PARCOURS Traitement informatique des corpus

Responsable : Sylvain Kahane sylvain@kahane.fr

Le parcours TIC propose un certain nombre de cours introduisant la problématique du traitement automatique d'un corpus de textes, ces textes pouvant être de domaines (histoire, philosophie, etc.) comme de genres (journalistique, scientifique, etc.) très variés.
L'objectif est à la fois appliqué (savoir traiter un corpus – i.e. lui appliquer les bons outils – pour extraire certains types d'informations) et théorique (comprendre comment fonctionnent de tels outils et ce qu'on peut attendre d'eux). Les connaissances théoriques touchent à quatre domaines :

- l'analyse du discours s'intéresse aux textes en tant que productions langagières dans un contexte donné, ce qui détermine le « genre » d'un texte ;
- la logique, initialement développée pour la modélisation du raisonnement, est à la base de la programmation informatique comme de la notion de grammaire formelle ;
- l'algorithmique s'intéresse à la définition, à la modélisation et au développement de solutions informatiques à des problèmes de classement de données, de recherche d'éléments ou de motifs, de résolution de contraintes, etc. ;
- les statistiques permettent, par un traitement numérique des occurrences des mots et de leurs appariements, de décider quels sont les éléments les plus pertinents d'un texte ou d'un corpus de textes.

3LIL506S **Documents électroniques**
Volume horaire : 24 h TD
Responsable : Delphine Battistelli Delphine.Battistelli@u-paris10.fr

Descriptif :
Pour permettre de mieux se servir du Web et des échanges électroniques, dans sa formation comme professionnellement, on détaillera ce qu'est un document électronique : les caractères utilisés pour les différentes langues, les formats facilitant ou non les transferts de documents, les conversions, etc. Le contexte d'utilisation sera également présenté : le Web et l'Internet, les manières de transférer de l'information, les moteurs de recherche.

Bibliographie :
Calederan C. et Laurent P. *Le document électronique à l'heure du Web*, INRIA, 2012
Michard A. *XML : Langage et applications*. Eyrolles, 2002

Modalités de contrôle :

- **Formule standard session 1 :** Contrôle terminal en temps limité (2h)
- **Formule dérogatoire session 1 :** Une épreuve en temps limité (2h)
- **Session 2** : Une épreuve en temps limité (2h)

3LIL606S **Linguistique informatique et linguistique de corpus 1**
Volume horaire : 18 h TD
Responsables : Delphine Battistelli Delphine.Battistelli@u-paris10.fr
Jean-Luc Minel : jean-luc.minel@u-paris10.fr

Description de l'enseignement, principaux contenus :
La numérisation des données langagières écrites et sonores a profondément bouleversé les méthodes de traitement et d'analyse en linguistique, en donnant notamment accès à des gros volumes de données, que l'on désigne sous le terme de *corpus*. De nouvelles méthodes d'observation des faits langagiers en sciences du langage ont ainsi vu le jour ces quinze dernières années et ont donné lieu à la pratique active de la linguistique de corpus. Ces méthodes nécessitent la constitution de ressources, *i.e.* données langagières et outils pour les traiter, tant à l'oral qu'à l'écrit (lexiques spécialisés, dictionnaires informatisés,

grammaires locales, annotation syntaxique et sémantique, outils de segmentation automatique et d'annotation prosodique de la parole, etc.). Elles exigent la mise en place de procédures rigoureuses afin de contrôler au mieux les résultats obtenus. Une première partie du cours sera consacrée à la présentation d'un panorama des outils et des méthodes actuellement utilisées en linguistique de corpus dans un premier temps sur des textes écrits, ensuite sur des données sonores. Une seconde partie du cours sera consacrée à la présentation des outils de programmation (le module NLTK notamment) en Python qui permettent de traiter des corpus.

Bibliographie :
Condamines A., (2005), *Sémantique et Corpus*, Paris : Lavoisier.
Habert B., Nazarenko A., Salem A. (1997), *Les linguistique de corpus*, Paris : Armand Colin.
Habert B., Fabre C., Issac F. (1998), *De l'écrit au numérique : constituer, normaliser et exploiter les corpus éléctronique,* Paris : InterEditions.
Habert B. (2006), *Instruments et ressources électroniques pour le français.* Paris/Gap : Ophrys (« L'Essentiel Français »).
Minel J.-L. (2009), *Filtrage sémantique : de l'annotation à la navigation textuelle*. Paris : Lavoisier.

Modalités de contrôle :
- **Formule standard session 1 :** Contrôle terminal en temps limité (1h30)
- **Formule dérogatoire session 1 :** Une épreuve en temps limité (1h30)
- **Session 2** : Une épreuve en temps limité (1h30)

3LIL607S **Linguistique informatique et linguistique de corpus 2**
Volume horaire : 18 h TD
Responsables : Delphine Battistelli : Delphine.Battistelli@u-paris10.fr
Jean-Luc Minel : jean-luc.minel@u-paris10.fr

Description de l'enseignement, principaux contenus :
Le cours présentera tout d'abord sommairement la notion d'automate telle qu'elle est abordée dans le domaine de la linguistique informatique. Il invitera ensuite à se familiariser avec Unitex, un logiciel d'annotation automatique de textes à l'aide d'automates. L'intérêt de ce logiciel réside dans son interface graphique qui permet de constituer et de maintenir des automates complexes. Deux objectifs sont alors visés : 1) apprendre à construire des ressources linguistiques qui permettent à leur tour l'annotation de textes (en particulier, une annotation sémantique) ; 2) se confronter à la complexité de la modélisation linguistique impliquant une réflexion sur les objets textuels manipulés et qui implique parfois de revenir sur ses premières intuitions.
Un deuxième logiciel, TXM, fondé sur des méthodes statistiques, sera ensuite présenté et donnera lieu à des analyses de corpus, notamment :
- production des concordances à partir de recherches de motifs lexicaux complexes construits à partir des propriétés des mots
- vocabulaire d'ensemble d'un corpus ou liste des valeurs attestées d'une propriété d'un mot donnée ;
- construction de différents tableaux de contingence croisant les mots, les textes et leurs structures ;
- calcul de la liste des mots apparaissant de façon préférentielle dans les mêmes contextes qu'un motif lexical complexe (cooccurrents statistiques).

Bibliographie :
Habert, Benoît. Portrait de linguiste(s) à l'instrument. Texto! [en ligne], décembre 2005, vol. X, n°4.
Marandin, Jean-Marie, Cori, Marcel. La linguistique au contact de l'informatique : de la

construction des grammaires aux grammaires de construction. Histoire Épistémologie Langage, 23-1, 2001, pp. 49-79.
Manuel d'Unitex http://www-igm.univ-mlv.fr/~unitex/
Présentation de TXM : http://textometrie.ens-lyon.fr/spip.php?rubrique96

Modalités de contrôle :

- **Formule standard session 1 :** Contrôle terminal en temps limité (1h30)
- **Formule dérogatoire session 1 :** Une épreuve en temps limité (1h30)
- **Session 2** : Une épreuve en temps limité (1h30)

3LILI406S **Statistiques**
Volume horaire : 18 h TD
Responsables : Frédéric Isel : fisel@u-paris10.fr
Sylvain Kahane : skahane@u-paris10.fr

Description de l'enseignement, principaux contenus :
L'objectif de cet EC sera de présenter aux étudiants divers outils statistiques qui permettent d'analyser des données quantitatives extraites de corpus linguistiques.

Tout d'abord, nous introduirons les quatre étapes impliquées dans l'étude de tout phénomène statistique : (1) le recueil des données (notion d'échantillon représentatif), (2) la présentation des données (représentations graphiques, tableaux), (3) l'analyse des données, et (4) la fiabilité des résultats. Nous évoquerons également la typologie de la statistique descriptive. Puis, nous définirons les notions de variables indépendantes (qualitatives, quantitatives), de variables dépendantes, et de facteurs (principaux, secondaires). Nous introduirons également les notions d'opérationnalisation des variables, de plan d'analyse et de vérification des hypothèses.

Enfin, nous étudierons en quoi consiste l'étape d'analyse statistique descriptive en nous appuyant sur ses principaux paramètres (moyenne, médiane, amplitude, écart-type, erreur standard de la moyenne) ainsi que celle d'analyse statistique inférentielle. À cet effet, une familiarisation aux logiciels d'analyse statistique Statistica, SPSS et R sera proposée.

Bibliographie :

Borsali, F. (2010). *Statistiques médicales et biologiques*. Collection L1 Santé, Editions Ellipses.

Bry, X. (1999). *Analyses factorielles simples*. Economia. Techniques Quantitatives Poche.

Corroyer, D., & Wolff, M. (2003). *L'Analyse statistique des données en psychologie*. *Concepts et Méthodes de base*. Paris : Armand Colin (Cursus).

Howell, D. C. (2008). *Méthodes statistiques en Sciences Humaines*. Paris : De Boeck.

Modalités de contrôle :

- **Formule standard session 1 :** Contrôle terminal en temps limité (1h30)
- **Formule dérogatoire session 1 :** Une épreuve en temps limité (1h30)
- **Session 2** : Une épreuve en temps limité (1h30)

3LSL105S **Introduction à la logique**
Volume horaire : 24 h TD
Responsable : François Métayer : francois.metayer@u-paris10.fr

Description de l'enseignement, principaux contenus :
Ce cours d'introduction à la logique abordera la formalisation des énoncés et des preuves du point de vue de la déduction naturelle. Cette présentation de la logique propositionnelle et du calcul des prédicats nous conduira très naturellement au seuil de la théorie des types, outil central de la linguistique computationnelle et de l'informatique théorique contemporaines.

Bibliographie :
Y. Delmas-Rigoutsos, R. Lallement, *La logique ou l'art de raisonner,* Editions du Pommier.
G. Dowek, *La logique*, Flammarion.
S. Kleene, *Logique mathématique*, Armand Colin.

Modalités de contrôle :
- **Formule standard session 1 :** Contrôle continu. Deux notes dont au moins un partiel sur table. Chaque note entre pour 50% dans la note finale.
- **Formule dérogatoire session 1 :** Une épreuve en temps limité (2h)
- **Session 2** : Une épreuve en temps limité (2h)

L3XXX **Algorithmique**
Volume horaire : 24 h TD
Responsable : Delphine Battistelli : Delphine.Battistelli@u-paris10.fr

Description de l'enseignement, principaux contenus :
L'objectif du cours est d'introduire les concepts fondamentaux de l'algorithmique et de la programmation. Il s'agit ainsi d'apprendre à manipuler les structures de données classiques (chaînes de caractères, tableaux, enregistrements, files et piles) pour concevoir des programmes structurés dans un langage de programmation. Un accent particulier sera mis sur l'étude d'algorithmes simples de manipulation de textes.

Bibliographie :
Claude Delannoy, *Initiation à la programmation*, Eyrolles, 2002.

Bruno Warin, *L'algorithmique : Votre passeport informatique pour la programmation*, Hors Collection, 2002.

Modalités de contrôle :
- **Formule standard session 1 :** Contrôle continu. Deux notes dont au moins un partiel sur table. Chaque note entre pour 50% dans la note finale.
- **Formule dérogatoire session 1 :** Une épreuve en temps limité (2h)
- **Session 2** : Une épreuve en temps limité (2h)

3LSM501S **Discours/texte 2** (dialogisme)

Volume horaire : 24h TD
Responsables : C. Mellet : caroline.mellet@wanadoo.fr
F. Sitri fsitri@u-paris10.fr

Description de l'enseignement, principaux contenus :
On envisagera ici l'analyse de discours comme un point de vue particulier sur les textes, qui s'intéresse aux productions langagières dans leur relation avec un « contexte ».
On se propose d'envisager les productions discursives du point de vue des « voix » qui les traversent, du dialogue qu'elles entretiennent avec d'autres discours. A partir de l'analyse de

textes divers, tant oraux qu'écrits, on présentera les différentes formes de discours rapporté et plus largement de représentation du discours autre. On pourra mettre en relation les formes rencontrées avec les genres discursifs dont relèvent les textes étudiés. Plus généralement, on proposera une définition de la notion de genre et on mettra en évidence les contraintes génériques qui pèsent sur la production et l'interprétation des discours.

Bibliographie

Authier-revuz J., 2001, « Le discours rapporté », in Grands repères culturels pour une langue: le français, Paris, Hachette, p.192.
Bakhtine M., 1984, Esthétique de la création verbale, Paris, Gallimard.
Bres J. et Vérine B., 2002, « Le bruissement des voix dans le discours: dialogisme et discours rapporté », *Le discours rapporté, Faits de langue, 19*, p. 159-169.
Maingueneau D., 1999, L'énonciation en linguistique française, Paris, Hachette (troisième partie : « Le discours rapporté »).

Modalités de contrôle :

- **Formule standard session 1 :** Contrôle continu. Deux notes dont au moins un partiel sur table. Chaque note entre pour 50% dans la note finale.
- **Formule dérogatoire session 1 :** Une épreuve en temps limité (2h)
- **Session 2** : Une épreuve en temps limité (2h)