CORPUS

Corpus

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

Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles

Catherine Boré et Marie-Laure Elalouf

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/2731
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
Catherine Boré et Marie-Laure Elalouf, « Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles », *Corpus* [En ligne], 16 | 2017, mis en ligne le 19 novembre 2017, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/2731

Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles

Catherine Boré et Marie-Laure Elalouf

1 Revenir en 2016, sur les principes méthodologiques et épistémologiques qui ont guidé la constitution d'un premier corpus d'écrits scolaires publié en 2005 (Elalouf, dir.) ne relève pas seulement de l'état de l'art. L'émergence d'une linguistique de corpus, se substituant à une simple « banque de données », met en lumière l'évolution épistémologique apparue dans le recueil et le traitement de ces données si spécifiques.

2 Pour en retracer l'historique, nous nous appuierons sur trois recherches inscrites dans une continuité. Entre les deux premières effectuées en 2005 (Elalouf, dir.) et 2007 (Boré, dir.), et les corpus réunis à ce jour par le groupe de Cergy[1], des constantes apparaissent, inhérentes à la spécificité des écrits scolaires. Mais le développement de logiciels de traitement automatique et les possibilités offertes par l'annotation[2] ouvrent des perspectives renouvelant leur analyse. Nous exposerons dans un premier temps les questions méthodologiques posées par la constitution et la transcription des premières recherches ; puis nous analyserons les problèmes récurrents des corpus actuels sous l'angle générique et textuel à partir de certains exemples d'annotation, avant de présenter les axes de travail et quelques pistes d'exploitation de la recherche actuelle.

1. Un historique des recherches

3 La première recherche en réponse à un appel d'offres de la commission recherche de l'IUFM de Versailles a réuni sept enseignants-chercheurs et formateurs tous responsables du recueil et de l'analyse d'un corpus contextualisé (Elalouf & Keraven, 2002, 2004 ; Boré, 2004 ; Elalouf, Bertucci, Boré, Corblin, dir., 2005). La seconde s'est développée au laboratoire Modyco (« Modèles Dynamiques Corpus »)[3] sous la direction de C. Boré, et a été publiée en 2007 (Boré, dir.).

4 Le principe d'un corpus recueilli dans des conditions écologiques a été maintenu et le cadre d'analyse des productions d'élèves a été retravaillé sur le plan théorique (notamment avec J.-P. Bronckart, 1996 et J.-P. Bernié, 2001), tandis que quelques analyses instrumentées étaient expérimentées (Boré, 2007a, 2007b, 2007c ; Elalouf 2007, 2007a, 2007b ; Elalouf & Boré, 2007).

5 Ces travaux se sont poursuivis au sein du laboratoire EMA (École Mutation Apprentissages)[4] sur la base de corpus collectés par les chercheurs eux-mêmes (Boré, 2015 a et 2015b) ou dans le cadre d'encadrement de mémoires et de thèses (Elalouf, 2011 ; Elalouf *et al.* 2012 ; Boré et Bosredon, 2013 ; Gerlaud, 2014).

6 L'objectif commun à ces projets successifs est de contribuer à la connaissance de l'écriture scolaire, en menant une analyse linguistique des productions recueillies à l'école et au collège (dans un premier temps) en relation avec une analyse didactique des dispositifs conçus pour l'enseignement et l'apprentissage de l'écriture, activité comprise depuis la graphie et la segmentation des énoncés jusqu'à la production de textes dans leurs différentes étapes. Une visée de formation sous-tend l'entreprise : mettre à la disposition des formateurs d'enseignants une banque de textes dans leurs différents états, formant un continuum entre le premier et le second degré, au service du développement d'une culture commune. Nous faisons l'hypothèse, après C. Bonnet (1998) et C. Fabre-Cols (2000), que la lecture d'un nombre important de textes d'élèves répondant à une même consigne permet au formateur – et à travers lui à l'enseignant – de se constituer une culture de ces textes qui ne prennent souvent du relief que par contraste avec d'autres, ce qui oblige à s'interroger sur les effets d'un dispositif d'enseignement-apprentissage sur chaque texte dans sa singularité et sur la classe dans son ensemble.

1.1 Retour sur les principes de la première recherche

7 Par rapport à des recherches pionnières en didactique de l'écriture (Garcia-Debanc, 1990 ; Fabre-Cols, 1991 ; Bucheton, 1995 ; David et Plane, dir., 1996), le corpus publié en 2005 comporte non pas une sélection de textes, à l'appui d'une analyse linguistique et didactique mais, pour chaque activité d'écriture observée, l'ensemble des textes produits par une classe dans leurs différentes versions. Une dimension ergonomique s'ajoute ainsi à la recherche : comprendre ce que fait l'enseignant dans un contexte institutionnel donné, dans une temporalité marquée par des contraintes, en fonction de son histoire, de sa formation, de son appréciation des possibles et des risques qu'il peut prendre. Cette double approche appelait des choix en termes de recueil et d'organisation des données.

1.2 Le choix des données et leur organisation

8 Le choix a été fait de présenter des situations écologiques, c'est-à-dire conçues et mises en œuvre par l'enseignant de la classe au moment où elles doivent prendre place dans la programmation prévue. Certes la présence du chercheur au fond de la classe, qui prend des notes, enregistre ou filme, constitue inévitablement un biais mais une observation régulière sur plusieurs séances tend à limiter cet effet. En contrepartie de cette latitude laissée à l'enseignant, des outils d'analyse communs devaient être élaborés pour permettre la comparaison entre sous-corpus. C'est ainsi que la séquence

d'apprentissage, constituée de plusieurs séances reliées par un objectif d'apprentissage commun, a été déterminée comme l'unité d'observation. En effet, cette unité pragmatique organise le travail enseignant, et son déroulement est scandé par des travaux individuels et collectifs, écrits et oraux, qui constituent le corpus.

9 Huit corpus ont ainsi été recueillis, soit un total d'environ 728 textes présentés sur un cédérom joint à l'ouvrage, comportant également les principes méthodologiques et l'ensemble des données sur lesquelles se fondent les analyses.

10 Voir en Annexe I Tableau 1.

1.3 La solidarité textes/ contextes

11 La collecte et l'organisation des éléments nécessaires pour appréhender le cheminement d'un collectif de sujets scripteurs déplace les frontières entre textes et contextes. En effet, les « formes d'intersubjectivité et d'intertextualité propres aux situations scolaires », que C. Fabre-Cols (2004) a décrits avec les outils de la génétique textuelle, supposent que les textes produits dans la séquence, avec leurs consignes, soient mis en relation avec les textes lus en classe, les référents culturels explicitement convoqués ou simplement évoqués dans la situation d'écriture, ainsi que les commentaires et interventions orales ou écrites de l'enseignant et/ou des pairs sur ces textes. La circulation et la reconfiguration des discours, en constante évolution, oriente vers une conception du corpus comme archive, selon une conception philologique-herméneutique, à rebours de la réduction du corpus à un réservoir d'exemples indifférenciés. On reprendra donc la définition positive donnée par F. Rastier :

12 Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés de manière théorique réflexive en tenant compte des discours et des genres et de manière pratique en vue d'une gamme d'applications. (2002, en ligne)

13 Le format du cédérom, autorisant différents parcours selon les liens hypertextuels nous est apparu alors comme un moyen approprié pour présenter les différentes versions des textes dans leur singularité, tout en leur permettant d'entrer dans des réseaux homogènes – par exemple, toutes les consignes, tous les textes jugés définitifs pour une même consigne – ou des réseaux hétérogènes – par exemple, un texte lu en classe, différents états de la production d'un même élève, enrichis des interventions orales ou écrites de l'enseignant et de ses propres retours réflexifs. L'Annexe III présente l'architecture du cédérom pour le sous-corpus F. On voit comment les textes et leur analyse, qui forment le cœur du corpus, sont mis en perspective par une caractérisation du contexte d'enseignement[5] (section 1 et 2), le recueil d'une évaluation diagnostique (section 3) et les textes lus (section 5). Parcours au sein d'un même corpus, comparaisons d'un sous-corpus à l'autre, les possibilités sont nombreuses en jouant sur différentes variables (genre, niveau d'enseignement, dispositif d'écriture, rapport à l'écrit, etc.) mais elles exigent des choix de transcriptions en cohérence avec objectifs visés.

1.4 Les choix de transcription

14 Le support numérique permet de conserver la trace de l'écriture manuscrite de l'élève et les modes d'interventions de l'enseignant. La version scannée présente l'iconicité

maximale : elle restitue la copie d'élève avec les éventuels jeux de couleurs, les graphismes, les interventions du professeur sur le texte lui-même et dans la marge. Le texte reproduit (Fig. 1) appartient au sous-corpus A : c'est l'avant-dernière version, avant la recopie définitive. On y voit comment l'enseignante intervient différemment pour l'orthographe (croix signalant l'absence d'accent), la syntaxe (« qui » mis entre parenthèses) et le lexique (remarque dubitative ne permettant pas une substitution pertinente).

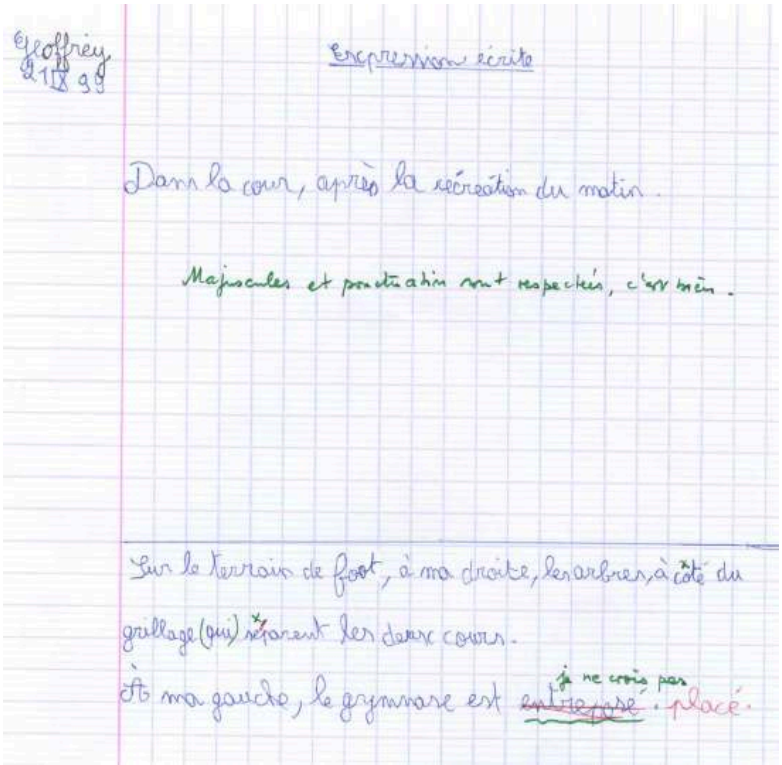


Figure 1. Version scannée corpus A (2005), texte de Geoffrey

15 La transcription que nous appelons iconique s'attache à rester le plus proche possible de cette version scannée en restituant la graphie, la mise en page, les ratures (passages barrés), les ajouts entre soufflets. Cette transcription suppose le recours à un minimum de signes et de conventions : son intérêt réside dans sa transparence.

16 Nous présentons (Fig. 2) la première version du texte de Geoffrey, avec un ajout et un remplacement en cours d'écriture, ainsi qu'un soulignement du professeur portant sur le relatif *qui* mais signalant une incompatibilité lexicale plus que syntaxique. La mise en espace de ce texte épouse son organisation par les verbes de perception et les indicateurs spatiaux, ce qui lui donne une silhouette bien différente de la linéarité de la version définitive :

Sur le terrain de foot <du haut>, je vois :
A ma droite il y a des arbres à coté du grillage qui mène à la cour du bas.
à ma gauche il y a le gymnase
devant moi il y a les classes
j'entent les maternel qui joue.
~~je sens la pollution~~ à toute heure et à midi on sent se quont va manger

Figure 2. Version 1 iconique – Geoffrey

17 Pour rendre compte à la fois des choix graphiques de l'élève et des interventions du professeur, la transcription diplomatique utilisée par l'ITEM[6] pour les manuscrits d'écrivains s'est révélée plus appropriée : elle permet en effet de délimiter entre crochets les portions de texte sur lesquels porte un commentaire, d'en préciser l'auteur et de décrire toutes les modifications qui affectent le texte. L'exemple (Fig. 3) est pris au corpus G. La transcription saisit ici une caractéristique des interventions du professeur : dans leur grande majorité, celles-ci ne sont pas verbales : les graphies erronées sont soulignées, les signes de ponctuation manquant du point de vue du professeur (P) sont ajoutés, les difficultés de compréhension sont marquées par un point d'interrogation.

> Alors le joueur de flûte quitta la ville et tous les enfants le suivirent
> Un matin, [virgule rajoutée par le professeur (P)] il se leva tôt en passant par la ruelle, avec sa flûte et les enfants [? de P en marge]. Il joua un petit morceau et les enfants en [chantaient souligné par P]. Une ou deux heures après 15 marke. [? de P en marge, point ajouté par P]. [Il "I" surchargé par P en "i"] se dit [« cet pa male » "cet" souligné par P, "s" rajouté à "pas" par P, "e" biffé par P] et ils se [partagère souligné par P] la mise [? de P en marge]. Et il [rentrer souligné par P] chez eux [chaqun souligné par P]. Le lendemain matin il alla autre [par "t" rajouté par P]. Il resta deux [heure souligné par P] et [ses souligné par P] l'[exploi "t" ajouté par P, ? de P] 70 mark [point virgule rajouté par P] il se [partaga souligné par P] la somme et tout le monde est [contemps souligné par P] [? rajouté par P au-dessous].

Figure 3. Version commentée – corpus G (2005)

18 Si cette transcription reproduit très exactement les opérations de différents scripteurs sur une même version, elle altère la lisibilité du texte par l'insertion de commentaires, à l'image de la lecture fractionnée de l'enseignant, lorsqu'elle se centre essentiellement sur la norme orthographique. C'est la raison pour laquelle la version orthographiée selon la norme en est le complément souvent nécessaire.

> Alors le joueur de flûte quitta la ville et tous les enfants le suivirent.
> Un matin, il se leva tôt en passant par la ruelle, avec sa flûte et les enfants. Il joua un petit morceau et les enfants en chantaient. Une ou deux heures après : 15 marks. Il se dit : « c'est pas mal » et ils se partagèrent la mise. Et ils rentrèrent chez eux chacun. Le lendemain matin, il alla autre part. Il resta deux heures et c'est l'exploit : 70 marks. Ils se partagèrent la somme et tout le monde est content.

Figure 4. Version orthographiée selon la norme– corpus G (2005)

19 En effaçant les graphies qui perturbent les habitudes de lecture, cette transcription permet une centration sur la textualité. L'élève (315g-arthur dans le cédérom) a suivi la consigne qui proposait d'écrire une fin heureuse ou malheureuse du conte *Le joueur de flûte de Hamelin*. Mais le choix d'une fin heureuse, avec pour seule perspective le gain, entre en conflit avec le genre du conte, ce qui suscite une appréciation négative de l'enseignant qui refuse de noter la copie : « Il manque des mots. Je comprends mal ton récit. As-tu bien lu le texte ? ». Chacune de ces observations peut être discutée au regard de la version orthographiée selon la norme. Il ne manque pas de mots, mais l'élève a fait le choix d'une phrase non verbale pour marquer la rapidité du gain. La progression du récit est scandée par les indicateurs temporels et une gradation dans la réussite. Et le choix de la monnaie, certes anachronique – le mark – laisse à penser que l'élève a bien *lu* le conte en le contextualisant. La confrontation des deux transcriptions fait donc apparaitre les écueils d'une lecture linéaire et fractionnée (Elalouf, 2005 : 128). Mais la version orthographiée selon la norme contraint le transcripteur à des choix : où s'arrêter dans la réécriture quand interfèrent des choix graphiques,

morphosyntaxiques, avec leur contrepartie sémantique ? Est-on autorisé à remplacer *il rentrer* [7] *chez eux chacun* par *ils rentrèrent chez eux chacun* ; *il se partaga la somme* par *Ils se partagèrent la somme* ? Et faut-il restituer une ponctuation forte avant ?

1.5 Les limites d'une transcription non annotée

20 C'est cependant sur la seule version orthographiée selon la norme qu'ont pu être tentées des analyses outillées de ce corpus. La recherche de récurrences lexicales entre les textes lus et écrits a pu montrer le degré d'appropriation d'un genre. Ainsi, Arthur est le seul élève de la classe à employer le terme d'*exploit*, ce qui pourrait signifier l'héroïsme du personnage de conte, mais il le fait dans un sens trivial, ce qu'une simple recherche d'occurrences ne permet pas de voir. Le même décalage entre l'emploi du terme *héros* dans les romans historiques et la prise en compte des valeurs chevaleresques a pu être montré à propos du corpus H (Elalouf, 2004-a). L'interprétation des co-occurrences serait plus riche s'il était possible, grâce à l'annotation, de remonter en deçà de la forme normée, aux opérations qui ont accompagné l'insertion d'un vocable – remplacement, ajout, déplacement – ainsi qu'aux négociations de sens suscitées par les relectures du scripteur lui-même, de ses pairs et du professeur.

21 De même, s'il était possible de mettre en relation, par le biais d'annotations, les différentes versions avec les interactions maitre-élèves, on pourrait observer comment une construction syntaxique propre à l'écrit, que l'enseignant introduit au moment de la réécriture, fait l'objet d'une appropriation différenciée. L'exemple est repris au corpus A ; il correspond au lancement de la réécriture sur le thème dans la cour, après la récréation du matin... :

> M1 : certains font des énumérations et des répétitions > comment pourrait-on éviter ça <
> [M écrit au tableau un exemple pris chez un élève : Je vois un parking, des voitures, un blouson, deux cages de foot.]
> M2 : qui pourrait me proposer de transformer cette phrase en plusieurs phrases > On pourrait garder le début : je vois un parking > Que pourrait-on écrire sur les voitures <
> E1 : je vois des voitures >
> M3 : Oui mais tu répètes je vois >
> E2 : il y a des voitures >
> M4 : oui, mais on a le droit de ne pas employer il y a > commence ta phrase par des voitures //
> E3 : //
> M5 : que font les voitures <
> E4 : des voitures sont garées sur le parking >
> M6 : oui / et comment faire pour ne pas répéter parking <
> E5 : des voitures sont garées dessus >
> M7 : [oui intonation d'insistance. Elle écrit : Dessus, sont garées des voitures.]

22 Par une sorte de coup de force discursif, le professeur valide une proposition d'élève (E5) et la transforme au moment de l'écrire au tableau, en antéposant *dessus* et en postposant le sujet *des voitures*.

23 Une seule élève reprend cette construction :

> Ewa : Sur un parking en face de notre école sont garées des voitures.

24 Les autres, aux prises avec la concurrence entre passif et pronominal, prépositions *sur* ou *dans*, et déterminants défini ou indéfini conservent l'ordre canonique :

> Salife : Plusieurs voitures ce gare sur un [?] parking.
> Jérôme : Des voiture se gare sont garées dans le parking.
> Romain : Les voitures sont garrer stoper stoper sur le parking.
> Julien : Des voitures se garent sur des places de parking.

25 Enfin, le recours au logiciel Tropes a permis de mettre en évidence l'hétérogénéité générique des textes produits à partir d'une même consigne (Elalouf, 2005 : 101-112), mais au prix d'une délimitation des textes en propositions, qu'ils soient sous ponctués ou qu'ils présentent une utilisation particulière des signes de ponctuation, ce qui oblige à des choix interprétatifs.

26 Les limites rencontrées dans l'exploitation de corpus normalisés nous ont conduites à poursuivre la collecte dans une perspective génétique tout en explorant les instrumentations possibles.

2. La recherche actuelle au Laboratoire EMA (Cergy) : récurrences et problématiques nouvelles

27 Le groupe de Cergy du laboratoire EMA travaille depuis l'ouverture du laboratoire[8] sur les productions d'écrits scolaires. Il détient à ce titre plusieurs sous-corpus composés des recueils de données des enseignants-chercheurs et des doctorants pour un total de 924 textes qui se répartissent ainsi :

> Sous-Corpus BOR. : [548][9] +122
> Sous-Corpus BOS. : 427
> Sous-Corpus ELA. : 110
> Sous-Corpus GER. : 103
> Sous-Corpus ROI. : 162

28 Voir en Annexe II, Tableau 2.

29 Les deux premiers corpus, dont la numérisation s'achève, sont exclusivement composés de textes fictionnels narratifs recueillis à l'école élémentaire. Une petite partie des textes (123) a été transcrite et annotée afin de figurer dans la base d'Ecriscol[10].

30 D'une manière générale, nous avons privilégié l'unité générique des sous-corpus, construits, en outre, de façon à permettre une étude longitudinale. Cependant la difficulté des corpus scolaires vient du caractère instable et hybride de leurs caractéristiques génériques : il n'est pas sûr que l'on puisse comparer des productions fictionnelles appartenant à des sous-genres vraiment distincts (récits policiers, contes, suites de narrations, etc.). Nous y revenons un peu plus loin. Aussi le premier travail est-il de sélectionner les sous-corpus comparables, notamment ceux qui ont été produits à partir de consignes identiques : cela peut concerner plusieurs classes et différents établissements, ou bien une seule classe à laquelle la même consigne est réitérée plusieurs fois dans l'année. On voit bien que dans les deux cas, le type de recherche mené sera différent.

31 Les questionnements et les choix qui ont marqué nos premières recherches en 2005 et 2007 se trouvent renouvelés par la mise à disposition, dix ans plus tard, de nombreux logiciels gratuits de traitements des corpus : des concordanciers comme AntConc et des outils de précision pour l'analyse textométrique comme les logiciels TXM (ENS Lyon) ou

Le Trameur (Paris 3) sont proposés aux chercheurs en linguistique et didactique qui ont pu se familiariser avec l'outil informatique. Mais les questions d'approche des corpus scolaires continuent de se poser.

32 Nous ferons état de deux principales : la classification des corpus en genres et la normalisation des corpus ; nous terminerons par quelques remarques sur le type de requêtes intéressant de grands corpus scolaires.

2.1 Difficultés récurrentes

2.1.1 Corpus, textes et « genres scolaires »

33 Dans nos travaux précédents, nous avions constaté la difficulté de définir des « genres scolaires » *a priori*. Cette difficulté tenait à l'absence d'une description textuelle de référence – la plupart des productions textuelles scolaires ayant pour modèles des types génériques issus de pratiques sociales différentes – et à la méconnaissance affectant la variabilité générique de ces textes.

34 • D'une part, en effet, ce sont des genres *dérivés* des genres communs (narration, compte-rendu, texte poétique formel, journal etc.) mais transposés à l'école après avoir été détachés de leur sphère sociale de référence. En outre, ces genres sont inculqués et s'imposent aux scripteurs[11]. Le genre le plus répandu à l'école élémentaire est le genre narratif avec de nombreuses combinaisons de séquences textuelles. La plupart des productions sont obtenues soit par la continuation et l'achèvement d'un épisode de narration fourni par la consigne, soit par l'organisation d'un récit suggéré par une suite d'images sans texte. Des normes implicites président à l'élaboration et la réception de ces textes, comme leur brièveté, et l'obligation d'un aboutissement final, contraintes qui façonnent les textes scolaires et leur donnent un « air de famille ».

35 Par ailleurs, ce sont aussi des textes multigénériques : en effet à ces textes s'incorporent, ou peuvent s'incorporer, d'autres genres textuels comme le texte de la consigne[12], qu'on peut considérer comme intertextuel : des phrases isolées tirées d'un intertexte assignable[13]sont parfois imposées par la consigne, réaménageant le genre textuel initial. On trouvera en Annexe V un exemple[14] de la complexité de ces dispositifs.

36 • D'autre part, au-delà de ces caractéristiques proprement génériques, ce sont des genres dialogiques et multi-auctoriaux car différents scripteurs se partagent l'espace textuel : en premier lieu, le scripteur et l'alloscripteur qu'est l'enseignant. En outre, s'intègre au texte de l'élève le commentaire métatextuel de l'enseignant. Mais l'alloscripteur est aussi le scripteur lui-même revenant sur son texte. Les ratures du scripteur qui intervient – immédiatement ou en différé sur son propre texte – contribuent à densifier la textualisation et peuvent avoir une incidence sur le genre. Tous deux – scripteur et alloscripteur – interviennent sur le plan du contenu comme de l'expression[15] : c'est le cas lorsque l'enseignant corrige à la place de l'élève et substitue matériellement sa trace scripturale à celle de l'élève. En revanche, le commentaire procède d'une nature linguistique différente : il est une réponse métadiscursive sur le lieu même du texte où il s'inscrit, et qu'il prend pour objet de son discours : les normes interdiscursives propres à la sphère scolaire font en effet du commentaire une composante obligatoire du texte scolaire. Il vient parfois aussi de l'élève (Annexe IV, où l'élève écrit *OK* pour commenter son propre travail).

37 On pourrait représenter la complexité du texte scolaire par l'ébauche suivante (Fig. 5) :

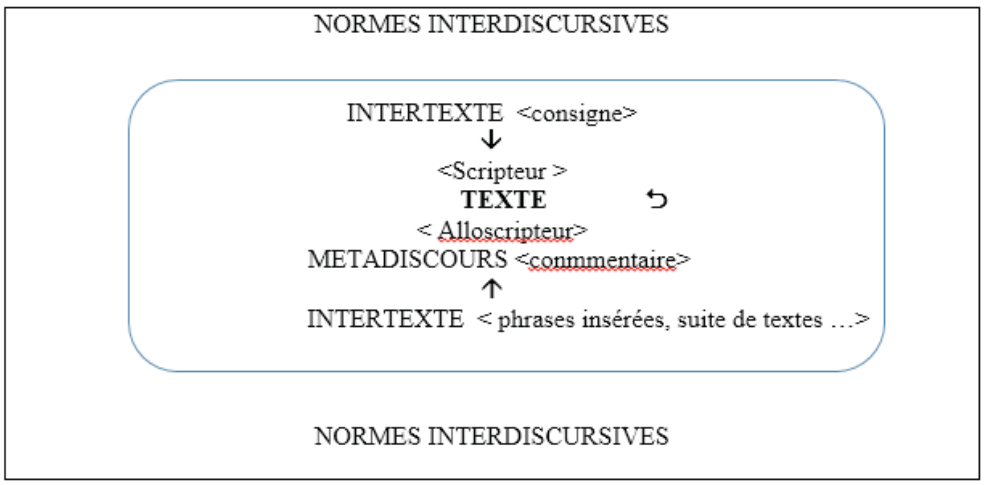


Figure 5. Le texte scolaire

38 Ainsi est-il difficile d'attendre une homogénéité générique entre sous-corpus dont la définition générique est constituée par du général – ce sont tous des textes assortis de commentaires – et du particulier : textes multisémiotiques[16] ou non.

39 Ces points ont à être théorisés pour les tâches d'annotation, en particulier dans les cas de présence de plusieurs scripteurs à des temporalités différentes, ou d'écriture collaborative.

2.1.2 Normalisation et état de langue

40 Dans les travaux de 2005 et 2007, nous avions été forcées de pratiquer une normalisation de l'orthographe et de la ponctuation afin de rendre possible un traitement automatique des copies[17] mais en perdant ainsi les graphies et choix originaux du scripteur. Ce point pouvait poser problème, par exemple, pour des graphies de verbes terminées en *-er* : ainsi, dans l'exemple suivant, le scripteur avait-il tenté de noter un verbe à l'imparfait, ou bien était-ce un passé simple pluriel, faussement induit par le pronom le précédant ?

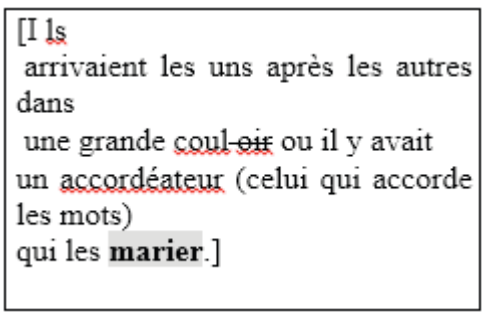

[I ls
arrivaient les uns après les autres dans
une grande coul~~oir~~ ou il y avait
un accordéateur (celui qui accorde les mots)
qui les **marier**.]

Figure 6. 1ELO6DB04, corpus BAR, 2004

41 Les décisions méthodologiques pour l'exploitation du corpus Ecriscol ont pris acte de ces difficultés. Le choix de noter la graphie correcte de manière acceptable pour le logiciel, tout en laissant apparaître celle de l'élève (Fig.7) présente le double avantage de permettre un traitement systématique des erreurs orthographiques et de leur évolution dans une même copie, quand il y a deux versions.

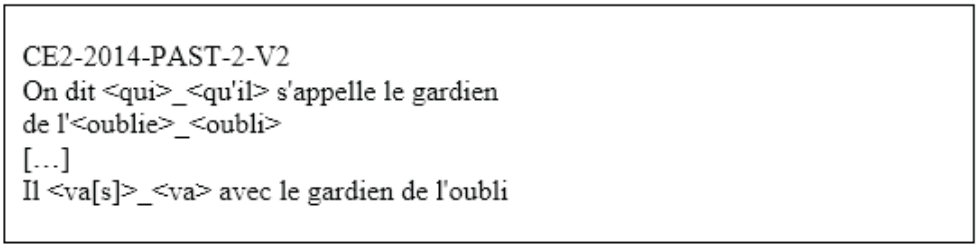

CE2-2014-PAST-2-V2
On dit <qui>_<qu'il> s'appelle le gardien
de l'<oublie>_<oubli>
[…]
Il <va[s]>_<va> avec le gardien de l'oubli

Figure 7. exemple de transcription et annotation (Sara Mazziotti)

42 La recherche des types d'erreurs récurrentes produites par les scripteurs et leur évolution selon un suivi longitudinal intéresse les spécialistes de l'acquisition, psycholinguistes et didacticiens. Le but est de faciliter l'interprétation des erreurs par la consultation de séries d'erreurs identiques dans un même contexte et la mise en évidence de co-occurrences.

43 Ce premier argument à l'encontre d'une normalisation des graphies[18] vaut d'ailleurs pour son principe même, car c'est la langue des apprenants qui doit être interrogée en fonction des buts que se donnent les utilisateurs, et celle-ci s'incarne dans des textes. En effet, nous n'avons pas simplement affaire à des *écrits ou des productions d'écrits* qu'il suffirait de mettre bout à bout en considérant qu'il s'agit d'un corpus indifférencié. Ce sont des textes courts, voire très courts, mais qui ont chacun une logique propre ; la mise en corpus de textes scolaires nous confronte à l'unicité des données, leur caractère non identique, non répétable, non-reproductible, non-interchangeable. Seule la décision méthodologique de les rassembler et de les ordonner permet d'en faire des corpus de *textes*, plutôt que des isolats disparates et idiosyncrasiques.

44 La *textualité* des corpus – ce qui va transformer des écrits en *textes* – réside ainsi autant dans leur position au sein du sous-corpus auquel ils appartiennent et de leur proximité avec d'autres textes regroupés intentionnellement – que de leur forme et de leur aptitude à la faire évoluer. Et, précisément, les corpus recueillis par le groupe de Cergy se présentent pour la plupart sous la forme de deux états successifs, ce qui permet l'analyse génétique de leur construction textuelle et de leur évolution. Nous continuons à privilégier l'analyse de F. Rastier[19] reposant sur les principes de la sémantique interprétative et des parcours sémantiques résumés en ces termes :

45 Le sens d'un texte ne se déduit pas d'une suite de propositions, mais résulte du parcours de formes sémantiques [...] liées à des formes expressives. [...]

46 La génération d'un texte consiste en une série de métamorphismes et de transpositions, qu'on peut mettre en évidence à l'oral par l'étude des reformulations, à l'écrit par celle des brouillons.

47 Ainsi, par exemple, si l'on s'intéresse au connecteur « et », fréquent dans les textes des plus jeunes scripteurs, et que l'on compare les deux versions du texte ci-dessous, on remarque d'abord que dans V1 (Fig. 8) « et » s'écrit « est » (l.1). Dans V2 (Fig. 9), la présence du connecteur « et » se trouve couplée au signe de ponctuation (points de suspension). Sa réitération avec une majuscule en tête de phrase ne peut pas être traitée de façon seulement quantitative, comme une occurrence supplémentaire par rapport à V1 ; ici l'intention textuelle – suspense et jeu pragmatique – l'emporte.

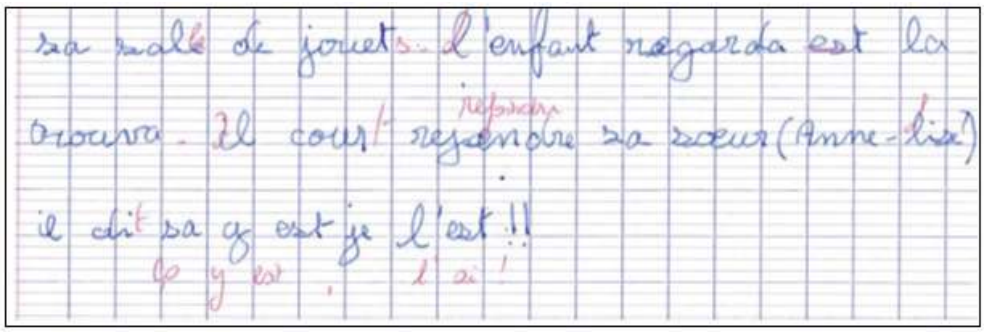


Figure 8. CE2A-2014-JO-1-V1 (corpus BOS)

Figure 9. CE2A-2014-JO-1-V2 (corpus BOS)

2.1.3 L'orthographe et la ponctuation : essais d'annotation

48 C'est pourquoi nous avons immédiatement adhéré au choix qui a été fait dans le projet Ecriscol dirigé par Claire Doquet (La Sorbonne nouvelle Paris 3, laboratoire Clesthia) de ne pas normaliser orthographe et ponctuation, au minimum dans la V1.

49 L'exemple suivant (Fig. 10) montre un extrait de la copie manuscrite présentée Figure 8, transcrit et annoté dans sa version 1. Le symbole P est utilisé pour noter les corrections de l'enseignant.

sa <sal//P#l//e>_<salle> de <jouet//P#s//>_<jouets>
<®[P#l]//P#L//®>_<L>'enfant regarda
<®[P#est]//P#et//®>_<et> la
trouva <®[P#i]//P#I//®l>_<Il> <cour//P#t//>_<court>
<rejoendre>_<rejoindre> //P#rejoindre// sa soeur (Anne-
<®[P#l]//P#L//®ise>_<Lise>)
il <di//P#t//>_<dit> <sa>_<ça> y est je l'<est>_<ai> ! !
//P#ça_y_est,_je_l'ai!//§

Figure 10. CE2A-2014-JO-1-V1
(corpus BOS, transcription et annotation Sara Mazziotti)

50 Pour la V2, la question se posait de savoir s'il fallait continuer à noter les nouvelles erreurs commises par le scripteur et leur rectification par l'enseignant, ou se contenter d'une version qui neutralisait les corrections alloscripturales, comme ici :

mais il la trouve il <ouva[s]>_<ouvre> la porte et...Et
<surpri>_<surpris> il trouve <tout>_<tous> les jouets que les
<élève>_<élèves>
ont perdu...Il retrouva sa toupie chinoise.§

Figure 11. CE2A-2014-JO-1-V2
(corpus BOS, transcription et annotation Sara Mazziotti)

51 En fonction de l'utilisation du corpus visée, il peut sembler utile de noter les modifications graphiques apportées par chaque scripteur (élève et enseignant), en dépit de la lourdeur de l'annotation, comme ici (Fig. 12) avec le même exemple :

mais il la trouve //P#POINT// <®[P#i]]//P#I//®l>_<Il>
<ouv®[P#a]//P#re//®[s]>_<ouvre> la porte
et…Et <surpri//P#s_VIRG//>_<surpris_VIRG>
il trouve <tou®[P#t]//P#s//®>_<tous> les jouets
que les <élève//P#s//>_<élèves> ont perdu.
Il <retrouv®[P#a]//P#e//®> _<retrouve> sa toupie chinoise.§

Figure 12. CE2A-2014-JO-1-V2
(corpus BOS, transcription et annotation Sara Mazziotti)

52 Dans la mesure où la ponctuation, lacunaire le plus souvent, délimite pour le scripteur[20] des blocs syntaxiques et énonciatifs qui font sens pour lui, on voit s'ouvrir un champ intéressant la conception de la phrase et du paragraphe de la part du scripteur, comme de celle de l'enseignant. En effet, quand la transcription note la ponctuation de la main de l'enseignant sur la copie qu'il corrige, le chercheur a accès de façon implicite à une partie des représentations de celui-ci sur la phrase et ses limites, ce qu'il considère comme un segment acceptable, etc.

53 Il faut cependant distinguer deux ensembles d'interventions : celles qui visent directement l'écriture du texte quand l'enseignant rature (corrige) le texte de l'élève, et celles qui se fondent sur ses commentaires. Une telle remarque ne va pas sans difficultés méthodologiques : elle prend acte du fait que les corpus puissent être de nature différente ; traiter des erreurs de l'élève en faisant des comptages de marques erronées peut certes être comparé au décompte des formes rectifiées, qu'elles soient le fait de l'élève ou de l'enseignant. Mais un corpus de commentaires métadiscursifs nécessite une analyse préalable avec l'émergence de mots-pôles pour être reconnu. Aussi la nature du traitement dépend-elle de ce qu'on veut/ peut mettre en corpus.

3. Axes et perspectives du groupe de Cergy

3.1 Rappels

54 L'utilisation de chiffres bruts et de requêtes hasardeuses est souvent le premier mouvement de l'utilisateur novice d'un logiciel d'analyse automatique des textes, alors que l'exploitation outillée d'un corpus prend du temps et tire son intérêt de la pertinence des requêtes. C'est encore plus vrai lorsqu'il devient nécessaire d'effectuer un balisage ciblé pour préciser une requête. Quelques tentatives avaient été esquissées. M.-L. Elalouf (*in* Boré, 2007 : 96-98 *sq*.) avait cherché à identifier les actes de paroles dans l'analyse d'interactions orales en formulant des requêtes à l'aide de la TEI. L'objectif principal visait l'extraction des énoncés injonctifs, ceux-ci étant préalablement définis en phrases dont le verbe principal est : à l'impératif ; à l'indicatif ; sous les formes du v. *aller + infinitif* ; ou encore : *il faut + infinitif, on va + infinitif*. Pour les énoncés interrogatifs (interrogation totale), un travail plus fin avait permis de distinguer les énoncés interrogatifs non marqués morphologiquement d'énoncés assertifs non-injonctifs. Mais les limitations concernaient la sélection du verbe principal dans les segments repérés.

55 De même, travaillant sur un corpus de 142 versions de contes didactiques en classe de 6ème, nous avions pu, avec l'aide de D. Malrieu qui avait effectué un balisage sur un échantillon mis aux normes orthographiques de 26 copies, extraire et analyser les discours rapportés et les mots métalinguistiques utilisés par les élèves. Le but de ce travail était de comparer ces résultats avec le conte didactique original[21] balisé par D. Malrieu (*in* Boré, 2007 : 132 *sq*.), dont les élèves avaient eu pour consigne de réécrire un épisode. Mais, outre les dimensions réduites du corpus d'élèves traité, et le biais introduit artificiellement par les segments répétés dus à la reprise par les élèves de mots de la consigne, la normalisation de l'orthographe et de la ponctuation privait l'interprète d'une comparaison avec les moyens spécifiques de signalement du discours rapporté utilisé par les élèves, en contraste avec le reste du texte. Les analyses manuelles montrent en effet un sur-encadrement du discours direct par la ponctuation et par des moyens syntaxiques spécifiques (comme la redondance du segment introducteur). C'est probablement à partir de constatations de ce type que pourraient être établies des requêtes pertinentes et, éventuellement le balisage de certains segments de texte.

3.2 Les régularités : dépasser l'opposition quantitatif/qualitatif

56 L'insertion obligatoire ou implicite des phrases intertextuelles provenant du cadre de production se prête particulièrement à des requêtes nécessitant d'aller au-delà du quantitatif.

57 C'est le cas du dispositif présenté en Annexe V qui impose au scripteur l'utilisation des phrases suivantes :

> Tu as perdu quelque chose ? lui demanda Anne-Lise. (P1)
> Ma toupie chinoise sur le chemin de l'école. (P2)

58 L'inscription, prescrite par la consigne, de P1 et P2 dans les textes ne prend sa signification que par contraste avec le cotexte : les concordanciers faisant apparaitre les environnements de ces segments renseignent alors sur l'écriture des textes.

59 Une rapide exploration[22] sur 50 textes de CE2 montre 35 occurrences de P1 et le même nombre de P2, mais il est beaucoup plus intéressant d'analyser les contextes gauche et droit dans lesquels sont insérées les phrases. Quand il existe, l'enchaînement de la paire de propositions est continu : aucune modification ne survient entre P1 et P2. En revanche, le contexte gauche mérite d'être exploré à partir de requêtes syntaxiques et lexicales : le dialogue peut ou non être anticipé en amont par un segment introducteur développé, comportant ou non un verbe, de dire ou « psychologique » etc.

60 Une autre mesure montre une haute fréquence du nom propre Gabriel dans ce même corpus (66 occurrences). Mais c'est la position syntaxique de ce pivot qu'il est intéressant d'explorer si l'on veut opposer contrastivement l'usage qu'en font les plus jeunes (CE2) et les plus âgés. Dans ce même corpus de CE2, sur 66 occurrences, 42 placent le nom propre Gabriel en position de sujet dans la phrase (dont 13 en position de sujet inversé). À partir d'un résultat aussi simple, il est sans doute possible d'élaborer des hypothèses sur l'organisation des constituants de la phrase chez les élèves confrontés aux contraintes de la même consigne.

3.3 Pourquoi se centrer sur le verbe ?

61 Le groupe de Cergy du laboratoire EMA a choisi de faire converger ses observations et analyses sur le verbe pour son caractère central et multidimensionnel. En effet, les travaux des psycholinguistes des trente dernières années ont montré son importance dans l'acquisition du langage et la conquête de l'écrit. Parallèlement, les recherches en didactique du français ont souligné la programmation des savoirs et pratiques liées à cette classe de mots dès le début de la scolarité, tout en signalant le caractère éclaté des approches pour l'étude de la langue – entre grammaire, conjugaison, lexique et orthographe, le caractère occasionnel des pratiques en lecture et écriture, et les obstacles au transfert entre les connaissances sur les verbes et leurs emplois. C'est à un changement de perspective, partant des usages pour dégager des régularités, qu'invite notre démarche. Les travaux sur les grands corpus oraux et écrits ont fait apparaitre des distributions différenciées des verbes dont l'école pourrait se saisir pour mieux cerner ce que les élèves maitrisent déjà et les remaniements qu'impose le passage à l'écrit. En tant que porteur de la prédication, le verbe constitue une sorte de plaque tournante entre le langage oral, le langage intérieur et l'écrit. Selon Vygotski :

62 Le langage intérieur est réduit au maximum, abrégé, sténographique. Le langage écrit est développé au maximum, plus achevé même dans sa forme que le langage oral. Il ne comporte pas d'ellipses. Le langage intérieur en est plein. Il est par sa structure presque exclusivement prédicatif. De même que dans le langage oral, la syntaxe devient prédicative lorsque le sujet et les membres de la proposition qui s'y rapportent sont connus des interlocuteurs, le langage intérieur, dans lequel le sujet de la conversation et l'ensemble de la situation sont connus de celui même qui pense, est presque composé des seuls prédicats. Le langage écrit a une deuxième particularité étroitement liée à son caractère volontaire, celle d'être plus conscient que le langage oral. ([1934] 1997 : 342)

63 Par l'étude des constructions verbales, à l'articulation du lexique et de la syntaxe, l'attention aux phénomènes énonciatifs dont le verbe est vecteur, l'examen des agencements spécifiques concernant les satellites de la prédication à l'écrit et à l'oral, la mise en évidence des récurrences au sein d'un texte et des occurrences singulières,

on devrait pouvoir mieux caractériser les dynamiques qui s'instaurent entre textes lus, produits et commentés au sein d'une classe.

Conclusion

64 Il est maintenant possible d'envisager des perspectives offertes par une annotation des corpus scolaires.

65 - Dans la mesure où le global détermine le local[23], étudier les caractéristiques d'un genre scolaire et ses reconfigurations dans l'écriture des élèves constitue une priorité parce que le découpage des unités textuelles en dépend. Ces recherches ne peuvent s'engager que sur des séries homogènes, c'est-à-dire construites à partir de consignes identiques ou de même type. De telles conditions ne sont pas toujours faciles à réunir : produire un texte narratif en insérant plusieurs propositions différentes, ou fabriquer une suite de texte à partir d'un texte d'auteur n'impliquent pas les mêmes contraintes, encore moins quand celles-ci sont multisémiotiques. La solution passe forcément par une sélection-contextualisation de sous-corpus et la recherche (inductive) des segments typiques à baliser.

66 - Des travaux issus de thèses en cours au laboratoire EMA vont entrainer de nouvelles recherches sur les liens entre les corpus oraux et leur implication dans les corpus écrits, comme suivre la circulation des mots (lexies) d'un discours magistral dans le processus d'élaboration des textes d'élèves, ou cerner les effets des emplois autonymiques à l'oral sur les productions écrites des élèves.

67 - Sur le plan didactique enfin, le balisage permettra d'identifier les points de résistance d'un sous-corpus à l'autre, comme les « zones floues » au regard de l'acceptabilité suggérées par Y. Reuter[24] : « quels seraient par exemple les observables pertinents à partir desquels on pourrait modéliser la structure floue [...] des formes verbales en /-E/ (imparfait, participe passé, infinitif) ? ».

68 La dimension et la portée de ces études ne sont pas de même échelle. C'est pourquoi, l'entrée « autour du verbe » constitue pour le Groupe de Cergy du laboratoire EMA un point de départ permettant d'unifier les travaux.

BIBLIOGRAPHIE

Anis J. & Boré C. (dir.) (2004). « Théories de l'écriture et pratiques scolaires ». *Linx, 51*, Université Paris 10-Nanterre.

Bernié J.-P. (2001). *Apprentissage, développements et significations*. Bordeaux : Presses univeristaires de Bordeaux.

Bonnet C., Corblin C. & Elalouf M.-L. (1998). *Les procédés d'écriture chez les élèves de 10 à 13 ans, un stade de développement*. Lausanne : LEP, Loisirs et Pédagogie.

Boré C. (2004). « L'écriture scolaire : langue, norme, "style", quelques exemples de discours rapporté », *Linx, 51*, pp. 91-106.

Boré C (2007a, dir.). *Construire et exploiter des corpus de genres scolaires*, Namur (B) : Presses universitaires de Namur, coll. « Diptyque ».

Boré C. (2007b). « Les genres scolaires comme corpus, construction d'une problématique ». In *Construire et exploiter des corpus de genres scolaires*. Namur : Presses universitaires de Namur, coll. « Diptyque », pp. 41-55.

Boré C. (2007c). « La métamorphose d'un genre : quelques descripteurs pour un genre scolaire de récit ». In *Construire et exploiter des corpus de genres scolaires.* Namur : Presses universitaires de Namur, coll. « Diptyque », pp. 141-165.

Boré C. (2015a). « Le conte étiologique chez des scripteurs français et brésiliens de 7-8 ans, aspects didactiques et linguistiques ». In *Recherches & Applications, Le Français dans le monde*, *58*, pp. 106-114.

Boré C. (2015b). « Lecture et interprétation des parodies de contes, un problème de genre ». In Ablali D., Bouhouhou A., Tebbaa O. (éd.) *Les genres textuels, une question d'interprétation*. Actes du Colloque International de Marrakech. Limoges : Lambert-Lucas, pp. 81-92.

Boré C. & Bosredon C. (2013). « La phrase selon les brouillons : un trajet de l'oral à l'écrit ». *Le français aujourd'hui, 181*, pp. 13-24.

Bronckart J.-P. (1996). *Activité langagière, textes et discours, Pour un interactionnisme socio-discursif*. Lausanne : Delachaux & Niestlé.

Bucheton D. (1995). *Écritures – réécritures : récits d'adolescents*. Bern, Berlin, Frankfurt/M., New York, Paris, Wien : Peter Lang.

David J. & Plane S. (dir.) (1996). *L'apprentissage de l'écriture*. Paris : Presses universitaires de France.

Elalouf M.-L. & Keraven J. (2002). « Une banque de données de textes d'élèves à l'épreuve ». *Pratiques*, *115-116*, pp. 107-124.

Elalouf M.-L. & Keraven J. (2004a). « L'acquisition du lexique à l'épreuve d'un grand corpus d'élèves ». In Calaque E. & David J., *Didactique du lexique*. Bruxelles : De Boeck, pp. 185-197.

Elalouf M.-L. (2004-b). « Constitution d'un grand corpus de textes d'élèves : problèmes méthodologiques et premiers résultats ». *Linx, 51*, pp. 129-146.

Elalouf M.-L. (dir.), Bertucci M.-M. & Boré C. *et al*. (2005). *Écrire entre 10 et 14 ans, un corpus, des analyses, des repères pour la formation*. Versailles : CRDP de l'académie de Versailles.

Elalouf M.-L. (2007a). « La fonction "statistiques" de *Tropes* : une aide au diagnostic de la cohésion textuelle ? ». Namur (B) : Presses universitaires de Namur, coll. « Diptyque 10 », pp. 167-183.

Elalouf M.-L. (2007b). « Les interactions orales en classe, un rôle essentiel dans la configuration de genres scolaires ». Namur (B) : Presses universitaires de Namur, coll. « Diptyque 10 », pp. 89-106.

Elalouf M.-L. & Boré C. (2007). « Construction et exploitation de corpus d'écrits scolaires ». *Revue Française de Linguistique Appliquée*, vol. XII-1, pp. 53-70.

Elalouf M.-L. (2011). « Constitution de corpus scolaires et universitaires, vers un changement d'échelle ? », *Pratiques*, *149-150*, pp. 56-70.

Elalouf M.-L., Beaumanoir-Secq M., Bornaz S. & Fort P.-L. (2012). « Enjeux de la constitution de corpus dans les écrits professionnels et de recherche du master “éducation et formation” : le cas de la didactique du français ». *Les didactiques en question(s)*. Bruxelles : De Boeck, pp. 382-403.

Fabre-Cols C. (1991). *Les brouillons d'écoliers ou L'entrée dans l'écriture*. Grenoble : Revue « Texte en main », coll. « Atelier du texte ».

Fabre-Cols C. (dir.) (2000). *Apprendre à lire des textes d'élèves.* Bruxelles : De Boeck-Duculot.

Fabre-Cols C. (2004). « Les brouillons et l'école : ce qu'a changé la critique génétique ». *Le français aujourd'hui, 144*, pp. 18-24.

Garcia-Debanc C. (1990). *L'élève et la production d'écrit*. Metz : Centre d'analyse syntaxique de l'université de Metz.

Garric N. & Longhi J. (dir.) (2012). « L'analyse de corpus face à l'hétérogénéité des données ». *Langages*, *187*.

Gerlaud B. (2014). « Comment considérer la ponctuation au lycée ? ». *Le français aujourd'hui, 187*, pp. 81-90.

Rastier F. (2009 [1987] [1996]). *Sémantique interprétative*. Paris : Presses universitaires de France.

Rastier F. (2002). « Enjeux épistémologiques de la linguistique de corpus », www.revue texto.net/ Inedits/Rastier/Rastier_Enjeux.html.

Vygotski L.S. (1934, 1997 pour la traduction française). *Pensée et langage*. Paris : La Dispute.

ANNEXES

Annexe I

	Genre	Classes	Nombre de travaux
Sous-corpus A	Prise de notes et description	1 classe de CM2	20 textes, 4 versions
Sous-corpus B	Récit fictionnel en suivant des consignes grammaticales	1 classe de CM2	20 textes, 2 versions
Sous-corpus C	Écrit intermédiaire	1 classe de CM2	24 textes en 3 versions
Sous-corpus D	Écrit intermédiaire	1 classe de CM2	23 textes en deux versions, la 3e version étant écrite en groupe (5 textes)

Sous-corpus E	Écrit sur un thème donné en choisissant le type de texte	1 classe de CM2	3 textes d'élèves dans deux versions+ 20 textes de juin selon la même consigne
Sous-corpus F	Récit en relation avec la lecture de l'*Odyssée*	1 classe de 6e SEGPA	7 versions (sur traitement de texte + évaluations nationales
Sous-corpus G	• Suite de conte • Variations sur une reconstitution de texte	1 classe de 6e à 1 trimestre d'intervalle	31 textes (version définitive) 47 textes (version définitive)
Sous-corpus H	Roman historique élaboré collectivement sur 5 séquences •Écrits diagnostiques sur lesquels se fonde l'enseignante pour construire la progression • Écrits préparatoires qui permettent la maturation du projet : incipit, lettrine, synopsis • Écrits-outils consultables pendant l'écriture longue : fiches de vocabulaire, dossiers documentaires • Différents états du roman historique écrit par groupes et interactions orales • Écrits destinés à l'évaluation sommative	1 classe de 5e	• 47 textes de 17 à 255 mots • 14 textes • 12 dossiers documentaires • 4 incipits + autoévaluation+ évaluation • 5 synopsis écrits en groupes • 4 ensembles de fragments descriptifs (bribes) • 4 romans de 585 mots à 1109 mots • 14 suites de textes avec certains brouillons • 10 descriptions de monstre avec certains brouillons TOTAL = 728

Tableau 1. Les corpus de la recherche IUFM. État des lieux en 2005

Annexe II

	Genre	Classes	Nombre de travaux
Sous-corpus BOR	Conte étiologique	1 classe de CP (25 élèves) 1 enseignant	• 8 • 7 sur 8 en écriture collaborative • 2 sur 8 comportant 2 versions
Sous-corpus BOS	Écriture d'un récit fictionnel à partir d'images + phrases extraites d'un album.	10 classes (4 écoles) •2 CE2 • 2 CM1 • 6 CM2	•20 (2x10) Toutes les productions comportent 2 versions
Sous-Corpus ELA	Écriture à partir d'une phrase inductrice : *La personne la plus ancienne de ma famille...* *Je me souviens* *J'ai oublié* Écriture d'un dialogue Acrostiches	1 classe en UPE2A	• 18 • 20 • 24 • 24 • 24 Toutes les productions comportent 2 versions 1 transcription d'entretien : étayage de la version 1 à la version 2

Sous–Corpus GER	Lecture analytique Argumentation Invention Description Résumé Confrontation de documents / synthèse	7 élèves de 2 de générale 7 él seconde et 10 étudiants 7 élèves de 2 de générale 7 élèves de 2 nde générale 6 étudiants de BTS 1 10 étudiants de BTS 1	5 5 1 1 1 2
Sous-Corpus ROI	Brèves de journal scolaire	3 classes 2 enseignants •1 CM1/ CM2 •1CM1 •1CM2	4 numéros de journaux

TOTAL = 924

Tableau 2. Les corpus du Laboratoire EMA (Cergy). État des lieux 2015

Annexe III

F. 6e SEGPA, juin, Val d'Oise (Péret)

Écrire en lisant l'Odyssée

À la fin de l'année scolaire, une institutrice spécialisée en AIS (adaptation et intégration scolaire), qui accueille des stagiaires, conçoit avec leur formatrice une séquence qui convoque le plus grand nombre de compétences travaillées en français dans cette section et s'inscrit dans une perspective d'intégration des élèves au collège, *L'Odyssée* figurant au programme de 6e, parmi les textes issus de l'héritage antique.

Au cours de la séquence, les élèves sont amenés à produire plusieurs travaux d'écriture et à utiliser celle-ci dans différentes fonctions : aide à compréhension, à la mémorisation et à la création.

10. F. Contexte

11. F. Établissement

12. F. Projet d'établissement

13. F. La SEGPA et ses acteurs

14. F. L'enseignante et ses élèves

20. F. La séquence « *Odyssée* »

21. F. Programmation des apprentissages en français sur l'année scolaire

22. F. Place de la séquence dans l'année

23. F. Description de la séquence

24. F. séance 1

25. F. Séance 2

26. F. Séance 3

27. F. Séance 4

28. F. Séance 5

30. F. Textes

31. F. Bérivan

311. F. Bérivan, écrit 1

312. F. Bérivan, écrit 2, 1er jet

313. F. Bérivan, écrit 2, 2e jet

314. F. Bérivan, écrit, 3e jet

315. F. Bérivan, écrit 2, 4e jet

316. F. Bérivan, écrit 2, 5e jet

317. F. Bérivan, écrit 3

318. F. Bérivan, évaluation nationale

32. F. Christelle

321. F. Christelle, écrit 1

322. F. Christelle, écrit 2, 1e jet

323. F. Christelle, écrit 2, 2e & 3e jet

324. F. Christelle, écrit 3

325. F. Christelle, évaluation nationale

33. F. Christopher

331. F. Christopher, écrit 1

332. F. Christopher, écrit 2, 1e jet

333. F. Christopher, écrit 2, 2e jet

334. F. Christopher, écrit 2, 3e jet

335. F. Christopher, écrit 2, 4e jet

336. F. Christopher, écrit 3

337. F. Christopher, évaluation nationale

34. F. Diégo

341. F. Diégo, écrit 1

342. F. Diégo, écrit 2, 1e jet

343. F. Diégo, écrit 2, 3e jet

344. F. Diégo, écrit 2, 3e jet

345. F. Diégo, écrit 2, 4e jet

346. F. Diégo, écrit 2, 5e jet

347. F. Diégo, écrit 3

348. F. Diégo, évaluation nationale

35. F. Émilie

351. F. Émilie, écrit 1

352. F. Émilie, écrit 2, 1e jet

353. F. Émilie, écrit 2, 2e & 3e jet

354. F. Émilie, écrit 2, 4e jet

355. F. Émilie, écrit 3

352. F. Émilie, évaluation nationale

36. F. Julien

361. F. Julien, écrit 1

362. F. Julien, écrit 2, 1e jet

363. F. Julien, écrit 2, 2e jet

364. F. Julien, écrit 2, 3e jet

365. F. Julien, écrit 3

366. F. Julien, évaluation nationale

37. F. Litès

371. F. Litès, écrit 1

372. F. Litès, écrit 2, 1e jet

373. F. Litès, écrit 2, 2e jet

374. F. Litès, écrit 2, 3e à 5e jet

375. F. Litès, écrit 3

376. F. Litès, évaluation nationale

40. F. Analyses

41. F. Choix des élèves

42. F. Première écriture

43. F. Deuxième écriture

44. F. 2e écriture. Litès

45. F. 2e écriture. Diego

46. F. 2e écriture. Bérivan

47. F. 2e écriture. Émilie

48. F. 2e écriture. Autres élèves

50. F. Annexes

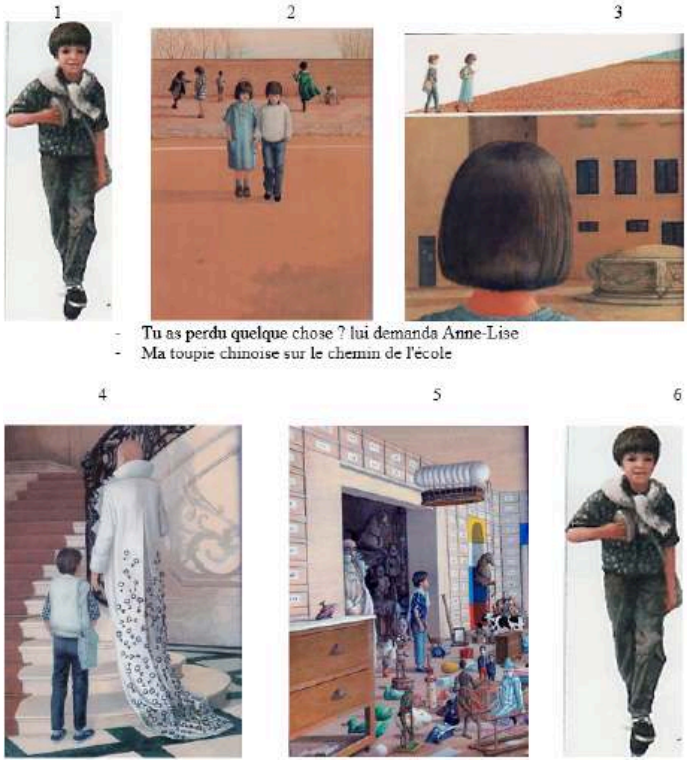
51. F. Tableau des compétences ciblées

52. F. Texte lu : *le Cyclope*, extrait de l'Odyssée

Annexe IV (corpus Bariatinski, 2004)

Annexe V (corpus Bosredon, 2013).

Images et phrases extraites de *Le gardien de l'oubli* de Jean-José Gisbert & Alphonse Ruano, Syros alternative, 1990.



Consigne de la première version (séance 1):

Voici 6 images racontant l'histoire de Gabriel qui a perdu sa toupie chinoise à laquelle il tient beaucoup. Elle est en métal, multicolore et quand elle tourne, elle lance des notes joyeuses.

Utilise les six images pour raconter cette histoire à des lecteurs qui, eux, n'ont pas les images.

Le titre de cette histoire sera «*Le gardien de l'oubli*»

Dans ton récit, tu devras utiliser les phrases qui se trouvent sous l'image 2.

Consigne de la réécriture (séance 2):

Tu vas écrire la version finale de l'histoire de Gabriel.

- Relis bien ton premier texte et vérifie que c'est vraiment une histoire qui ressemble, aux contes ou aux récits d'aventures ou encore aux récits fantastiques que tu connais. Tu peux apporter des améliorations et changer des choses.

- Vérifie la ponctuation, puis recopie ton histoire sur une nouvelle feuille.

- Fais attention à l'orthographe.

NOTES

1. Les auteures coordonnent ce groupe qui a réuni de nouveaux corpus entre 2010 et 2015 au Laboratoire EMA (École, Mutation, Apprentissages), EA 4507 de l'Université de Cergy-Pontoise. Voir Annexe II, Tableau 2.

2. En particulier dans le cadre de l'opération de recherche Ecriscol, dirigée par C. Doquet à Paris 3 au laboratoire Clesthia.

3. UMR 7114, Université Paris Ouest Nanterre La Défense.

4. EA 4507, Université Cergy-Pontoise.

5. Informations sur l'établissement, la classe et l'enseignant. Ce qui se passe effectivement en classe : description précise de chaque séance avec consignes, démarches, supports ; enregistrement audio soutenu par une prise de notes, quelques enregistrements vidéo ; et en compléments : ce qui sous-tend le travail en classe : préparations, ajustements ultérieurs, représentations de l'enseignant, notamment son rapport à l'écrit.

6. ITEM : Institut des textes et manuscrits modernes.

7. Ce qui pourrait aussi se lire : « ils rentraient ».

8. En 2010, et auparavant, au laboratoire MoDyCo.

9. Construits entre 1990 et 2010, ils constituent des archives consultables à titre de comparaison.

10. Nous remercions particulièrement Catherine Bosredon, doctorante au Laboratoire EMA, qui a réuni ces corpus, ainsi que Sara Mazziotti, étudiante à Paris 3, qui les a transcrits et annotés. Les 123 copies de ces sous-corpus, issues de deux classes de CE2, et d'une classe de CM1, avec leurs deux versions, ont été produites selon un protocole identique permettant une comparaison. Seules les options pédagogiques des enseignants diffèrent, notamment lors de la réécriture de leur texte par les élèves.

11. Nous ne traiterons pas ici la question de la contrainte d'écriture, inhérente au statut d'élève, et nous bornerons à rappeler que les textes produits sont avant tout destinés à l'apprentissage de l'écriture.

12. La consigne est elle-même un genre discursif hétérogène qui n'est ni forcément ni totalement injonctif, et qui eut s'incorporer des extraits narratifs intertextuels.

13. Quand il s'agit d'intégrer une phrase de roman, ou quand l'énoncé de la consigne prend la forme du début d'un conte, d'un poème, ou de séquences textuelles comme une description, un dialogue etc. À ces contraintes, il convient d'ajouter les éléments interdiscursifs comme les collocations, stéréotypes, citations ouvertes ou cachées, allusions, etc. relevant de formations discursives des discours scolaires.

14. Exemple issu du sous-corpus Bosredon.

15. Ajoutons cependant qu'il est des cas où l'enseignant n'intervient pas immédiatement sur la copie, comme on le voit dans l'Annexe II où l'élève revient seul, en rouge, sur une première version de son texte.

16. C'est le cas lorsque la rédaction du texte repose sur la prise en compte d'images.

17. Expériences ponctuelles avec Cordial et Tropes.

18. Nous verrons que d'autres problèmes se posent pour la ponctuation, et pour la délimitation des phrases.

19. Rastier F., *Passages et parcours dans l'intertexte*, Texto! octobre 2008, vol. XIII, n° 4 [en ligne].

20. Notamment le point, présent dès les premiers écrits, qui représente une borne délimitant des blocs syntaxiques, excédant souvent la phrase canonique.

21. Orsenna E. (2001), *La Grammaire est une chanson douce*, Paris, Stock.

22. Avec le logiciel AntConC.

23. F. Rastier résume sa démarche dans la « Sémantique interprétative » (1987) reprise en ces termes dans www.hermeneutika.cz/soubor/semantique--interpretative/. « Dans son programme, formulé au milieu des années 1980, la sémantique interprétative synthétise ce courant [conception non référentielle et non-compositionnelle du langage], en reconnaissant les lacunes du paradigme logico-grammatical pour proposer une théorie -unifiée, du mot au texte et au corpus. Puisque le global détermine le local, (*nous soulignons*) le corpus de description a une incidence sur le sens du texte, qui à son tour détermine le sens de ses unités, jusqu'au morphème.

Comme dans les grammaires de construction, le problème de la sémiosis (appariement des contenus et des expressions) revêt ainsi une valeur critériale ».

24. Reuter Y. (éd.) (1998), « La description théories, recherches, formation, enseignement », Lille, Presses universitaires du Septentrion, p. 150.

RÉSUMÉS

Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles

L'article met en perspective des recherches qui ont pour objectif commun la constitution de corpus scolaires recueillis dans des conditions écologiques. Il expose les questions méthodologiques qui se sont posées au départ concernant le choix et l'organisation des données, la définition d'une unité d'observation, les différentes possibilités de transcription. Les limites rencontrées dans l'exploitation de transcriptions non annotées ont conduit à l'identification de problèmes récurrents : la caractérisation des genres scolaires, genres dialogiques et multi-auctoriaux, la tension entre normalisation et prise en compte de la textualité des écrits. L'analyse d'un sous corpus numérisé et transcrit montre comment les décisions méthodologiques prises pour l'exploitation du corpus Ecriscol prennent acte de ces difficultés.

Tow steps towards building school writing corpora. Recurrent problems and new perspectives

This paper builds on researches which aim at a common purpose: the construction of school writing corpora collected in ecological conditions. It presents initial issues as far as methodology is concerned: the choice and organisation of data, the definition of a basic observation unit, the various transcription standards. But there are limits to non annotated transcriptions, which allow identification of recurrent problems: how to characterize school genres, which are dialogical genres with many authors? how to manage the tension which prevails between standardisation and textuality. The paper analyses a small digitised corpus; it shows how methodological decisions take into account these difficulties in the annotation of the corpus Ecriscol.

INDEX

Mots-clés : corpus, écrit scolaires, dialogisme, textualité, annotation.

AUTEURS

CATHERINE BORÉ

ÉMA (EA 4507) – Université de Cergy-Pontoise – F-95000

MARIE-LAURE ELALOUF

ÉMA (EA 4507) – Université de Cergy-Pontoise – F-95000