LIDILEM
UNIVERSITÉ Grenoble Alpes

Ateliers de formation – élaboration et exploitation de corpus

Consortium CORLI

Lundi 3 juillet 2017 – 14h-17h

Atelier Lexicoscope

Olivier Kraif

Le Lexicoscope est un outil accessible en ligne dédiée à l'interrogation de corpus de grand volume annotés syntaxiquement. Tout comme le Sketch Engine, il permet d'étudier les profils combinatoires des mots en s'appuyant sur un modèle de cooccurrence syntaxique, basé sur l'extraction des dépendances. A la différence de celui-ci, il intègre la possibilité d'étudier le profil combinatoire d'expressions complexes, voire de constructions lexicosyntaxiques plus abstraites. Ce type d'interrogation s'appuie sur un langage de requête complet et puissant (comme CQP ou TigerSearch), mais il n'est pas indispensable de connaître ce langage, ni les jeux d'étiquettes associés, pour la prise en main et l'utilisation de l'outil : en effet, une fonctionnalité originale de construction des requêtes à partir d'exemples permet d'amorcer la recherche sans connaissances préalables.

Lien vers le Lexicoscope :

http://phraseotext.u-grenoble3.fr/lexicoscope

Kraif et Diwersy (2012). « Le Lexicoscope: un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques ». In *19e conférence TALN*.

Atelier Anatext

Agnès Tutin

Anatext est un outil simple textométrique, multilingue, utilisable en ligne. Il suffit de copier/coller son texte pour procéder à une analyse de corpus. L'outil permet d'effectuer des recherches dans les textes à partir d'un étiquetage réalisé avec TreeTagger et comporte plusieurs fonctionnalités statistiques :

- comptage des phrases, des mots et des syllabes ;
- étude du vocabulaire et des spécificités lexicales;
- extraction et tri des fréquences des lemmes (par catégories : nom, verbe, adjectif, etc.) ;
- extraction et tri des fréquences des formes (par catégories : nom, verbe, adjectif, etc.) ;
- extraction des segments répétés ;
- recherche de patterns (du type : DET ADJ NOM) ;
- extraction des cooccurrences.

L'atelier présentera les principales fonctionnalités de l'outil et des exemples d'applications sur des corpus du français et de l'anglais.

Lien vers Anatext :

http://olivier.kraif.u-grenoble3.fr/anaText/

Mardi 4 juillet 2017 – 9h-12h

Atelier Scienquest / Scientext

Achille Falaise

ScienQuest, initialement développé dans le cadre du projet Scientext, est un outil d'exploitation de corpus annotés (en parties du discours, lemmes et dépendances syntaxiques) simple à utiliser, pour des utilisateurs non spécialistes du TALN. Il s'agit d'un outil en ligne, ne nécessitant aucune installation, mais limité aux corpus présents sur la plateforme (on ne peut pas importer son propre corpus). Actuellement, la plupart d'entre-eux sont des corpus de textes scientifiques, en anglais et en français.

Lien vers la plateforme, avec une liste des corpus :

http://corpora.aiakide.net

Achille Falaise, Agnès Tutin, Olivier Kraif (2011). « Une interface pour l'exploitation de corpus arborés par des non informaticiens : la plate-forme ScienQuest du projet Scientext », Revue TAL, Volume 52 – n° 3/2011, pp. 103-128.

http://www.atala.org/Une-interface-pour-l-exploitation

Atelier Expressions régulières

Marie-Paule Jacques

Divers outils d'exploration de corpus, tels qu'AntConc ou TXM, permettent d'augmenter la puissance des requêtes qui leur sont adressées par l'utilisation d'expressions régulières. Ces expressions reposent sur des « jokers » et des caractères réservés. Il est par exemple possible de rechercher en une seule requête tous les mots se terminant par « able » ou commençant par « re » ou une combinaison de ces deux contraintes, de définir si un caractère ou une série de caractères sont optionnels, etc.

La matinée sera consacrée à une découverte de ces expressions et à leur mise en œuvre pour des requêtes en corpus. L'atelier ne suppose pas de connaissance spécifique autre qu'une connaissance des fonctions de recherche dans les textes. Il débutera avec le logiciel AntConc :

http://www.laurenceanthony.net/software/antconc/

mais pourra aborder les requêtes à construire dans TXM (selon le temps disponible) :

http://textometrie.ens-lyon.fr/?lang=fr

Atelier Métadonnées

Loïc Liégeois, Christophe Parisse ou Carole Etienne (à définir)

Les métadonnées, associées à des corpus de linguistique, sont un outil incontournable de la linguistique de corpus. Elles permettent en particulier :

— de repérer des données sur le web et d'y accéder ;

— de savoir quel type de matériel linguistique les données contiennent et ainsi éviter d'avoir à explorer manuellement les données retrouver ces informations ;

— de sélectionner des corpus ou des enregistrements dans le but de mener des analyses poussées (par exemple en sociolinguistique ou en psycholinguistique) en fonction de critères précis (âge des locuteurs du corpus, disponibilité des données primaires, type de situation d'interaction etc.).

Toutefois, l'existence et l'utilisation des métadonnées à des fins de recherche sont fortement dépendantes de deux conditions liées et complémentaires :

— la structuration des métadonnées dans des formats standard partagés et facilement disponibles, à la fois riches (pour permettre des analyses poussées) et souples d'utilisation ;

— l'existence d'outils permettant de saisir les métadonnées.

Le groupe de travail « Interopérabilité et exploration des corpus de linguistique » du consortium CORLI, qui travaille depuis plusieurs années sur le sujet, présentera un format et un outil utilisable pour encoder les métadonnées. Il s'agira en particulier de présenter et de faire manipuler par les participants :

— un niveau « zéro » de codage des métadonnées de corpus oraux et multimodaux suffisamment riche pour le travail scientifique et suffisamment simple pour être facilement partagé ;

— un outil de saisie des métadonnées suivant le format de la TEI. Cet outil fonctionne dans un navigateur web et peut être utilisé en local ou à distance. Il est modulaire et basé sur un format de description de métadonnées permettant de créer autant de versions que nécessaire pour couvrir tout type de métadonnées, y compris des métadonnées hors de la TEI (OLAC et Dublin Core par exemple).