Un corpus de référence pour l'écriture de l'école à l'université : la ressource É-Calm

Claire Doquet, Lydia-Mai Ho-Dac, Claude Ponton

Un corpus de référence pour l'écriture de l'école à l'université : la ressource É-Calm

Claire Doquet, Lab-E3D ; Lydia-Mai Hodac, CLLE ; Claud Ponton, Lidilem

Introduction

Les écrits d'élèves intéressent depuis longtemps les linguistes, et singulièrement aujourd'hui, la linguistique de corpus [Doquet *et al.*, 2017a]. Ils font partie des écrits non standards qui posent à l'analyse automatique de redoutables problèmes mais dont l'intérêt est évident pour travailler sur l'écriture manuscrite, la production spontanée de texte et l'utilisation de l'écrit chez des scripteurs non experts [Steuckardt et Collette, 2019]. Recueillis tout au long de la scolarité, ces ensembles de données tracent des trajets développementaux de l'acquisition de la langue écrite et de ses usages. La mise au jour de ces trajets est pour la linguistique un défi technologique et théorique. Technologique, parce que l'outillage de la linguistique de corpus ayant été pensé pour des textes proches des normes de l'écrit, les écrits scolaires mettent à l'épreuve les approches et les ressources classiques du traitement automatique des langues. Théorique, parce que la créativité langagière des élèves va au-delà de ce que prévoient les modèles élaborés sur des écrits standards, obligeant parfois à reconsidérer des catégories qui semblaient aller de soi. La situation d'apprentissage de l'écriture et les difficultés qu'elle révèle permettent en effet de mettre au jour les zones les plus résistantes de la langue qui apparaissent aussi, mais sous forme atténuée, chez les scripteurs disposant d'un haut degré de maîtrise de l'écrit.

À partir d'un corpus d'écrits d'élèves et d'étudiants rendu accessible sur une plateforme dédiée, le projet É–Calm (Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactique), financé par l'ANR, a permis de caractériser certaines compétences scripturales (orthographe et cohérence textuelle) et de mieux comprendre la manière dont les enseignants, par leurs interventions sur les copies, orientent l'écriture, afin d'étayer l'accompagnement de la réécriture de l'école à l'université. L'objectif principal de la recherche était de mettre à disposition en *open access* un grand nombre d'écrits produits dans des contextes d'apprentissage variés afin d'en analyser les caractéristiques linguistiques et discursives. L'ensemble ainsi constitué donne une visibilité aux textes de scripteurs à différents niveaux d'apprentissage et de maîtrise de l'écrit. La figure 1 donne un exemple de texte composant ce corpus.

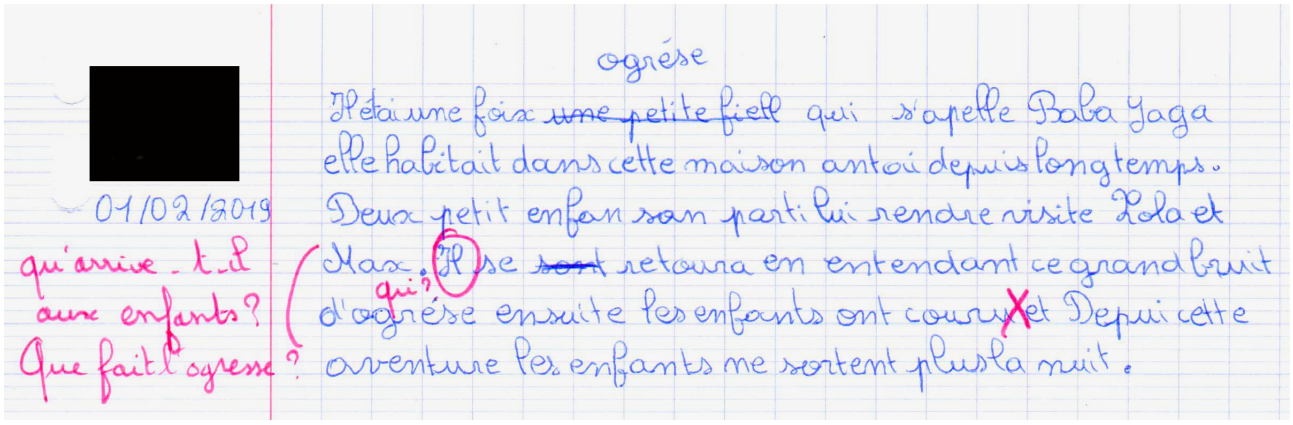

FIGURE 1 – Exemple de production d'un enfant de CM1

L'enjeu scientifique et sociétal d'une telle ressource est d'objectiver le regard sur les écrits scolaires. Les spécificités de ces écrits ont guidé les traitements et les analyses du corpus :

- Les écrits scolaires s'inscrivent dans une situation de communication particulière : celle d'un apprenant répondant à une commande de son enseignant, dans un contexte didactique donné (niveau de classe, consigne, environnement de la tâche...). L'organisation du corpus et l'accès aux écrits sont guidés par ces métadonnées contextuelles.

- Les écrits scolaires portent la trace de leur élaboration par l'élève auteur : le choix est fait de reprendre, pour la transcription de ces écrits, les catégories de la génétique textuelle. Elles permettent de mettre l'accent sur la dynamique de l'écriture et pas seulement sur les écrits terminés, elles donnent à lire, à travers les écrits des élèves, le trajet de l'écriture, outillant une didactique de l'écriture au sens plein du terme (comme l'hésitation entre parler d'une petite fille ou d'une ogresse dans la figure 1).

- Les écrits scolaires portent la trace des réactions de l'enseignant lecteur et correcteur, qui portent sur au moins deux dimensions de l'écrit : l'orthographe et la cohérence/cohésion textuelle (voir les interventions en rose dans la figure 1). É–Calm propose une typologie des interventions enseignantes, y compris non verbales (soulignements), ainsi qu'une catégorisation des corrections orthographiques et des marques de cohérence - ou d'incohérence - repérées sur les copies.

Le corpus É–Calm

La ressource É–Calm compte aujourd'hui près de 4500 textes (pour plus d'un million de mots) recueillis entre le début de l'école élémentaire et l'université et mis à disposition sur Ortolang (`https://www.ortolang.fr/market/corpora/e-calm`) et une plateforme dédiée sur Huma Num (`http://e-calm.huma-num.fr/`). Conçue à la fois comme une vitrine et un lieu de diffusion des résultats mais aussi des regards que des scientifiques peuvent porter sur les écrits des élèves, cette plateforme s'adresse aux chercheurs mais aussi au monde enseignant, notamment les formateurs.

La constitution de corpus scolaires numériques est relativement récent. Le premier de ce type est sans doute le corpus Lancaster [Smith *et al.*, 1998] qui propose des textes d'enfants manuscrits retranscrits et accessibles en ligne. Toujours en anglais, l'*Oxford Children's Corpus* [Banerji *et al.*, 2013] propose depuis 2006 plus de 70.000 textes courts écrits par des enfants âgés de 4 à 13 ans dans le cadre de concours publics de rédaction en ligne. En 2011, l'université de Karlsruhe diffuse un corpus de textes spontanés d'enfants allemands du grade 1 à 8 [Lavalley *et al.*, 2015]. Le premier corpus scolaire français remonte à 2005 et comporte 500 textes écrits d'enfants de CM2 à 5ème [Elalouf, 2005]. Depuis 2010, plusieurs projets de corpus scolaires sont en cours ou achevés [Garcia-Debanc et Bonnemaison, 2014, Doquet *et al.*, 2017b, Boré et Elalouf, 2017, Vogüé *et al.*, 2017, Wolfarth *et al.*, 2017] dont certains sont à l'origine du corpus É–Calm.

Les textes composants le corpus É–Calm sont issus de 4 projets pré-existants qui montrent certaines différences en fonction (a) de la consigne d'écriture qui, dans certains cas, a été proposée par les chercheurs du projet ; (b) du niveau d'étude considéré ; (c) de l'intervention ou non de l'enseignant sur la copie avec possibilité de réécriture.

Les corpus *EcriScol*[1] [Doquet *et al.*, 2017b] et Littératie Avancée[2] [Jacques et Rinck, 2017] sont composés de textes d'élèves et d'étudiants produits en réponse à une demande de leurs enseignants. Une grande partie de ces textes contient des interventions de ces enseignants et parfois même, pour Ecriscol, les brouillons et les versions intermédiaires. à noter que le corpus Littératie Avancée est le seul ne contenant que des textes informatisés. Les corpus *ResolCo*[3] [Garcia-Debanc *et al.*, 2017] et *Scoledit*[4] [Wolfarth *et al.*, 2017] ne comportent que des textes provoqués par la recherche.

Le corpus *ResolCo* se caractérise par une consigne élaborée pour confronter l'élève à une "tâche-problème" nécessitant la gestion de liens de cohésion e.g. relations coréférentielles, relations de discours, etc. pour construire un texte cohérent. Cette "tâche-problème" consiste à rédiger une histoire incluant 3 phrases prédéfinies incluant, entres autres, des anaphores, des temps verbaux spécifiques et un enchaînement d'événements impliquant certaines relations de discours.

Scoledit est un corpus longitudinal permettant l'étude de l'évolution des compétences en littéracie durant l'école primaire [Wolfarth *et al.*, 2018]. Ce corpus est composé de textes produits selon les mêmes consignes par les mêmes 373 élèves durant toute leur scolarité primaire. Des productions narratives et des dictées ont ainsi été recueillies du CP au CM2 entre 2014 et 2018.

1. `http://www.univ-paris3.fr/ecriscol`
2. `https://www.ortolang.fr/market/corpora/litteracieavancee`
3. `http://redac.univ-tlse2.fr/corpus/resolco.html`
4. `http://www.scoledit.org/scoledit`

La table 1 fournit un aperçu quantitatif du nombre de textes de la version actuelle du corpus É–Calm.

Niveau	#Textes	#Mots	#Mots/Textes	Corpus
Primaire	3133	330214	105	[E][R][S]
Secondaire	555	115866	209	[E][R]
Supérieur	607	671056	1106	[E][R][LA]
Total	**4295**	**1117136**	**260**	**[E][R][S][LA]**

TABLE 1 – état quantitaif de la version actuelle du corpus É–Calm, avec [LA] pour *Littéracie avancée*, [E] pour *Ecriscol*, [R] pour *Resolco* et [S] pour *Scoledit*

Traitements manuels et automatiques des données

Tous les textes ont été récoltés sous la forme de copies manuscrites (ou tapuscrites pour ceux récoltés à l'Université) avec autorisation de diffusion. En plus de l'étape de transcription requis pour les copies manuscrites, les données ont nécessité un long processus d'encodage pour permettre une homogénéisation et une structuration des données et des méta-données associées. La norme TEI-P5 a été appliquée afin d'assurer le partage entre chercheurs de différentes disciplines, la pérénité des données et la possibilité d'appliquer directement des traitements automatiques prêts à l'emploi pour les corpus encodés selon la TEI. Cette norme a également été très précieuse pour guider l'encodage des métadonnées et des aspects génétiques comme les traces de révision (rature, ajout...) [5].

Une fois l'encodage des données brutes au format TEI-P5, l'étape suivante a consisté à proposer une normalisation de l'orthographe et à aligner les transcriptions et les versions normalisées. Cette étape a été réalisée manuellement selon deux méthodes selon les ressources : par la création en parallèle d'une version normalisée (comme en traduction) pour *Scoledit* ou via une interface d'annotation e.g. GLOZZ [6] [Mathet et Widlöcher, 2009], considérant alors la normalisation orthographique comme une première couche d'annotation. La première méthode requiert une étape d'alignement supplémentaire qui a nécessite le développement d'un outil spécifique : AliScol [Wolfarth *et al.*, 2018].

La mise à disposition des copies transcrites alignées à leur version normalisée permet l'application de traitements automatiques pour enrichir les données par l'annotation des lemmes, des étiquettes morphosyntaxiques et des relations de dépendances syntaxiques. L'ensemble de ces annotations a permis un ensemble d'analyses caractérisant les erreurs observées dans la ressource [Belinda Lavieu-Gwozdz, 2021, Claude Ponton, 2021, Claire Wolfarth, 2018].

Un dernier pan de traitements manuels a consisté à annoter différents aspects de l'organisation discursive des textes d'élèves. Ainsi, un extrait du corpus *Resolco* a été segmenté en Unités Minimales de discours et annoté en relations de discours permettant d'éprouver les modèles du discours à ce type de données [Myriam Bras, 2021]. Enfin, près du quart de la ressource a été annotée en continuités référentielles, ce qui permet d'analyser en détail les stratégies mises en oeuvre par les élèves pour tisser des liens référentiels dans leurs narrations [Claudine Garcia-Debanc, 2021].

La communication permettra de présenter la ressource É–Calm et les résultats de ces nombreuses analyses.

Références bibliographiques

5. Voir le tableau 2 en annexe.
6. `http://glozz.free.fr/`

Bibliographie

[Banerji *et al.*, 2013] Banerji, N., Gupta, V., Kilgarriff, A. et Tugwell, D. (2013). Oxford children's corpus : A corpus of children's writing, reading and education. *Corpus Linguistics*, pages 315–317.

[Belinda Lavieu-Gwozdz, 2021] Belinda Lavieu-Gwozdz, Élise Vinel, V. G. C. B. (2021). Cartographie des usages et des erreurs orthographiques sur les verbes dans des récits écrits par des élèves de 6 à 15 ans. *Langue Française*, 211(3):51–65.

[Boré et Elalouf, 2017] Boré, C. et Elalouf, M.-L. (2017). Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles. *Corpus*, 16:31–64.

[Claire Wolfarth, 2018] Claire Wolfarth, C. P. e. C. B. (2018). Gestion de la morphographie verbale en production d'écrits : que peut nous apprendre un corpus longitudinal ? *Repères [En ligne]*, 57.

[Claude Ponton, 2021] Claude Ponton, Rafaela Gutiérrez-Cáceres, L. T. E. F. C. B. C. W. (2021). Scolinter : un corpus trilingue. l'exemple de la segmentation en mots. *Langue Française*, 211(3):37–50.

[Claudine Garcia-Debanc, 2021] Claudine Garcia-Debanc, Josette Rebeyrolle, L.-M. H.-D. (2021). La continuité référentielle dans le corpus rÉsolco : Méthode d'annotation et premières analyses. *Langue Française*, 211(3):99–114.

[Doquet *et al.*, 2017a] Doquet, C., David, J., Fleury, S. et (Eds) (2017a). Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement. *In Corpus [Online]*, volume 16 (Special Issue). OpenEdition.

[Doquet *et al.*, 2017b] Doquet, C., Enoiu, V., Fleury, S. et Maziotti, S. (2017b). Problèmes posés par la transcription et l'annotation d'écrits d'élèves. *Corpus*, 16:133–156.

[Elalouf, 2005] Elalouf, M.-L. (2005). *Ecrire entre 10 et 14 ans : Un corpus, des analyses, des repères pour la formation*. Canopé - CRDP de Versailles.

[Garcia-Debanc et Bonnemaison, 2014] Garcia-Debanc, C. et Bonnemaison, K. (2014). La gestion de la cohèsion textuelle par des élèves de 11-12 ans : Réussites et difficultés. *In Actes du 4e Congrès Mondial de Linguistique Française*, pages 961–976.

[Garcia-Debanc *et al.*, 2017] Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M. et Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*, 16.

[Jacques et Rinck, 2017] Jacques, M.-P. et Rinck, F. (2017). Un corpus de littéracie avancée : résultat et point de départ. *Corpus*, 16.

[Lavalley *et al.*, 2015] Lavalley, R., Berkling, K. et Stücker, S. (2015). Preparing children's writing database for automated processing. *In Proceedings of LTLT@SLaTE*, pages 9–15.

[Mathet et Widlöcher, 2009] Mathet, Y. et Widlöcher, A. (2009). La plate–forme GLOZZ : environnement d'annotation et d'exploration de corpus. *In Actes de TALN 2009*, Selnis. ATALA, LIPN.

[Myriam Bras, 2021] Myriam Bras, Laure Vieu, M. J.-A. P.-B. C. P. C. R. (2021). Vers un corpus de textes d'élèves annoté en relations de discours. *Langue Française*, 211(3):115–129.

[Smith *et al.*, 1998] Smith, N., McEnery, T. et Ivanic, R. (1998). Issues in transcribing a corpus of children's handwritten projets. *Literacy and Linguistic Computing*, 13:217–225.

[Steuckardt et Collette, 2019] Steuckardt, A. et Collette, K. (2019). *Écrits hors-normes*. Les Éditions de l'Université de Sherbrooke (ÉDUS).

[Vogüé *et al.*, 2017] Vogüé, D., S. Espinoza, N., Garcia, B. ad Perini, M. et Marzena Watorek, F. (2017). Constitution d'un grand corpus d'écrits émergents et novices : Principes et méthodes. *Corpus*, 16:65–86.

[Wolfarth *et al.*, 2018] Wolfarth, C., Ponton, C. et Brissaud, C. (2018). Gestion de la morphologie verbale en production d'écrits : que peut nous apprendre un corpus longitudinal ? *Repères*, 57.

[Wolfarth *et al.*, 2017] Wolfarth, C., Ponton, C. et Totereau, C. (2017). Apports du tal à la constitution et à l'exploitation d'un corpus scolaire. *Corpus*, 16.

Annexes

La table 2 décrit les éléments TEI-P5 spécifiques qui ont été appliqués [7].

Tags TEI-P5	**Description**
settingDesc	Région et caractéristiques sociales de l'institution comme, par exemple : zone d'éducation prioritaire, populations rurales *vs* urbaines, etc.
textDesc	Informations sur les consignes données aux étudiants ainsi que sur la préparation et la rédaction, à savoir si le texte est un projet, un travail préparé ou un travail révisé
particDesc	- caractéristiques des scripteurs comme l'âge, la langue maternelle, les troubles du langage (ie. dyslexie, apraxie), etc. - Caractéristiques de l'enseignant comme le nombre d'années d'expérience
mod	- trace d'une rature dans le document avec mention du texte raturé et/ou du texte ajouté
lb et pb	- indication des fins de ligne et de page imposées par le support papier
gap	- Caractères illisibles
unclear	- Portion de texte pour laquelle le transcripteur est peu sûr de sa transcription
metamark	- Commentaire en marge de l'enseignant ou remarque sur la transcription

Table 2 – Informations encodées selon la TEI-P5 dans le corpus É–Calm

7. https://tei-c.org/Guidelines/P5/