Analyses linguistiques de productions écrites d'élèves de 9 à 12 ans : problèmes méthodologiques à résoudre pour la constitution et le traitement de grands corpus scolaires

Claudine GARCIA-DEBANC et Véronique PAOLACCI, avec les contributions de Myriam BRAS et Mai HODAC, CLLE, UMR 5263, CNRS et Université de Toulouse2-Le Mirail

L'étude de l'apprentissage de l'écriture requiert la collecte et l'analyse de grands corpus de productions écrites, afin de garantir objectivité et scientificité pour ce type d'études. Or, dans l'approche des textes d'enfants, les traditions des communautés scientifiques sont différentes : les psychologues procèdent à un traitement quantitatif de certaines marques dans un ensemble important de textes d'apprenants, les linguistes et les didacticiens procèdent le plus souvent à des études de cas sur un nombre limité de productions écrites. En matière d'acquisition et d'enseignement, on dispose de corpus internationaux, principalement en langue anglaise, d'une part sur les premières acquisitions du langage (CHILDES), d'autre part sur les erreurs d'apprenants d'anglais langue seconde (Granger, 2009) ou de Français Langue Etrangère (FRIDA, French Interlanguage Database ; Granger, 2007). Malgré quelques tentatives pour constituer de grands corpus scolaires d'école primaire (Elalouf et alii, 2005, 2012 ; Leblay, Auriac, 2011), les chercheurs, jusqu'ici, n'ont pas à leur disposition de corpus de productions écrites, édités et balisés, qui permettent de mesurer des évolutions dans l'acquisition des processus rédactionnels et la maitrise des marques linguistiques de l'écrit.

L'enjeu de l'atelier est de confronter plusieurs approches d'un corpus de textes recueillis en situation scolaire et de définir quelques indicateurs de réussite et de progressivité dans la maitrise de compétences syntaxiques et de cohésion textuelle par des élèves de 9 à 12 ans.

Le corpus de référence, d'où sont extraits les textes qui seront analysés au cours de l'atelier, comporte 400 textes d'élèves de 9 à 12 ans, recueillis dans des classes françaises, en réponse à une consigne demandant (a) d'intégrer dans un récit trois phrases comportant des pronoms personnels anaphoriques (*il*, *elle*) et deux déterminants démonstratifs anaphoriques (*cette maison, ce grand bruit*) et des marques temporelles (*en entendant ce grand bruit*) et (b) de planifier un récit à partir d'une phrase de clôture (*depuis cette aventure, les enfants ne sortent plus la nuit)* (Garcia-Debanc, à paraitre ; Roubaud, Garcia-Debanc, à paraître). Cette tâche-problème, inspirée par des travaux portant sur la résolution de problèmes de cohésion textuelle par des enfants (Charolles, 1988) permet d'évaluer les compétences de planification, de mise en texte et de révision (Hayes et Flower, 1980) d'élèves des trois dernières années d'école primaire et de début de collège. Le travail de l'atelier portera sur un échantillon très limité de 7 textes d'élèves, du CE2 (3° année primaire) à la sixième (1° année collège).

Après avoir rappelé les enjeux scientifiques de l'étude de grands corpus oraux et présenté la méthodologie de collecte des données, l'atelier permettra de poser les problèmes méthodologiques relatifs au traitement de ces textes, qui sont non normés, et à en proposer des modes de résolution.

Comment procéder à l'édition de ces textes à la fois pour ne pas perdre d'information et permettre un traitement automatisé ? Quel système de balises et d'annotations mettre en place ? La transcription que nous proposons vise, à long terme, la constitution d'un corpus normé d'écrits d'apprenants. Les outils de traitement automatique de corpus ont pour point commun d'avoir été principalement développés à partir de textes écrits normés et en vue du traitement de l'écrit normé. Les transcriptions proposées sont pensées dans le but de préserver

les particularités des données primaires (les copies d'élèves comportent un grand nombre de ratures, jeux typographiques et autres marques du processus d'écriture) tout en permettant l'extraction d'une version "normée" sur laquelle des outils de traitements automatique de corpus peuvent être appliqués (analyse syntaxique, projection de lexiques, segmentation discursive...). Nous proposons ici une adaptation de la norme CHAT (Codes for the Human Analysis of Transcripts) qui offre un ensemble de normes de transcription au départ dédiées à l'étude de situations naturelles de dialogue et aujourd'hui adaptées à un grand nombre de langues naturelles, orales ou gestuelles. Cette adaptation se base sur l'idée que les particularités des copies d'élèves se rapprochent en partie de celles des données orales, les ratures, fautes d'orthographes et absence de certains mots pouvant être définies comme des disfluences. La transcription selon la norme de CHAT des écrits d'apprenants permet à la fois d'accéder à des versions normées permettant l'application d'un certain nombre d'outils du TAL et de modèles linguistiques, tout en permettant une approche quantitative des "disfluences écrites ».

Après avoir analysé la tâche et explicité les conventions de transcription retenues, en conformité avec les conventions de CHILDES et les travaux antérieurs d'autres équipes dans des projets concernant des oraux et écrits d'adolescents scolarisés, en plusieurs langues (Jisa et alii 2009), les exposés et les mises en situation des stagiaires proposeront des analyses portant sur les procédés linguistiques utilisés pour la résolution des problèmes de cohésion textuelle (Bèguelin, 1988, 1994 ; Garcia-Debanc, 2010, 2013 ; Roubaud et Garcia-Debanc, sous presse) et la gestion des personnages, les marques de ponctuation et le découpage des phrases (Paolacci, Rossi, 2012), les connecteurs (Bèguelin, 1992) et les indices de structuration temporelle, notamment les adverbiaux cadratifs (Charolles, 2002, 2005, 2006). Ces différentes perspectives sont choisies pour rendre compte de l'articulation entre la résolution des problèmes de formulation locaux, à l'intérieur de la phrase, tels que l'insertion de relatives ou de gérondifs, et les problèmes de cohésion textuelle, à l'échelle du texte. Elles permettront également de questionner les différents modes de segmentation d'un texte, selon que l'approche s'intéresse à la syntaxe, à la cohésion textuelle ou à la sémantique du temps.

Les analyses croisées de ce corpus de productions écrites permettent de dessiner une cartographie de la maîtrise progressive d'un certain nombre de marques linguistiques par des sujets en situation d'apprentissage de la langue écrite.

Plan de l'atelier

- Enjeux de la collecte et du traitement linguistique de grands corpus de textes d'élèves
- Présentation de la méthodologie et des données collectées
- Analyse des modes de résolution des anaphores dans les textes d'élèves : modalités de traitement et résultats
- La segmentation en phrases syntaxiques et l'emploi des marques de ponctuation et connecteurs : étude de cas (Texte : L'horloge maudite) et résultats quantifiés.
- Les conventions de transcription pour préparer une analyse automatisée : étude de cas sur quelques textes du corpus
- Etude sémantique de la structuration temporelle. La segmentation pour une étude sémantique. Etude de quelques marques de structuration temporelle : marqueurs linguistiques de temps (temps verbaux, adverbiaux temporels, connecteurs), modalités d'organisation temporelle du discours (encadrement, enchainement, connexion).
- Synthèse : de l'analyse linguistique au repérage de compétences en production écrite. Quelle convergence entre les indicateurs repérés dans les diverses approches ?

Les textes d'élèves qui feront l'objet des analyses seront mis à disposition sur le site de l'Ecole d'été.

Eléments bibliographiques

Auriac-Slusarczyk, E., Leblay, C. (2010). Acquisition et enseignement en production écrite, *Synergies Pays Scandinaves* n° 5, 17-30

[Reichler-]Béguelin, M.-J. (1988). Anaphore, cataphore et mémoire discursive, *Pratiques* 57, pp. 15-43.

[Reichler-]Béguelin, M.-J. (1992). «L'approche des 'anomalies' argumentatives». *Pratiques*, n° 73, pp. 51-78.

[Reichler-]Béguelin, M.-J. (1994). «L'encodage du texte écrit. Normes et déviances dans les processus référentiels et dans le marquage de la cohésion». In L. Verhoeven & A. Teberosky (éds), *Proceedings of the Workshop "Understanding early literacy in a developmental and cross-linguistic approach"*, Network on Written Language and Literacy, Strasbourg, European Science Foundation, 1994, 175-204.

Charolles M., (1988). La gestion des risques de confusion entre personnages dans une tâche rédactionnelle, *Pratiques, 60*, 75-97.

Charolles M., (2002). *La référence et les expressions référentielles en français*, Paris : Ophrys.

Charolles, M., Pery-Woodley M-P. (2005a). *Langue Française* 148, Paris, Larousse.

Charolles, M., Le Draoulec, A., Pery-Woodley, M.-P., Sarda L. (2005b). « Temporal and spatial dimensions of discourse organisation », *French Language Studies* 15, Cambridge University Press, 115-130.

Charolles M., (2006), "*Un jour* (*one day*) in narratives", in I.Korzen & L.Lundquist (eds.), *Comparing Anaphors. Between Sentences, Texts and Languages*, Copenhagen, *Copenhagen Studies in Language*, 34, Copenhagen, Samfundslitteratur Press: 11-26.

Elalouf, M.-L. (dir.). (2005) *Ecrire entre 10 et 14 ans. Un corpus, des analyses, des repères pour la formation*, Paris, Scérén, CRDP Versailles, CDDP Essonne.

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

Granger, S. (2007). Corpus d'apprenants, annotation d'erreurs et ALAO : une synergie prometteuse. Cahiers de Lexicologie 91, 2007-2, 115-130.

Granger, S. (2009), « Learner corpora». In LÜDELING A. & KYTÖ M. (2009). *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, Berlin-New York, 259-275.

Hayes, JR, Flower, L.S (1980). « Identifying the organization of writing processes » dans Gregg L.W, Steinberg E.R. (eds) *Cognitive processes in writing*, Hillsdale, N.J, LEA, p. 3-30.

Paolacci, V. & Rossi-Gensane, N. (2012). Quelles images de la phrase dans les écrits d'élèves de fin d'école primaire française ? Description linguistique et réponses didactiques aux difficultés des élèves. *3e Congrès Mondial de Linguistique Française*. Lyon, 5-6 juillet 2012, 341-359.

Roubaud, M.-N. (2005). Reconsidérer l'erreur. *Les cahiers pédagogiques*, n° 438, pp. 31-32 [www.cahiers-pedagogiques.com].