Annotation syntaxique de corpus oraux
Projets récents et perspectives

Appel à communication
Journée d'étude Conscila (ENS Paris)

Vendredi 7 décembre 2012

À l'heure actuelle, de plus en plus de corpus de français parlé sont librement mis à la disposition de la communauté scientifique (corpus PFC, Corpus du Français Parlé Parisien, Valibel, CRDO, TCOF, etc.). Or, ces données présentent des particularités non prises en compte par la plupart des outils de traitements de corpus. Ainsi, il est difficile d'employer ces instruments directement sur le français parlé. De même, les données issues de l'oral posent des problèmes pour leur intégration dans les cadres traditionnels. Les logiciels et les approches linguistiques ont pour point commun d'avoir été principalement développés à partir de textes écrits (ou à partir d'exemples inventés) et en vue du traitement de l'écrit. Ainsi, afin d'adapter les systèmes actuels ou, tout simplement, d'approfondir notre connaissance du français, il est indispensable de produire des annotations sur les ressources orales.

Cependant, les initiatives dans ce domaine en sont encore au stade embryonnaire pour le français. On peut citer les travaux (sur l'annotation morphosyntaxique) de Eshkol et al. (2010), le projet PERCEO (http://cnrtl.fr/corpus/perceo/), la récente journée ATALA *Annoter les corpus oraux* (Paris, avril 2011) ou encore l'école thématique CNRS sur l'annotation de données langagières (sept. 2011). Pour la syntaxe plus spécifiquement, on peut, entre autres, signaler le projet FNRS de L. Degand et A.-C. Simon (2011-2013) portant sur la *Périphérie gauche des unités de discours* ainsi que le projet ANR Rhapsodie (2008-2012) sous la direction d'A. Lacheret. Un nouveau projet ANR ORFEO (Outils et Recherches sur le Français Ecrit et Oral) de constitution et d'annotation de corpus va également démarrer début 2013 sous la direction de J.-M. Debaisieux. Malgré ces travaux, à l'heure actuelle, aucun corpus de français parlé annoté en syntaxe n'est disponible, à notre connaissance.

L'un des objectifs de cette journée thématique sera de faire le point sur les initiatives récentes, en cours et futures dans le domaine de l'annotation syntaxique de corpus de français parlé, en montrant notamment comment l'annotation systématique fait émerger des questions fondamentales pour la description du français en général. Il s'agira également de voir dans quelle mesure on peut/doit développer de nouveaux modèles et outils pour intégrer les phénomènes présents à l'oral. Les communications pourront aussi bien porter sur des protocoles d'annotation, des outils que des études ciblées, des problèmes rencontrés, etc., et soulèveront une série de questions : quel standard d'annotation pour l'oral ? De quels outils dispose-t-on pour exploiter les annotations ? Par ailleurs, les démonstrations de logiciels pour l'annotation/exploitation seront aussi les bienvenues.

La journée se terminera par une table ronde, à laquelle tous les participants seront invités, et qui devrait permettre à la fois de faire une synthèse des présentations mais aussi de lister quelques-unes des bonnes pratiques et de lancer des pistes à explorer dans le cadre de projets futurs.

Organisation

Christophe Benzitoun – ATILF CNRS & Université de Lorraine
Noalig Tanguy – Lattice UMR 8094 ENS/Paris 3 & Valibel / Université Catholique de Louvain

Comité scientifique

Frédéric Béchet (Aix-Marseille Université / LIF UMR 7279)
Marie-José Béguelin (Université de Neuchâtel)
Alain Berrendonner (Université de Fribourg)
Mireille Bilger (Université de Perpignan)
Sandrine Caddéo (Aix-Marseille Université / Laboratoire Parole et Langage UMR 7309)
Paul Cappeau (Université de Poitiers)
Christophe Cerisara (Loria UMR 7503)
Jeanne-Marie Debaisieux (Université Paris 3 Sorbonne Nouvelle / Lattice UMR 8094)
Liesbeth Degand (Université catholique de Louvain / Valibel)
José Deulofeu (Aix-Marseille Université / LIF UMR 7279)
Anne Dister (Facultés universitaires Saint-Louis, Bruxelles)
Iris Eshkol (Université d'Orléans / Laboratoire Ligérien Linguistique UMR 7270)
Françoise Gadet (Université Paris Ouest Nanterre La Défense / Modyco UMR 7114)
Kim Gerdes (Université Paris 3 Sorbonne Nouvelle / LPP / Institut d'Automation / Académie de Sciences Chinoise)
Eva Havu (Université de Helsinki)
Sylvain Kahane (Université Paris Ouest Nanterre La Défense / Modyco UMR 7114)
Anne Lacheret (Université Paris Ouest Nanterre La Défense / Modyco UMR 7114)
Florence Lefeuvre (Université Paris 3 Sorbonne Nouvelle / Clesthia)
Michel Pierrard (Université Libre de Bruxelles)
Paola Pietrandrea (Université Roma Tre / Lattice UMR 8094)
Thierry Poibeau (Lattice UMR 8094 ENS/Paris 3)
Sophie Prévost (Lattice UMR 8094 ENS/Paris 3)
Nathalie Rossi-Gensane (Université Toulouse 2 / CLLE ERSS UMR 5263)
Frédéric Sabio (Aix-Marseille Université / Laboratoire Parole et Langage UMR 7309)
Catherine Schnedecker (Université de Strasbourg / Lilpa)
Anne-Catherine Simon (Université catholique de Louvain / Valibel)
Sandra Teston-Bonnard (Université de Lyon 2 / ICAR UMR 5191)
Véronique Traverso (ICAR UMR 5191)
Dan Van Raemdonck (Université Libre de Bruxelles)
Dominique Willems (Université de Gand)

Les propositions de communication (de deux pages maximum, bibliographie comprise) sont à adresser **avant le 20 octobre** aux adresses suivantes : Christophe.Benzitoun@univ-lorraine.fr/ noalig.tanguy@uclouvain.be