Transcrire et *normer* un corpus scolaire, pour quelles analyses ?

Claire WOLFARTH, Catherine BRISSAUD &

Claude PONTON

Résumé : L'apprentissage de l'écriture et de la production d'écrit est devenu un enjeu majeur pour l'école dans une société où les compétences à l'écrit sont rendues toujours plus nécessaires. La recherche en didactique de l'écrit a donc ici un rôle majeur à jouer pour accompagner les enseignants dans la mise en œuvre de cet apprentissage. De la même manière que l'introduction de corpus de grandes tailles dans les sciences du langage a changé les méthodes et les objets de ces sciences, l'introduction de tels corpus dans la didactique du français peut contribuer à un réel changement. Dans cette optique, nous développons le corpus *Scoledit*, qui vise l'exploitation de 7000 productions d'apprenants. Dans cet article, nous nous pencherons sur la numérisation et les voies de diffusion de ce corpus, ainsi que sur l'intérêt d'en proposer une version « normalisée » pour pouvoir l'analyser à l'aide de méthodes issues du traitement automatique des langues.

L'écrit occupe de plus en plus de place dans nos sociétés, c'est pourquoi l'apprentissage de l'écriture et de la production d'écrit est un enjeu majeur pour l'école d'aujourd'hui. Le nombre conséquent de travaux de recherche qui y est consacré (Doquet *et al.*, 2016) souligne bien l'importance de cet enjeu. Il est ainsi nécessaire d'outiller la recherche afin de mieux cerner cet objet d'étude et d'opérer un renouvellement et un changement d'échelle comparable à celui qui s'est produit en sciences du langage suite à l'emploi de grands corpus

(Elalouf, 2011). Nous proposons d'accompagner ce changement par l'élaboration d'un nouveau corpus d'apprenants.

À l'heure actuelle, les grands corpus d'apprenants concernent plutôt les langues secondes (Granger, Vandeventer, Hamel, 2001 ; Agren, 2008) et sont orientés vers l'analyse des erreurs. Des études en français langue première ont été conduites depuis longtemps sur un nombre de textes parfois important : P. Clanché (1988), sur plus de 7000 textes libres écrits par 200 élèves de 6 à 10 ans durant une année ; C. Fabre (1990), sur 300 brouillons d'écoliers ; M.-L. Elalouf (2005), quelque 500 textes rassemblés dans huit classes de CM2 et de 6e ; H. Andersen, C. Leblay et E. Auriac-Slusarcyk (2010), sur plusieurs centaines de récits et comptes-rendus scientifiques, recueillis en CE2, CM2, 6e et 4e ; C. Garcia-Debanc et K. Bonnemaison (2014), sur près de 400 textes autour d'une tâche de cohésion textuelle ; J. David et C. Doquet (2016), sur plus de 800 productions d'apprenants à l'école primaire.

Ces corpus sont souvent peu homogènes et pour la plupart inaccessibles ou pas encore accessibles en ligne : ils ne sont ni transcrits numériquement, ni enrichis, et ne sont donc pas utilisables directement par la communauté enseignante ou scientifique. Loin de cantonner le processus de constitution de corpus à la tâche de recueil, nous proposons de considérer l'élaboration d'un corpus comme un processus plus large qui comprend à la fois le recueil et la numérisation de ce corpus, la construction d'outils d'exploitation à l'aide de méthodes issues du traitement automatique des langues (TAL) et la diffusion de ces données et outils.

L'objet du projet que nous décrivons dans cet article est la constitution d'un corpus numérique longitudinal de textes scolaires et de dictées produits par des élèves de 6 à 11 ans rencontrés à plusieurs reprises lors de leur scolarité élémentaire. L'ensemble du corpus devrait rassembler plus de 7 000 productions permettant l'étude longitudinale des procédés d'écriture de différents niveaux (orthographe, syntaxe, ponctuation, cohérence textuelle).

L'enjeu d'un tel projet est multiple. Il s'agit, dans un premier temps, de permettre une description linguistique des écrits

d'apprenants à l'école primaire à la fois en synchronie et en diachronie, du CP au CM2. Ceci devrait apporter une connaissance plus fine des phénomènes d'acquisition de l'écriture comme l'évolution des compétences orthographiques et syntaxiques, l'évolution de l'acquisition de la morphosyntaxe, celle de la cohérence des temps ou encore l'évolution de l'usage de la ponctuation. Pour ce projet, ce sont ces critères linguistiques qui nous intéressent plus particulièrement mais le corpus et les outils élaborés devraient également permettre l'étude de critères textuels et langagiers, et ainsi s'intéresser à la construction du texte ou à l'évolution du geste graphique, pour ne citer que ces exemples. Nous faisons l'hypothèse, tout comme C. Bonnet (1998), C. Fabre-Cols (2000), C. Boré et M.-L. Elalouf (2016), que l'accès à un grand nombre de textes d'élèves répondant à une consigne commune permettra aux enseignants et formateurs d'acquérir une culture de ces textes en les mettant en regard les uns avec les autres. Ainsi, il serait possible de connaitre de manière plus juste les compétences que peuvent acquérir les élèves à un niveau donné. Dans un deuxième temps, il devrait donc être possible, à partir de ce projet, d'élaborer des séquences et des dispositifs didactiques à destination des enseignants.

Précisons également que ce projet représente un véritable enjeu pour le TAL. En effet, il s'agit d'un type de corpus encore peu étudié et, en raison de son éloignement à la norme, son traitement automatique constitue un véritable défi.

Dans une première partie, nous présenterons les différentes étapes de la constitution du corpus : la méthodologie du recueil et les étapes de numérisation, ainsi que les voies de diffusion de ce dernier. Puis, nous nous attarderons sur les méthodes d'analyse et d'exploitation de ce corpus, en nous penchant sur l'exemple de la segmentation en propositions et en phrases, au travers de l'usage de la ponctuation et des connecteurs. Nous terminerons par l'exposé des axes de travail actuels et futurs du projet *Scoledit*.

1. Les premières étapes de la constitution du corpus

Avant de nous pencher plus particulièrement sur la spécificité de notre travail situé au croisement de la didactique du français et du TAL, nous présenterons les étapes initiales qui sous-tendent la constitution d'un corpus, à savoir le recueil des données, leur numérisation et leur support de diffusion.

1.1. Recueil des productions

Le corpus que nous sommes en train de construire contiendra dans sa version finale des productions d'élèves suivis tout au long de leur scolarité à l'école primaire. Le recueil se déroule dans 40 écoles dont 37 issues du projet « Lire-écrire au CP » de l'Institut Français de l'Éducation[1]. Ces écoles se répartissent dans cinq académies : Grenoble, Lyon, Clermont-Ferrand, Toulouse et Bordeaux. Pour chaque élève, ont été recueillies :

- une dictée à l'entrée en CP ;
- une dictée et une production de texte à la fin du CP ;
- une dictée et une production de texte à la fin du CE1 ;
- une production de texte à la fin du CE2[2] ;
- une dictée et une production de texte à la fin du CM1.

Aux mois de mai et juin 2018, nous recueillerons également des productions de texte et des dictées réalisées en CM2.

Pour l'heure, le corpus recueilli représente plus de 6 500 productions[3], dont plus de 3 000 dictées et plus de 3 500 productions

---

1 Cette recherche a été financée par l'Institut Français de l'Éducation, la DGESCO et le laboratoire ACTé de l'Université Clermont Auvergne.
Goigoux, R. (2015). *Lire et Écrire. Étude de l'influence des pratiques d'enseignement de la lecture et de l'écriture sur la qualité des premiers apprentissages.* Institut Français de l'Éducation. Rapport de recherche.
Goigoux, R., Jarlégan, A., & Piquée, C. (2015). Évaluer l'influence des pratiques d'enseignement du lire-écrire sur les apprentissages des élèves : enjeux et choix méthodologiques. Recherches en Didactiques. *Les Cahiers Théodile,* (17), 33-52.

2 En raison des difficultés liées à l'élaboration d'un corpus longitudinal, aucune dictée n'a pu être recueillie en classe de CE2.

de textes narratifs. La répartition des productions est détaillée dans la Figure 1. Les contraintes de recueil d'un corpus longitudinal sont lourdes et ne permettent pas d'envisager de suivre l'ensemble des élèves sur les cinq années : élèves changeant d'école, élèves malades, changements d'enseignants, etc. Ainsi, nous ne disposons actuellement de l'ensemble des productions CP /CE1 / CE2 que pour 259 élèves sur les 1 649 élèves au total. Les productions de ces 259 élèves constituent le *corpus longitudinal* (voir Figure 1). La recherche « Lire-écrire au CP » de l'Institut Français de l'Éducation, de laquelle est issue la majorité des productions de CP et de CE1, ne s'était intéressée qu'à certaines classes d'enseignants volontaires, en général une seule par école. Pour le recueil de CE2, nous avons décidé de faire participer l'ensemble des élèves de chaque école. L'ajout de ces élèves, ainsi que de trois écoles, forment le *corpus complémentaire*, soit 518 productions pour le CE2.

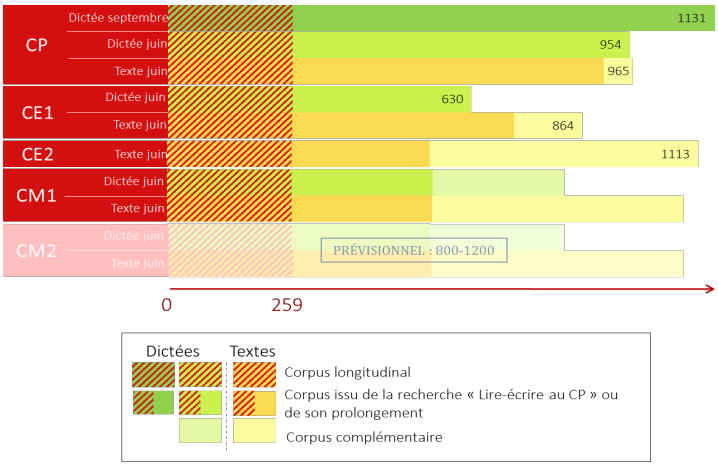


Figure 1 : Structure du corpus[4]

[3] Ne sont incluses ici que les productions pour lesquelles nous disposons d'une autorisation de diffusion.

[4] Les données concernant l'année de CM1 sont encore provisoires.

La méthode de recueil de ce corpus revêt deux intérêts particuliers. Le premier intérêt réside dans le fait que les productions sont réalisées par les mêmes élèves, tout au moins pour une partie des productions, à différents moments de leur apprentissage. Le deuxième avantage n'est autre que l'unicité de la consigne donnée aux élèves. En effet, pour chaque type de production, la même consigne a été donnée à l'ensemble des élèves à partir du CE1.

À l'entrée au CP, il a été proposé aux élèves, entre autres épreuves qui ne concernent pas notre corpus (voir la recherche « Lire-écrire au CP » de l'Institut Français de l'Éducation), une dictée de trois mots *lapin*, *rat* et *éléphant* et une phrase *Tom joue avec le rat*. Ces mêmes mots et phrase ont à nouveau été dictés en fin de CP ainsi que la phrase *Les lapins courent vite*, afin de tester le marquage en nombre nominal et verbal. Pour l'épreuve de production de texte, il s'agissait d'une production de texte narratif à partir de 4 images séquentielles (annexe I) et de la consigne suivante « Aujourd'hui vous allez écrire chacun l'histoire d'un petit chat. Je vais vous montrer ce qui arrive à ce petit chat. Regardez bien les images. […] Vous allez écrire cette histoire ici. […] Vous avez 15 minutes pour ce travail[5]. […] ».

En fin de CE1, une nouvelle dictée a été proposée contenant six mots *patin*, *pâtisson*, *capuchon*, *récréation*, *charitable* et *magnifique*, et deux phrases : *En été, les salades vertes poussent dans les jardins* et *Les jeunes canetons picorent le blé avec la poule noire,* provenant des évaluations nationales de la DEPP. Après avoir choisi un ou deux personnages parmi quatre (annexe I), les élèves de CE1 devaient également produire un texte narratif, à partir de la consigne « Voici 4 personnages. Choisis un ou deux personnages et raconte une histoire. Entoure le ou les personnages que tu as choisis. »La même épreuve a

[5] Pour une présentation détaillée de l'épreuve, se référer à SOULÉ, Y., KERVYN, B. GEOFFRE, T. & CHABANNE, J.-C. (2016). « Évaluer la production d'écrit en fin du cours préparatoire (première primaire). De l'élaboration d'une épreuve de test à l'analyse des résultats obtenus ». In J. Dolz , J.-L. Dumortier, É. Falardeau et P. Lefrançois, L'évaluation en classe de français, outil didactique et politique, Namur, Presses Universitaires de Namur, p. 85-107

été proposée en CE2 et CM1 à la différence que ces années-là les élèves disposaient de 30 minutes au lieu de 20 en CE1.

Les données recueillies à l'aide de ces diverses consignes correspondent toutes à un « premier jet » ; ces productions n'ont fait l'objet d'aucune relecture ou correction de la part de l'enseignant ou d'un adulte. De plus, l'aide d'un adulte ou d'un outil, que ce soit un dictionnaire ou un manuel de conjugaison, n'était pas autorisé. Un exemple de production est donné en annexe II.

1.2. Numérisation et diffusion du corpus

Collecter et rassembler les productions n'est que la première étape du processus d'élaboration d'un corpus. Il faut ensuite numériser ces productions, c'est-à-dire les scanner, les transcrire et les enrichir (les décrire à l'aide de métadonnées et les annoter), avant de pouvoir les analyser et les diffuser. L'étape de transcription et les problèmes qu'elle pose ont fait l'objet d'une publication récente (Wolfarth *et al.*, 2016), nous ne les développerons donc pas ici. Précisons cependant que nous avons fait le choix, lors de cette transcription, de ne pas nous attarder sur les aspects qui relèvent du domaine de la génétique du texte telles que les traces de révision, mais de nous concentrer sur la production finale de l'élève.

Les étapes de numérisation, c'est-à-dire à la fois le scan et la transcription du corpus, sont des étapes indispensables pour l'analyse et la diffusion des productions. Nous avons choisi de rendre accessible le corpus via un site internet[6] à destination des enseignants, linguistes et didacticiens qui voudraient s'y intéresser. Ce site permet d'afficher l'ensemble des productions d'un élève année par année (Figure 2). Ce site permet également d'effectuer diverses recherches dans les données de ce corpus et constitue un espace de dialogue avec

[6] Ce site est disponible à l'adresse suivante : http://otus.u-grenoble3.fr/scoledit/. Il comporte pour le moment uniquement les productions de CP et de CE1 ; celles de CE2 sont en cours d'ajout. Son utilisation est ouverte mais soumise à un enregistrement préalable.

les enseignants et les divers utilisateurs du corpus. Nous reviendrons sur ces fonctionnalités à la fin de cet article.

Figure 2 : Site de présentation du corpus Scoledit, élève 49, CP

Toutefois, pour pouvoir exploiter finement ce corpus à des fins linguistiques ou didactiques, il est nécessaire d'enrichir ses données. En effet, les seules transcriptions « brutes » ne permettent pas des recherches précises comme, par exemple, trouver toutes les variantes de tel ou tel mot. Plusieurs approches sont possibles et divers traitements vont donc être nécessaires.

1.3. Préparation à l'analyse du corpus

La plupart des outils issus du TAL, développés à l'aide de corpus standards, tels que les corpus journalistiques, les corpus littéraires ou les corpus oraux transcrits, utilisent pour unité le mot, appelé *token*, éventuellement le caractère ou la phrase. L'aspect peu normé des productions scolaires tant d'un point de vue syntaxique, qu'orthographique ou de la construction du récit ne permet pas

l'utilisation des outils usuels de traitement de corpus. En effet, la plupart de ces outils se basent sur la reconnaissance des formes produites pour engendrer les diverses analyses lexicales, syntaxiques et textométriques. Face à des formes à l'orthographe irrégulière, cette reconnaissance n'est plus possible. Le premier enjeu de notre analyse va donc être d'identifier les formes qui composent notre corpus.

Cette tâche peut être vue comme une tâche de « correction » de notre corpus pour le faire tendre vers une norme plus proche des corpus standards. L'objectif d'obtenir cette approximation de la norme n'est pas de la présenter comme modèle aux enseignants qui souhaiteraient consulter notre corpus mais bien de l'utiliser comme support à l'outil d'analyse. Pour tendre vers cette norme, plusieurs méthodes sont envisageables.

Premièrement, il est possible d'utiliser les outils classiques de correction tels ceux qu'utilisent les outils de traitements de textes ou encore des outils plus spécialisés comme Antidote [7]. Malheureusement, aucun de ces outils n'obtient de résultat satisfaisant pour notre corpus, notamment en raison du nombre élevé d'erreurs orthographiques et de segmentation que les systèmes ne sont pour la plupart pas capables d'identifier.

Deuxièmement, il est possible d'annoter manuellement notre corpus erreur par erreur, à l'image du travail de l'équipe de J. David et C. Doquet (2016) pour le projet *Ecriscol*. Cependant, cette méthode est très coûteuse en temps et en moyens humains. Nous souhaitons de fait développer une troisième alternative que nous espérons plus adaptée et moins coûteuse. Le développement de celle-ci et des difficultés qui en découlent sera l'objet de cette deuxième partie.

1.4. Méthode et résultats attendus

Afin d'identifier mots, signes de ponctuations, propositions et autres composants des productions d'élèves, nous proposons de comparer ces productions à une version dite normée, proche des corpus standards.

[7] https://www.antidote.info/

Dans le cadre des dictées, la norme est la même pour tous les apprenants, il s'agit des mots ou des phrases qui leur ont été dictés. En observant la production de l'élève à la lumière des éléments dictés, il va donc être possible, à l'aide d'un algorithme d'alignement, d'identifier automatiquement le contenu de cette production. Par exemple, pour la dictée de fin de CP de l'élève 72 (Figure 3), cette méthode va nous permettre d'identifier la séquence produite « la piun » comme étant la transcription de *lapin*, ce que ne nous aurait pas permis un correcteur classique.

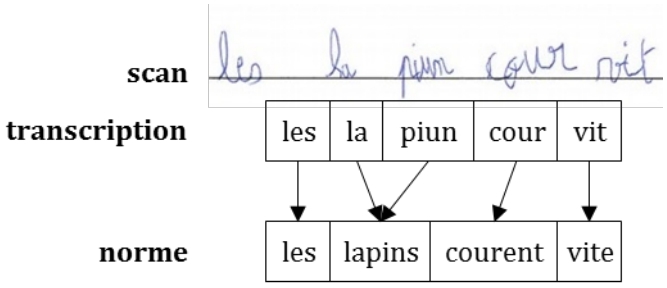


Figure 3 : Extrait de la dictée réalisée en fin de CP par l'élève 72

Comme nous l'avons mentionné, à partir de cette comparaison nous pouvons identifier la majorité des formes produites par les élèves. Il nous est alors possible d'intégrer des outils de recherche à la plateforme de diffusion. Sur l'exemple suivant (Figure 4), la recherche de la forme *canetons* dans les dictées de CE1 donne les résultats suivants : cette forme a été produite 625 fois, à l'aide de 77 variantes graphiques différentes. La forme attendue « canetons » a été produite 160 fois, tandis que la forme au singulier « caneton » a été produite 239 fois.

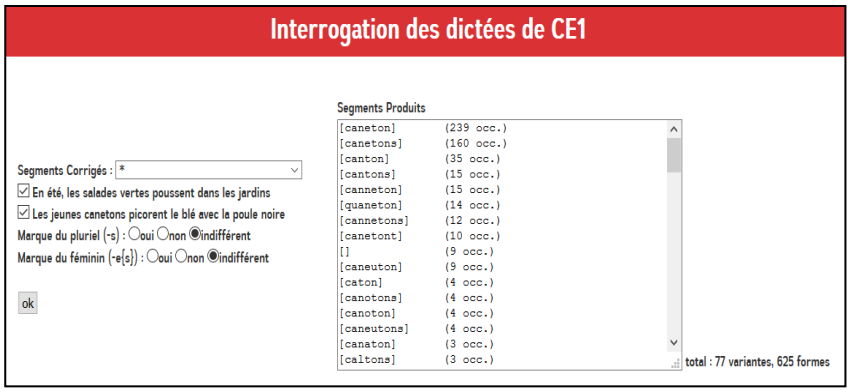


Figure 4 : Exemple de requête, variantes de la forme canetons dans les dictées de CE1

Aligner production attestée et norme attendue permet donc d'améliorer les possibilités d'interrogation de notre corpus. Cependant, si la norme est facilement identifiable pour les mots et phrases produits sous dictée, il n'en est pas de même pour les textes produits de manière moins contrainte. Pour ces dernières productions, nous allons devoir construire ce qui s'apparentera à une approximation de la norme à partir de laquelle, après alignement, il sera plus simple d'analyser notre corpus.

Néanmoins, le premier problème que pose la construction de cette approximation de la norme, que nous nommerons *version normée* par la suite, est sa définition même. Nous n'aboutirons pas à une version de la production de l'élève entièrement normée et acceptable pour un scripteur adulte, tant d'un point de vue orthographique que stylistique. La définition de cette étape de normalisation est en cours mais nous donnons ci-après quelques éléments de réflexion sur lesquels nous travaillons. Rappelons avant tout que l'objectif premier de la normalisation est de servir de support à l'interrogation du corpus.

1.5. Élaborer une version normée, orthographe et segmentation

Dans le cas des dictées, comparer les formes produites aux formes attendues permet d'identifier les mots produits par les enfants en normalisant les variantes d'orthographe et de segmentation. Dans le cas des productions de textes, normer ces variantes n'est pas un problème aussi trivial qu'il n'y parait. Ainsi, de la même manière que transcrire implique différents choix de transcription, proposer une *version normée* des productions implique des choix de *normalisation* corrélés à la définition de norme que l'on adopte et aux phénomènes que l'on veut pouvoir repérer et analyser dans notre corpus. Dans l'optique de ne pas considérer l'élève écrivant comme un scripteur adulte incomplet commettant des erreurs mais plutôt comme un scripteur apprenant écrivant avec une logique propre à l'apprentissage, nous avons choisi de normer les productions au plus près de ce qu'a écrit l'apprenant. Selon cette logique, entre autres exemples, le choix a été fait de ne pas normaliser le temps des verbes mais de les considérer tels que produits par l'élève, en modifiant cependant l'orthographe et la morphologie flexionnelle pour permettre une analyse outillée. Pour illustrer nos propos, nous pouvons citer notamment les extraits de production « Le petit chat et partie de son lit. Mai boum il tombe et il pleurer. Sa maman et saifraire se raivaya et il le voillér pleurer. Et miantenansai le matin. La mamans les porte pour lessortir du lit.Fin» (production 586, CP), pour laquelle on proposera la *version normée* « Le petit chat est parti de son lit. Mais boum il tombe et il pleurait. Sa maman et ses frères se réveilla et ils le virent pleurer. Et maintenant c'est le matin. La maman les porte pour les sortir du lit. Fin ».

Dans cet exemple, la normalisation de la segmentation en mots et des variantes orthographiques a permis à la fois d'identifier les segments correctement réalisés comme « petit », « chat », « matin », etc., les segments mal orthographiés comme « sai » (*ses*) et « fraire » (*frères*) mais aussi les segmentations non conformes comme dans « sai » (*c'est*). Ces choix de normalisation sont toujours faits au plus près de la production de l'enfant, de sorte qu'ils entrainent le moins de modifications possibles. Notons ici que ces choix sont assez

similaires aux choix réalisés par M.-L. Elalouf (2005), M.-N. Roubaud et P. Cappeau (2005), J. David et C. David (2016) pour élaborer une version orthographiquement normée des productions de leurs corpus respectifs.

Cependant, ces choix ne sont pas les seuls problèmes que pose la normalisation des productions de textes. Normer l'orthographe et la segmentation en mots va permettre le même résultat que pour les dictées, à savoir l'identification des formes utilisées, mais ne va pas permettre de traiter les difficultés d'analyse rencontrées à des niveaux plus macroscopiques comme celui de la syntaxe. Pour traiter ce genre de problèmes, il faudra également s'attarder sur les marqueurs de segmentation en phrases ou en propositions.

1.6. Élaborer une version normée, le cas de la segmentation

Pour illustrer les propos qui vont suivre, nous proposons de considérer la production de texte en fin de CP de l'élève 93 (Figure 5), que l'on peut transcrire ainsi : « un petti chat ne dor pluiltonbe de les sicille et il C'estfest male est ilplere est sa mér le ra mene. » et dont on peut produire la *version normée* suivante : « Un petit chat ne dort plusil tombe de l'escalier et il s'est fait mal et il pleure et sa mère le ramène ». En observant les lieux de segmentation entre les propositions (soulignés dans l'exemple), il semble probant que les principes énoncés au paragraphe précédent ne permettent pas de prendre en compte l'absence de ponctuation ou de marque de segmentation entre les segments « Un petit chat ne dort plus » et « il tombe de l'escalier ». Cet exemple nous montre que, tout comme nous l'avons fait au niveau du mot, il peut être intéressant de rétablir les marqueurs de segmentation en propositions, quelle que soit la forme qu'ils prennent, afin de permettre une analyse plus globale des productions. Nous pouvons également noter la répétition du connecteur « et », mais en vertu du principe qui nous amène à rester au plus proche du texte produit par le scripteur apprenant nous maintenons l'ensemble de ces connecteurs dans la *version normée*, nous contentant d'en normer l'orthographe. Nous parlerons ici de marqueurs de propositions, englobant à la fois signes de ponctuation

et connecteurs, ces derniers pouvant remplir la même fonction. De même, nous parlerons de segments ou de propositions et non de phrases, la phrase étant une unité souvent dépendante des signes de ponctuation employés.

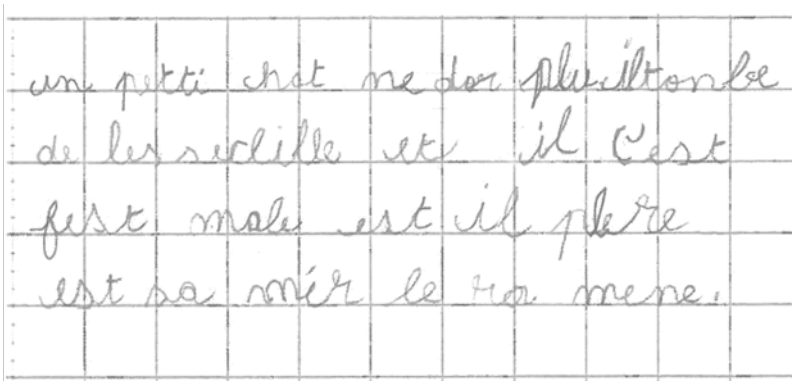


Figure 5 : Extrait d'une production de texte réalisée en fin de CP

Au vu de l'exemple présenté ici, il semble que les marqueurs de segmentation en propositions peuvent être absents dans certaines productions. Afin de permettre une analyse syntaxique de celles-ci, il va être intéressant de rétablir ces marqueurs. Cependant, comme nous l'avons évoqué, ces marqueurs peuvent être de différentes natures puisqu'ils peuvent à la fois être de nature typographique (signes de ponctuation) ou lexicale (connecteurs). À l'intérieur même de ces catégories, les possibilités sont souvent multiples, particulièrement dans les productions d'apprenants qui sont souvent une succession de propositions simples ou relativement simples, à l'image de l'exemple ci-dessus (Figure 5). La question majeure est donc ici : si nous voulons diviser le segment « Un petit chat ne dort plusil tombe de l'escalier » en deux propositions distinctes, quel marqueur allons-nous utiliser ?

Une première solution pour répondre à cette question pourrait être de se pencher sur un manuel de ponctuation. Cependant selon les mots de J. Popin et C. A. Thomasset (1998, p. 13), la ponctuation « échappe au concept de norme » et « entre dans le domaine du

standard ». D. Bessonnat (1991) distingue également une ponctuation prescriptive, régie par une norme, d'une ponctuation facultative, appartenant au domaine de la stylistique. La ponctuation est donc sujette à variation selon les scripteurs. Il est donc fort à parier qu'elle est également sujette à variation entre scripteur expert et scripteur débutant. Si l'on ajoute à cette variance la possibilité de substituer des connecteurs à la ponctuation, il devient impossible de faire un choix argumenté.

Une autre possibilité serait de se référer à ce que produisent les scripteurs débutants eux-mêmes. L'observation de l'emploi de ces marqueurs dans 70 productions de CP est reportée dans le tableau suivant (Tableau 1). Au vu de celui-ci, il semble que le marqueur le plus fréquemment utilisé par les scripteurs débutants est le connecteur *et* ; cependant de nombreuses études (Chanquoy, et Fayol, 1991 ; Passerault, 1991 ; Fayol *et al.*, 2014 ; Paolacci et Gensane, 2014 ; Paolacci et Rossi-Gensane, 2016) nous amènent à penser que ces marqueurs évoluent au fil de l'acquisition et tendent à une plus grande diversité. Un même marqueur ne peut donc être utilisé pour l'ensemble des années de recueil. Dans une perspective d'analyse longitudinale, l'idée de marquer une segmentation en propositions absente par un marqueur unique ne peut donc être retenue.

Marqueurs interpropositionnels								217
Ponctuation								**49**
point	45	retour à la ligne	16	point d'exclamation	4	virgule		0
Connecteurs								**152**
coordination								**145**
et	112	puis	6	alors	1	maintenant	1	
après	10	tout à coup /	3	car	1	où	1	
mais	8	tout d'un coup		ensuite	1	soudain	1	
subordination								**7**
parce que	4	pendant que	3					
Aucune marque								**69**

Tableau 1 : Utilisation de la ponctuation et des connecteurs au CP (70 productions)

En cas d'absence de marqueur de segmentation, le choix d'un marqueur (ponctuation ou connecteur) serait arbitraire. Par conséquent, nous avons fait le choix d'utiliser une balise neutre <segmentation / > indiquant cette absence.

Toutefois, la balise <segmentation / > n'épuise pas toutes les questions. Par exemple, il faudrait se demander si cette balise n'est pas trop générique et s'il y a une pertinence linguistique ou didactique à distinguer les cas de segmentation. Dans l'exemple de la production de l'élève 2522 (classe de CP) : « […] la maman chat se reveiremei son petit é il sen dors. », que l'on peut normaliser : « […] la maman chat se réveille **<segmentation / >** remet son petit et il s'endort. », la balise <segmentation / > pourrait être remplacée par le connecteur *et* ou par une virgule, mais ne pourrait pas l'être par un point du fait de l'ellipse du sujet, contrairement à l'exemple 93, étudié plus haut (Figure 5). Cette question renvoie également à la question des différentes fonctions (Chanquoy et M Fayol, 1991) de ces marqueurs et des différences d'intensité (Fayol *et al.*, 2014), selon qu'ils marquent une rupture faible ou forte entre les propositions, un enchainement temporel ou une opposition. De plus, les fonctions de ces marqueurs invoqués par les scripteurs experts (Popin et Thomasset, 1998 ; Drillon, 1991 ; Jaffré, 1991) ne sont souvent pas les mêmes que celles invoquées par les scripteurs débutants (Passerault, 1991). Cette question nécessite une plus large investigation, tant sur sa pertinence pour les traitements que pour sa résolution, mais nous pouvons tout de même retenir une piste qui serait de classer ces marqueurs selon les critères didactiques proposés par D. Bessonnat (1991) : portée du marqueur, sa forme, son degré hiérarchique, etc.

À travers cet exemple, nous avons voulu montrer que marquer un élément absent des productions à l'aide d'un remplaçant générique, en l'occurrence d'une balise pour les marqueurs de segmentation, permet de rendre la normalisation moins dépendante du transcripteur et de sa sensibilité. Bien que cette solution ne résolve pas tous les problèmes, elle permet néanmoins une plus grande uniformité des données. Nous proposons donc d'adapter ce système à d'autres phénomènes relevant plus souvent de choix stylistiques que

d'une norme comme les marques de ponctuation à l'intérieur des propositions (marques d'énumération, conjonction dans un groupe nominal, etc.), qui ne sont pas toujours obligatoires, ou encore les marques de dialogue souvent absentes en classe de CP.

Cet exemple nous permet également de prendre la mesure des choix de *normalisation* et de leur implication dans la suite des traitements qui pourront être effectués sur notre corpus.

2. Perspectives

Le travail présenté ici n'en est qu'à son début et permettra de prolonger la réflexion au-delà des aspects que nous venons d'exposer. Comme nous l'avons mentionné dans cet article, si les choix de transcriptions sont désormais fixés pour l'ensemble du corpus et que la majorité des productions de CP et de CE1 a pu être transcrite selon ces choix, il n'en est pas de même pour les choix d'élaboration des *versions normées* de ces productions. Ces choix sont dépendants de l'objectif fixé, c'est-à-dire des analyses qui seront effectuées sur le corpus. Ces choix varient donc selon l'évolution de la réflexion et les analyses projetées. D'un point-de-vue didactique, les questions de normalisation sont fortement corrélées avec la question de l'évaluation ou de la correction d'un texte d'élève et la question du respect de ce texte.

L'avancée de ces réflexions va nous permettre de mettre en place différents outils dans le but de visualiser l'évolution des compétences à l'écrit. *In fine,* il s'agit de mieux cibler les attendus en matière de production d'écrit. Nous en avons vu un exemple (paragraphe 0.), certains de ces outils sont déjà disponibles ou en cours de diffusion sur le site internet du projet. Nous pouvons également mentionner ici la recherche de formes à l'intérieur des productions de textes de CP et de CE1 (annexe III**Erreur ! Source du renvoi introuvable.**) et l'accès à trois lexiques (annexe IV**Erreur ! Source du renvoi introuvable.**) :

- la liste des formes produites dans notre corpus ;
- la liste des formes corrigées, c'est-à-dire dont l'orthographe a été normée ;

- la liste du vocabulaire, cette liste n'affichant pas les formes fléchies mais les regroupant sous un même lemme s'il y a lieu.

Le repérage de réussites à un moment donné ou, au contraire, d'écueils récurrents permettra de proposer des activités ciblées. Pour plus de clarté, prenons l'exemple du lemme $CHAT_N$. La liste de formes produites affichera entre autres : *ch*– 3,*cha*– 145, *chas* – 25*, chat* – 759 et *chats* – 15. La liste des formes corrigées affichera : *chat* – 989 et *chats* – 80. La liste du vocabulaire affichera : *chat* – 1052.

Ces listes laissent transparaitre encore de nombreuses erreurs et incohérences parce que, comme nous l'avons montré dans cet article, l'identification des formes normées, et par conséquent les lemmes, ne peut se faire sans établir une *version normée* au préalable. Le travail présenté ici n'a pas pour vocation à être utilisé tel quel pour le moment, mais doit permettre d'entrevoir l'utilité de nos outils. C'est donc en améliorant cette première étape que nous améliorerons nos outils et en construirons de nouveaux.

Cependant, nous n'envisageons pas cette construction d'outils comme une tâche solitaire mais comme une tâche de dialogue avec les utilisateurs ou futurs utilisateurs de notre plateforme, qu'ils soient enseignants, linguistes ou didacticiens. À cette fin, différents espaces de dialogue ont été ouverts sur notre site qui permettent de laisser des commentaires divers, tant sur la qualité des données, que sur la structure de la plateforme, les besoins en outillage spécifiques, etc. De plus, dans une perspective d'amélioration de la qualité des données, nous avons également permis pour chaque transcription la possibilité de laisser un commentaire et ainsi relever les erreurs ou les points de désaccord.

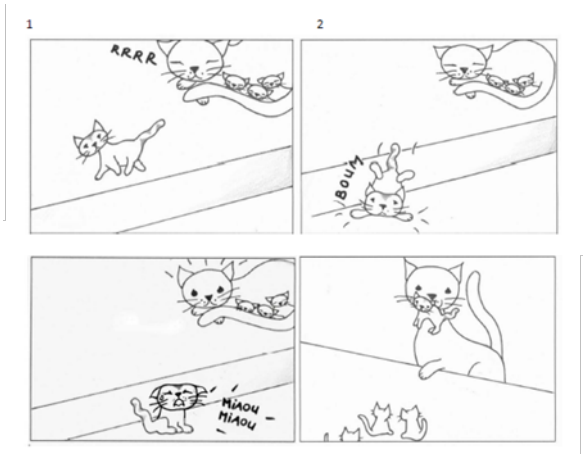
Si cette plateforme est encore un outil en construction, elle ouvre déjà certaines perspectives d'exploitation du corpus *Scoledit*. Elle permet ainsi de ne pas envisager la constitution d'un corpus comme la simple accumulation de textes ou de productions mais révèle également la nature des traitements qui peuvent enrichir l'exploitation et les sources d'usage d'un corpus.

Remerciements

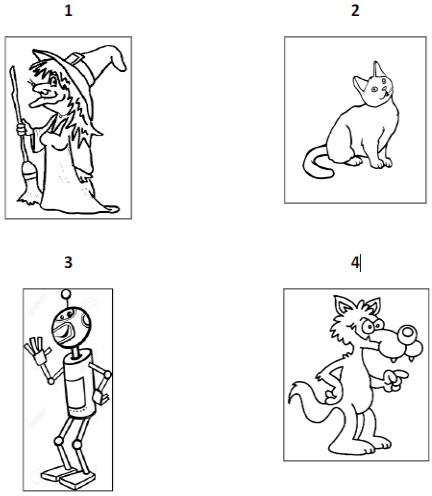
Ce travail est soutenu par Démarre SHS ! de l'Institut des Données de Grenoble et a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-15-IDEX-02.

Annexe I

Supports pour les tâches de productions de textes, issus de la recherche « Lire-écrire au CP » de l'Institut Français de l'Éducation



Images séquentielles présentées aux enfants en fin de CP



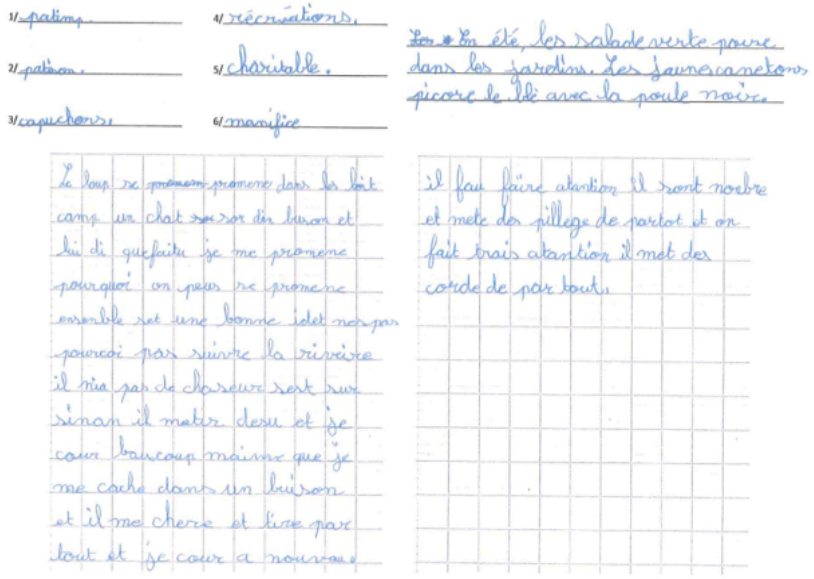
Images présentées aux enfants en fin de CE1 et de CE2

Annexe II

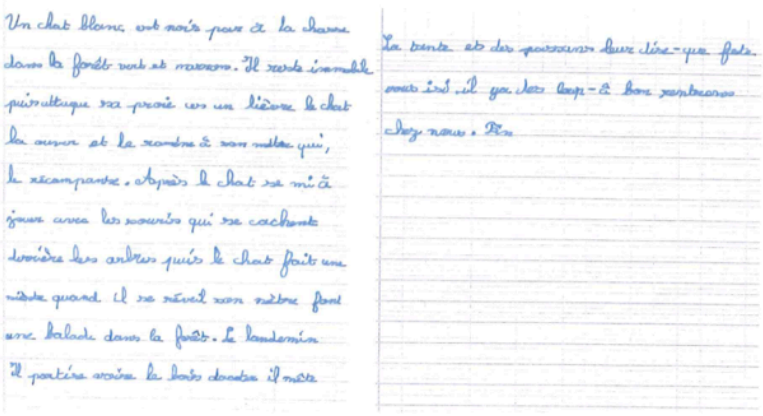
Exemples de productions produites par l'élève 1558.

Juin (fin CP)

Exemple de productions réalisées en classe de CP (élève 1558)

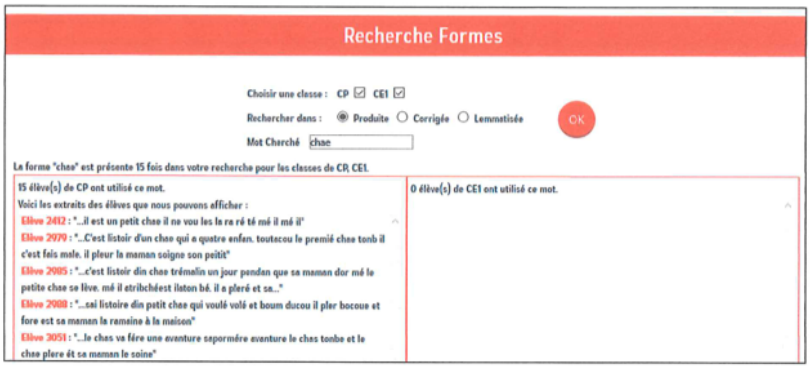


Exemple de productions réalisées en classe de CE1 (élève 1558)



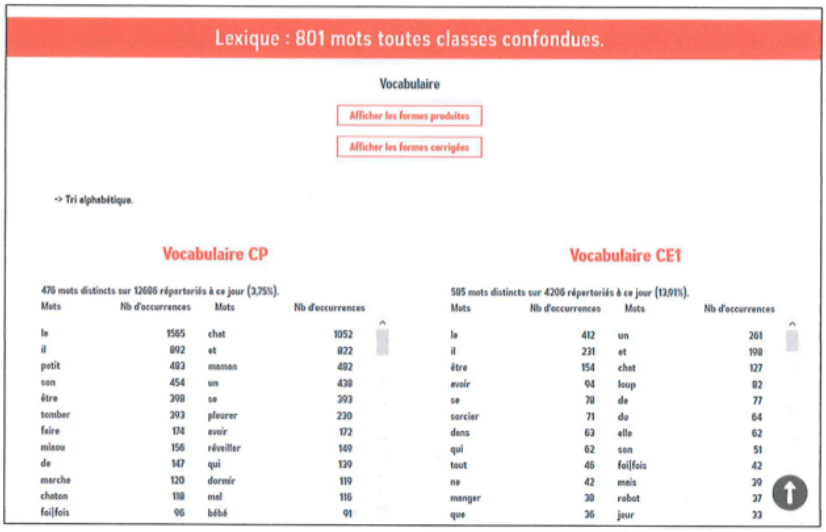
Exemple de production réalisée en classe de CE2 (élève 1558)

Annexe III



Exemple de recherche de forme possible, résultat pour la forme *chae*

Annexe IV



Affichage du vocabulaire contenu dans les productions de CP et de CE1

Bibliographie

AGREN, M. (2008). À la recherche de la morphologie silencieuse : Sur le développement du pluriel en français L2 écrit (Vol. 84). Lund University.

ANDERSEN, H.-L., LEBLAY, C. & AURIAC-SLUSARCZYK, E. (2010). Pourquoi travailler sur un corpus commun ? Pourquoi travailler de manière pluridisciplinaire ? Synergies Pays Scandinaves, (5), 17-30.

BESSONNAT, D. (1991). Enseigner la..."ponctuation" ? (!). Pratiques : théorie, pratique, pédagogie, (70), 9-45.

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.

BORE, C. & ELALOUF, M.-L. (2016). Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles. In C. Doquet, J. David & S. Fleury (2016), *Spécificités et contraintes des grands*

corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement. Corpus (16, p. 31-63).

CAPPEAU, P. & ROUBAUD, M.-N. (2005). *Enseigner les outils de la langue avec les productions d'élèves : cycles 2 et 3*. Bordas pédagogie.

CHANQUOY, L. & FAYOL, M. (1991). Étude de l'utilisation des signes de ponctuation et des connecteurs chez des enfants (8-10 ans) et des adultes. Pratiques : théorie, pratique, pédagogie, (70), 107-124.

CLANCHE, P. (1988). L'enfant écrivain : génétique et symbolique du texte libre. Le Centurion.

DAVID, J. & DOQUET, C. (2016). Les écrits d'élèves : un corpus de référence pour le français contemporain. In SHS Web of Conferences (Vol. 27, 11001). EDP Sciences. J. David & S. Vaudrey-Luigy (2014), Enseigner la ponctuation. Le français aujourd'hui, 187.

DOQUET, C., DAVID, J. & FLEURY, S. (2016) (coordonné par). Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement. Corpus (16).

DRILLON, J. (1991). Traité de la ponctuation française. Gallimard.

ELALOUF, M.-L. (2011). Constitution de corpus scolaires et universitaires : vers un changement d'échelle ? Pratiques. Linguistique, littérature, didactique, (149-150), 56-70.

ELALOUF, M.-L. (2005). Écrire entre 10 et 14 ans : un corpus, des analyses, des repères pour la formation. SCEREN-CRDP de l'Académie de Versailles.

FABRE-COLS, C. (dir.) (2000). Apprendre à lire des textes d'élèves. Bruxelles : De Boeck-Duculot.

FABRE, C. (1990). Les brouillons d'écoliers, ou, L'entrée dans l'écriture. Ceditel, Université de Grenoble-Stendhal.

FAYOL, M., CARRE, M. & SIMON-THIBULT, L. (2014). Enseigner la ponctuation : comment et avec quels effets ? Le français aujourd'hui, *(4)*, 31-40.

GARCIA-DEBANC, C. & BONNEMAISON, K. (2014). La gestion de la cohésion textuelle par des élèves de 11-12 ans : réussites et difficultés. In SHS Web of Conferences (Vol. 8, pp. 961-976). EDP Sciences.

GRANGER, S., VANDEVENTER, A. & HAMEL, M.-J. (2001). Analyse de corpus d'apprenants pour l'ELAO basé sur le TAL, linguistique de corpus. *TAL*, 42 (2) :609–621.

JAFFRE, J.-P. (1991). La ponctuation du français : études linguistiques contemporaines. Pratiques : théorie, pratique, pédagogie, *(70),* 61-83.

PAOLACCI, V. & ROSSI-GENSANE, N. (2016). La question de la progressivité des apprentissages en production écrite à l'école élémentaire française : le cas de la segmentation en phrases. In SHS Web of Conferences (Vol. 27, p. 07011). EDP Sciences.

PAOLACCI, V. & GENSANE, N.-R. (2014). Ponctuation et écrits d'élèves. Le français aujourd'hui, *(4),* 115-125.

PASSERAULT, J.-M. (1991). La ponctuation. Recherches en psychologie du langage. Pratiques, *70* (85-103).

POPIN, J. & THOMASSET, C.-A. (1998). La ponctuation. Nathan.

WOLFARTH, C., Ponton, C. & TOTEREAU, C. (2016). Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire. In C. Doquet, J. David S. Fleury, Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement, Corpus, 16 | 2017, 185-214.