Sens Texte Informatique Histoire EA 4509

Recherches linguistiques et corpus

Responsable Franck Neveu

Le thème *Recherches linguistiques et corpus* mis en place au sein du laboratoire STIH de l'Université Paris-Sorbonne a pour objet le développement d'une réflexion commune et croisée, d'ordre épistémologique et méthodologique, sur la notion de corpus telle qu'elle est exploitée aujourd'hui dans les **sciences du langage**, et, plus largement, dans les **sciences humaines et sociales**. On y traite des questions relatives aux notions de donnée, d'observable, d'empiricité, de théorie (lien type/occurrence), de variable contextuelle, d'annotation, de codage, de catégorisation, etc. On s'interroge sur la fonction des corpus dans l'activité de recherche. Ce thème transversal est organisé en **séminaires-ateliers** ouverts notamment aux *chercheurs*, aux *enseignants-chercheurs*, aux *ingénieurs*, aux *doctorants* et aux *étudiants de master*. Les séances sont constituées de deux conférences, suivies d'un atelier d'observation et d'application.

4e séance Mercredi 13 mai 2015, 14h-17h30
Université Paris-Sorbonne, salle J636 (Esc. G/3ème étage)

Faire produire des données langagières de qualité en masse : jouir de la foule est un art

Karën Fort
Université Paris-Sorbonne, STIH (EA 4509)

Le tournant quantitatif de l'histoire littéraire

Alexandre Gefen
CNRS & Université Paris-Sorbonne, CELLF (UMR 8599)

Entrée libre

L'ordinateur portable peut être un élément de confort.

1, rue Victor Cousin, 75005, Paris
Contact : franck.neveu@paris-sorbonne.fr

Faire produire des données langagières de qualité en masse : jouir de la foule est un art

Karën Fort*

Le Traitement Automatique des Langues (TAL) est un domaine aujourd'hui en pleine expansion. Ses applications envahissent nos vies, avec ou sans notre accord. Mails, blogs, tweets, commentaires sur un produit ou un film, tout est source potentielle d'analyse automatique, pour nous proposer le meilleur service... et la meilleure surveillance.

La plupart des outils de traitement de ces grandes masses de données sont mis au point et testés sur des corpus annotés par des humains. Annoter, c'est associer à des segments de signal (d'énoncé) des notes ou des catégories. Par exemple aux mots leur catégorie morpho-syntaxique. Or, cette annotation manuelle originelle est extrêmement coûteuse et, encore aujourd'hui, mal maîtrisée.

La cohérence de l'attribution de catégories par les annotateurs doit être contrôlée. Les mesures de qualité de l'annotation ne sont pas toujours fiables. La multiplicité des cibles d'annotation (parties du discours, noms de personnes, marqueurs d'opinion positive ou négative...) rend les comparaisons difficiles et complique la compréhension en profondeur des obstacles et des solutions.

Je présenterai lors de cet atelier une partie de mes travaux sur ces sujets. Je rapprocherai la tâche d'annotation du jugement d'acceptabilité en linguistique. Je montrerai comment, une fois le processus d'annotation mieux « encadré », il est possible de faire réaliser par la foule une tâche d'annotation réputée complexe, l'analyse syntaxique, via un jeu (http://zombilingo.org).

* Karën Fort est Maîtresse de Conférences en Informatique à l'Université Paris-Sorbonne. Elle est membre du laboratoire STIH (équipe Linguistique computationnelle). Ses recherches portent : (i) sur le *crowdsourcing* (la myriadisation), en particulier dans sa dimension science citoyenne (*citizen science*), notamment les jeux ayant un but (*Games With A Purpose*) pour la création de ressources langagières ou l'apprentissage des langues ; (ii) sur l'annotation collaborative en général, ses dimensions de complexité et l'apprenabilité des humains en lien avec l'apprenabilité des systèmes ; (iii) sur les mesures d'évaluation de l'annotation manuelle et leur signification. Elle s'intéresse également à l'analyse du discours pathologique.

Le tournant quantitatif de l'histoire littéraire

Alexandre Gefen**

En sciences humaines comme en sciences dures, les numérisations massives des textes et des données produisent depuis plus d'une décennie des big data ou long data ouvrant des pistes de recherche novatrices. Mais les méthodes critiques qui les accompagnent ont des enjeux épistémologiques, institutionnels et pédagogiques considérables. Dans ce qu'on appelle désormais les « humanités numériques », la lecture à distance de corpus constitués par des cartes et graphes offre une forme spécifique de savoir et un paradigme méthodologique et épistémologique qu'il importe de saisir dans toute sa puissante heuristique, sans se laisser entraîner par l'idée naïve d'une production transparente de savoirs par moissonnage des corpus, masse de données qui restent des artéfacts muets en l'absence d'une herméneutique spécifique. À ce titre on réfléchira sur les méthodes disponibles et les bilans des premières expériences d'analyse quantifiée de l'histoire littéraire.

** Chargé de recherche au Centre d'Étude de la Langue et de la Littérature Française (CNRS-Université Paris 4), Alexandre Gefen travaille sur des questions de théorie littéraire appliquées en particulier à la littérature française contemporaine. Fondateur du site Fabula.org et membre du Labex OBVIL (Sorbonne université), il s'intéresse parallèlement aux champs des Humanités numériques, à travers des travaux portant sur le web scientifique, la philologie numérique et ses enjeux épistémologiques, les écritures en réseau et les cultures numériques. Il est Porteur, avec Franco Moretti (Stanford), du projet « Pour une histoire empirique de la littérature », Transatlantic program for collaborative work in the field of digital humanities, FMSH Paris-Fondation Mellon (2015-2018).