

ÉOLE

Échelle d'acquisition en Orthographe LExicale — B. Pothier & Ph. Pothier

Présentation de la base

ÉOLE (Échelle d'acquisition en Orthographe LExicale) indique, pour 11 694 mots les plus fréquents du français écrit contemporain, le pourcentage d'élèves qui les orthographient correctement à chacun des cinq niveaux de l'école élémentaire (CP, CE1, CE2, CM1, CM2). À la différence des bases qui décrivent la fréquence d'exposition aux mots écrits (Manulex, Novlex), ÉOLE renseigne sur l'état réel des acquisitions lexicales des élèves, par niveau de classe.

Genèse et démarche

ÉOLE est née d'un constat de terrain : au fil d'un travail conduit avec des enseignants, des futurs enseignants et des conseillers pédagogiques sur la didactique de l'orthographe, il est apparu que les pratiques d'enseignement de l'orthographe lexicale manquaient d'assise rationnelle, non par absence de textes ou de prescriptions, mais parce que les démarches essayées jusque-là avaient été abandonnées faute de résultats convaincants. Les enseignants engagés dans ces recherches-actions demandaient des outils permettant de hiérarchiser l'apprentissage de l'orthographe lexicale : à partir de quelles données décider qu'un mot relève de tel niveau plutôt que de tel autre ? Est-il plus simple, pour un enfant, d'écrire tel mot fréquent que tel mot rare ? Pourquoi les échelles disponibles situaient-elles si tardivement des mots d'usage courant ?

Les échelles existantes, Dubois-Buyse, Vocabulaire Orthographique de Base, dataient en effet de plusieurs décennies et reposaient sur des corpus d'écrits spontanés recueillis dans un tout autre état de la langue. Plutôt que de les refondre, les auteurs ont fait le choix d'une réactualisation complète, fondée sur des critères scientifiques de linguistique quantitative (ou computationnelle) : constituer un corpus contemporain, le traiter systématiquement, puis tester la totalité des mots retenus auprès d'élèves réels. ÉOLE est ainsi le produit d'une recherche-action de plusieurs années associant chercheurs et praticiens, conçue pour permettre un enseignement raisonné et hiérarchisé de l'orthographe lexicale.

Constitution du corpus

Le point de départ est un corpus d'environ un million de mots collecté dans la presse nationale française au tournant des années 2000, choisi comme le reflet du vocabulaire écrit de l'adulte, l'objectif visé étant que l'élève, devenu adulte, sache écrire les mots qu'il aura le plus de chances d'utiliser. Le choix de la presse s'explique par sa diversité : les titres retenus (quotidiens nationaux et régionaux, hebdomadaires, magazines spécialisés) couvraient des lectorats variés en âge, en tendances et en centres d'intérêt, garantissant un échantillon représentatif de l'écrit adressé au grand public. Plusieurs techniques d'extraction ont d'abord été explorées puis écartées (recopie manuelle, dictée à un tiers, numérisation page à page, reconnaissance vocale), avant que la collecte automatisée des versions en ligne des journaux ne permette de rassembler le corpus. De celui-ci ont été dégagées 20 378 formes différentes, ensuite confrontées à trois listes de référence antérieures : l'échelle Dubois-Buyse (3 700 mots), le Vocabulaire Orthographique de Base (9 130 mots) et la Liste Orthographique de Base (2 000 mots). Un critère de récurrence minimale (quatre occurrences au moins, tous les corpus confondus, avec reprise des termes

présents dans les trois listes antérieures) a conduit à retenir 11 979 mots pour l'expérimentation ; après plusieurs révisions successives, la liste actuelle en compte 11 694.

Chaque terme retenu a fait l'objet d'un classement grammatical systématique, vérifié à la lumière de deux dictionnaires de référence (le Trésor de la Langue Française et le Grand Robert), avec des choix explicites : élimination des formes conjuguées et des participes passés adjectivés, conservation des seuls féminins dont la graphie ne se déduit pas mécaniquement du masculin, exclusion des marques redondantes de pluriel.

Recueil des données

Le protocole a d'abord été validé par un pré-test conduit auprès d'une centaine d'enfants par niveau de classe (environ 500 élèves au total), qui a permis d'ajuster les consignes de passation et de correction. Chaque mot a ensuite été dicté en contexte, en fin d'année scolaire : les noms avec un déterminant au singulier, les adjectifs accompagnés d'un nom au singulier, les verbes à l'infinitif. Les 11 694 mots ont été répartis en 240 listes de 50 mots, dictées dans 464 écoles réparties sur l'ensemble du territoire français selon les critères INSEE, auprès de 48 902 élèves du CP au CM2. Le protocole garantissait qu'un même mot soit écrit par au moins 40 élèves de chacun des cinq niveaux, soit un minimum de 200 productions par mot ; au total, 2 368 885 graphies ont été recueillies et analysées, d'un point de vue quantitatif comme qualitatif. Plus tard, un recueil complémentaire, mené en France et au Canada avec des modalités identiques, a porté sur les mots concernés par les rectifications orthographiques de 1996.

Données fournies

Pour chacune des entrées, la base indique : le taux de réussite orthographique à chaque niveau de classe (compétences constatées en fin d'année scolaire) ; la ou les natures grammaticales, selon une nomenclature détaillée (nom masculin singulier, adjectif invariant en genre, etc.) ; ont été ajoutés, ensuite, les renseignements pour les 136 mots touchés par les rectifications orthographiques de 1990, dont l'enseignement est entré à l'Ecole en France, en 2016, avec les taux de réussite propres à la graphie rectifiée.

Par convention docimologique, une notion est considérée comme acquise lorsqu'elle est réussie par 75 % de la population testée : ce repère structure la lecture de l'échelle, mais son interprétation appelle les précautions détaillées ci-dessous.

Quelques enseignements de l'enquête

Le résultat le plus contre-intuitif de l'enquête est que l'acquisition orthographique n'est pas linéaire : pour certains mots, les taux de réussite diminuent au fil de la scolarité. Ainsi “tabou”, transcritible par simple application des correspondances phonème-graphème, est mieux réussi au CP (85 %) qu'au CM2 (51 %) : les plus jeunes encodent phonologiquement, tandis que leurs aînés, instruits de la diversité des graphies possibles du français, produisent des formes comme « tabout » ou « taboux ». Loin de signaler une régression, ces erreurs attestent au contraire de la progression de l'apprentissage : plus s'étend le répertoire de graphèmes que l'élève sait susceptibles de transcrire un même phonème, plus la sélection de la graphie conventionnelle devient incertaine et plus le risque d'erreur s'accroît. À l'inverse, un mot comme “adverbe” suit une progression régulière (4 % au CP, 100 % au CM2), conforme à l'image classique d'un apprentissage cumulatif. La coexistence de ces deux profils interdit toute lecture simpliste de l'échelle.

L'enquête chiffre aussi l'étendue du lexique orthographique disponible à chaque niveau : en retenant le seuil de 75 % de réussite, 347 mots sont utilisables au CP, 1 013 au CE1, 2 188 au CE2, 3 606 au CM1 et 5 054 au CM2 soit respectivement 2 189, 3 612 et 5 064 pour les trois derniers niveaux si l'on retient l'orthographe rectifiée). Un nombre bien suffisant, dont les enseignants prennent la mesure lorsqu'ils composent des textes de dictée ajustés aux compétences lexicales réelles de leurs élèves (ÉOLE, nouvelle édition, préface, § « Le nombre de mots acquis par niveau », p. 12).

Elle bat enfin en brèche plusieurs idées reçues. Non, un mot ne « s'écrit pas comme il se prononce » : l'information n'est d'aucun secours à l'élève, et les tentatives d'écriture par codage phonologique produisent des formes très éloignées de la norme. Non, connaître le sens d'un mot ne garantit pas de savoir l'écrire, ni l'inverse : des élèves de CP orthographient correctement des termes dont ils ignorent manifestement la signification, et échouent sur des mots parfaitement familiers. Non, la lecture assidue ne suffit pas à fixer l'orthographe : le lecteur lit pour comprendre, en s'appuyant sur la silhouette des mots, sans en mémoriser la graphie exacte. Ces constats, documentés par l'analyse qualitative des quelque 12,5 millions de graphies recueillies, plaident pour un enseignement explicite et hiérarchisé de l'orthographe lexicale ce que ÉOLE rend précisément possible.

Avant d'utiliser ÉOLE : mises en garde préalables

ÉOLE recense le niveau scolaire auquel les mots les plus fréquents du français écrit sont correctement orthographiés par la majorité des enfants. C'est un outil descriptif remarquable, mais dont l'interprétation exige de garder à l'esprit les précautions suivantes, conformes aux avertissements formulés par les auteurs.

1. Le seuil de 75 % : un constat, jamais une exigence

Un mot est rattaché à un niveau de classe dès lors que 75 % des élèves testés à ce niveau, soit trois quarts d'entre eux, l'écrivent sans erreur en fin d'année scolaire. Ce pourcentage est un constat descriptif, jamais une prescription : il indique que les enfants de ce niveau peuvent écrire ce mot, non qu'ils doivent savoir l'écrire. L'échelle rend compte d'un état moyen des acquisitions tel qu'il est statistiquement observé dans les classes ; elle ne fixe ni programme, ni exigence, ni norme d'évaluation individuelle. Par définition même du seuil retenu, jusqu'à un quart des élèves d'un niveau donné n'orthographient pas correctement un mot pourtant classé « acquis » à ce niveau : une erreur n'a donc rien d'alarmant et ne saurait fonder ni un jugement scolaire ni une inquiétude clinique.

On ne saurait donc tenir rigueur à un élève de ne pas orthographier un mot « acquis à 75 % » à son niveau ce qui constituerait un contresens sur la nature même de l'outil : par construction, jusqu'à un quart des élèves de ce niveau ne l'orthographient pas correctement, sans que cela soit pathologique ni même préoccupant.

2. Le cas particulier des archiphonèmes

Certains mots de l'échelle contiennent un archiphonème, c'est-à-dire, selon la définition de B. Pothier, le résultat de l'assimilation d'un phonème sourd par un phonème sonore, et inversement (“abcès” prononcé [aPse], “anecdote” [anɛGdɔt], “paquebot” [paGbɔ]). Les erreurs observées sur ces items (« apsès », « anegdote », « pagbot », « opscur »...) ne relèvent pas d'une difficulté de conversion grapho-phonétique : la transcription de l'enfant est fidèle à la forme orale réellement produite. Il s'agit d'erreurs témoignant d'une

représentation orthographique non encore stabilisée. Les taux de réussite d'ÉOLE sur ces mots doivent donc être interprétés avec précaution. Pour un approfondissement du système phonologique du français et de ces phénomènes d'assimilation, on se reportera à la Mallette de jeux phonologiques, cycles 1 et 2 (B. Pothier, Les Ateliers Retz), dont le guide pédagogique en propose une présentation synthétique.

3. Les mots transparents : une réussite qui ne dit pas tout

Un taux de réussite élevé sur un mot dont la graphie entretient une relation grapho-phonétique univoque ("tipi", par exemple) ne signifie pas que les élèves en connaissent l'orthographe : il signifie qu'ils ne peuvent guère se tromper, la simple application des correspondances phonème-graphème suffisant à produire la forme normée. La réussite reflète alors une compétence d'encodage phonographique, non la présence du mot dans le lexique orthographique de l'enfant. Inversement, les auteurs relèvent que les erreurs des élèves plus âgés sur ces mots transparents (« tabout », « taboux »...) témoignent d'un apprentissage en cours : plus l'enfant connaît de graphies possibles d'un même phonème, plus le risque d'erreur s'accroît, l'erreur est ici une démonstration du degré d'apprentissage, non un recul.

4. Des compétences de fin d'année, mesurées grâce aux mots dictés en contexte.

Les pourcentages d'ÉOLE correspondent à des compétences constatées en fin d'année scolaire, grâce à des mots dictés en contexte minimal. Ils ne peuvent être appliqués tels quels à un élève évalué en cours d'année, ni transposés directement à la production de textes, situation où la charge cognitive (gestion simultanée de l'idéation, de la syntaxe, de l'orthographe grammaticale) abaisse mécaniquement les performances orthographiques.

5. Une échelle d'orthographe lexicale, exclusivement

ÉOLE renseigne sur l'orthographe lexicale (la graphie des mots pris isolément) et non sur l'orthographe grammaticale (accords, morphologie verbale, homophones grammaticaux). Les auteurs ont d'ailleurs écarté du recueil les formes conjuguées et les marques d'accord redondantes. Toute inférence sur les compétences orthographiques globales d'un élève à partir de la seule échelle serait abusive.

6. Un corpus relevé dans le vocabulaire de la presse des années 2000

Le corpus source reflète le vocabulaire écrit de l'adulte tel qu'il apparaissait dans la presse nationale au début des années 2000. Certains termes ont pu vieillir, d'autres, apparus depuis, sont absents. L'absence d'un mot dans ÉOLE signifie seulement qu'il ne figurait pas parmi les formes les plus fréquentes de ce corpus, et non qu'il serait rare ou difficile aujourd'hui.

7. Les rectifications orthographiques de 1990

Depuis 2016, les deux graphies, traditionnelle et rectifiée, doivent être acceptées comme exactes. Les 136 mots concernés font l'objet d'un repérage spécifique dans l'échelle : l'utilisateur veillera à ne compter comme erronée aucune production conforme à l'une ou l'autre norme.

8. Un outil de référence, pas un instrument de diagnostic

ÉOLE est un étalonnage descriptif de l'acquisition orthographique en population scolaire tout-venant. Elle permet de situer une production, de choisir des mots pour un entraînement ou une ligne de base, de

relativiser une erreur ; elle ne constitue pas, à elle seule, un instrument de diagnostic des troubles du langage écrit, lequel relève d'une évaluation normée et pluridimensionnelle.

Conditions d'utilisation et citation

Les données d'ÉOLE sont la propriété intellectuelle de leurs auteurs, Béatrice Pothier et Philippe Pothier, titulaires des droits d'auteur. Leur mise à disposition en ligne s'effectue avec leur autorisation expresse et écrite.

Les données d'ÉOLE (ce qui inclut les résultats des requêtes) sont mises à disposition selon les termes de la licence Creative Commons Attribution — Pas d'Utilisation Commerciale — Partage dans les Mêmes Conditions 4.0 International (CC BY-NC-SA 4.0). Toute utilisation en dehors du champ de cette licence doit faire l'objet d'une autorisation préalable des auteurs.

Toute utilisation des données doit citer la référence suivante : *Pothier, B. & Pothier, Ph. (2020). ÉOLE. Échelle d'acquisition en Orthographe LExicale, cycles 1, 2 et 3 (nouvelle édition). Paris : Retz.*