Introduction

Claire Doquet, Jacques David, Serge Fleury

CORPUS

Corpus

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

Introduction

Claire Doquet, Jacques David et Serge Fleury

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/2727
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

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

Référence électronique
Claire Doquet, Jacques David et Serge Fleury, « Introduction », *Corpus* [En ligne], 16 | 2017, mis en ligne le 18 novembre 2017, consulté le 09 janvier 2018. URL : http://journals.openedition.org/corpus/2727

Introduction

Claire Doquet, Jacques David et Serge Fleury

1 Les écrits des élèves suscitent un intérêt grandissant chez de nombreux chercheurs dont les travaux s'inscrivent dans des domaines aussi variés que la linguistique, la psycholinguistique, la sociolinguistique et la didactique du français. Cet intérêt s'explique à la fois par la singularité de l'objet discursif qu'ils constituent et par la rareté des études empiriques appuyées sur des ensembles des textes de grande envergure (voir cependant Garcia-Debanc, 1990). Si des corpus importants ont pu être publiés notamment dans de nombreuses thèses en sciences du langage, l'analyse des textes qui les composent reste partielle, et leur accès difficile. Le travail précurseur mené à Versailles par C. Boré et M.-L. Elalouf (Elalouf *et al.* 2005) trouve aujourd'hui des voies de prolongement dans d'autres lieux et avec d'autres outils. Ce travail, à visée d'abord didactique a constitué la première tentative de communication, par voie de CD-Rom, d'un corpus d'écrits d'élèves permettant d'observer un certain nombre de faits de langue et de discours de l'écrit entre 10 et 14 ans. Elle a été suivie d'un travail collectif, dans le cadre du Groupe de Recherche CNRS *Approches Pluridisciplinaires de la Production Verbale Écrite*[1], où des écrits produits selon deux consignes différentes ont été recueillis dans des contextes variés (Auriac-Slusarczyk *et al.* 2008).

2 Les premières publications témoignent de la co-construction de ce corpus lui-même et des règles de sa constitution (Boré, dir., 2007), tant il est évident que l'écriture scolaire ne peut être appréhendée indépendamment de son contexte de production. Comme l'ont montré par ailleurs les sciences de l'éducation (Sensevy & Mercier, 2007 ; Bautier & Rayou, 2013), la situation didactique imprime aux écrits scolaires des normes, des contraintes interlocutives et des fonctions spécifiques qui travaillent, au même titre que le thème ou le type d'écrit produit, le matériau langagier. Une des interrogations au moment de constituer un corpus d'écrits scolaires est donc le nombre et la nature des métadonnées à recueillir pour permettre l'analyse.

3 En corollaire se pose la question des traces de l'écriture : brouillons, notes, ébauches, l'ensemble de l'avant-texte est à considérer pour appréhender non seulement le texte final, mais aussi la manière dont il s'est écrit et l'ensemble des opérations qui ont contribué à sa constitution. Les éléments d'un texte ayant fait l'objet d'une réécriture

peuvent être considérés comme objets d'une réflexion méta-discursive spontanée de la part du scripteur : il apparaît en effet que toute rature – au sens d'opération portant sur le déjà écrit : suppression, ajout, remplacement, déplacement – est la trace d'un aller-retour entre le discours en train de se produire et la langue (Fabre, 1990) ; il correspond ainsi à un retour sur le déjà écrit dont le scripteur évalue l'adéquation, pour éventuellement le rectifier. Le relevé systématique des ratures permet donc d'appréhender les niveaux de réflexion méta-discursive qui accompagnent une écriture.

4 En référence à son propre travail, M.-C. Elalouf soulignait en 2011 le changement d'échelle que constitue une analyse de grand corpus pour l'exploration de l'écriture scolaire. Un pas de plus est franchi aujourd'hui avec la mise à disposition en *open access*, sur internet, de corpus d'écrits d'élèves et plus largement d'apprenants (*Learner Corpora*) ; l'université de Louvain a entrepris de les recenser[2] mais la liste qui en résulte ne fait apparaître que cinq bases de données en français écrit. De fait, malgré les avancées considérables des outils informatiques d'analyse de textes et les méthodologies liées aux grands corpus oraux, le traitement quantitatif des données langagières émanant de scripteurs débutants ou en cours d'apprentissage est difficile, en particulier du fait du caractère linguistiquement peu normé (ou autrement normé) de leurs productions (Lüdeling 2008 ; Hirschmann *et al*., 2013).

5 C'est à partir de ces constats qu'est né le projet de ce numéro de *Corpus*, qui souhaite faire le point sur les avancées en cours et donner la parole à des équipes francophones travaillant à la constitution, au traitement et à l'exploitation d'ensembles d'écrits d'élèves qui, s'ils ne sont pas encore systématiquement organisés en « grands corpus », constituent des tentatives de changer d'échelle pour décrire le développement des pratiques scripturales. Ce type de données permettra également, à partir du recensement systématique des éléments faisant l'objet d'erreurs et/ou de tâtonnements scripturaux, de faire émerger un schéma de l'évolution des compétences des élèves et de mettre en évidence les acquisitions tardives ; à partir de là, un retour sur le système de la langue devrait permettre de mettre au jour des zones linguistiquement « résistantes », c'est-à-dire qui ne se laissent pas aisément maîtriser.

6 Le numéro se compose de trois parties : la première est consacrée à l'exposition des ancrages historiques et théoriques des corpus d'apprenants et de leurs analyses ; la deuxième propose différentes méthodes de transcription, d'annotation et d'exploitation des données ; la troisième présente les résultats de travaux réalisés ou en cours.

7 Dans la première partie, Catherine Boré et Marie-Laure Elalouf reviennent sur leur expérience de constitution d'un corpus d'écrits d'élèves (publié en 2005) pour mettre en évidence les principes de constitution de corpus scolaires recueillis dans des conditions écologiques (Elalouf & Boré, 2007). Elles exposent les questions méthodologiques qui se sont posées lors de l'élaboration de ce corpus, concernant le choix et l'organisation des données, la définition d'une unité d'observation, les différentes possibilités de transcription et les perspectives d'analyses qu'elles offrent. Dans l'article suivant, Natacha Espinoza, Brigitte Garcia, Marie Perini, Frédérique Sitri, Sarah de Vogüé et Marzena Watorek proposent une réflexion sur la construction d'un vaste corpus d'écrits qui permet d'approfondir la compréhension des processus en jeu dans l'accès à la littératie dans sa diversité, au travers de la pluralité des genres et des types discursifs qui la constituent et chez des apprenants de profils divers : enfants / adultes, langue 1 / langue 2, entendants / sourds. Cette réflexion s'inscrit dans une perspective fonctionnaliste et énonciative, et s'appuie sur deux ensembles théoriques

complémentaires, car les productions d'apprenants ne constituent pas une déviance par rapport à la norme mais sont des manifestations de systèmes linguistiques dont il s'agit de dégager les normes propres ; l'hétérogénéité est en effet constitutive de toute production verbale. Enfin, dans une réflexion globale sur la nature et les enjeux de la transcription des manuscrits, Pierre-Yves Testenoire met en perspective deux champs disciplinaires qui explorent des manuscrits : la philologie et la génétique textuelle. Il examine les caractéristiques épistémologiques de ces deux approches pour réfléchir aux modalités de la transcription correspondantes.

8 Dans la deuxième partie, consacrée aux méthodologies de traitement des manuscrits d'élèves, Marie-Noëlle Roubaud propose de comparer les protocoles de recueil de bases de données orales avec ceux d'un corpus écrits de jeunes apprenants, au début du primaire. Elle cerne ainsi plusieurs problèmes de transcription propres à leurs écrits (choix des pronoms, emploi de connecteurs, formes parataxiques, failles macrosyntaxiques...), problèmes qui pourraient trouver des solutions dans l'annotation et le codage des corpus oraux d'élèves du même âge. En revanche, elle s'interroge sur les différences liées aux propriétés linguistiques de ces productions d'oral *vs* d'écrit, et engage une discussion sur les transferts méthodologiques possibles. Dans leur contribution, Claire Doquet, Vanda Enoiu, Serge Fleury et Sara Mazziotti exposent les solutions avancées pour transcrire, annoter et analyser quantitativement les écrits d'élèves de fin de primaire, à l'aide du logiciel *Le Trameur*, qui permet un traitement textométrique et linguistique de ces écrits scolaires, tel qu'il est envisagé dans le programme Écriscol. L'étude montre la nécessité de mettre au point un protocole de recueil et d'analyse ajusté pour ces écrits d'élèves, afin d'offrir un éventail étendu d'analyse qualitatives, notamment sur les (dys)fonctionnements linguistiques les plus caractéristiques : écarts à la norme orthographique et syntaxique, stéréotypies lexicales et discursives, modes d'énonciation singuliers... C'est dans une perspective analogue, et sur d'autres composantes linguistiques, que Claudine Garcia-Debanc, Lydia-Mai Ho-Dac, Myriam Bras & Josette Rebeyrolle mettent en évidence les problèmes épistémologiques et méthodologiques posés par l'annotation discursive de productions écrites d'élèves et les choix techniques qu'implique sa mise en œuvre. En s'appuyant sur la démarche ayant conduit à la création d'un corpus de textes en français écrit standard, annoté discursivement, les auteurs évaluent les adaptations nécessaires pour traiter des textes d'élèves. Elles présentent deux chantiers d'annotation portant sur des textes d'élèves de fin d'école primaire en réponse à une consigne mettant en jeu la cohésion textuelle, qui demande d'insérer dans un texte narratif trois phrases comportant des pronoms et des syntagmes nominaux anaphoriques. L'annotation des *Relations de Discours* procède successivement par segmentation du texte en *Unités de Discours Élémentaires*, analyse de la cohérence temporelle et annotation des *Relations de discours*. Pour clore provisoirement cette deuxième section, Claire Wolfarth, Claude Ponton et Corinne Totereau présentent une recherche qui a pour but la constitution d'un corpus scolaire et le développement d'un outil d'aide à son exploitation, à partir de l'annotation de phénomènes linguistiques saillants. L'objet de ce travail est d'explorer les possibilités qu'offre le traitement automatique des langues pour appréhender ces écrits particulièrement éloignés de la norme. L'article propose un exposé détaillé du module d'annotation de certaines erreurs orthographiques. Il en expose la méthode d'identification et de correction au moyen d'une ressource lexicale de formes phonologiques ainsi que le modèle d'annotation élaboré.

9 Dans la troisième partie, plus orientée vers des comptes rendus d'analyse et des considérations didactiques, Marie-Paule Jacques et Fanny Rinck exposent les travaux accomplis et en cours sur un corpus de « littéracie avancée » (écrits universitaires et professionnels d'étudiants de niveau licence 1 à master 2). Les auteures décrivent ainsi les compétences rédactionnelles de ces étudiants et les besoins en formation révélés dans les problèmes qu'ils rencontrent. Associée aux métadonnées recueillies, l'analyse de ce corpus écrit offre des possibilités de réponses coordonnées aux plans linguistique et didactique, d'une part, en décrivant les énoncés défaillants, d'autre part, en manipulant des procédés syntaxiques élargis à la textualité. Le « Forum d'élaboration des connaissances », que présentent Marie-Eve Desrochers et Godelieve Debeurme, constitue un espace de collaboration où élèves et enseignants sont amenés à écrire des textes, de longueur et de type divers, à les annoter, à les améliorer et à les enrichir à l'intérieur d'une base de connaissances qui conserve les traces de leurs échanges sous forme de toile interactive. L'archivage de différentes données permet l'analyse statistique et le profilage d'élèves conduisant à mieux évaluer leurs besoins en tant qu'apprenants. L'analyse des interactions (*posting behaviours*) et les suivis de fréquentation sur la plateforme (*non-posting behaviours*) peuvent amener les enseignants et les élèves à réfléchir sur leurs pratiques et diriger leurs interventions pédagogiques. Enfin, Solveig Lepoire-Duc et Jean-Pierre Sautot présentent une base de données fonctionnant sur les principes de l'informatique décisionnelle pour décrire la hiérarchie et les diverses dimensions de ces données. Cette structuration permet d'observer la diversité des relations temporelles dans des textes enfantins. L'analyse s'appuie sur un corpus de 200 récits d'élèves de cycle 3 de l'école élémentaire française et questionne le découpage séquentiel du texte, l'analyse interne des séquences et la normalité des productions des élèves.

10 Ce numéro se veut une synthèse des recherches sur des grands corpus d'écrits d'apprenants en français langue maternelle. Plusieurs des recherches présentées ici convergent d'ailleurs pour élaborer des principes de recueil, de transcription et d'annotation de ces corpus.

BIBLIOGRAPHIE

Auriac-Slusarczyk E., Chanquoy L., Cogis D., Garcia-Debanc C., Gunnarsson C., Largy P., Leblay C., Slusarczyk B. (2008). « Étude pluridisciplinaire du processus de révision chez de jeunes rédacteurs du primaire au collège ». In D. Alamargot, J. Bouchand, E. Lambert, V. Millogo, & C. Beaudet (éd.), *Proceedings of the International Conference « De la France au Québec : l'Écriture dans tous ses états »*, Poitiers, France, 12-15 November 2008 [http://www.ecritfrancequebec2008.org/].

Bautier E. & Rayou P. (2013). « La littératie scolaire : exigences et malentendus. Les registres de travail des élèves ». *Éducation & Didactique*, vol. 7, n° 2, pp. 29-46.

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

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

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.

Fabre C. (1990). *Les brouillons d'écoliers ou l'entrée dans l'écriture.* Grenoble, Ceditel.

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

Hirschmann H., Lüdeling A., Rehbein I., Reznicek M. & Zeldes A. (2013). « Underuse of Syntactic Categories in Falko. A Case Study on Modification ». In : S. Granger, G. Gilquin & F. Meunier (Hrsg.) *20 years of learner corpus research. Looking back, Moving ahead. Corpora and Language in Use* - Proceedings 1. Louvain la Neuve : Presses universitaires de Louvain.

Lüdeling A. (2008). « Mehrdeutigkeiten und Kategorisierung : Probleme bei der Annotation von Lernerkorpora ». In : M. Walter & P. Grommes (Hrsg.) *Fortgeschrittene Lernervarietäten*. Niemeyer, Tübingen, pp. 119-140.

Sensevy G. & Mercier A. (dir.) (2007). *Agir ensemble : l'action didactique conjointe du professeur et des élèves*. Rennes : PUR, coll. « Paideia ».

NOTES

1. http://www.gdr-pve.fr/GDR2657_bilan_projet.pdf.
2. http://www.uclouvain.be/en-cecl-lcworld.html.

AUTEURS

CLAIRE DOQUET

Université de la Sorbonne Nouvelle – EA 7345 Clesthia

JACQUES DAVID

Université de la Sorbonne Nouvelle – EA 7345 Clesthia

SERGE FLEURY

Université de la Sorbonne Nouvelle – EA 7345 Clesthia