CORPUS

Corpus

16 | 2017
Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement

Problèmes posés par la transcription et l'annotation d'écrits d'élèves

Claire Doquet, Vanda Enoiu, Serge Fleury et Sara Maziotti

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/2776
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 1 janvier 2017
ISBN : 16638-9808
ISSN : 1638-9808

Référence électronique
Claire Doquet, Vanda Enoiu, Serge Fleury et Sara Maziotti, « Problèmes posés par la transcription et l'annotation d'écrits d'élèves », *Corpus* [En ligne], 16 | 2017, mis en ligne le 06 janvier 2018, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/2776

Problèmes posés par la transcription et l'annotation d'écrits d'élèves

Claire Doquet, Vanda Enoiu, Serge Fleury et Sara Maziotti

Introduction

1 L'analyse linguistique d'écrits scolaires a déjà été largement pratiquée, en particulier dans le cadre de l'Institut National de la Recherche Pédagogique où elle était posée à ses débuts comme devant construire une image de la pratique réelle de la langue écrite par les élèves qui donnerait aux enseignants « la possibilité de passer d'évaluations normatives à des évaluations objectives faites selon des critères positifs » (Romian, 1976 : 10). Ce sont donc, à l'origine, des questions didactiques qui pilotent les travaux linguistiques sur les textes d'élèves. Les moyens utilisés alors, aux balbutiements de *l'analyse automatique du discours* (Pêcheux, 1969), ont permis des investigations outillées par l'informatique naissante et devant déboucher, par exemple, sur « une description précise de la langue des élèves que les instituteurs classent comme « bons élèves » ou comme « élèves médiocres ». La question de la norme enseignante, toujours d'actualité, est centrale dans les travaux sur l'évaluation des écrits qui émaneront de l'étude des textes (Mas *et al.*, 1991).

2 Quatre décennies plus tard, l'écrit produit dans le système scolaire est mieux connu et l'outillage informatique, très performant, permet des investigations multiples. Dans le double mouvement d'une didactique de l'écriture qui exploite certaines voies ouvertes par la linguistique et d'un outillage linguistique qui s'adapte aux productions de scripteurs débutants (Masseron, 2011), des équipes de linguistes, dont le présent numéro reflète la diversité, s'intéressent aux écrits des élèves non seulement pour les décrire en vue d'outiller la didactique de l'écriture mais pour y repérer des traits du système linguistique faisant difficulté, que l'aisance des scripteurs avancés a enfouis mais que mettent au jour les tâtonnements et les erreurs des débutants. Dans le cadre

large d'une linguistique descriptive des usages, il importe de mettre au jour, comme l'ont déjà fait certains travaux concernant l'orthographe (Brissaud, 1998 ; Cogis, 1999), les régularités de la mise en fonctionnement du système de la langue par des scripteurs en apprentissage. Ce repérage ne peut se passer de la constitution de corpus qui, par leur taille, puissent prétendre à une représentativité. C'est ce type de corpus que le laboratoire Clesthia (EA 7345) a commencé de constituer dans le cadre de son opération de recherche *Analyse linguistique de l'écriture scolaire. ECRISCOL*[1] (Écriture Scolaire) est un corpus d'écrits d'élèves prélevés à différents âges de la scolarité correspondant aux trois paliers du socle commun de connaissances, de compétences et de culture et à la fin de la scolarité secondaire :

3 - cours Élémentaire 1, élèves de 7-8 ans ;

4 - cours Moyen 2, élèves de 10-11 ans ;

5 - classe de 3^{e} / début 2nde ;

6 - élèves de terminale / entrée à l'université.

7 L'objectif est de recueillir un grand nombre d'écrits produits dans des contextes d'apprentissage variés pour en analyser les caractéristiques linguistiques et discursives. La recherche s'appuie sur le prélèvement, dans des classes d'école, de collège et de lycée ainsi qu'à l'entrée à l'université, d'écrits produits selon des consignes comparables, afin de tracer une cartographie des compétences selon les différentes variables (socio-culturelles, didactiques, identitaires) prises en compte. L'ensemble ainsi constitué donnera une visibilité aux textes de scripteurs à différents niveaux d'apprentissage et de maîtrise de l'écrit. Ce corpus devrait permettre de décrire d'une part l'évolution des habiletés scripturales au cours de l'apprentissage, d'autre part la construction même de ces habiletés et le lien qu'elles peuvent avoir les unes avec les autres, pour interroger les catégories traditionnelles de l'étude de la langue (orthographe, morphosyntaxe, lexique) dans leurs interactions à différents stades de maîtrise de l'écrit.

8 La procédure se décompose en quatre temps :

9 - Délimitation et constitution d'un corpus de référence permettant d'observer, à partir de données quantitativement représentatives, les caractéristiques des écrits selon l'âge et les compétences des élèves.

10 - Création d'une plate-forme d'accès (*Open Access*) aux données (les sources, les corpus construits) et aux analyses, avec possibilité pour les chercheurs de télécharger tout ou partie des données.

11 - Analyse systématique, à l'aide de logiciels de textométrie et en particulier *Le Trameur* (présenté ci-dessous), des caractéristiques linguistiques des textes (analyse contrastive selon les compétences des scripteurs et les consignes d'écriture) concernant les différents aspects de l'écriture : orthographe, lexique, morpho-syntaxe, continuité thématique, ponctuation, connecteurs, discours rapporté, etc.

12 - Corrélation avec les caractéristiques pédagogiques et situationnelles de l'écriture (accompagnement par l'enseignant, écriture individuelle ou collective, outils à disposition, etc.) et recommandations sur les variables didactiques les plus efficientes selon le type de texte à produire et les savoir-faire à acquérir.

13 Notre recherche repose sur la mise en œuvre de protocoles spécifiques permettant aux logiciels de textométrie de traiter de manière efficiente les textes des élèves, qui ont

des caractéristiques linguistiques et discursives spécifiques (constructions asyntaxiques, stéréotypes lexicaux, écarts orthographiques, modes d'énonciation idiosyncrasiques…). Après une brève présentation du matériau en cours de réunion, nous présentons ici trois volets du travail de traitement : la transcription, l'annotation et les modes d'analyse des corpus.

1. Constitution du corpus et principes de traitement

1.1 Corpus, données, matériau

14 Pour les écrits d'élèves comme pour tout type de matériaux, la notion de corpus elle-même est au centre du travail de recueil des données. Il eût été facile, tout chercheur en didactique de l'écriture en a dans ses tiroirs, de réunir au petit bonheur des écrits d'élèves que nous aurions juxtaposés pour former un ensemble propre à la lecture, voire à l'exploration textométrique. Le premier numéro de *Corpus* visait à problématiser cette question : si « *corpus* renvoie, en un premier sens, à une collection de textes présentant une certaine unité de genre ou bien d'époque » (Dalbera, 2002), le corpus scientifique peut être défini comme un « ensemble de données sélectionnées et rassemblées pour intéresser une même discipline » (Mellet, 2002) et, en sciences du langage, comme « un ensemble d'éléments sur lequel se fonde l'étude d'un phénomène linguistique » (Dalbera, *ibid.*). Comme le soulignait alors S. Mellet, les corpus sont généralement construits et utilisés dans deux cadres complémentaires : les corpus clos et exhaustifs permettent l'étude d'un ensemble de données qui constitue le corpus lui-même, ce dernier ne reflétant rien d'autre que lui-même ; les corpus échantillonnés, au contraire, sont conçus comme reflétant des ensembles beaucoup plus vastes, souvent non *finis*, ce qui pose le problème de leur représentativité. C'est un corpus du second type qui est visé par notre travail, un « objet intermédiaire entre les faits empiriques et le modèle théorique » (Mellet, 2002) destiné à servir, pour différentes recherches, de corpus de référence de l'écriture scolaire.

15 Le corpus réuni jusqu'ici comporte 1 225 textes accompagnés de leurs avant-textes (notes, brouillons, essais divers ayant contribué à l'écriture), ce qui constitue un ensemble de près de 2 300 textes, soit 35 000 mots environ. Le travail de traitement, pour l'instant effectué manuellement, fait correspondre à chaque élément, trois fichiers :

16 - fichier image reproduisant le manuscrit ;

17 - fichier texte comprenant la transcription du manuscrit, dont les principes sont explicités ci-après (partie 2) ;

18 - fichier texte comprenant la transcription annotée, également explicitée plus loin (partie 3).

19 Les deux premiers types de fichier sont destinés à être lus par les utilisateurs de la base de données projetée ; le dernier, la transcription annotée, ne sera pas montré aux utilisateurs mais servira de socle à l'élaboration des fichiers à traiter informatiquement (cf. ci-après, partie 4).

1.2 Un corpus écologique

20 Les écrits recueillis doivent donc, d'abord, refléter la réalité de ce qui se produit dans le cadre scolaire. Il ne s'agit pas, par conséquent, de mettre en œuvre un protocole expérimental mais d'aller chercher dans les classes des écrits tels que le quotidien de l'enseignement permet de les faire émerger. Entre données expérimentales et données écologiques, nous avons choisi les secondes, avec tout de même des aménagements permettant de rendre comparables des éléments qui proviennent de lieux, de niveaux, de contextes socio-économiques et didactiques différents. Dans le souci de donner une certaine homogénéité à nos données, que nous avons souhaité recueillir à différents âges de la scolarité et à différents niveaux de la compétence scripturale, nous proposons à tous une tâche similaire, à savoir une continuation de récit. Cette tâche, très courante en fin d'école primaire et au collège, permet d'évaluer nombre de savoir-faire, y compris lorsqu'ils excèdent la simple rédaction : capacité à insérer son écriture dans un écrit existant, continuités thématique et énonciative, appréhension et appropriation d'un style d'auteur… Surtout, cette consigne constitue un point de départ à la fois contraignant en terme de genre d'écrit (de la narration) et laissant libres les enseignants de choisir le texte narratif de départ et l'ensemble des conditions didactiques.

21 Le recueil de textes selon un protocole aussi peu contraint pose des problèmes d'hétérogénéité. Il est très difficile, dans des conditions contextuelles différentes, de savoir à quoi imputer les écarts constatés lors des analyses. Pour autant, nous persistons à ménager la liberté des enseignants non seulement par souci de recueillir des productions conformes à celles qui seraient produites hors recherche, mais aussi pour pouvoir, à terme, faire intervenir la didactique, avec toutes ses variations, comme variable explicative de ces écarts. C'est la raison pour laquelle nous avons ajouté deux épreuves test, à faire passer en début et en fin d'année, qui sont également des continuations de récits que nous demandons aux enseignants de faire produire à leurs élèves en énonçant la nature de ces tests, proposé sans environnement didactique particulier. Ces productions servent à étalonner les élèves et les classes, pour évaluer de manière plus juste les productions réalisées dans le contexte habituel des séquences d'enseignement.

1.3 Traces de l'écriture

22 Au-delà des textes eux-mêmes, nous souhaitons accéder à la reconstitution de l'écriture telle qu'elle se donne à lire à partir des ratures et de l'ensemble des interventions, verbales ou non, opérées sur les différents états du texte. Il devient alors possible d'analyser automatiquement ces traces de réécriture, et plus largement de commentaires (soulignements de segments par exemple), en travaillant systématiquement, par exemple, sur les segments supprimés, ou bien sur ceux qui ont été ajoutés a posteriori, ou encore sur les écarts entre les brouillons et les textes finaux. Nous nous situons, pour l'ensemble de cette étude, dans la lignée des travaux de C. Fabre-Cols, qui est la première à avoir appliqué aux brouillons d'élèves la méthode d'analyse élaborée par les généticiens du texte (Fabre, 1990 ; Fabre-Cols, 2002). Par la suite, C. Boré (1998) et C. Doquet (2011) ont également travaillé dans ce sens, parfois sur des matériaux différents.

23 La base de données en construction met en relation les différents états d'un même texte, les différents textes d'un même élève, les textes des différents élèves d'une même classe… mais elle permettra aussi, grâce aux métadonnées collectées, d'opérer des tris selon le sexe, l'âge, la catégorie socio-professionnelle des parents, les langues parlées à la maison, etc. Deux versions de chaque texte sont présentées à l'utilisateur : la copie scannée et sa transcription. Nous présentons et transcrivons les textes tels qu'ils ont été écrits, c'est-à-dire avec les erreurs. Ce choix est lié à la nécessité, selon nous, d'accéder à un matériau authentique, y compris dans ses écarts à la norme et dans les difficultés de lecture qui peuvent en découler. Les erreurs sont constitutives de la langue écrite des élèves et nous pensons que les gommer fausserait son appréhension. De plus, puisque notre analyse est aussi génétique, il est très important de pouvoir examiner, outre les erreurs, la manière dont les élèves les repèrent et les corrigent. Comme l'a souligné C. Fabre, les rectifications dites formelles sont souvent, pour des élèves qui sont en train d'apprendre à écrire, la seule manifestation possible de leur activité métalinguistique :

> si l'écriture se pratique comme un tout, il importe de ne pas mésestimer le signifiant graphique, certes souvent survalorisé en situation scolaire : des continuités peuvent exister entre les modifications « superficielles » et celles qui le sont moins. […] Plutôt que d'évacuer les ratures orthographiques comme extérieures aux fonctionnements « profonds » de l'écriture, nous croyons qu'il serait pertinent d'éclairer davantage leur liens avec ceux-ci, et de poser comme hypothèse large que la « conscience du texte » […] bute ou prend appui sur la mise en graphie. (Fabre, 1987 : 579)

24 À partir de l'ensemble de ces principes un des aspects les plus importants de notre travail a été jusqu'ici, outre le recueil lui-même, l'élaboration collaborative d'un modèle de transcription et d'annotation des écrits ; nous avons aussi amorcé la mise en œuvre de ces choix dans le codage des données à explorer par des logiciels d'analyse textuelle. Puisque notre choix initial consiste à mémoriser les traces de l'écriture dans les productions écrites visées, cela nous conduit à mettre en place un modèle de représentation de ces différentes strates, qui soit compatible avec les formats permis dans les outils permettant leur exploration informatique ultérieure.

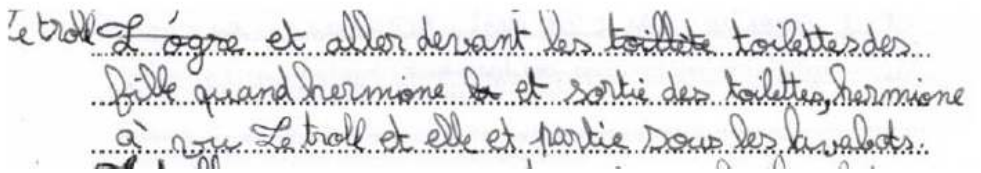
25 La transcription et l'annotation, dont nous rendons compte dans les lignes qui suivent, ont déjà fait l'objet de maintes réflexions et modifications mais restent à améliorer.

2. La transcription

2.1 Modèles théoriques pour la transcription des manuscrits

26 Comme le rappelle P.-Y. Testenoire ici-même[2], il existe théoriquement trois modèles de transcriptions : la transcription diplomatique, la transcription linéaire (ou *linéarisée*) et la transcription chronologique (ou *chronologisée*). La transcription diplomatique est la moins interprétative des trois puisqu'elle se contente de restituer dans un espace graphique ce qui figure sur un autre espace graphique, en tentant de respecter l'ensemble des marques, telle une photographie. Au contraire, la transcription linéarisée repose sur l'interprétation en remettant sur un axe linéaire la succession des opérations d'écriture, selon une chronologie reconstruite.

27 Considérons, à titre d'exemple, cet extrait de copie :



28 Ici, une transcription linéaire pourrait indiquer sans risque qu'à la fin de la première ligne, que l'élève a écrit « aller devant les toillète », biffé « toillète » et inscrit à sa place « toilettes ». Mais comment traiter la première opération visible, la biffure de « L'ogre » et l'insertion de « Le troll » ? Si l'on est certain, du fait de la position des GN sur la ligne, que « L'ogre » figurait avant et qu'il s'agit bien d'un remplacement, rien ne certifie du moment auquel il a eu lieu ; toute transcription linéaire serait, par conséquent, un choix interprétatif du transcripteur. C'est la raison pour laquelle J.-L. Lebrave (1990 : 148) a reproché à la transcription linéaire de ne proposer qu' « une interprétation univoque de la chronologie du manuscrit, qui devient contraignante si l'utilisateur n'a pas simultanément accès au document source ». Pour tenter de pallier ce problème, il a élaboré une méthode de transcription, dite *chronologique*, privilégiant la restitution des données temporelles et mettant au jour les différentes strates de l'écriture ; lui-même a présenté récemment ce travail en expliquant ses limites, mais aussi toute sa pertinence : « dès qu'on s'attaque à des manuscrits complexes, comme ceux de Flaubert, il devient rapidement évident qu'une reconstitution exhaustive de toutes les opérations, dans l'ordre où elles sont apparues, est impossible. [...] En revanche, la reconstitution partielle d'opérations « locales » dans un fragment de manuscrit reste parfaitement possible » (Lebrave, 2009 : 16). C'est ce qu'a fait I. Fenoglio, par exemple, dans son étude sur les manuscrits autobiographiques d'Althusser où elle compare précisément la genèse d'extraits de deux textes du même auteur (Fenoglio 2002), c'est également ce que nous pourrons faire pour aider à la lecture de brouillons courts et très surchargés, mais de manière générale, nous suivons les linguistes généticiens pour privilégier la transcription diplomatique, à laquelle F. Masai assignait la mission de « reproduire fidèlement le document tel qu'il est sorti de l'officine productrice » (Masai, 1950 : 187). De la transcription diplomatique philologique telle que défendue par F. Masai à celle que pratique la génétique textuelle (Grésillon, 1994), si les objets de recherche changent, la technique demeure, même si les travaux contemporains relativisent la « fidélité » de la reproduction des manuscrits : d'après A. Crasson et J.-D. Fedeke (2007) « la transcription diplomatique *photographie* le document en rapportant, avec les outils qui le permettent, malgré leurs limites, tous les événements du manuscrit ». Par *événement du manuscrit*, il faut entendre l'ensemble des traces que laisse l'activité d'écriture, y compris par exemple des dessins ou la couleur de l'encre, mais aussi, bien entendu, les ratures : biffures, segments textuels hors ligne ou en marge, qui permettent au lecteur de reconstituer des opérations scripturales : l'ajout, la suppression et leurs composés, le remplacement et le déplacement (Grésillon, 1994). En l'absence de convention partagée pour les transcriptions, nous avons choisi de nous conformer au codage souvent utilisé par l'Institut des Textes et Manuscrits (ITEM), qui dérive de celui des philologues :

29 - un segment ajouté est présenté <entre chevrons>

30 - un segment supprimé est présenté [entre crochets][3]

31 La combinaison de ces deux opérations de base permet de coder les remplacements et les déplacements.

2.2 Complexité des écrits d'élève

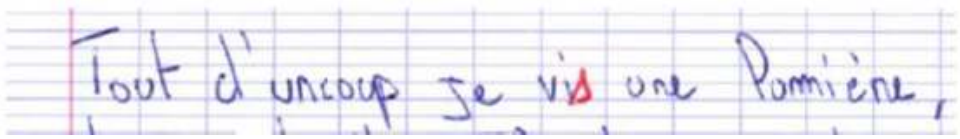
32 Ces principes posés, la transcription des écrits d'élèves n'est pas aisée pour autant. L'ajout, la biffure et le remplacement d'une lettre, d'un mot ou d'une phrase ne sont que quelques-uns des processus que nous pouvons rencontrer dans un texte d'élève, et il n'est pas rare qu'ils se présentent combinés ou bien accumulés. Par exemple, un élément peut être biffé ou ajouté par l'élève, mais aussi par l'enseignant ; il peut l'être au moment de l'écriture, mais aussi ultérieurement, comme le signale par exemple une encre différente ; un remplacement peut se composer d'une suppression par l'enseignant suivi d'un ajout par l'élève, etc. Le protocole de transcription doit donc coder chaque processus selon une syntaxe facilement identifiable et qui demeure claire, même en cas d'accumulation de modifications sur le même segment.

33 Voyons cet extrait d'une copie d'élève de fin d'école primaire (10-11 ans) :

34 *[Tout a] Nadine et Nathacha <T2#Natacha> avait de plus en plus peur,*

35 Le segment biffé en début de ligne est entre crochets. La mention T2# entre les chevrons marquant l'ajout indique qu'il s'agit d'une opération différée dans le temps par rapport à l'écriture initiale.

36 Dans l'exemple suivant, c'est le professeur qui, en corrigeant, utilise l'encre rouge :

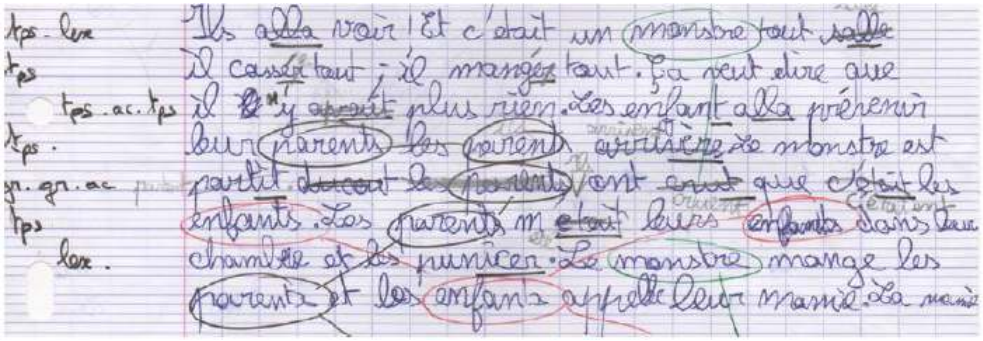


37 Ce segment sera transcrit :

38 *Tout d'uncoup je vi<P#s> une lumière,*

39 Les interventions de l'enseignant sur la copie de l'élève, très fréquentes, constituent souvent une co-écriture : la plupart du temps, en particulier dans les niveaux les plus bas de l'enseignement, les enseignants interviennent sur les brouillons et les élèves se livrent ensuite à une deuxième écriture, sur la base des modifications suggérées.

40 La transcription relativement simple proposée ici, qui revient à marquer à la fois les interventions de l'enseignant et les interventions en différé, trouve ses limites dans des écrits très annotés, qui ne sont pour autant pas rares, comme celui-ci, où l'enseignant lui-même utilise plusieurs couleurs d'encre pour mettre de l'ordre dans ses annotations :



41 Dans ce cas, il est pratiquement impossible de transcrire l'ensemble des annotations sans nuire – le mot est faible – à la lisibilité du texte.

2.3 Une transcription qui prépare l'annotation

42 Malgré les efforts pour choisir dans le protocole une syntaxe spécifique à chaque processus, plusieurs questionnements continuent à se poser. Les modifications fréquentes, qui ont affecté notre protocole à distance de quelques mois, témoignent une sorte de prolongement de l'étape que nous croyons exclusivement antécédente à la transcription. En effet, lors du processus de transcription, le nombre de cas particuliers, qui exigeaient une réflexion plus spécifique, nous a amené à définir plus clairement nos priorités face aux processus que nous voulions retrouver dans la transcription. Néanmoins, le maniement du protocole n'était pas dû exclusivement aux désirs des linguistes, car notre but étant celui de produire des transcriptions avec une syntaxe la plus proche possible de la syntaxe des annotations, afin de parvenir, *via* un processus automatique, à une base analysable par le logiciel de textométrie *Le Trameur*.

43 La transcription a l'objectif de rester fidèle au texte et de reproduire tous les éléments et les étapes que nous pouvons repérer dans le texte, y compris les erreurs d'orthographe, car la forme normée ne sera introduite que dans les annotations. Si nous analysons un segment de texte écrit en temps 1 par l'élève dans une copie de CM2, nous nous apercevons de la richesse de processus qui s'enchaînent et qui correspondent au temps d'écriture 1 de l'élève, suivi par le temps d'intervention de l'enseignant, qui souligne une partie de certains mots, et un temps 2 pendant lequel l'élève essaie de corriger ses fautes d'orthographe. La transcription ressemblera à :

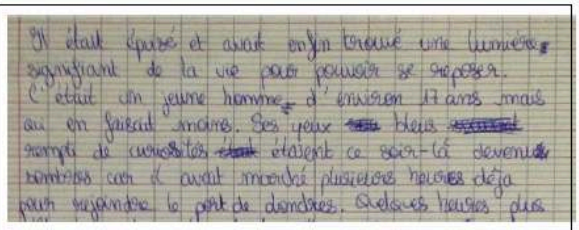
44 *Gégé et Max se retoun{Pσ#e} <T2#retournent> et puis vo{Pσ#i}[s] <T2#voient>*

45 où « P » se réfère à une intervention accomplie par l'enseignant, le sigma au processus de soulignement et les parenthèses au mot ou aux lettres soulignés par l'enseignant. Le dièse a été choisi comme délimiteur qui sépare l'action ou l'indication du temps d'écriture du mot ou des lettres concernés par l'action. Dans l'exemple, donc, l'enseignant souligne le « -e » final du verbe mal conjugué et dans un deuxième temps, l'élève se corrige en ajoutant au-dessus de « retoune » le verbe au pluriel « retournent ». Le même processus se répète avec le verbe « vois » où, en réalité, nous ne savons pas si l'élève a effacé le « -s » en temps 1 ou en temps 2, car il était écrit avec un stylo mais gommé ensuite.

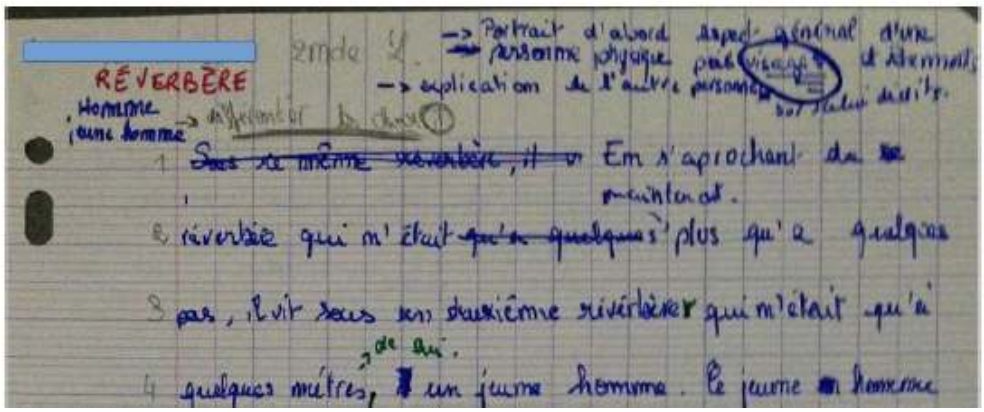
46 Le passage de la transcription à l'annotation est à l'origine de la plupart de nos questionnements concernant la façon dont nous pouvions rester fidèles au texte et, en même temps, respecter les contraintes du logiciel *Le Trameur*, afin que ce dernier puisse

exploiter les données. Certains éléments, finalement exclus du traitement automatique, ont été au cœur de longs débats : les dessins, les essais orthographiques en marge, les tableaux et les listes de mots, présents surtout dans les brouillons, ne sont pas transcrits et ne feront donc pas partie de la base de données exploitable par le *Trameur*, même si dans les transcriptions le type d'élément à repérer directement dans le scan de la copie originale sera indiqué dans les crochets.

47 À titre illustratif, nous présentons ici deux exemples de transcriptions, accompagnées de la reproduction des manuscrits correspondant. Le premier est un texte peu modifié, donc facile à transcrire :

Manuscrit	
Transcription	Il était épuisé et avait enfin trouvé une lumière [VIRGULE] signifiant de la vie pour pouvoir se reposer.§ C'était un jeune homme [VIRGULE] d'environ 17 ans mais qui en faisait moins. Ses yeux [#XXX#] bleus [#XXX#] <rempli>_<remplis> de curiosités ®[était]//étaient//® ce soir-là devenu[s] sombres car il avait marché plusieurs heures <déja>_<déjà> pour rejoindre le port de Londres. Quelques heures plus

48 Voici à présent un brouillon beaucoup plus complexe, en particulier à cause des annotations marginales qui sont des auto-consignes, à la fois présentes sur la page et non destinées à s'intégrer au texte. Ce type d'énoncé est noté entre accolades. On trouve également dans cet extrait des segments inscrits par l'enseignant (en vert) et des segments inscrits par l'élève au moment d'une relecture (au crayon gris).



49 Ce brouillon sera transcrit ainsi :

<REVERBERE>
{<Homme$
jeune homme><T2#→différencier_les_deux>}
{<→Portrait_d'abord_aspect_général_d'une_
Personne_physique_puis_visage_ et_vêtements
→explication_de_l'autre_personne_sur_XXX>}
[Sous_ce_même_réverbère,_il_v] En s'approchant d[e_ce]
réverbère qui n'était [qu'a_quelques_pas] //maintenant// plus qu'a quelques
pas, il vit sous un deuxième réverbère[P#s], qui n'était qu'à
quelques mètres, [P#de_lui] un jeune homme. Ce jeune [#XXX#] homme

3. L'annotation

50 L'annotation correspond à la deuxième étape dans le traitement d'une copie d'élève. Elle consiste dans l'ajout d'une couche supérieure d'informations – en fait, la correction des erreurs d'orthographe de l'élève. Il s'agit donc d'une étape qui se situe dans la continuité de la transcription. On peut même se demander dans quelle mesure ce que nous appelons *transcription* n'est pas déjà une annotation, dans la mesure où – cela a été montré à travers les exemples précédents – la transcription est toujours le résultat de choix du transcripteur.

51 Comme nous l'avons signalé, le fichier annoté sert de base au fichier XML qui sera exploité par les logiciels d'analyse textuelle ; il est donc important de respecter les codages habituels en XML. Cela pose un problème concernant les chevrons, qui sont à la fois utilisés dans les transcriptions diplomatiques des généticiens et dans l'écriture XML, avec deux significations différentes : signes d'ajout en génétique, ils sont en XML des marqueurs de balises. Ce signe double va donc changer de signification avec le passage de la transcription à l'annotation. Les fichiers annotés n'étant pas destinés à être lus par les utilisateurs des textes, nous avons affecté aux chevrons la valeur de balise et trouvé un autre signe, en l'occurrence une double barre oblique, pour signifier les ajouts. À titre d'exemple, le segment suivant dans une transcription :

52 *Portrait général <mais précis> de l'héroïne*

53 s'écrira, dans le fichier annoté correspondant, de la manière suivante :

54 *Portrait général //mais précis// de l'héroïne*

55 Le passage d'une copie transcrite à une copie annotée comporte donc les étapes suivantes :

56 - Premièrement, remplacer des chevrons utilisés pour l'ajout par des barres obliques (<mots ajoutés> par //mots_ajoutés//, comme dans l'exemple :

57 *Gégé et Max se retoun{Pσ#e} //T2#retournent// et puis vo{Pσ#i}[s] //T2#voient//.*

58 - Deuxièmement, ajouter le symbole ® pour mettre en évidence un cas de remplacement, comme dans l'exemple :

59 *veu[P#t]<P#x> tu qui devient : veu®[P#t]//P#x//® tu.*

60 Le remplacement des chevrons par des barres obliques dans l'annotation nous permet d'employer les chevrons pour les balises ajoutées manuellement, qui interviennent :

61 pour indiquer la forme normée en cas d'erreur d'orthographe ;

62 pour rétablir des majuscules après un signe de ponctuation démarquant des phrases ou dans le cas des noms propres, ou encore sur les signes de ponctuation lorsqu'on identifie la fin de la phrase, par exemple en présence d'une majuscule non précédée d'un point final.

63 Dans cette tâche nous nous sommes confrontés à quelques contraintes techniques liées surtout au traitement de texte opéré par logiciel. Par exemple, la nécessité de placer un tiret bas à l'intérieur d'un segment annoté qui est composé de plusieurs unités ou de transcrire en lettres capitales tout symbole présent dans l'annotation (VIRGULE, POINT, POINT_D_INTERROGATION) :

64 *qu[P#']il annoté : <qu[P#APOS]il>_<qu_APOS_il>*

65 L'annotation permet de mettre en évidence deux segments délimités par des chevrons qui correspondent respectivement à l'élément à corriger et à la proposition de forme normée de l'annotateur. Dans le cas d'accumulation de processus à l'intérieur du premier segment, la présence d'une correction de la part de l'annotateur prive de tout autre processus, permet au logiciel d'identifier la forme normée :

66 *<veu®[P#t]//P#x//®_tu>_<veux_TIRET_tu>*

67 D'autres difficultés sont posées par les copies, pour lesquelles nous avons dû établir différentes possibilités de transcriptions :

68 L'enseignant, en corrigeant, ne propose pas la forme attendue : toupi[P#s] sera annoté :

69 *<toupi[P#s]>_<toupie>.*

70 L'élève produit une série de paradigmes ou essais orthographiques : il est nécessaire de maintenir les différentes propositions, entre lesquelles l'élève n'a pas choisi, et de permettre l'analyse automatique. Ce cas n'est pas encore totalement réglé.

71 L'élève et l'enseignant proposent plusieurs formes en temps (avec des soulignements ou ajouts). Dans ces cas, le segment de gauche doit réunir les différentes propositions qui doivent ultérieurement se retrouver sous une forme normée dans le segment de droite. Nous sommes ainsi confrontés à un choix, en fonction de ce que nous voulons mettre en évidence, surtout dans le cas des paradigmes. Il faut également préciser que les corrections peuvent accumuler le symbole de l'ajout, d'une rature ou d'une modification de l'enseignant qui se retrouve dans la séquence de gauche :

72 *<//T2#voules//>_<voulait>*

73 La stabilisation du protocole et la mise en correspondance de la transcription et de l'annotation sont encore en cours. L'apparition, au gré du recueil des copies, de cas problématiques qui n'ont pas été traitées auparavant conduit régulièrement à adapter, voire à reconfigurer le protocole initial. Même si chaque copie a ses propres particularités, l'intérêt de ce projet réside notamment dans la création et l'harmonisation des outils qui nous permettront d'analyser linguistiquement les écrits d'élèves.

4. Analyse outillée du corpus ECRISCOL avec Le Trameur

74 *Le Trameur* (http://www.tal.univ-paris3.fr/trameur/) est, au départ, un logiciel de textométrie. Traditionnellement les objectifs de la textométrie sont les suivants :

compter des unités dans les textes, dans les différentes parties d'un texte pour mesurer, contraster leurs comportements respectifs. La textométrie ne donne pas le sens, elle permet au mieux de le construire ; elle permet d'élaborer des parcours interprétatifs guidés par l'examen des différentes configurations des unités visées dans leurs réalisations textuelles. La démarche textométrique repose principalement sur l'observation des variations de fréquence d'unités textuelles (formes, lemmes, etc.) appelées *contenus* textuels, dans les différentes parties d'un ensemble de textes, considérées comme des *contenants* textuels (parties, sections, zones, chapitres, paragraphes, phrases, séquence, etc.) (Söze-Duval, 2008). Une description formelle de ces deux systèmes d'unités (*contenants* et *contenus*) permet d'obtenir, à l'aide de procédures informatisées, des décomptes sous forme de vastes tableaux statistiques. La textométrie mobilise des méthodes d'Analyse de Données Textuelles afin d'étudier la répartition statistique des *contenus* au sein des *contenants* des corpus textuels. Les synthèses statistiques qui en résultent sont des points de départ pour la mise en évidence des principales dimensions des corpus analysés.

75 La mise au jour des unités à compter se réalise par la segmentation de la chaîne textuelle : à partir d'un texte découpé en items, on constitue un système de coordonnées (*Trame*) où chaque item est repéré par son numéro d'ordre dans la chaîne textuelle. Chaque *item* isolé peut recevoir une étiquette (lemme, catégorie grammaticale, syntaxique, sémantique, stylistique, etc.) au cours d'une ou plusieurs opérations d'annotation. Les items semblables sont rattachés à un même *type* (à partir des propriétés de leur forme intrinsèque ou sur la base de certaines annotations). Les *types* deviennent ainsi des unités génériques dont on peut recenser les occurrences. Les empans textuels (parties ou contenants) sont indexés sur la *Trame* comme suites d'*items* consécutifs, entre la position x_1 et la position x_2. Les systèmes de *contenants* du corpus sont regroupés dans une structure de données appelée *Cadre*. Une ressource textuelle constituée sous la forme de *Trame/Cadre* est une *base textométrique* (Fleury, 2013).

76 *Le Trameur* met en avant la possibilité de traiter des données annotées : (1) la création par le logiciel d'une nouvelle base textométrique peut se faire en invoquant un étiquetage morphosyntaxique automatique ; (2) l'importation d'une base textométrique (déjà construite par le logiciel ou en recueillant des sources externes) permet de lire un nombre aléatoire d'annotations projetées sur une *Trame* prédéfinie. Le logiciel permet en outre de créer dynamiquement de nouvelles annotations sur une *Trame* ou de corriger ces annotations. Les ressources stockées dans cette architecture sont donc modulables au regard des nécessités mises au jour par des parcours exploratoires donnés. Une fois la base chargée, toutes les annotations peuvent être utilisées dans les calculs permis par le logiciel.

77 Dans le cadre du projet *ECRISCOL*, les données traitées ont conduit à construire des bases textométriques dans un format respectant l'architecture *Trame/Cadre* du logiciel *Le Trameur*. Les différentes couches d'information mises au jour dans les étapes de transcription et d'annotation décrites précédemment sont réutilisées pour produire certaines strates (la forme initiale, la forme normée) de la base textométrique importable dans *Le Trameur*.

78 Une base *ECRISCOL* est un fichier au format XML : elle contient une description de la segmentation des textes en mot et pour chaque mot la liste de ses annotations. La figure précédente donne à voir un extrait de la base. Une telle base peut être vue comme un « mille-feuille », elle concatène ici 4 couches d'annotation, chacune d'elle constituant un flux textuel particulier. Dans le cadre de ce projet, ces 4 couches d'annotation sont les suivantes :

> Annotation n° 1 : forme initiale (avant correction éventuelle par exemple)
> Annotation n° 2 : soit le lemme construit *via TreeTagger*, soit l'opération de transformation réalisée sur la forme initiale
> Annotation n° 3 : catégorie construite *via TreeTagger*
> Annotation n° 4 : si la forme initiale est modifiée, cette annotation porte la forme finale (après correction par exemple), sinon elle porte la même valeur que l'annotation n° 1

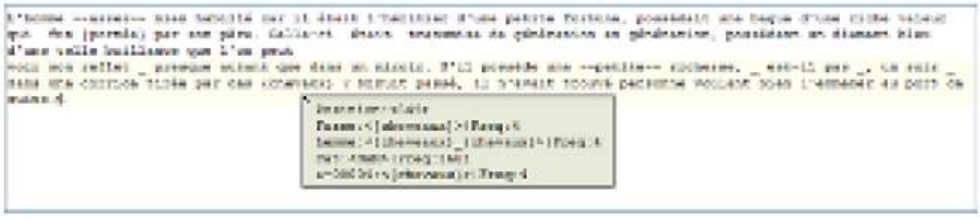
79 *Le Trameur* permet ensuite de donner à voir ces différentes « couches » textuelles dans ces différents modules d'édition du corpus. L'annotation n° 1 met au jour le texte dans sa version « initiale », ci-dessous une section du corpus :

```
L'homme _ bien habillé car il était l'héritier d'une petite fortune, possédait une bague d'une riche valeur qui
[était_avant] (porté) par son père. Celle-ci  [est]  transmise de génération en génération, possédant un diamant bleu
d'une telle brillance que l'on peut
[se_voir_#XXX#] son reflet [#XXX#_au] presque autant que dans un miroir. S'il possède une _ richesse,
[pourquoi_n_APOS_#XXX#] est-il pas (dans_accompagné_de_quelqu_APOS_un_et_pourquoi_il_lui_avait_fallu_marcher_seul], un
soir [#XXX#] dans une corride tirée par des (chevaux) ? Minuit passé, il n'avait trouvé personne voulant bien
l'emmener au port de suite.§
```

80 Cette même zone de texte peut-être vue du point de vue de l'annotation n° 4 (le texte final) :

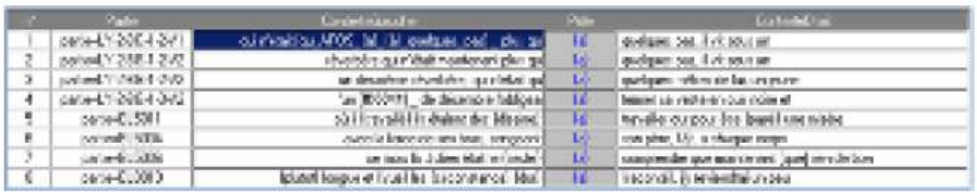
```
L'homme ~~assez~~ bien habillé car il était l'héritier d'une petite fortune, possédait une bague d'une riche valeur
qui  fut (portée) par son père. Celle-ci  était  transmise de génération en génération, possédant un diamant bleu
d'une telle brillance que l'on peut
voir son reflet _ presque autant que dans un miroir. S'il possède une ~~petite~~ richesse, _ est-il pas _, un soir _
dans une corride tirée par des (chevaux) ? Minuit passé, il n'avait trouvé personne voulant bien l'emmener au port de
suite.§
```

81 Chaque mot de cette section donnant à voir en permanence les différentes annotations qui lui sont associées :



82 Ci-dessous, même genre de contraste entre 2 couches textuelles de la base *ECRISCOL* (forme initiale vs forme finale) au regard de la concordance des unités dont la forme initiale est {a} (les parenthèses indiquent que le mot est initialement mal orthographié). La concordance de ces items peut être matérialisée via la couche textuelle (annotation n° 1) associée aux formes initiales des mots :

83 **Concordance (annotation n° 1)** : forme initiale {a}



84 ou via la couche textuelle portant leur forme finale (annotation n° 4) :

85 **Concordance (annotation n° 4)** : forme finale {à}

86 Les différents modules de calcul disponibles dans le logiciel permettent ensuite de travailler sur les différents niveaux d'annotation.

87 La démarche mise en œuvre pour le traitement des données du corpus ECRISCOL via *Le Trameur* a conduit dans un premier temps à consolider le processus de transformation des données transcrites et annotées vers des bases textométriques importables dans le logiciel. Cette première étape permet ainsi d'explorer les différentes couches d'information disponibles (pour par exemple quantifier les formes correctes *vs* incorrectes dans un texte du corpus) : les concordances ci-dessus illustrent ce type d'exploration. Nous souhaitons maintenant poursuivre les explorations des données du corpus en élaborant un jeu de bases textométriques issues du corpus (une classe par niveau par exemple) puis en les analysant sur des thématiques encore à définir.

5. Conclusion

88 Le travail exposé ici est en cours, et les conventions utilisées ne constituent en aucun cas des modèles : elles sont en effet amenées à évoluer, en fonction des cas rencontrés dans les écrits mais aussi de l'utilisation que l'on souhaitera faire de ce corpus. Tel qu'il est à l'heure actuelle, le corpus ECRISCOL permet d'observer différentes dimensions de l'écriture scolaire, des premiers brouillons des élèves aux écrits finaux, en passant par les commentaires des professeurs. Le traitement, progressif, de l'ensemble des documents recueillis permettra d'automatiser des explorations que nous avons pour

l'instant réalisées en grande partie manuellement, et d'associer à chaque texte les métadonnées correspondantes. Il sera donc possible, par exemple, de formuler des requêtes permettant de recenser, dans un ensemble d'écrits, l'ensemble des écarts à la norme orthographique, et de les trier selon la classe de mots concernées, et si c'est jugé utile, certaines caractéristiques sociologiques ou didactiques. On peut aussi observer, c'est le travail de Master de Sara Maziotti, des régularités entre d'une part la teneur et l'abondance des commentaires des enseignants sur les brouillons, et d'autre part les écarts entre ces mêmes brouillons et les textes finaux. Macro-syntaxe, orthographe, morphologie, reformulation... les voies d'exploration sont nombreuses et s'ouvrent au fur et à mesure des traitements effectués sur les données collectées.

89 L'exploration des écrits des élèves constitue un enjeu actuel de premier ordre qui concerne plusieurs disciplines :

90 - la didactique de l'écrit est la plus spontanément concernée par ce type de travail, la connaissance des performances réelles des élèves corrélée à certaines caractéristiques externes (données sociologiques, environnement didactique, etc.) devant permettre d'ajuster les interventions enseignantes aux questionnements et besoins des élèves ;

91 - l'outillage de l'analyse de corpus, confronté à la diffusion régulière d'écrits non standards[4], doit pouvoir s'appuyer sur des modules d'analyse adaptés aux écarts à la norme qui spécifient les différents corpus ;

92 - la linguistique française a également à apprendre des usages non normés de la langue ; ces usages reflètent souvent, comme l'ont déjà montré plusieurs études (Paolacci & Rossi-Gensane, 2012 ; David & Doquet, 2016), des faits de langue qui passent inaperçus des experts mais qui constituent des points d'interrogation des apprentis scripteurs.

BIBLIOGRAPHIE

Anis J. (1983). « Pour une graphématique autonome », *Langue française*, n° 59, 31-44.

Anis J. (1989). « De certains marqueurs graphiques dans une linguistique de l'écrit », *DRLAV*, n° 41, 33-52.

Anis J., Chiss J.-L. & Puech C. (1988). *L'Écriture : théories et descriptions*. Bruxelles, De Boeck.

Brissaud C. (1998). *Acquisition de l'orthographe du verbe au collège : le cas des formes en /E/. Invariants et procédures*, thèse de doctorat de Sciences du langage de l'Université Stendhal - Grenoble 3.

Charpin F. (1976). « Analyse automatique de textes *libres* ». *Repères*, n° spécial Analyse de textes d'enfants, 11-20.

Crasson A. & Fekete J.D. (2007). « Structuration des manuscrits : Du corpus à la région », *Item* [En ligne : http://www.item.ens.fr/index.php?id=173027].

Catach N. (1980). *L'Orthographe française – traité historique et pratique*, Paris, Nathan.

Cogis D. (1999). *Production graphiques et procédures dans l'acquisition des marques de genre en français par les enfants entre huit et onze ans*, thèse de doctorat de l'université Paris 3 - Sorbonne nouvelle.

Dalbera J.-P. (2002). « Le corpus entre données, analyse et théorie », *Corpus*, 1 [En ligne : http://corpus.revues.org/10].

David J. (2008). « Les explications métagraphiques appliquées aux premières écritures enfantines ». *Pratiques*, n° 139-140, 163-187.

David J. & Doquet C. (2016). « Les écrits d'élèves : un corpus de référence pour le français contemporain ». *Actes du Congrès Mondial de Linguistique*, juillet 2016.

Doquet C. (2011). *L'Écriture débutante. Analyse linguistique des pratiques scripturales d'élèves à l'école élémentaire*. Rennes, Presses Universitaires de Rennes, coll. « Paideia ».

Fabre C. (1987). *Les Activités métalinguistiques dans les écrits scolaires*. Thèse de Doctorat d'État ès Lettres, Université Descartes Paris 5.

Fabre C. (1990). *Les Brouillons d'écoliers ou l'entrée dans l'écriture*, Grenoble, Ceditel / L'atelier du texte.

Fleury S. (2013). « Le Trameur. Propositions de description et d'implémentation des objets textométriques » ; [en ligne], http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf, consulté le 25 juin 2015.

Fleury S. et Zimina M. (2014). « Trameur : A Framework for Annotated Text Corpora Exploration », in Tounsi L. *et al.* (éd.) *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : System Demon-strations*, Dublin, p. 57-61 ; [en ligne] http://www.aclweb.org/anthology/C14-2013.pdf, consulté le 25 juin 2015.

Fleury S., *Le Trameur* 2015 ; [en ligne] http://www.tal.univ-paris3.fr/trameur/, consulté le 25 mars 2015.

Grésillon A. (1994). *Éléments de critique génétique : lire les manuscrits modernes*, Paris, Presses universitaires de France.

Lamothe-Boré C. (1998). *Choix énonciatifs dans la mise en mots de la fiction : le cas des brouillons scolaires*, Thèse de doctorat de Sciences du Langage, Université Stendhal - Grenoble 3.

Lebrave J.-L. (1990). « Déchiffrer, transcrire, éditer la genèse », in A. Grésillon, J.-L. Lebrave & C. Viollet (éd.) *Proust à la lettre : les intermittences de l'écriture*, Tusson, Du Lérot, pp. 141-205.

Lebrave J.-L. (2009). « Manuscrits de travail et linguistique de la production écrite », *Modèles linguistiques*, n° 59, 13-21.

Mas M., Garcia-Debanc C., Romian H., Séguy A., Tauveron C., Turco G., *Comment les maîtres évaluent-ils les écrits de leurs élèves en classe ?* Paris, INRP.

Masai F. (1950). « Principes et conventions de l'édition diplomatique », *Scriptorium*, t. 4, n° 2, 177-193 [en ligne : http://www.persee.fr/doc/scrip_0036-9772_1950_num_4_2_2294].

Masseron C. (2011). « L'analyse linguistique des écrits scolaire », *Pratiques*, n° 149-150, 129-162.

Mellet S. (2002). « Corpus et recherches linguistiques », *Corpus* [En ligne], 1 | 2002, mis en ligne le 15 décembre 2003. URL : http://corpus.revues.org/7.

Paolacci V. & Rossi-Gensane N. (2012). Quelles images de la phrase dans les écrits d'élèves de fin d'école primaire française ? *Description linguistique et réponses didactiques aux difficultés des élèves. Congrès Mondial de Linguistique Française*. Lyon, 5-6 juillet 2012, 341-359.

Pêcheux M. (1969). *Analyse automatique du discours*. Paris, Dunod.

Romian H. (1976). « L'analyse de textes d'enfants. Pourquoi ? Pour quoi faire ? ». *Repères*, n° spécial *Analyse de textes d'enfants*, 7-10.

Ros-Dupont M. (1995). « La segmentation non normée de l'écrit de l'enfant de CE1 : erreur ou étape obligée de l'apprentissage ». *Liaisons-HESO*, n° 25-26, 97-117.

Söze-Duval K., 2008, « Pour une textométrie opérationnelle » ; [en ligne] http://www.tal.univ-paris3.fr/trameur/RTI6provisoire.doc, consulté le 25 juin 2015.

NOTES

1. http://www.univ-paris3.fr/ecriscol.

2. Testenoire P.-Y., « Transcrire des écrits scolaires : entre philologie et génétique textuelle ». *Corpus*, n° spécial, janvier 2017.

3. L'Item ne tranche pas entre deux modes de présentation des segments supprimés : [entre crochets] ou ~~barrés~~. Nous avons choisi le premier pour faciliter le repérage automatique des suppressions.

4. Par exemple, les écrits des Poilus, disponibles sur Ortolang (http://www.univ-montp3.fr/corpus14/).

RÉSUMÉS

Les écrits scolaires posent des problèmes d'analyse automatique à cause des nombreux écarts à la norme langagière qu'ils comportent. Dans le but de constituer et de rendre exploitable un corpus significatif d'écrits d'élèves, le groupe de recherche ECRISCOL (Écrits Scolaires) de l'université de la Sorbonne Nouvelle a élaboré des solutions techniques pour préserver l'accès aux manuscrits des élèves tout en rendant possible une analyse automatique par des logiciels d'analyse textuelle. Le corpus en construction comporte une dimension développementale – il est constitué de textes produits par des élèves d'âges et de niveaux différents – et il rend compte de l'écriture même des textes puisque l'ensemble de ses traces – notes, brouillons, versions finales – sont présentées et rendues analysables.

Issues in Transcribing and Annotating Student Writing

Student writing is not easy to process automatically because of numerous deviations from standard language. To build and make available a significant corpus of student writing, the ECRISCOL group (Ecrits Scolaires = school writings) of the Sorbonne Nouvelle University has designed technical solutions to preserve access to handwritten student texts while making automatic linguistic processing still possible by means of text-analysis software tools. The corpus being built includes a developmental dimension –it is made up of texts from different age groups and levels– and a production-time one since different versions of the texts are included in the corpus –notes, drafts, final versions.

INDEX

Keywords : school writing, transcription, annotation, textual analysis, metadata
Mots-clés : écriture scolaire, transcription, annotation, analyse de données textuelles, métadonnées

AUTEURS

CLAIRE DOQUET

Université de la Sorbonne Nouvelle - EA 7345 Clesthia

VANDA ENOIU

Université de la Sorbonne Nouvelle - EA 7345 Clesthia

SERGE FLEURY

Université de la Sorbonne Nouvelle - EA 7345 Clesthia

SARA MAZIOTTI

Université de la Sorbonne Nouvelle - EA 7345 Clesthia