot signalé par l'enseignant.

18 Les dictées ont permis d'obtenir les productions orthographiques de chaque mot selon le niveau scolaire des élèves et de calculer leur taux de réussite orthographique. Ces informations ont ensuite été ajoutées aux caractéristiques des 1 200 mots de la base lexicale ayant été donnés à orthographier aux élèves.

3.1. Évaluation du degré de spécificité de la base de données ÉQOL

19 Afin d'attester de la pertinence de l'élaboration de la base de données ÉQOL en ce qui concerne les écrits scolaires auxquels les élèves du primaire québécois sont exposés, nous avons procédé à une comparaison des items lexicaux de notre base de données à ceux de trois bases de données existantes (Lexique 3.80, Manulex-infra et Novlex). En premier lieu, nous avons cherché à observer la part d'entrées lexicales communes à ÉQOL et aux trois bases de données que nous venons de citer. Ces comparaisons sont représentées dans le tableau 1 ci-dessous.

Tableau 1. Degré de recouvrement lexical entre ÉQOL et Lexique 3.80, Novlex et Manulex

Corpus	**Nombre de mots communs : valeurs absolues (pourcentage)**	**Nombre de mots uniques : valeurs absolues (pourcentage)**
Lexique 3.80	13 085 (88,37 %)	1 722 (11,63 %)
Novlex	4 566 (30,84 %)	10 241 (69,16 %)
Manulex	11 246 (75,95 %)	3 561 (24,05 %)

20 Ces premiers résultats permettent de constater que les trois bases de données considérées pour les comparaisons ne sont pas équivalentes quant au degré de correspondance entre les mots qu'elles contiennent et ceux de la base ÉQOL. On observe que 88,37 % et 75,95 % des mots de ÉQOL figurent respectivement dans les bases de données Lexique 3.80 et Manulex contre seulement 30,84 % pour la base de données Novlex. Le faible taux de recouvrement de la base de données Novlex pourrait s'expliquer par la taille de son corpus source, contenant nettement moins de mots que les deux autres bases de données (voir tableau 2). En effectuant un bref survol qualitatif des mots qui distinguent ÉQOL des autres, nous avons pu remarquer qu'un grand nombre d'entre eux sont des mots spécifiques à l'enseignement des mathématiques et de la discipline univers social. À titre d'exemple, plusieurs des mots non représentés dans les autres bases de données renvoient à l'histoire et aux spécificités de la géographie québécoise ou, plus largement, canadienne. Plusieurs noms propres (prénoms, villes, rivières, provinces, etc.) figurent aussi dans notre liste. Ce premier état des lieux gagnera à être approfondi par des mesures quantitatives.

21 Dans un second temps, nous avons cherché à évaluer le degré de corrélation entre les fréquences des mots communs présents dans ÉQOL et dans les trois bases de données déjà

citées. Pour ce faire, les fréquences ont toutes été rapportées à la même valeur relative (fréquence par million). Les résultats sont présentés dans le tableau 2 ci-dessous.

Tableau 2. Indices de corrélation entre ÉQOL et Lexique 3.80, Novlex et Manulex

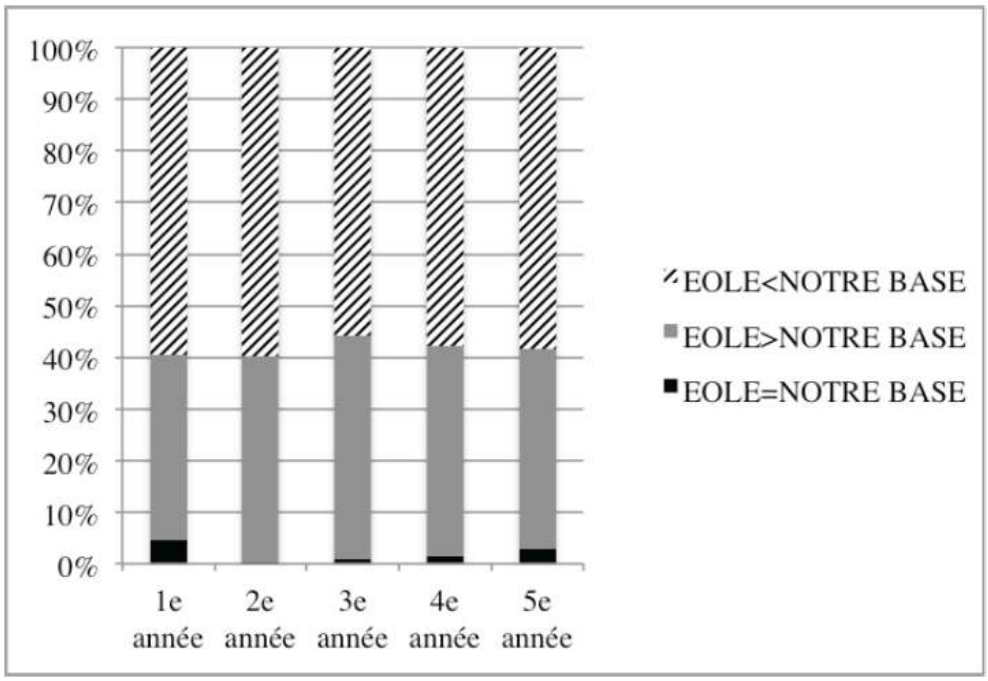
Corpus	Nombre de mots communs	Corrélation	Significativité
Lexique 3.80	13 085	0,94	< 0,001
Novlex	4 566	0,86	< 0,001
Manulex	11 246	0,97	< 0,001

22 Les résultats concernant les comparaisons des fréquences indiquent clairement que, quelle que soit la base de données considérée, les fréquences de la base ÉQOL corrèlent fortement avec celles des autres bases de données (cf. Tableau 2). Il nous semble intéressant de relever que la paire ÉQOL-Manulex est celle dont la corrélation est la plus importante. Ceci pourrait s'expliquer par le fait que Manulex utilise, en plus des livres de littérature jeunesse, un corpus de manuels scolaires, mais uniquement de manuels de français, langue d'enseignement. Cette base de données dépeindrait donc une réalité d'exposition orthographique un peu plus proche de celle de notre corpus, tant au niveau du contenu que de la fréquence. Tout comme Manulex, la base de données Novlex est également basée sur un corpus de livres de littérature jeunesse et de manuels de français, mais cette base de données ne prend en compte que des manuels de CE2 (soit l'équivalent de la 3e année dans le système scolaire québécois), ce qui pourrait expliquer la plus faible corrélation observée entre les corpus Novlex et ÉQOL.

23 La série de tests *t* à mesures répétées que nous avons effectuée montre des différences significatives sur le plan des fréquences rapportées (ÉQOL *vs* Lexique : $t(13084) = 5.28$, $p<.001$; ÉQOL *vs* Novlex : $t(4565) = 5.65$, $p<.001$; ÉQOL *vs* Manulex $t(11245) = 6.75$, $p<.001$). Ces résultats indiquent que même si les patrons fréquentiels des bases de données sont fortement comparables, les fréquences pour chacun des items sont suffisamment différentes d'une base de données à l'autre pour justifier l'élaboration d'une base spécifique au lexique scolaire auquel sont exposés les élèves québécois.

24 En dernier lieu, nous avons souhaité comparer notre échelle d'acquisition de l'orthographe lexicale à l'échelle ÉOLE de Pothier et Pothier (2003). La comparaison a été effectuée pour tous les niveaux scolaires séparément, de la 1re à la 5e année du primaire pour notre échelle québécoise, et du CP au CM2 pour ÉOLE. Toutefois, ne pouvant commenter ici les comparaisons pour chaque année scolaire, nous ne présenterons que les résultats toutes années confondues. La figure 1 illustre la comparaison des résultats de dictées faites sur les mots homologues, c'est-à-dire des mots qui apparaissent dans les deux échelles d'acquisition, à travers les années scolaires. En hachuré, ce sont les items mieux réussis dans les dictées qui ont permis de constituer notre corpus (par exemple, dans notre échelle, le taux de réussite du mot *ami* est de 83 % au CP alors qu'il est de 18 % dans ÉOLE), en gris les items mieux réussis lors des dictées qui ont permis de constituer ÉOLE (par exemple, le taux de réussite du mot *avant* est de 50 % dans ÉOLE au CP alors qu'il est de 11 % dans notre échelle) et, finalement, en noir les items réussis de manière égale dans les deux dictées (par exemple, le mot *date*).

Figure 1. Différences (<, >) et similarités (=) de résultats de dictées faites sur les mots homologues dans les échelles d'acquisition ÉOLE et ÉQOL selon les années scolaires



25 Cette figure illustre que la variation est stable entre les différentes années avec une tendance favorable pour ÉQOL (environ 60 % des mots étudiés). Si le plus souvent l'orthographe des mots tend à être acquise plus précocement chez les élèves québécois que chez les élèves français, il faut toutefois noter que pour 40 % des mots, les résultats sont plus élevés dans ÉOLE. Par ailleurs, une analyse des écarts entre les productions dans les deux bases indique que si pour une portion importante des mots comparés les écarts restent à ±15 %, les différences peuvent dépasser des écarts de 70 %.

26 Ces comparaisons témoignent de la pertinence de la création de cette nouvelle échelle. Ces résultats doivent être interprétés à la lumière de plusieurs facteurs, dont ceux méthodologiques et environnementaux. Les performances plus élevées des élèves québécois pourraient s'expliquer, entre autres, par le fait que les participants de notre recherche, qui montraient des difficultés orthographiques, ont été exclus de l'étude, alors que l'étude de Pothier et Potier (2003) ne comportait pas ce facteur d'exclusion. D'autre part, ÉOLE (Pothier & Potier 2003) a été élaborée depuis déjà plus d'une décennie et les approches didactiques ont pu changer.

27 Bien que des différences aient été observées dans l'acquisition de l'orthographe lexicale entre les deux échelles, les deux études montrent qu'il faut environ cinq ans afin que les élèves soient en mesure d'orthographier les mots selon la norme, car plus de 75 % de l'ensemble des mots des élèves de CM2 sont orthographiés adéquatement par les élèves. Ces échelles corroborent les études interlangues qui indiquent que, dans des systèmes d'écriture opaques, comme le français, l'apprentissage de l'orthographe lexicale est beaucoup plus lent que dans des systèmes d'écriture plus transparents (Caravolas 2004 ; Landerl & Wimmer 2008 ; Sprenger-Charolles & Béchennec 2004).

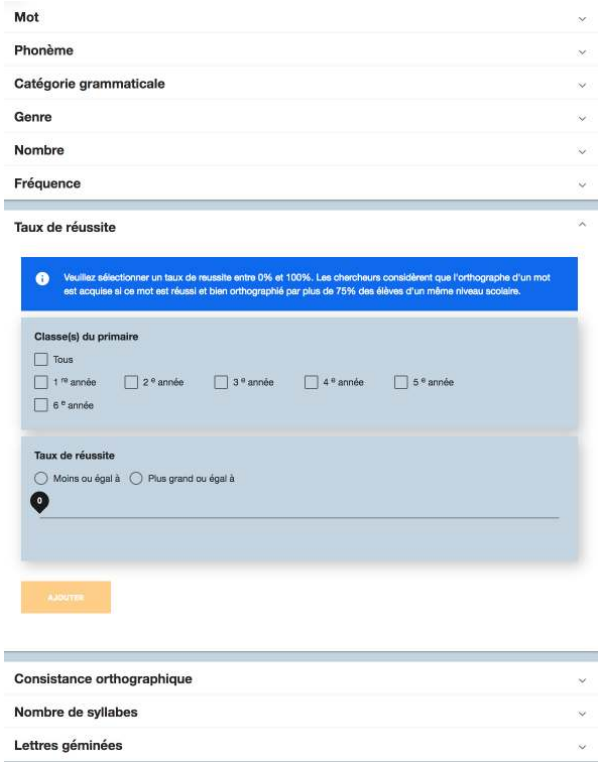
3.2. Applications de la base de données ÉQOL

28 La base de données lexicale ÉQOL est désormais disponible en ligne (à l'adresse suivante : https://appligogiques.com/eqol) et sera régulièrement enrichie. Deux versions de la base sont utilisables : une version brute sous forme de grille Excel (voir Figure 2) et une version conviviale (voir Figure 3) disposant d'une interface d'interrogation en ligne. Toutefois, la version interrogeable ne comporte pas encore toutes les caractéristiques sous-lexicales des mots. En la rendant accessible à tous, nous souhaitons que cette ressource puisse servir à des fins de recherche, mais aussi d'apprentissage et d'enseignement.

Figure 2. Exemple de données tirées de ÉQOL illustrant un sous-ensemble de l'information disponible pour les lexèmes de la base de données. Dans la dernière colonne, on peut noter le taux moyen de réussite de l'orthographe de chaque item pour la première année scolaire

	A	C	D	E	G	H	I	J	K	S	X	Y
1	mot	catégorie_grammaticale	genre	nombre	Nb_lettres_géminées	nb_phonèmes	nb_syllabes_oral	nb_syllabes_écrit	consistance_PG_token_moyenne	fréquences_cumulées_1-2_moy_	total_fréquence_forme_million	taux_réussite_a1
3	à	pre			0	1	1	1	31,74	15,80	18381,86	0,22
34	abord	nom	m	s	0	4	2	2	72,95	0,23	217,68	0,02
03	accord	nom	m	s	1	4	2	2	53,55	0,02	71,52	0,02
52	acheter	ver			0	5	3	3	72,72	0,08	164,82	0,63
75	acte	nom	m	s	0	3	1	2	79,59	0,00	62,20	0,77
80	action	nom	f	s	0	5	2	2	63,63	0,11	146,16	0,09
92	activités	nom	f	p	0	8	4	4	80,59	0,54	348,29	0,38
05	addition	nom	f	s	1	6	3	3	59,48	0,28	108,84	0,00
07	additionne	ver			2	7	3	3	50,97	0,00	93,29	0,02
17	adjectif	adj,nom	m	s	0	8	3	3	74,53	0,88	354,51	0,10
49	adore	ver			0	4	2	3	78,50	0,42	105,73	0,52
85	affiche	nom,ver	f	s	1	4	2	3	64,48	0,11	99,51	0,09
09	afin	adv			0	3	2	2	54,17	0,00	410,49	0,69
15	âge	nom	m	s	0	2	1	3	43,55	0,05	136,83	0,18
37	agit	ver			0	3	2	2	67,36	0,34	317,20	0,38
54	agricole	adj		s	0	7	3	4	75,71	0,00	65,31	0,41
57	agriculteurs	nom	m	p	0	10	4	4	81,05	0,00	87,07	0,27
59	agriculture	nom	f	s	0	10	4	5	80,68	0,00	267,44	0,21
65	aide	nom,ver		s	0	2	1	2	45,61	1,26	1374,52	0,42
88	ailes	nom	f	p	0	2	1	2	43,45	0,02	102,62	0,25
90	ailleurs	adv			1	4	2	2	61,08	0,02	155,49	0,00
98	aime	ver			0	2	1	2	39,01	2,09	472,69	0,62
13	ainsi	adv			0	3	2	2	47,61	0,05	398,05	0,12
14	air	nom	m	s	0	2	1	1	53,63	0,06	251,89	0,37
15	aire	nom	f	s	0	2	1	2	44,18	0,00	174,15	0,14
24	ajoute	nom,ver	f	s	0	4	2	2	61,68	0,59	491,34	0,66

Figure 3. Exemple de l'interface Web de ÉQOL illustrant le module de requête pour l'information sur le taux moyen de réussite à l'orthographe de chaque item pour la première année scolaire



29 Bien que ÉQOL ait tout d'abord été conçue pour répondre aux besoins spécifiques de notre recherche sur l'apprentissage de l'orthographe lexicale, nous espérons que l'originalité de son contenu, un corpus issu de manuels de différentes disciplines scolaires et de littérature jeunesse, incitera d'autres chercheurs, qui travaillent sur l'écrit, à y avoir recours dans le cadre de leurs études.

30 Nous espérons par ailleurs que celle-ci pourra être utilisée par les enseignants, orthopédagogues (enseignants spécialisés) et orthophonistes afin, notamment, de faciliter l'enseignement explicite du lexique scolaire aux élèves, tant à l'oral qu'à l'écrit. Plusieurs recherches ont montré que grâce à un enseignement explicite de ce lexique, les connaissances lexicales des élèves tendent à s'améliorer, ceux-ci développant de manière concomitante une meilleure compréhension des textes disciplinaires (Cormier, Pruneau & Rivard 2004 ; Elleman, Lindo, Morphy & Compton 2009 ; Kelley, Lesaux, Kieffer & Faller 2010 ; Lesaux *et al.* 2010 ; Marin 2007 ; Taboada & Rutherford 2011 ; Weinberg, Boukacem & Burger 2012).

31 ÉQOL permettra en outre la création de matériel didactique ou de rééducation prenant appui sur le lexique scolaire que les élèves doivent apprendre à orthographier dans chacune des disciplines scolaires et à chaque niveau du primaire.

Conclusion

32 Les premiers résultats de notre travail tendent à confirmer d'une part la nécessité de créer une nouvelle base de données propre au lexique scolaire auquel sont exposés les élèves de la 1re à la 6e année du primaire et, d'autre part, la pertinence de l'élaboration

d'une nouvelle échelle d'acquisition de l'orthographe lexicale spécifique aux compétences orthographiques des élèves québécois.

33 ÉQOL devra toutefois être enrichie pour permettre une analyse plus complète des facteurs susceptibles d'influencer l'acquisition de l'orthographe lexicale. Celle-ci sera complétée prochainement par des informations concernant les familles lemmatiques et les voisins orthographiques.

34 En ce qui concerne l'échelle d'acquisition, rappelons que celle-ci ne porte que sur un corpus de 1 200 mots. Bien que l'investissement et le financement nécessaires pour l'enrichir soient considérables, il serait tout de même souhaitable d'augmenter le volume des mots pris en compte dans cette échelle. Par ailleurs, puisque des différences importantes ont été observées entre le niveau de compétence orthographique des élèves québécois et celui des élèves français, d'autres recherches devront être menées pour expliquer quels peuvent être les facteurs à l'origine de ces écarts.

35 Enfin, il nous semble que les apports potentiels pour les enseignants, les orthopédagogues ou les orthophonistes sont nombreux, mais pour le confirmer, il sera nécessaire que des groupes de discussion de professionnels soient constitués afin de vérifier le fonctionnement de la base de données, aussi bien sur le plan de sa présentation (son format) qu'en ce qui concerne ses modalités d'interrogation et d'application.

BIBLIOGRAPHIE

Bara F., Gentaz É. et Colé P. (2008). « Littératie précoce et apprentissage de la lecture : comparaison entre des enfants à risque, scolarisés en France dans des réseaux d'éducation prioritaire, et des enfants de classes régulières », *Revue des sciences de l'éducation* 34(1) : 27-45.

Billard C., Fluss J., Ducot B., Warszawski J., Ecalle J., Magnan A., Richard G. et Ziegler J. (2008). « Étude des facteurs liés aux difficultés d'apprentissage de la lecture. À partir d'un échantillon de 1 062 enfants de seconde année d'école élémentaire », *Archives de pédiatrie* 15(6) : 1058-1067.

Bosse M. L. (2005). « De la relation entre acquisition de l'orthographe lexicale et traitement visuo-attentionnel chez l'enfant », *Rééducation orthophonique* 222 : 9-30.

Caravolas M. (2004). « Spelling development in alphabetic writing systems : A cross-linguistic perspective », *European Psychologist* 9(1) : 3-14.

Caravolas M., Hulme C. et Snowling M. J. (2001). « The foundations of spelling ability : Evidence from a 3-year longitudinal study », *Journal of memory and language* 45(4) : 751-774.

Caravolas M., Lervåg A., Defior S., Seidlová Málková G. et Hulme C. (2013). « Different patterns, but equivalent predictors, of growth in reading in consistent and inconsistent orthographies », *Psychological science* 24(8) : 1398-1407.

Casalis S. et Colé P. (2018). « Le morphème, une unité de traitement dans l'acquisition de la litéracie », *Langue française* 199(3) : 69-81.

Cormier M., Pruneau D. et Rivard L. (2004). « S'approprier un vocabulaire scientifique en milieu minoritaire », *Cahiers francocanadiens de l'ouest* 16 (1 et 2) : 175-197.

Desrochers A. (2006). « OMNILEX : Une base de données sur le lexique du français contemporain », *Cahiers Linguistiques d'Ottawa* 34 : 25-34.

Ecalle J. et Magnan A. (2015). *L'apprentissage de la lecture et ses difficultés* – 2e éd. Paris : Dunod.

Ecalle J. et Magnan A. (2002). *L'apprentissage de la lecture : Fonctionnement et développement cognitifs*. Paris : Armand Colin.

Elleman A. M., Lindo E. J., Morphy P. et Compton D. L. (2009). « The impact of vocabulary instruction on passage-level comprehension of school-age children : A meta-analysis », *Journal of Research on Educational Effectiveness* 2(1) : 1-44.

Fayol M., Bonin P. et Collay S. (2008). « La consistance orthographique en production verbale écrite : une brève synthèse », *L'année psychologique* 108(3) : 517-546.

Fayol M. et Jaffré J.-P. (2016). « L'orthographe : des systèmes aux usages », *Pratiques - linguistique, littérature, didactique* 169-170 : 1-15.

Fluss J., Ziegler J., Ecalle J., Magnan A., Warszawski J., Ducot B., Richard G. et Billard C. (2008). « Prévalence des troubles d'apprentissages du langage écrit en début de scolarité : l'impact du milieu socioéconomique dans 3 zones d'éducations distinctes », *Archives de pédiatrie* 15(6) : 1049-1057.

Gillon G. T. (2017). *Phonological Awareness. From Research to Pratice,* 2nd edition. New York, NY : Gilford Press.

Gingras M. et Sénéchal M. (2017). « Silex : A database for silent-letter endings in French words », Behavior Research Methods 49(5) : 1894-1904.

Hirsch E. D. (2003). « Reading comprehension requires knowledge of words and the world », *American Educator* 27(1) : 10-13.

Kelley J. G., Lesaux N. K., Kieffer M. J. et Faller S. E. (2010). « Effective academic vocabulary instruction in the urban middle school », *The Reading Teacher* 64(1) : 5-14.

Lambert E. et Chesnet D. (2001). « Novlex : une base de données lexicale pour les élèves de primaire », *L'Année Psychologique* 101(2) : 277-288.

Landerl K. et Wimmer H. (2008). « Development of word reading fluency and spelling in a consistent orthography : An 8-year follow-up », *Journal of educational psychology* 100(1) : 150-161.

Lété B., Peerman R. et Fayol M. (2008). « Consistency and word-frequency effects on spelling among first- to firth-grade French children : A regression-based study », *Journal of Memory and Language* 58(4) : 952-977.

Lété B., Sprenger-Charolles L. et Colé P. (2004). « MANULEX : A grade-level lexical database from French elementary school readers », *Behavior Research Methods, Instruments & Computers* 36(1) : 156-166.

Lesaux N. K., Kieffer M. J., Kelley J. G. et Harris J. R. (2014). « Effects of academic vocabulary instruction for linguistically diverse adolescents : Evidence from a randomized field trial », *American Educational Research Journal* 51(6) : 1159-1194.

Marin B. (2007). « Ressources lexicales et savoirs scolaires », *Dilbilim* 1(1) : 13-26.

Morin M. F. et Montésinos-Gelet I. (2007). « Effet d'un programme d'orthographes approchées en maternelle sur les performances ultérieures en lecture et en écriture d'élèves à risque », *Revue des sciences de l'éducation*, 33(3) : 663-683.

Nagy W. et Townsend D. (2012). « Words as tools : Learning academic vocabulary as language acquisition », *Reading Research Quarterly* 47(1) : 91-108.

National Reading Panel (2000). *Teaching children to read*. Rockville : National Institute of Child Health and Human Development.

New B. (2006). « Lexique 3 : une nouvelle base de données lexicales », *Actes de la 13e Conférence sur le Traitement Automatique des Langues Naturelles* 2 : 892-900.

Noble K. G. et Mccandliss B. D. (2005). « Reading development and impairment : behavioral, social, and neurobiological factors », *Journal of Developmental & Behavioral Pediatrics* 26(5) : 370-378.

Pacton S., Fayol M. et Lété B. (2008). « L'intégration des connaissances lexicales et infralexicales dans l'apprentissage du lexique orthographique », *ANAE* 96-97 : 213-219.

Peereman R., Sprenger-Charolles L. et Messaoud-Galusi S. (2013). « The contribution of morphology to the consistency of spelling-to-sound relations : A quantitative analysis based on French elementary school readers », *Topics in Cognitive Psychology/L'Année Psychologique* 113(1) : 3-33.

Perfetti C. et Stafura J. (2014). « Word knowledge in a theory of reading comprehension », *Scientific Studies of Reading* 18(1) : 22-37.

Pothier B. et Pothier P. (2003). *Échelle d'acquisition de l'orthographe lexicale.* Paris : Retz.

Seymour P. H., Aro M. et Erskine J. M. (2003). « Foundation literacy acquisition in European orthographies », *British Journal of psychology* 94(2) : 143-174.

Schmitt N., Jiang X. et Grabe W. (2011). « The percentage of words known in a text and reading comprehension », *The Modern Language Journal* 95(1) : 26-43.

Sprenger-Charolles L. et Béchennec D. (2004). « Variability and invariance », *Written Language & Literacy* 7(1) : 9-33.

Stanké B., Royle P. Rezzonico S., Moreau A. C., Dumais C. et Le Mené M. (2018). *ÉQOL. Une banque de 15 000 mots à votre disposition*. Récupéré de https://appligogiques.com/eqol.

Taboada A. et Rutherford V. (2011). « Developing reading comprehension and academic vocabulary for English language learners through science content : A formative experiment », *Reading Psychology* 32(2) : 113-157.

Ters I., Mayer G. et Reichenbach D. (1995). *L'échelle Dubois-Buyse : l'orthographe usuelle française*. Paris : Éditions MDI.

Townsend D., Filippini A., Collins P. et Biancarosa G. (2012). « Evidence for the importance of academic word knowledge for the academic achievement of diverse middle school students », *The Elementary School Journal* 112(3) : 497-518.

Weinberg A., Boukacem D. et Burger S. (2012). « L'enseignement d'un vocabulaire disciplinaire dans deux contextes d'immersion universitaire : Quelle approche favoriser ? », *Canadian modern language review* 68(1) : 1-27.

ANNEXES

Annexe 1

Dictée – TEST- 1er cycle (CP et CE1)

1. Nous devons inscrire une **virgule** dans la phrase pour marquer une pause.

2. Mme **Poiti** n'aime pas les ordures traînant dans les rues.

3. Mon petit frère a très hâte à sa **fête**.

4. Malgré les recherches, il n'y a aucun signe de **Parfo** près de l'école.

5. Boire du **lait** est important pour la croissance des os.

6. Julie a un **poisson** dans son aquarium.

7. La maman lit un **conte** à ses enfants.

8. Ils ont une **famille** très nombreuse.

9. La ville de **Vanlu** n'est plus très loin. (la syllabe *Van* doit être prononcée comme le mot *vent*)

10. Son rêve est de faire le tour du **monde**.

11. Le chien **qui** jappe s'ennuie de son maitre.

12. La crème **Modin** sera excellente pour guérir son pied.

13. Le bébé du voisin a pleuré toute la **nuit**.

14. Finir l'école est une **étape** importante dans la vie.

15. Être **grand** a ses avantages.

16. Sa **chemise** préférée est celle avec des pois.

17. Julie **donne** son jouet préféré parce qu'elle est généreuse.

18. Le jardin botanique de **Noufo** est rempli de fleurs multicolores.

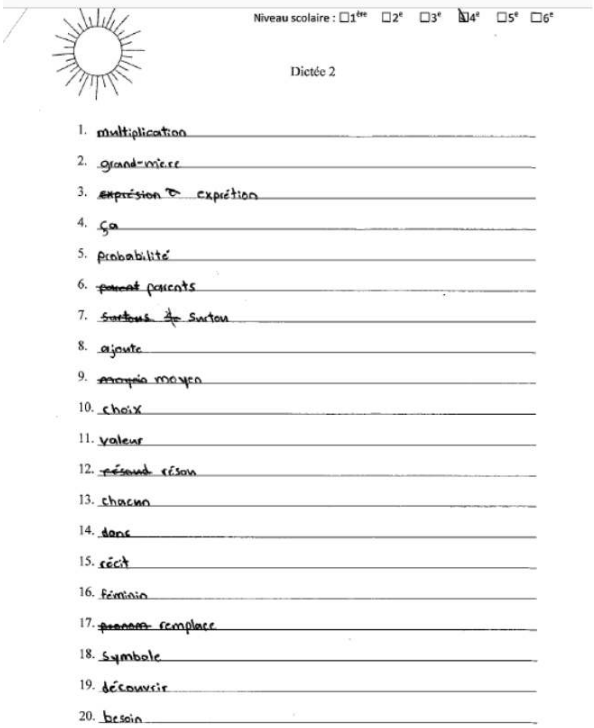
19. Mathieu a emprunté son **livre** préféré à la bibliothèque.

20. Il va dans sa chambre **afin** d'étudier dans le calme.

Seuls les mots en caractères gras ont été orthographiés par les élèves.

Annexe 2

Exemple d'une dictée réalisée par un élève de CM1


Niveau scolaire : ☐1ère ☐2e ☐3e ☒4e ☐5e ☐6e

Dictée 2

1. multiplication
2. grand-mère
3. ~~expression~~ → expréition
4. ça
5. probabilité
6. ~~pourcent~~ pourcents
7. ~~surtous~~ → surtou
8. ajoute
9. ~~moyen~~ moyen
10. choix
11. valeur
12. ~~résoud~~ résou
13. chacun
14. dans
15. récit
16. féminin
17. ~~remplace~~ remplace
18. symbole
19. découvrir
20. besoin

NOTES

1. Autrefois désigné sous l'appellation « sciences humaines », « univers social » regroupe au Québec ce qui correspond en France aux domaines fondamentaux « éducation morale et civique » et « questionner le monde » au cycle des apprentissages fondamentaux avec l'ajout de « géographie » et d'« histoire » au cycle des consolidations des connaissances.

2. Un mot consistant est un mot ne comportant pas de phonème pouvant correspondre à plusieurs graphèmes.

3. L'orthographe phonologique consiste à orthographier des mots à partir des correspondances phonèmes-graphèmes sans tenir compte de l'orthographe de la norme du système d'écriture. Le mot *bateau* peut ainsi s'écrire *bato*, mais peut aussi s'orthographier de plusieurs autres façons selon la forme orthographique retenue pour représenter chacun des phonèmes de ce mot.

4. L'orthographe lexicale des mots correspond à la norme du système d'écriture (p. ex. : l'orthographe lexicale du mot /bato/ est *bateau*).

RÉSUMÉS

Par son rôle déterminant dans la réussite scolaire et professionnelle, ainsi que dans l'insertion sociale, l'apprentissage de l'orthographe lexicale représente un défi majeur pour les élèves du primaire. Dans ce contexte, nombreux sont les enseignants, orthophonistes et chercheurs à s'intéresser à la question des outils utiles à son enseignement et à son apprentissage, et à avoir

recours notamment à des bases de données lexicales. Bien qu'elles constituent un apport considérable pour le domaine, les ressources existantes souffrent de plusieurs insuffisances. D'une part, elles s'appuient sur des corpus de référence peu représentatifs du lexique scolaire auquel les élèves sont exposés au primaire, et d'autre part, elles ne permettent pas aux utilisateurs de déterminer la façon dont se développe l'acquisition de l'orthographe lexicale. À la lumière de ces constats, nous avons élaboré une ressource lexicale hybride, rendant compte à la fois de la fréquence et des caractéristiques linguistiques du lexique scolaire présent dans les manuels du primaire au Québec et, pour une partie des mots de ce lexique, de la capacité des élèves à les orthographier à chaque niveau du primaire. Dans cet article, nous revenons dans un premier temps sur la méthodologie d'élaboration de cette nouvelle base de données - nommée ÉQOL - dotée d'une Échelle Québécoise d'acquisition de l'Orthographe Lexicale, à l'origine de la création de la base de données. Dans un deuxième temps, nous évaluons l'apport de cet outil. La comparaison de notre base de données (Stanké, Royle, Rezzonico, Moreau, Dumais & Le Mené 2018) à des outils existants (trois bases de données et une échelle d'acquisition lexicale) témoigne de la pertinence de la création d'un outil propre au lexique scolaire adressé aux élèves du primaire et de la nécessité d'utiliser une nouvelle échelle d'acquisition de l'orthographe lexicale, spécifique aux compétences orthographiques des élèves québécois.

Because lexical orthography has a determining role in academic and professional success, as well as social integration, learning lexical orthography is a major challenge for primary school children. With this in mind, numerous teachers, speech-language pathologists and researchers are interested in accessing tools for learning and teaching lexical orthography, and in particular, lexical databases. Although they are an important source of information in the field, current resources suffer from a number of shortcomings. On the one hand, their reference corpora are not representative of the primary school lexicon. Secondly they do not allow users to explore how lexical orthography is acquired. For these reasons we developed a hybrid lexical resource that can simultaneously account for frequency and linguistic characteristics of the academic lexicon used in primary-school textbooks in Quebec, and for a subset of these lexical items children's ability to spell them at each primary-school level. First, we present the methodology used to create this new lexical database (ÉQOL) and second, we assess its usefulness. A comparison of our database (Stanké, Royle, Rezzonico, Moreau, Dumais & Le Mené 2018) with existing tools (three databases and one vocabulary acquisition scale) supports its relevance as a specific tool for primary school vocabulary targeting grade-school students and the importance to use a new scale for the acquisition of lexical orthography adapted to the orthographic competencies of Québec students.

INDEX

Mots-clés : Base de données lexicale, échelle d'acquisition, orthographe lexicale, lexique scolaire
Keywords : Lexical database, acquisition scale, lexical orthography, academic lexicon

AUTEURS

BRIGITTE STANKÉ

Université de Montréal, Centre de recherche interdisciplinaire en réadaptation du Montréal métropolitain

MARINE LE MENÉ

Université de Strasbourg

STEFANO REZZONICO

Université de Montréal, Centre de recherche interdisciplinaire en réadaptation du Montréal métropolitain

ANDRÉ MOREAU

Université du Québec en Outaouais

CHRISTIAN DUMAIS

Université du Québec à Trois-Rivières

JULIE ROBIDOUX

Université de Montréal

CAMILLE DAULT

Université de Montréal

PHAEDRA ROYLE

Université de Montréal, Centre de recherche sur le langage, l'esprit et le cerveau