Analyse linguistique de grands corpus d'écrits scolaires : problèmes de transcription, d'annotation et de traitement

Journée d'études organisée par le groupe *Écriture Scolaire*
du laboratoire Clesthia (EA 7345)

Mercredi 18 mars 2015
9h30-17h00
Université Paris-Sorbonne - Salle Bourjac
17 rue de la Sorbonne - 75005 Paris

Contact : Yilun Li (fantiyu001@hotmail.com)

Présentation de la journée

Les écrits des élèves suscitent un intérêt grandissant chez de nombreux chercheurs appartenant à des domaines ou des paradigmes de recherche 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 corpus de grande envergure. 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 du fait du caractère linguistiquement peu normé (ou autrement normé) de leurs productions. Il est urgent de remédier à cette lacune : le travail sur grand corpus remet en question de l'étude de cas dits "exemplaires" au profit d'une vision panoramique révélant, grâce à l'informatique statisticienne, de grandes tendances scripturales invisibles à l'œil nu. Les retombées dans le domaine de l'éducation de masse sont importantes, à commencer par la possibilité d'aider les programmes scolaires de la nation à coller à la réalité de besoins quantifiés à très grande échelle.

Dans le cadre de son opération de recherche *Analyse linguistique de l'écriture scolaire* (http://www.univ-paris3.fr/ecriscol-300509.kjsp) le laboratoire Clesthia de la Sorbonne Nouvelle (EA 7345) propose une journée de travail sur la question de la mise à disposition et du traitement informatique des écrits scolaires. Cette journée se déroulera en deux temps :

- Matinée : interventions axées sur les spécificités des corpus d'écrits d'élèves et leur diffusion. Structuration des corpus, visée des recherches, études longitudinales.
- Après-midi : interventions axées sur le traitement informatique des données. Annotations, traitements lexicaux et morphosyntaxiques.

Programme

Matinée
Modérateur : Jacques David (Univ. Cergy-Pontoise - CRTF)

9h45 - Ouverture de la journée par Franck Neveu (Univ. Paris Sorbonne - STIH, et ILF)

10h00 - Marie-Laure Elalouf, (Univ. Cergy-Pontoise - ÉMA)
Constitution d'un grand corpus de textes d'élèves, retour sur les questions méthodologiques posées par un corpus publié en 2005.

10h25 - Marie-Noëlle Roubaud (Univ. Aix-Marseille - ADEF)
Principes méthodologiques pour l'établissement d'un corpus de textes scolaires

10h50 - Questions et pause.

11h20 - Thierry Chanier (Univ. Blaise Pascal - Clermont-Ferrand 2 - LRL).
Concevoir la diffusion d'une banque de corpus dès le début du projet de recherche.

11h45 - Fanny Rinck & Marie-Paule Jacques (Univ. Grenoble 3 - LIDILEM)
Corpus de littéracie avancée : structuration et métadonnées.

12h10 - Questions

12h30-13h50 : pause repas

Après-midi
Modératrice : Claire Doquet (Univ. Paris 3 Sorbonne Nouvelle – Clesthia)

14h00 - Claire Wolfarth , Claude Ponton & Corinne Totereau (Univ. Grenoble 3 – LIDILEM)
Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire longitudinal

14h25 - Claudine Garcia-Debanc, Karine Perez-Bonnemaison, Josette Rebeyrolle, Myriam Bras, Mai Hodac, Sophie Mayras-Cauchois (CLLE, UMR 5263, CNRS & UT2 Jean Jaurès)
Problèmes méthodologiques posés par l'annotation discursive de textes d'élèves.

14h50 - Trang Luong (Univ. Paris Ouest Nanterre La Défense - MoDyCo)
Problèmes posés par la transcription/annotation des copies d'étudiants.

15h15 - Questions et pause.

15h35 - Céline Poudat (Univ. Nice Sophia Antipolis - BCL,)
Eléments de méthode pour explorer des contrastes et des hypothèses en corpus.

16h00 - Serge Fleury (Univ. Paris 3 Sorbonne Nouvelle - Clesthia)
Exploration textométrique de la base ECRISCOL avec le trameur.

16h25 - Bilan et perspectives.

Comité d'organisation :

Jacques David (Université de Cergy, EA 1392 CRTF)
Claire Doquet (Université Paris 3, EA 7345 Clesthia)
Serge Fleury (Université Paris 3, EA 7345 Clesthia)
Li Yilun (Université de Paris 3, EA 7345 Clesthia)

Comité scientifique :

Sonia Branca (Université Paris 3, EA 7345 Clesthia)
Catherine Boré (Université de Cergy, EA 4507 EMA)
Catherine Brissaud (Université Stendhal Grenoble 3, EA 609 Lidilem)
Marie-Laure Elalouf (Université de Cergy, EA 4507 EMA)
Claudine Garcia-Debanc (Université de Toulouse 2, UMR 5263 CLLE)
Olivier Lumbroso (Université Paris 3, EA 2288 DILTEC)
Franck Neveu (Université Paris Sorbonne, EA 4509 STIH)
Sylvie Plane (université Paris 4 Sorbonne, EA 4509 STIH)
Céline Poudat (Université de Nice, BCL UMR 7320)
Marie-Noëlle Roubaud (Université de Aix-Marseille, EA 4671 ADEF)
Agnès Steuckardt (Université Montpellier 3, UMR 5267 Praxiling)