

Université de Cergy-Pontoise - Maison des Sciences de l'Homme de Paris-Nord - Institut National de l'Audiovisuel
Centre de Recherche Textes et Francophonies, pôle LaSCoD, pôle interdisciplinaire de recherche en Sciences du Langage
et Sciences de l'Information et de la Communication

Journée d'étude

6 mai 2010

9h15-17h30

INA

Salle Cognac-Jay

85-87 rue de Patay

75013 Paris

Métro ligne 14 : Arrêts *Olympiades* ou *Bibliothèque François Mitterrand*

RER C : Arrêt *Bibliothèque François Mitterrand*

« L'analyse des corpus face à l'hétérogénéité des données: réflexions méthodologiques et pratiques interdisciplinaires »

Coordination

Julien Longhi (julien.longhi@u-cergy.fr)

Aurélien Le Foulgoc (aurelien.lefoulgoc@u-cergy.fr)

En collaboration avec Denis Maréchal (dmarechal@ina.fr)

Cette journée d'étude organisée en partenariat avec la Maison des Sciences de l'Homme de Paris-Nord et l'Institut National de l'Audiovisuel, par le pôle LaSCoD du Centre de Recherche Textes et Francophonies de l'université de Cergy-Pontoise, est la première d'une série de manifestations, destinées à la réalisation d'un projet de recherche, dont l'objectif est la constitution d'un corpus informatisé, composé de matériaux de nature différente, consacré aux jeunes de banlieues ; en partant de l'hypothèse que ceux-ci participeraient à la construction d'une mythologie urbaine contemporaine, hypothèse qui, si elle se vérifiait, permettrait d'éclairer un certain nombre de processus de sémiotisation de l'espace social.

On fondera cette journée sur l'idée qu'un corpus est un ensemble raisonné de textes ou de matériaux audiovisuels, structuré par une cohérence interne. On soulignera d'emblée qu'il existe différentes modalités de constitution des corpus et qu'il n'y a pas une seule conception scientifique reconnue mais une pluralité de méthodes et de démarches, qui posent un certain nombre de questions d'ordre épistémologique, dont la première est qu'on ne peut postuler l'existence d'un modèle idéal, qui serait d'emblée représentatif et herméneutique. La recherche étant interdisciplinaire et dialogique, le premier objectif de la journée sera donc comparatif et différentiel, étant entendu que les disciplines représentées ici postulent qu'un corpus est nécessairement un construit, que les données qu'il propose ont fait l'objet d'une sélection préalable et que la question de la légitimité de sa représentativité est à l'horizon de la réflexion et peut constituer un écueil qu'il ne faut pas négliger. Enfin, la confrontation des approches visera à croiser dans une perspective heuristique les démarches propres à la constitution des corpus audiovisuels et les acquis des linguistiques de corpus en tentant d'instaurer les bases d'une recherche dont la question centrale porte sur le fait de savoir si les textes et les discours collectés contribuent à la constitution d'une problématique cohérente.

Les chercheurs invités montreront chacun à leur manière comment les corpus peuvent être problématisés et utilisés afin de faire émerger l'homogénéité ou l'hétérogénéité des données et des phénomènes appréhendés, ainsi que la complexité des phénomènes à décrire. L'hétérogénéité des corpus, tout comme la diversité des pratiques proposées, mettra en valeur la richesse, et la difficulté, à renouveler l'analyse des corpus selon leur hétérogénéité et leurs diverses modalités d'appréhension.

Programme

(Présentations : 45min ; Questions : 15min)

- 9h15-9h45 : **Julien Longhi** – Maître de conférences à l'Université de Cergy-Pontoise – CRTF-LaSCoD

Introduction

- 9h45-10h45 : **Pascal Marchand** – Professeur à l'Université de Toulouse 3 – LERASS

L'approche lexicométrique : recherche improbable d'une homogène diversité?

Le débat récent à propos de l'identité nationale, sur le site Web du Ministère de l'Immigration, de l'intégration, de l'identité nationale et du développement solidaire, permettra d'illustrer les différentes phases d'analyse de données textuelles, les difficultés d'hétérogénéité dans l'expression des opinions et les différentes hypothèses que l'on peut poser sur un tel corpus.

Lien utile : <http://pascal-marchand.fr>

- 10h45-11h : **Pause**
- 11h-12h : **Nathalie Garric** – Maître de conférences à l'Université de Tours – LLL

Maîtriser l'hétérogénéité des corpus par la variation des corpus, des outils et des méthodes

Les nouveaux contextes d'intervention de l'analyse de discours ouvrent ses champs d'application à des domaines qui introduisent des modalités de recherche différentes. Les données, les corpus, les outils, les méthodes et plus généralement les attentes scientifiques ne sont plus les mêmes que dans les contextes initiaux, notamment dans ceux du discours politique. Ces changements induits par les contextes d'analyse nécessitent la recherche de solutions théoriques et pratiques que nous aborderons à partir d'une étude de cas, l'analyse de discours d'apprentis en formation en alternance. Cette étude, construite dans l'interdisciplinarité, est singulière tout particulièrement en raison des données qu'elle utilise et produit à la fois. Elle a été réalisée à partir d'entretiens semi-directifs et de textes, apparentés au bilan de savoirs, et s'insère dans le cadre de la linguistique de corpus en s'appuyant sur un traitement quantitatif textométrique et des analyses qualitatives énonciatives. Son objectif est d'identifier et d'analyser les processus à l'origine des ruptures des contrats d'apprentissage dans le parcours de formation d'apprentis de niveau V.

Lien utile : <http://www.cavi.univ-paris3.fr/lexicométrica/jadt/jadt2006/PDF/038.pdf>

- 12h-14h : **Déjeuner**
- 14h-15h : **Corinne Gauthier, Anne Pavis** – INA – Inathèque de France

La constitution de corpus de recherche à partir des bases de données documentaires de l'INA

Cette intervention portera sur la question de l'histoire et de la structure des sources audiovisuelles et écrites de la radio-télévision, de leur appréhension par le chercheur en sciences humaines et des outils d'aide au traitement et à l'analyse de corpus constitués (via *Hyperbase* et *MediaCorpus*) proposés au centre de consultation de l'Inathèque de France.

Lien utile : <http://www.ina-sup.com/collections/comment-consulter-0>

- 15h-16h **Josquin Debaz** – Postdoctorant - EHESS – GSPR

Socio-informatique des controverses et des conflits. Saisir les jeux d'acteurs et d'arguments dans de grands corpus évolutifs

Les travaux menés sous l'appellation de « socio-informatique » ont conduit à une importante accumulation de corpus, d'outils, de modèles, de textes et de discussions. Ils ont ainsi ouvert un espace pour l'expérimentation et le croisement des démarches d'enquêtes sur de grands dossiers où la réflexivité doit l'emporter sur les fonctions d'utilité instrumentale.

Lien utile : <http://socioargu.hypotheses.org/1>

- 16h-16h15 : **Pause**
- 16h15-17h15 : **Guilhem Fouetillou** - CTO et cofondateur de Linkfluence

Ecologie du web social

Les méthodologies développées par *Linkfluence* depuis 3 ans maintenant visent à établir le web social comme un terrain d'expérimentation à grande échelle, plastique aux interactions sociales. Cette plasticité (caractère stigmergique du web) permet de suivre le social en action à travers ses artefacts techniques. *Linkfluence* a développé des méthodes d'échantillonnage des espaces sociaux du web en s'appuyant sur les propriétés structurelles de l'hypertexte. Cet échantillonnage permet aujourd'hui de lier approches quantitatives et qualitatives en sciences sociales.

Lien utile : <http://fr.linkfluence.net/>

- 17h15-17h30 : **Clôture de la journée Aurélien Le Foulgoc**