COLDOC 2012

COLLOQUE DES DOCTORANTS ET JEUNES CHERCHEURS
DU LABORATOIRE MoDyCo

Traitement de corpus linguistiques : outils et méthodes

MoDyCo CNRS (UMR 7114)
Université Paris Ouest Nanterre la Défense
Université Paris Descartes

4 - 5 octobre 2012

Amphithéâtre Durkheim, Université Paris Descartes, 7 rue de la Sorbonne

Appel à communications

Le COLDOC est le colloque annuel organisé par les doctorants et jeunes chercheurs du laboratoire MoDyCo (UMR 7114 – CNRS/Université Paris Ouest Nanterre/Université Paris Descartes). Cette année notre objectif est de nous interroger sur les outils et méthodes qui émergent du travail de corpus. Dans les dernières décennies, la linguistique a vécu une évolution certaine de l'**objet d'étude** de ses recherches : c'est dorénavant moins souvent la **langue** (objet a priori illimité et introspectif) et plus souvent le **corpus** (en tant qu'échantillon attesté de réalisations de cette langue) qui devient l'objet immédiat et central à étudier. Aujourd'hui, cette position centrale du corpus dans la recherche concerne sans doute la majeure partie de la communauté des chercheurs et doctorants en sciences du langage.

Cet **essor des problématiques liées aux corpus** alimente un débat latent. De façon informelle, le changement est souvent présenté sous deux aspects opposés : soit sous un angle exagérément négatif (comme une « mode », trop réductrice sur le fond, qui convient mal à la nature de la langue et finira par inhiber le débat théorique), soit sous un angle exagérément positif (comme une révolution qui rendra les sciences du langage plus «scientifiques» parce que plus empiriques, en étant plus proches du « réel »).

Refusant de nous arrêter à ces clivages, nous proposons aux doctorants du laboratoire MoDyCo et à tous les doctorants et jeunes chercheurs qui le souhaitent, de prendre le temps d'**examiner l'éventail des outils et des méthodes** liés à cette « vague » d'études de corpus, pour mieux mettre en lumière les points d'articulation entre l'observation et l'analyse. Nous nous inscrivons pleinement dans l'esprit la complémentarité de l'empirique et du théorique, exprimée jadis par Francis Bacon :

> *L'empirique, semblable à la fourmi, se contente d'amasser et de consommer ensuite ses provisions. Le dogmatique, telle l'araignée, ourdit des toiles dont la matière est extraite de sa propre substance. L'abeille garde le milieu ; elle tire la matière première des fleurs des champs, puis, par un art qui lui est propre, elle la travaille et la digère.*
>
> Novum Organum (1620), Livre I, 95

Le coeur de notre sujet est donc cet **« art de l'abeille »**, ce travail face au corpus linguistique qui, du moment de la collecte des énoncés/textes jusqu'à l'interprétation théorique finale et à ses applications, apparaît bien comme une étape de « digestion » des données du corpus, autrement dit un travail de « traitement » du corpus.

Concrètement, cette évolution semble liée à un développement des outils informatiques (navigation, collecte,outils d'aide à la transcription, outils d'analyse) qui ont **transformé l'accès aux sources et affecté la démarche d'étude linguistique**. Nous faisons l'hypothèse

que les mouvements et technologies influent non seulement sur la linguistique, mais aussi sur les autres sciences humaines et sociales de façon trans-disciplinaire. Les autres communautés de recherche en sciences humaines et sociales ont apparemment, elles aussi, vu leur facette «expérimentale » s'intensifier durant la période récente. Par ailleurs, l'essor d'internet et des ordinateurs a entraîné une différenciation au niveau des traitements des corpus. Une partie des linguistes a continué à se pencher sur les corpus, en utilisant les possibilités d'un traitement plus fin des données. D'autres linguistes ont préféré travailler sur l'amélioration des outils de traitement en TAL. La question de la **mutualisation de ce type de travaux** se pose ici comme là. Cela s'accompagne d'une grande diversité d'approches selon les thèmes et les écoles, et d'une tendance des instances à accompagner ce mouvement (projets de constitutions de « grands » corpus, groupes de travail d'annotation).

Selon la tradition COLDOC qui a fait siennes les questions méthodologiques larges, et qui s'attache à des problématiques ouvertes à l'ensemble des sciences du langage, nous appelons à **toute présentation** qui intéresse le thème du traitement de corpus linguistique, de sa conception à son résultat. Les problématiques associées comprennent, sans limitation, les points suivants :

- point de vue sur les textes ou énoncés selon le champ des sciences du langage
- niveau d'analyse pertinent et nature du corpus :
    - corpus oraux en phonologie, syntaxe, prosodie, études des troubles de la parole, etc.
    - corpus textuels en lexicographie, analyse de discours, syntaxe, «info-com»
    - et les corpus multimodaux en acquisition, etc.
- constitution de corpus, corpus clos vs. ouvert, représentativité, taille du corpus
- transcription, alignement, structuration et organisation du corpus
- définition des phénomènes ou procédés linguistiques étudiés
- annotations ou autres traitements effectués, décomptes ou mesures et leur pertinence
- choix des entrées dans l'analyse : occurrences, constructions, catégories, environnements, etc.
- choix de formalisation des résultats : tableaux statistiques, graphiques, typologies, schémas, etc.
- interprétation des résultats (selon hypothèses ou question posée)
- extractions opérées, modèles formels, apprentissage automatique
- mutualisation des corpus, des traitements et/ou des résultats
    - exploitation des bases existantes (grands corpus consultables)
    - au-delà de la publication, vers un partage des données et des résultats sur le travail de corpus

Nous invitons donc les doctorants et jeunes chercheurs à venir exposer leur réflexion sur un de ces aspects à partir de leur propre pratique et ce **quel que soit le degré d'avancement de leur recherche.**

Conférenciers invités

Bernard COMBETTES (ATILF, Université de Lorraine)

Anne CONDAMINES (CLLE-ERSS/CNRS, Université Toulouse Le Mirail)

Modalités de soumission

Les soumissions seront envoyées à l'adresse coldoc2012@gmail.com. Autant pour les communications que pour les posters, il est possible pour les non-francophones de soumettre les projets et les communications en anglais.

- **Communications :** Les doctorants et chercheurs intéressés par le sujet doivent soumettre un projet de deux pages comprenant un titre, un résumé, une bibliographie de cinq titres maximum et une liste de cinq mots clés (police 12, marge 2,5, interligne 1,5). Les communications orales comprendront vingt minutes de présentation suivie de dix minutes de discussion.

- **Posters :** Le colloque organisera aussi une séance de posters de format A1 pour présenter des recherches qui sont dans une phase initiale ou ne se prêtant pas au format long. Pour ces posters, la soumission à envoyer est un projet explicatif d'une page comprenant un titre, un résumé, une bibliographie de cinq titres maximum et une liste de cinq mots clés (police 12, marge 2,5, interligne simple).

Calendrier

Déroulement du colloque les jeudi 4 et vendredi 5 octobre 2012, Amphithéâtre Durkheim, Université Paris Descartes, 7 rue de la Sorbonne.

Date limite d’envoi des propositions : 4 Mai 2012

Notification d’acceptation : 22 Juin

Programme disponible : Mi-juillet

Remise des articles pour relecture : 3 septembre 2012

Comité scientifique

Jean-Michel ADAM (Université de Lausanne), Annie BERTIN (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Caroline BOGLIOTTI (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Bernard COMBETTES (ATILF, Université de Lorraine), Anne CONDAMINES (CLLE-ERSS/CNRS, Université Toulouse Le Mirail), Marcel CORI (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Flore COULOUMA (CREA-EA 370, Université Paris Ouest Nanterre La Défense), Guillaume DESAGULIER (MoDyCo - Université Paris 8, Université Paris Ouest Nanterre La Défense), Brigitte JUANALS (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Simon KREK (Institut Jozef Stefan, Ljubljana), Anne LACHERET (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Bernard LAKS (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Denis LE PESANT (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Danielle LEEMAN (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Sabine LEHMANN (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Sarah LEROY (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Sylvain LOISEAU (Laboratoire Lexiques, Dictionnaires, Informatique/CNRS, Université Paris 13-Nord), Dominique MAINGUENEAU (CEDITEC, Université Paris 12, IUF), Philippe MARTIN (Département d'Études Françaises, Université de Toronto), Jean-Luc MINEL (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Colette NOYAU (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Christophe PARISSE (MoDyCo / CNRS, INSERM, Université Paris Ouest Nanterre La Défense), Christiane PRENERON (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Sandrine REBOUL-TOURE (SYLED / CEDISCOR, Université Paris III-Sorbonne nouvelle), Fanny RINCK (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Clara ROMERO (MoDyCo / CNRS, Université Paris Ouest Nanterre La Défense), Frédérique SITRI (MoDyCo / CNRS, Université Paris 10, UPRES SYLED, université Paris III-Sorbonne nouvelle), Sylvie MELLET (Université Nice Sophia Antipolis, Ana ZWITTER VITEZ ('Institut de linguistique slovene appliquee Trojina, Ljubljana)

Comité d'organisation

Marine Damiani, Kaja Dolar, Carmen Lucia Florez-Pulido, Romain Loth, Julien Magnier et Anne Pegaz Paquet.