Stage au GREYC
en Traitement Automatique des Langues

L'équipe Données, Documents et Langue du GREYC – CNRS UMR 6072, Caen, propose un stage d'une durée de 2 mois en collaboration avec le laboratoire de linguistique CRISCO.

Sujet

Mise en place d'outils d'annotation linguistique pour textes poétiques et théâtraux du 17e au 20e.

Encadrement

Stéphane Ferrari, MC informatique, responsable du groupe *Sémantique et TAL* de l'équipe Données, Documents et Langue.

Durée:

2 mois

Objectifs

L'étudiant participera à un projet d'élaboration d'outils d'analyse automatique des formes métriques en vue de la constitution d'un corpus annoté d'oeuvres poétiques et théâtrales du XVIIe au début du XXe siècle. Parmi les traitements envisagés, on retiendra notamment ceux concernant l'analyse métrique et l'analyse syntaxique.

> Concernant l'analyse métrique, le principe repose sur un repérage des noyaux syllabiques (voyelles) afin de permettre la détermination du mètre et des rimes sans passer par le calcul des syllabes. En aval de ce traitement, l'analyse métrique consiste à préciser les regroupements en hémistiches, en vers, en modules de strophes puis en strophes.
>
> Concernant l'analyse syntaxique, le problème majeur est celui de l'adaptation d'outils et de ressources pour l'assignation de catégories grammaticales (*tagging*) aux formes très variables de l'ancien français.

Un travail en amont consistera à proposer des formats de représentation des ressources permettant leur utilisation optimale sur l'ensemble des textes de la collection, notamment en tenant compte de l'évolution de la langue sur la période concernée. De même, une réflexion sera menée sur les formats d'annotation du corpus en sortie. Le format XML, et en particulier la norme TEI, est envisagé pour le stockage et la consultation ultérieure du corpus, mais des ajustements ou ajouts seront nécessaires, pour prendre en considération les possibilités de chevauchement entre les différentes structures étudiées : la structure syntaxique et les vers ou les hémistiches ne sont pas en rapport de hiérarchie.

L'étudiant s'intégrera dans un projet en collaboration pluridisplinaire entre informaticiens (GREYC), linguistes (CRISCO) et littéraires (THL) de Caen.

Contact :

Stephane.Ferrari@info.unicaen.fr

02 31 56 73 97