



Sujet :

Développement de ressources pour l'analyse linguistique du Français

Niveau : Bac + 5 – Master

Durée du stage : 4 à 6 mois

Contexte

Le laboratoire LIC2M a développé un logiciel d'analyse linguistique multilingue nommé LIMA. Il inclut pour le Français un étiqueteur, un analyseur syntaxique, un moteur de résolution de coréférences, etc. Plusieurs de ces modules utilisent un moteur de reconnaissance d'expressions régulières. LIMA utilise des ressources pour effectuer l'analyse: dictionnaires, fichiers de règles de reconnaissances, listes d'entités prédéfinies, etc. Ces ressources sont construites manuellement pour la plupart.

Objectifs du stage

Le travail du stagiaire consistera à enrichir les ressources du français utilisées par LIMA de manière à améliorer ses performances. Cela consistera à exploiter les résultats de l'analyseur pour repérer des erreurs et à compléter les ressources pour supprimer ces erreurs. En sus de cette pure expertise linguistique, il sera demandé au stagiaire de participer à l'amélioration des procédures et outils permettant de faciliter ces développements de ressources.

Domaines de spécialité requis

Informatique, Linguistique, Morphologie, Syntaxe

Moyens informatiques mis en œuvre

Langages : langages de scripts (bash, perl...)

Logiciels : LIMA, GNU/Linux, éditeurs de ressources

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT - LIST

Dépt/Service/Labo : DTSI/SRCI/LIC2M :

Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue

Web : <http://www-list.cea.fr/>

Encadrement

Nom de l'ingénieur responsable : Gaël de Chalendar **Tél.** : +33 1 46 54 80 18

E-mail : gael.de-chalendar@cea.fr **Fax** : +33 1 46 54 75 80

Secrétariat : DTSI/SRCI **Tél.** : +33 1 46 54 91 17

Nom du chef de laboratoire : Olivier Mesnard **Tél.** : +33 1 46 54 73 38