Analyser linguistiquement l'écriture à l'école: EcriScol, un corpus génétique

Claire Doquet

Analyser linguistiquement l'écriture à l'école : EcriScol, un corpus génétique

Claire Doquet
Université Sorbonne Nouvelle,
laboratoire Clesthia, équipe EcriScol
claire.doquet@sorbonne-nouvelle.fr

Abstract
Cet article pose la question de l'exploration outillée de productions verbales non standard, en l'occurrence des écrits d'élèves. Il présente des possibilités d'investigation des écrits scolaires à partir de la base EcriScol en constitution au laboratoire Clesthia de la Sorbonne Nouvelle. Le corpus EcriScol donne une vue développementale des écrits d'élèves du début de l'école primaire au seuil de l'université. Grâce à un système de transcription des opérations d'écriture, il permet d'accéder non seulement au texte, mais à ses différents états et aux événements d'écriture qui l'ont constuit. Il constitue ainsi un observatoire des lieux d'interrogation linguistique et textuelle des élèves, selon le niveau scolaire et la consigne scripturale.

Les *Learner Corpora*, constitués en grand nombre en langue anglaise, restent rares en français et encore plus en Français Langue Maternelle.[1] Le développement de ces corpus est lié à celui de logiciels d'analyse textuelle qui permettent d'analyser quantitativement un grand nombre de données, produisant un résultat d'autant plus significatif qu'il s'appuie sur un matériau abondant. Ce type de données est propice à deux types d'exploitation :

- renseigner sur les caractéristiques de la langue écrite des élèves pour tracer des portraits développementaux des compétences mises en œuvres lors de l'écriture ;
- diagnostiquer des éléments lacunaires *vs* acquis selon l'âge et les conditions d'exercice de l'apprentissage, et construire, à partir de ce diagnostic, des programmations adaptées.

Le présent article s'inscrit dans le travail du groupe de recherche EcriScol (*Ecriture Scolaire*) du laboratoire Clesthia de la Sorbonne Nouvelle (http://www.univ-

[1] Un recensement récent figure dans la thèse de Claire Wolfarth : *Apports du TAL à l'exploitation linguistique d'un corpus scolaire longitudinal.* Grenoble, décembre 2019 (http://www.theses.fr/s139881)

paris3.fr/ecriscol-300509.kjsp), dont l'objectif principal est de mettre à disposition en *Open Acces* un ensemble d'écrits d'élèves provenant de différents niveaux de l'apprentissage entre le début de l'école primaire[2] et la fin du lycée, ainsi qu'un outillage d'analyse linguistique de l'écriture scolaire. Il s'agit d'étudier l'écriture et pas seulement les écrits, en recueillant et en analysant des textes mais également l'ensemble de ce que la génétique textuelle[3] appelle les avant-textes (notes, brouillons, etc.). Le recueil est effectué dans les différents niveaux concernés (CE1-CE2, CM2-6ème, 3ème-2nde, seuil de l'université) dans des conditions de passation écologiques : chaque enseignant, dans sa classe, organise l'écriture et les éventuelles réécritures. Les textes sont assortis de métadonnées descriptives.

La constitution d'un tel corpus suppose de régler des questions qui font toujours débat au sein de la communauté des chercheurs travaillant sur ce type de données, en particulier la tension entre la volonté de donner à lire les écrits d'élèves en en respectant la matérialité linguistique, y compris dans les plus bas niveaux de l'apprentissage, et le risque que les nombreux écarts à la norme nuisent à la lisibilité des écrits ; cette question rejoint celle de l'exploitation du corpus par des logiciels d'analyse textuelle, basée sur une procédure de normalisation qui risque d'altérer le matériau, voire de le dénaturer.

J'exposerai pour commencer les caractéristiques du corpus EcriScol et les choix faits en matière de normalisation orthographique ; puis je reviendrai sur une de ses spécificités, la mise à disposition pour analyse automatique des opérations d'écriture interprétées grâce aux ratures, pour en expliciter l'utilité et les possibilités heuristiques.

1. Le corpus EcriScol : constitution et développement

C'est en septembre 2013 qu'est né au laboratoire Clesthia (*Langue, Système, Discours* - Sorbonne Nouvelle Paris 3) le programme EcriScol[4] qui constitue une base de données comportant, pour chaque texte, l'ensemble des traces écrites de son élaboration (plan, notes, brouillon, etc.), ainsi que des métadonnées institutionnelles, sociologiques et didactiques permettant des recherches critériées[5]. Actuellement, 1500 écrits transcrits et annotés sont disponibles sur le site EcriScol, tout comme les bases de données analysables par le logiciel de textométrie iTrameur, développé au laboratoire Clesthia par Serge Fleury (http://www.tal.univ-paris3.fr/trameur/iTrameur/). Il est donc possible d'explorer l'ensemble des copies à l'aide de cet outil textométrique, soit globalement, soit par niveau scolaire.

[2] L'école primaire française est organisée en 5 niveaux : le Cours préparatoire (CP) accueille des élèves de 6-7 ans, les Cours Elémentaires 1 et 2 (CE1 et CE2) accueillent des élèves entre 7 et 9 ans, les Cours Moyens 1 et 2 (CM1 et CM2) accueillent des élèves entre 9 et 11 ans.

[3] La génétique textuelle, née en France dans les années 1970 sous le nom de *critique génétique*, est une approche des manuscrits d'écrivains visant à mettre au jour les mécanismes de l'écriture. Elle a été appliquée à partir de 1990, en France, à l'étude des brouillons d'écoliers (Fabre, 1987 et 2002 ; Doquet, 2011 ; *inter al*.). Voir la partie 3 du présent article.

[4] EcriScol est soutenu par les consortiums *Corpus Écrits* puis *Corpus Linguistiques* (CORLI) de la Très Grande Infrastructure de Recherche (TGIR) *Huma-Num*, en lien avec l'EquipEx *Ortolang*, ainsi que par le LabEx *Empirical Foundations of Linguistics.*

[5] Ce grand corpus d'écrits d'élèves est destiné à entrer dans le corpus de référence pour le français contemporain (Siepmann et *al*., 2016).

EcriScol n'est pas le premier corpus d'écrits scolaires constitué en France en vue d'études linguistiques. Le premier du genre fut produit au laboratoire EMA (*Ecole, Mutations, Apprentissage*) de l'université de Cergy-Pontoise (Boré & Elalouf : 2017) pour analyser conjointement des écrits scolaires et les dispositifs d'enseignement et d'apprentissage ayant présidé à leur production. Quelques années plus tard, dans le cadre du groupe de recherche national « Production Verbale Ecrite » du CNRS (http://www.gdr-pve.fr), le corpus expérimental dit « Grenouille », composé de textes narratifs et explicatifs produits entre le milieu de l'école primaire et le milieu du collège, a servi de support d'études à des chercheurs d'origines disciplinaires diverses : linguistique, psycholinguistique descriptive, psychologie expérimentale, didactique (Gunnarsson-Largy & Auriac-Slusarczyk, 2013). On pourrait bien sûr citer d'autres références ; j'ai choisi celles-ci pour leur caractère *princeps* : chacun avec ses spécificités, ces deux corpus ont fait date dans l'analyse, outillée ou non, des écrits scolaires. Leurs modes de constitution distincts mettent en évidence un problème récurrent dès lors que l'on veut sérier les écrits scolaire, qui tient à une hétérogénéité liée à leur instabilité générique : échappant bien souvent à la norme, linguistique comme textuelle, les écrits d'élèves entrent très difficilement dans la logique des corpus, qui implique que chaque écrit pris isolément (*token*) puisse être considéré comme un cas particulier d'un *type* commun. De fait, ces écrits sont rarement réductibles à une modélisation, qui est pourtant un des aboutissements des analyses textuelles souhaités par les auteurs des corpus. Boré & Elalouf (2017) notent cette difficulté à comparer les productions fictionnelles qui constituent leur matériau, tant est saillante la disparité de ces ensembles.

Collecté dans des conditions expérimentales très précises, le corpus « grenouille » aboutit à des productions d'une plus grande homogénéité et, de toute évidence, il répond mieux que le premier aux critères d'identification d'un *corpus* ainsi énoncés par le dictionnaire du Centre National de Ressources Textuelles et Linguistiques (CNRTL) : « recueil réunissant ou se proposant de réunir, en vue de leur étude scientifique, la totalité des documents disponibles d'un genre donné, par exemple épigraphiques, littéraires, etc. » et « ensemble de textes établi selon un principe de documentation exhaustive, un critère thématique ou exemplaire en vue de leur étude linguistique ». Dans le domaine des écrits d'élèves, la classification des écrits en genres ou types de textes est une véritable question de recherche, d'autant plus prégnante que le recueil est écologique[6]. L'extrême variabilité évoquée par Boré & Elalouf oblige, si l'on veut réunir un ensemble de données assez homogène pour prétendre à l'exhaustivité et à la représentativité d'un phénomène, à contraindre considérablement la production des écrits, au point que l'on pourra alors se questionner sur leur capacité à rester représentatifs des écrits produits au quotidien de la classe.

Dès l'abord, la question d'un corpus d'écrits d'élèves, au sens plein du terme *corpus*, est donc épineuse. L'équipe EcriScol a fait le choix du recueil écologique, tout en suggérant aux enseignants participants des consignes d'écriture qui permettent, pour les recueils de début et de fin d'année scolaire, une homogénéité des consignes. C'est la

[6] Le recueil écologique des données s'effectue dans les conditions habituelles d'exercice de l'activité, contrastivement aux approches expérimentales qui construisent une situation de recueil reproductible d'un milieu à l'autre. L'approche écologique en sciences humaines et sociales considère que le sujet construit son environnement qui, par voie de retour, influe sur la construction du sujet lui-même (Bronfenbrenner, 1979). Selon cette conception, il n'est pas pertinent d'étudier un sujet hors de son milieu.

seule intervention sur les pratiques des enseignants, qui sont par ailleurs complètement libres : nous n'imposons pas de protocole de séance, la seule exigence est d'interdire l'utilisation du crayon-gomme et du blanco (correcteur liquide) puisque nous nous intéressons particulièrement aux ratures. En effet, la transcription de l'ensemble des ratures et traces d'opérations scripturales est un élément spécifique du corpus EcriScol. Nous recueillons pour chaque copie l'ensemble des éléments de ce que la génétique textuelle appelle l'*avant-texte*, à savoir les différents écrits préalables à l'écrit final : notes, plans, ébauches, brouillons associés à une copie seront recensés comme constituant son *dossier de genèse* (Grésillon 1994 : 109). Ce dossier, composé de plusieurs éléments disjoints, est présenté exhaustivement dans le corpus. En général, un devoir d'élève comporte deux éléments : le brouillon et la copie finale ; parfois, on trouve également des notes prises par exemple dans un cahier de brouillon, ou bien un état intermédiaire entre le premier brouillon et la copie finale. Ces éléments sont présentés en liste et chacun peut être consulté, assorti des métadonnées qui s'y adjoignent et de sa transcription. Voici un exemple de l'écran de consultation (fig. 1) :

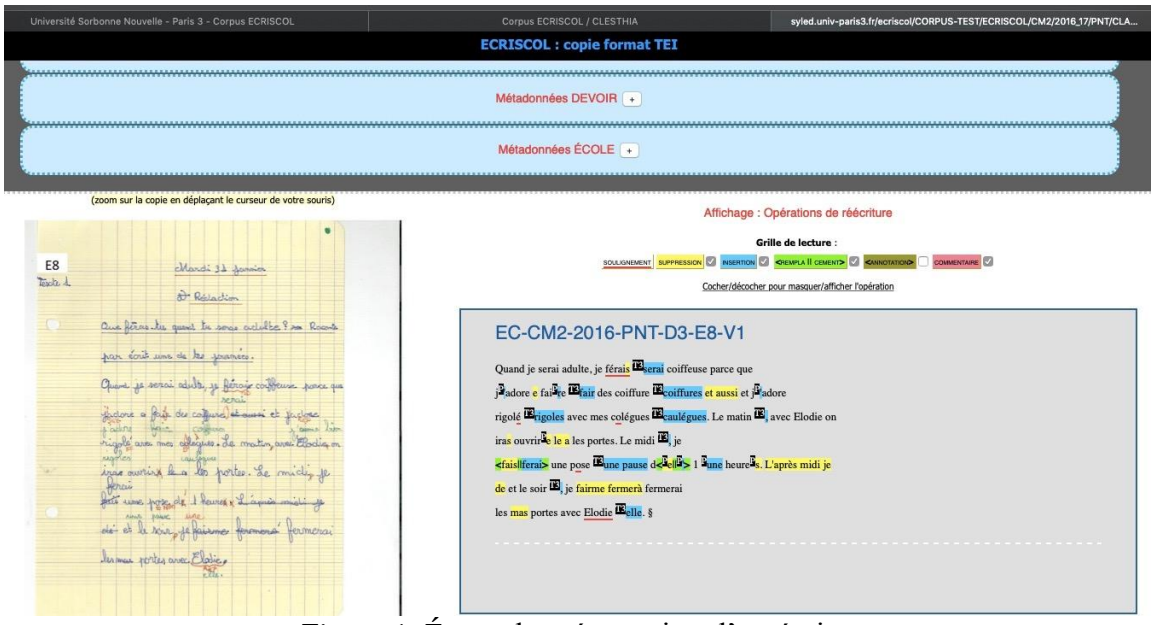

Figure 1. Écran de présentation d'un écrit.

Dans la partie gauche se trouve le fac-similé du manuscrit[7]. A droite, sa transcription, chaque rature étant rendue visible par une convention spécifique (surlignement jaune = suppression ; bleu = ajout ; vert = remplacement ; indication de la temporalité avec l'étiquette T3 qui indique l'intervention sur le texte après la première scription et la correction de l'enseignant, représentant respectivement les temps 1 et 2 de l'écriture). Au-dessus se trouvent des menus donnant accès aux métadonnées, qui concernent : l'établissement scolaire (rural/urbain, bénéficiant ou pas d'un dispositif d'enseignement prioritaire…), le scripteur (âge, scolarité, situation socio-professionnelle des parents…) et le devoir (consigne, ressources à disposition pendant l'écriture, durée d'écriture, taille du texte…).

[7] Sur le fac-similé apparaissent des commentaires de l'enseignant dont la transcription n'a pas été reproduite ici pour des questions de lisibilité du document.

Les 1500 copies du corpus sont aujourd'hui accessibles sous cette forme. Sur la plateforme EcriScol peuvent également être téléchargées les bases de données correspondantes, au format XML, permettant l'exploration outillée du corpus.

2. La normalisation des écrits : quelques investigations

Une des difficultés à traiter informatiquement les copies d'élèves est l'écart entre ces productions et les usages standards du langage. L'élément le plus saillant est, en français, l'orthographe ; on relève également divers écarts liés à la syntaxe et, de manière quasi systématique dans les petites classes, la ponctuation démarcative.

L'illustration donnée ci-dessus présente des énoncés non conformes à l'usage standard, par exemple :

- Orthographe : des erreurs de morphogrammes grammaticaux de temps (*ferais* indiquant le futur simple ; *rigolé* indiquant l'infinitif =) ou de nombre (*des coiffure*) ; des erreurs liées à l'homophonie (*pose* pour *pause*) ; des erreurs de phonogrammes (*je fairme*).
- Lexique : collocation non standard (*je ferais coiffeuse*)

Comment va réagir un étiqueteur tel que *TreeTagger* face à de tels écarts ? Et au-delà de l'étiquetage, comment faire travailler des procédures automatiques sur des écrits qui s'éloignent des règles de construction des phrases et des textes ? La question posée est celle de la normalisation de ces écrits, en tension avec le projet de travailler sur la langue écrite des élèves et non sur sa correction. Le groupe EcriScol a choisi une normalisation minimale permettant aux logiciels de catégoriser chaque forme sans modification ponctuationnelle ni syntaxique. Pour pouvoir lemmatiser les écrits, il faut en effet soumettre à l'étiqueteur des formes par lui identifiables, c'est-à-dire orthographiquement normées. Souhaitant en même temps donner à voir les écrits des élèves dans l'intégralité de leurs caractéristiques, nous avons choisi de produire sur la plateforme *EcriScol*, à côté du fac-similé, une version transcrite qui reproduit au plus près ce que le manuscrit rend visible : les formes dans leur graphie originelle. A côté de chaque graphie non standard, la transcription propose une graphie normalisée. Voici ce que donnent les premières lignes du texte que nous avons pris comme exemple (fig. 2) :

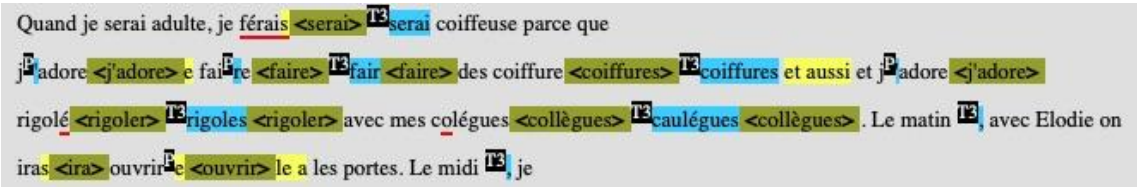

Figure 2. Transcription avec annotation orthographique.

Chaque forme non standard est ainsi annotée, cette procédure permettant la constitution d'un fichier XML qui est ensuite reformaté en vue de son traitement par le logiciel de textométrie *iTrameur* (http://www.tal.univ-paris3.fr/trameur/iTrameur/). Avec cette manière de procéder, il est possible de repérer, dans tout le corpus, des graphies non standard correspondant à une graphie donnée. Choisissons par exemple les formes verbales, dont on sait que les finales en [e] demeurent des sources d'erreurs fréquentes parfois jusqu'à l'âge adulte (David, Brissaud & Guyon, 2006). Notre texte-exemple (fig. 2) en comporte deux occurrences : un futur réalisé en *–ais* et un infinitif réalisé en *–é*

puis en *–es*. Dans l'ensemble des écrits de niveau CM2 (fin d'école primaire), une recherche sur les erreurs de graphie de l'infinitif en -er fait apparaître *aller* comme le verbe le plus fréquemment erroné (24 occurrences).[8] Une fois cela repéré, on peut regarder quelles sont les graphies erronées des homophones réalisés phoniquement [ale].[9] Voici le relevé (fig. 3) :

Item	Fq	Concordance	Ventilation
#aller#	24		
#allée#	17		
#allé#	16		
#allait#	15		
#allés#	11		

Figure 3. Erreurs sur les homophones de [ale] en fin d'école élémentaire.

Les formes (colonne 1) sont ordonnées par fréquence (colonne 2) ; le concordancier (colonne 3) permet d'obtenir les formes en contexte. La ventilation (colonne 4) représente sous la forme d'une courbe la répartition des éléments observés dans la portion du corpus choisie. Peu significative quand les résultats concernent comme ici un niveau scolaire, elle le devient quand on travaille sur le corpus entier, à savoir des écrits produits du début de l'école élémentaire au seuil de l'université. Voici le résultat pour les formes homophones en [ale] (courbes de fréquence relative) sur l'ensemble du corpus, de la 2ème année d'école primaire à l'université (fig. 4) :

[8] C'est aussi l'infinitif en –er le plus fréquent, devant *manger* et *chercher*. Les deux verbes les plus fréquents à l'infinitif sont *voir* et *dire*.

[9] Nous n'avons pas tenu compte de la différence phonématique des finales en [e]/[□], distinction qui tend aujourd'hui à se réduire.

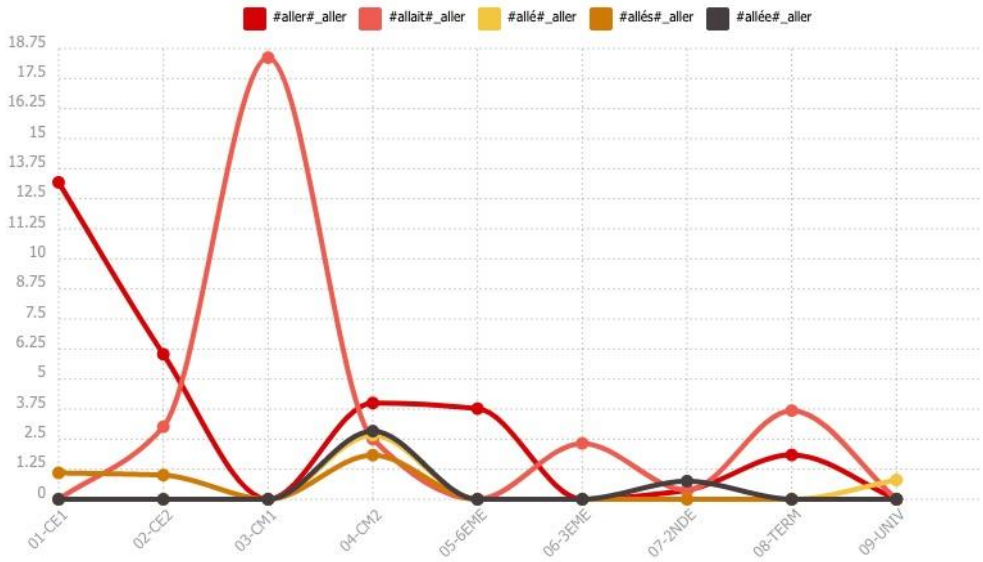

Figure 4. Erreurs sur les homophones de [ale] : vue développementale.

Ce graphique est basé sur une recherche croisée de formes et de lemmes[10]. Le premier terme de la légende est la graphie normée de formes qui, dans les copies des élèves, sont erronées. Ainsi, le graphique rend compte pour chaque forme de nombre de graphies erronées dans les copies.

Le niveau CM1 (4ème année d'école primaire) est atypique parce qu'il comporte un seul devoir avec une consigne très particulière qui mobilise le système énonciatif du passé (alternance imparfait/passé simple) au détriment des passés composés ; le verbe *aller* y est sur-représenté mais les seules formes qu'on y rencontre sont *allait* (12 occurrences) et *alla* (9 occurrences) ; même si elles sont peu nombreuses dans l'absolu, elles représentent une partie importante des verbes présents dans ce sous-corpus : c'est ce qui explique le pic de la forme *allait*. Si l'on fait abstraction de ce niveau CM1, les courbes suivent un chemin cohérent, particulièrement visible pour *aller* qui part d'un taux important d'erreurs en CE1 (2ème année d'école primaire) pour arriver à un taux nul à l'université. L'ensemble des formes suit cette évolution.

Dans cette base de données en constitution, les effets de corpus sont encore nombreux. Par exemple, on ne peut expliquer autrement, outre le cas déjà signalé du CM1, que le taux d'erreurs sur *allait* chute en classe de 2nde (début du lycée) pour remonter en 1ère alors que (1) l'imparfait est a priori tout autant utilisé en 2nde, par exemple dans le cadre de l'écriture d'invention, et (2) il n'y a pas de raison que les élèves graphient moins bien l'imparfait en 2nde qu'en 1ère.

Nonobstant ce défaut lié au nombre relativement faible de copies présentes dans la base, celle-ci donne accès à des informations précieuses sur l'évolution des erreurs commises, par lemme ou forme comme nous l'avons vu mais aussi par catégorie grammaticale par exemple. Voici le graphique des erreurs sur des formes verbales représentées selon leur degré de spécificité entre le CE1 et l'université (fig. 5) :

[10] Chaque légende indexée à une couleur suit la syntaxe suivante : #X#_X' où X' est le lemme correspondant à la forme X. Le signe # qui encadre X signale, dans cette base, que la forme proposée par l'élève a dû être corrigée lors de la normalisation.

Figure 5. Spécificité des erreurs sur les formes verbales selon le niveau de classe.

Les temps ou modes verbaux sur lesquels les erreurs sont, dans l'ensemble du corpus, les plus fréquentes, sont :

- le participe passé : en français, sa graphie est très souvent erronée du fait de la complexité de l'accord, variable selon l'auxiliaire, à quoi s'ajoute très souvent le non-marquage phonique de la flexion en genre et en nombre ;
- le présent de l'indicatif, qui est le temps le plus usité ;
- l'imparfait de l'indicatif ;
- l'infinitif.

Une exploration plus fine permettrait de caractériser les erreurs pour chaque temps verbal. Concernant l'évolution des compétences, le graphique montre les éléments suivants :

- de manière générale, les erreurs d'orthographe sur les verbes sont caractéristiques des premiers niveaux de l'apprentissage ; leur absence caractérise les niveaux les plus élevés ;
- c'est le présent de l'indicatif qui marque le plus nettement cette tendance, avec les écarts les plus importants entre le début et la fin de la scolarité ;
- la même évolution est constatée sur le participe passé, mais elle commence plus tard : seulement à la fin de l'école primaire (CM2) ;
- c'est aussi à la fin du primaire que les erreurs sur l'imparfait apparaissent nettement, sans doute concomitantes à son usage même.

Beaucoup d'autres pistes d'exploration pourraient bien entendu être suivies. Il s'agit ici de montrer quelques possibilités d'exploitation de la base, que les lecteurs peuvent consulter et explorer eux-mêmes en ligne (http://syled.univ-paris3.fr/ecriscol/CORPUS-TEST/).

3. La transcription génétique : horizons de recherche

Les éléments présentés ci-dessus donnent quelques exemples des investigations que permet EcriScol en termes d'écarts à la norme. Ce n'est pas une originalité : la plupart des corpus d'écrits d'élèves comportent une normalisation, au moins orthographique, parfois allant jusqu'à la ponctuation ou la syntaxe.[11]

Comme explicité en partie 1, le corpus *EcriScol* donne lieu à des transcriptions de type génétique, c'est à dire que sont transcrits non seulement les textes, mais l'ensemble des opérations d'écriture que l'on peut y observer. Ces opérations peuvent être recensées et analysées par *iTrameur*, qui permet donc de travailler de manière quantitative sur l'écriture elle-même, à partir de ses traces matérielles.

L'importance des ratures, traces des opérations scripturales, pour la compréhension de l'écriture, est liée aux travaux de génétique textuelle et de linguistique de l'énonciation. L'intérêt porté dans le champ linguistique aux manifestations spontanées de l'activité métalinguistique (Rey-Debove 1978 et 1982) et méta-énonciative (Authier-Revuz 1995 et 2020) des locuteurs et des scripteurs a permis de théoriser les retours sur le déjà-dit / déjà-écrit comme marques de cette activité (Doquet, 2011). Toute rature est métalinguistique, au sens où la rature « travaille sur un discours déjà là » (Rey-Debove 1982 : 111), impliquant donc une activité sur le discours et non seulement une expansion de ce discours. Pour Rey-Debove, « supprimer un mot en le barrant [est] une activité métalinguistique comparable à supprimer un mot en niant le signe par la parole. Car le modèle de la négation de signe est métalinguistique « *Considérez que je n'ai pas dit X* » et n'a rien à voir avec la négation de choses « *Ce n'est pas un X* » qui utilise la négation linguistique ne portant que sur le signifié. » (*ibid*.)

Dans son étude des brouillons d'écoliers, Fabre (1987, 2002) a montré que les élèves, dès le plus jeune âge, rectifient leurs écrits ; elle suit Rey-Debove pour considérer que toute rectification a une valeur métalinguistique. Une opération de rectification de l'écrit suppose selon elle que le scripteur établisse un rapport paradigmatique entre deux éléments : celui dont il ne veut plus et celui qu'il lui préfère. Ce rapport paradigmatique se constitue par la procédure suivante : le scripteur porte sur son écrit un regard d'ordre métalinguistique en le jaugeant par rapport à ce qu'il souhaite en faire ; repérant un segment à modifier, il cherche quels moyens lui offre la langue pour cette modification, il choisit une solution et effectue le remplacement. Selon Fabre, qui donne en exemple des corrections orthographiques et des remplacements de mots par d'autres, « c'est cette incursion dans l'axe du "système" qui fait sortir la rature du plan du langage "premier", de dénotation, et relève de la fonction métalinguistique : traitement du signifiant seul, modification de la relation signifiant/signifié, concurrence entre deux signes du système…"» (Fabre, 1987 : 47).

Lorsque la rature a pour objet de proposer une autre graphie pour un même mot, c'est bien clairement d'activité métalinguistique qu'il s'agit, le scripteur ne pouvant éviter de « passer dans le système » (système morphologique, système de correspondance grapho-phonétique) pour comparer des graphies possibles et en choisir une. La réflexion orthographique, où il ne s'agit plus de choisir un signe mais d'identifier le signifiant graphique d'un signe donné, met en jeu une exploration fine des règles qui régissent le système signifiant.

Pour se donner une première idée de la proportion des opérations d'écriture par rapport aux erreurs orthographiques, on peut comparer pour chaque niveau scolaire les

[11] Par exemple, la normalisation du corpus *Scoledit* développé à l'université Grenoble-Alpes intervient sur la ponctuation, dans le souci de discriminer les phrases pour faciliter la segmentation des textes.

opérations de base (ajout *add* en rose et suppressions *del* en jaune) et les erreurs *corr* en rouge (courbe de fréquence relative – fig. 6) :

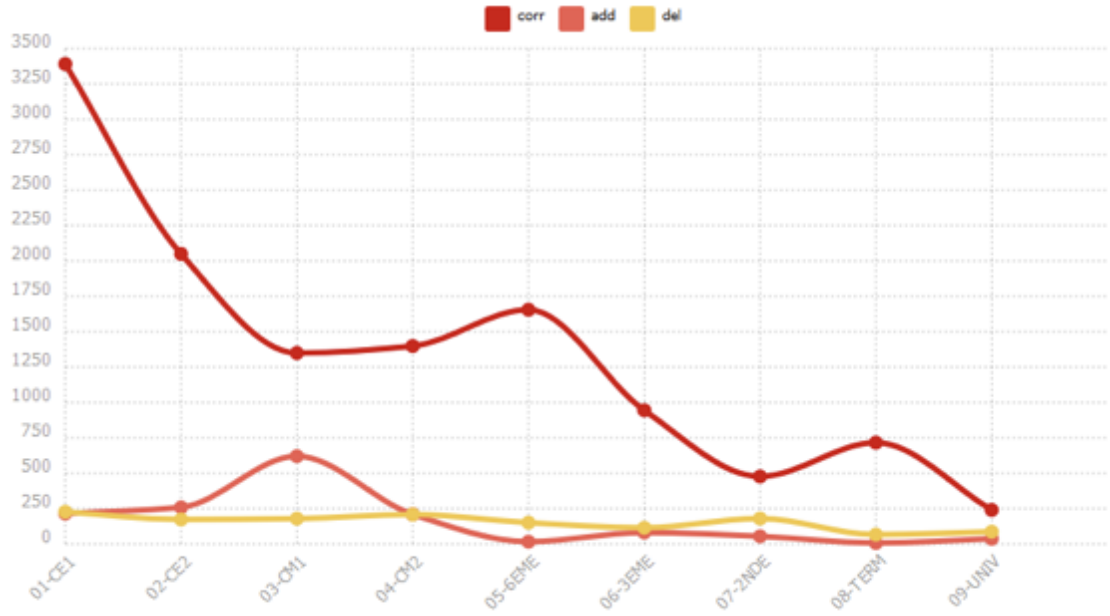

Figure 6. Opérations de réécriture et erreurs d'orthographe au cours de la scolarité.

Ce graphique montre que si les erreurs baissent tendanciellement, malgré une remontée entre le CM2 et la 6ème qui est sans doute partiellement imputable aux changements d'habitude scolaires, les opérations d'écriture réalisées par les élèves restent systématiquement – et même à l'université ! – en nombre inférieur. Comme il porte sur l'ensemble des opérations et erreurs, ce graphique ne permet pas de faire le rapport entre une forme retouchée et le taux d'erreurs qui l'affecte. Voici un exemple d'une telle investigation, sur la préposition *à* qui, en français, est une source majeure de confusions homophoniques, représentée par les spécificités par niveau (fig. 7) :

Figure 7. Spécificités des erreurs sur a/à au cours de la scolarité.

Les données représentées sont au nombre de trois : l'ajout de *à* par le professeur (#à#_add_P, en rouge), l'ajout de *à* par l'élève (#à#_add, en rose), l'erreur sur *à* (#à#_corr, en jaune) ; les ajouts de *à* sont presque toujours des corrections orthographiques, puisqu'ils signifient en fait que le scripteur a ajouté non pas la préposition *à*, mais un accent grave sur le *a* déjà présent[12]. Le taux d'erreurs suit l'évolution habituelle (Brissaud & Chevrot, 2000) et il est rassurant de constater qu'il est spécifique aux niveaux les plus bas (jusqu'au CM1) et au contraire singulièrement absent du niveau universitaire. Ce qui m'intéresse davantage ici, c'est la répartition des ajouts de *à* selon que c'est le professeur ou l'élève qui les ajoute. On observe que l'ajout de *à* par le professeur est spécifique au CE2, niveau auquel les enseignants cherchent à faire différencier les homophones (les niveaux inférieurs sont centrés sur la correspondance grapho-phonétique ; cela n'exclut pas des corrections de *à*, bien entendu, mais comme elles sont rares elles ne sont pas spécifiques à ces niveaux). Par la suite, *à* n'apparaît plus comme un indice spécifique du geste de correction professoral. En revanche, les corrections de *à* par les élèves spécifient le CM2, et simultanément les erreurs sur *à* y apparaissent comme non spécifiques. On pourrait donc faire l'hypothèse que les élèves, à la fin de l'école primaire, corrigent leurs erreurs sur *à*, et donc que les textes en comportent moins. En 6ème, c'est l'inverse qui s'observe : les erreurs redeviennent des spécifications de ce niveau scolaire, pour entrer ensuite dans une spécificité inverse : c'est leur petit nombre qui caractérise les productions entre la 3ème et la Terminale. La charnière entre l'école primaire et le collège apparaît, ici comme dans la courbe générale des erreurs d'orthographe (fig. 6), comme un moment où les compétences fragilisées donnent lieu à une chute des performances ; on voit ici qu'elle

[12] Notre système de transcription, sans doute à améliorer de ce point de vue, n'a pas prévu l'ajout d'accent seul – alors que pour le binôme *a/à*, c'est ce qui se passe matériellement le plus souvent. Dès qu'un *a* devient *à*, quelle que soit la procédure effectivement adoptée par l'élève (ex : biffer *a* et écrire *à*, ou simplement ajouter un accent à *a*), cette transformation est traitée comme un remplacement où *a* est supprimé et *à*, ajouté.

se double d'une diminution des traces de la vigilance orthographique que constituent les ratures sur *a/à*. La même diminution s'observe à la charnière du collège et du lycée, où l'absence de ces traces est signalée comme spécifique (niveau 2nde).

Il est intéressant de constater empiriquement, sur les écrits eux-mêmes, ce que les spécialistes du système éducatif observent depuis longtemps dans les résultats des élèves et les entretiens avec les acteurs de l'enseignement (Bonnery, 2007 ; Manesse, 2009) : le passage de l'école primaire au collège mais aussi celui du collège au lycée fragilisent certaines compétences en construction depuis plusieurs années. Sans doute les modalités de l'enseignement de la langue sont-elles à questionner dans leur continuité – ou son absence – d'un cycle à l'autre (Bishop & Cadet, 2011). Sur l'exemple que nous venons d'étudier, il est très net que ces passages correspondent à des moments de déstabilisation de la graphie des homophones *a* et *à*, qui sont pourtant parmi les premiers à être travaillés à l'école primaire[13].

On ne saurait évidemment tirer des conclusions définitives d'une étude qui porte sur un matériau de taille encore réduite au regard de l'ensemble des niveaux scolaires considérés, et ne peut à ce titre être considéré comme représentatif. Les dimensions de ce travail n'ont pas permis, en outre, d'excéder le traitement de l'orthographe. Cette brève analyse aura, je l'espère, donné un aperçu des investigations ouvertes par la transcription génétique que double une annotation orthographique des écrits scolaires, et de l'intérêt pour la didactique de s'intéresser au geste de rectification de l'écrit chez les élèves et à la visibilité de leur activité métalinguistique que donnent à voir les ratures.

Références bibliographiques

Authier-Revuz, Jacqueline. 1995. *Ces Mots qui ne vont pas de soi. Boucles méta-énonciatives et non-coïncidences du dire*. Paris : Larousse. Rééd. Lambert-Lucas, 2013.

Authier-Revuz, Jacqueline. 2020 *La Représentation du Discours Autre : principes pour une réflexion*. Berlin : De Gruyter.

Bellemin-Noël, Jean. 1972. *Le texte et l'avant-texte*. Paris : Larousse.

Bishop, Marie-France & Cadet, Lucile (a cura di). 2011. *Continuités et ruptures dans l'enseignement de la langue. Le français aujourd'hui*. 173

Bonnery, Stéphane. 2007. *Comprendre l'échec scolaire. Elèves en difficulté et dispositifs pédagogiques*. Paris : La Dispute.

Boré, Catherine & Elalouf, Marie-Laure. 2017. Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles. *Corpus* [En ligne], 16. URL : http://journals.openedition.org/corpus/2731

Brissaud, C. & Chevrot, J.-P. (2000). « Acquisition de la morphographie entre 10 et 15 ans : le cas du pluriel des formes verbales en /E/ ». Verbum, XXII(4), p.425–439.

Bronfenbrenner Urie (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, Massachusetts, and London: Harvard University Press.

[13] Une interprétation des erreurs pourrait être la traditionnelle faute d'attention. Ce serait une explication plausible pour une ou deux copies, mais cela ne peut pas expliquer les tendances lourdes que montre notre graphique réalisé sur 1500 copies.

Cori Marcel, David Sophie, Léon Jacqueline. 2008 Présentation : éléments de réflexion sur la place des corpus en linguistique. Langages 171. 5–11.

David, Jacques & Brissaud, Catherine & Guyon, Odile. 2006. Apprendre à orthographier les verbes : le cas de l'homophonie des finales en /E/. *Langue française*. 151. 109–126.

David, Jacques & Doquet, Claire & Fleury, Serge (a cura di). 2016. *Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement*. *Corpus*, n°16.

Doquet, C. (2011) *L'Ecriture débutante. Pratiques scripturales à l'école élémentaire.* Rennes : Presses Universitaires de Rennes.

Eshkol-Taravella, Iris & Lefeuvre-Halftermeyer, Anaïs. 2017. Linguistique de corpus : vues sur la constitution, l'analyse et l'outillage. *Corela* [En ligne] HS-21. URL : http://journals.openedition.org/corela/4800.

Fabre, Claudine. 1987. *Les Activités métalinguistiques dans les écrits scolaires*. Paris : Université Descartes Paris V. (Thèse de Doctorat d'Etat ès Lettres)

Fabre-Cols, Claudine. 2002. *Réécrire à l'école et au collège*. Paris : ESF.

Grésillon Almuth. 1994. *Éléments de critique génétique. Lire les manuscrits modernes*. Paris, PUF.

Gunnarsson-Largy, Cecilia & Auriac-Slusarczyk, Emmanuelle. 2013. Présentation du corpus Grenouille. C. In Gunnarsson-Largy, & E. Auriac-Slusarczyk (a cura di) *Ecriture et réécriture chez les élèves. Un seul corpus, divers genres discursifs et méthodologies d'analyse*. Louvain-la-Neuve : Academia-Bruylant.7–14.

Manesse, Daniele. 2009. L'orthographe des adolescents : le cas des élèves en grande difficulté au collège. *Langage et Pratiques*. 19–29.

Rey-Debove, Josette. 1982. Pour une lecture de la rature. In Fuchs, Catherine et *al*. (a cura di) *La Genèse du texte : les modèles linguistiques*. Paris : Hachette-CNRS. 21–72.

Siepmann, D., Bürgel, C., Diwersy, S. (2016) Le Corpus de référence du français contemporain (CRFC), un corpus massif du français largement diversifié par genres. SHS Web of Conferences. 27. 11002. 10.1051/shsconf/20162711002.