

Stage en linguistique de corpus

Dans le cadre du projet PESACA (post-édition semiautomatique à partir du corpus agroalimentaire) au sein de la COMUE UBFC, l'équipe de chercheurs souhaite renforcer son équipe des linguistes.

Ce projet PESACA vise, à travers l'entraînement d'un modèle de post-édition, à améliorer la qualité des sorties de systèmes de traduction automatique neuronale en français langue de spécialité, en prenant comme cas d'usage pilote le domaine vitivinicole.

La mission du stagiaire consistera en nettoyer et analyser un corpus en français langue de spécialité, d'une part et de l'autre, à en extraire des données structurées et non structurées en se basant sur des concepts sémantiques, statistiques et linguistiques.

PROFIL

Le stagiaire en Master TAL avec de solides bases en linguistique de corpus devra faire preuve d'autonomie, de curiosité et d'esprit analyse afin de mener sa mission à bien.

Connaissances et/ ou intérêts pour :

- Logiciels : Python, OpenCV, Sketch Engine, AntConc.
- XML EAD, TXT, machine learning, BERT.
- Linguistique de corpus.
- Principes de l'apprentissage statistique et applications.
- Méthodes de traitement automatique des langues.
- aisance en programmation.
- Manipulation de bases de données volumineuses.

Durée Trois mois. Ce stage est à pourvoir dès que possible.

Gratification Le taux horaire de la gratification est égal au minimum à 3,90 € par heure de stage.

Lieu

Université de Bourgogne, Centre Interlangues – Texte, Image, Language (EA 4182)

Université de Franche-Comté, Centre de Recherches Intertextuelles et Transculturelles, CRIT (EA 3224)

Contact

Luz MARTINEZ – Luz.martinez-villermosa@u-bourgogne.fr

Izabella THOMAS - izabella.thomas@univ-fcomte.fr