Corpus écrits et transcrits

Niveau d'étude
Bac +4

ECTS
2 crédits

Composante
UFR Langage, lettres et arts du spectacle, information et communication

- **Langue(s) d'enseignement:** Français
- **Forme d'enseignement :** Cours magistral
- **Ouvert aux étudiants en échange:** Non

Présentation

Description

Les recherches en linguistique s'appuient maintenant de façon incontournable sur des outils de traitement de corpus – et les applications du TAL sont également très consommatrices de corpus langagiers finement annotés.

Après avoir présenté différentes utilisations des corpus langagiers dans une double perspective, à la fois scientifique (linguistique outillée) et industrielle (IDL), le cours aborde des aspects méthodologiques liés à la constitution et à l'exploitation de ces ressources.

Objectifs

Objectifs

On montrera dans un premier temps l'intérêt de la démarche de la linguistique de corpus, et on donnera un aperçu de l'usage des corpus en Traitement automatique des langues (TAL). Les notions suivantes seront abordées :

- typologie des corpus écrits et transcrits

- principes de constitution des corpus, métadonnées et annotations

- fonctionnalités des outils d'exploration de corpus

- aspects techniques liés à la constitution : encodage des caractères, formats et normes d'encodage (XML-TEI)

- recherche de patterns et requêtes complexes (expressions régulières, xpath)

- données textométriques de base : fréquences, spécificités, mesures d'association.

Heures d'enseignement

Corpus écrits et transcrits - CMTD	Cours magistral - Travaux dirigés	20h

Contrôle des connaissances

Contrôle continu, examen sur table.

Période : Semestre 7

Compétences visées

Référentiel des compétences RNCP :

- Identifier les usages numériques et les impacts de leur évolution sur le ou les domaines concernés par la mention

- Se servir de façon autonome des outils numériques avancés pour un ou plusieurs métiers ou secteurs de recherche du domaine

A l'issue du cours, les étudiant.e.s seront capables de compiler un corpus au format XML TEI comportant les métadonnées nécessaires et des annotations simples, d'ouvrir et de manipuler un corpus avec un logiciel adapté, d'utiliser des expressions régulières pour effectuer des opérations de recherche et/ou de nettoyage, d'élaborer des requêtes pour rechercher des patterns, d'interpréter les données textométriques de base (fréquences, spécificités, mesures d'association).

Bibliographie

Néé, E. (2018, sous la dir. de) *Méthodes et outils pour l'analyse des discours*, PUR.

Poudat, C., Landragin Frédéric (2017) *Explorer un corpus textuel. Méthodes, pratiques, outils*, Bruxelles, De Boeck.

Supports élaborés dans le cadre du MOOC Linguistique de corpus :

https://www.fun-mooc.fr/fr/cours/introduction-a-la-linguistique-de-corpus/

Infos pratiques

Lieu(x) ville

Grenoble

Campus

Grenoble - Domaine universitaire