INGÉNIEUR
EN DÉVELOPPEMENT ET OUTILLAGE DE CORPUS COMPLEXES

Recrutement CDD 11 mois
École Normale Supérieure de Lyon

Affectation : Laboratoire Interactions Corpus Apprentissages Représentations CNRS/Université de Lyon UMR5191 – Équipes CÉDILLES et Axe transversal CCC (Cellule Corpus Complexes).

Directrice laboratoire : Sandra Teston-Bonnard (sandra.teston-bonnard@ens-lyon.fr)

Contact : **merci d'adresser au plus vite CV + lettre de motivation à** Denis Vigier (denis.vigier@ens-lyon.fr)

Durée du contrat : 11 mois
Date de début du contrat : Date de la prise de fonction

Lieu de travail : Ecole Normale Supérieure de Lyon, Bâtiment Recherche, 15 parvis René Descartes, 69007 Lyon.

Diplôme demandé : Master ou Doctorat

Contexte. Dans le cadre du Labex ASLAN, le laboratoire ICAR développe des bases de données textuelles destinées à l'étude de la langue française, dans des perspectives diachroniques et synchroniques. Ces bases comportent notamment des corpus textuels diachroniques (corpus Presto, XVI^{e} - XXI^{e} siècles), des corpus multimodaux (oral/transcription) et des corpus d'écrit médié par ordinateur (*tweets*). Ces bases continuent d'être enrichies, à l'occasion de projets débutés récemment (pour les *tweets*, projet ANR SoSweet, octobre 2015 − octobre 2019) ou terminés depuis peu (pour le français des XVI^{e} - XXI^{e} siècles, projet ANR-DFG PrESTo, avril 2013 – avril 2017 ; pour le français parlé, projet ANR Orféo, février 2013 − février 2017).

Mission. Le travail de l'ingénieur recruté s'inscrira principalement (75%) dans la suite du projet Presto, et secondairement (25%) sur des objectifs transversaux liés aux autres corpus du laboratoire (notamment, projets Orféo et SoSweet) ; dans ce cadre, l'ingénieur se mettra aussi à la disposition de l'unité pour répondre aux besoins éventuels qui émergeraient à partir de travaux menés dans les différentes équipes (CLAPI, ViSA...).

Les missions de l'ingénieur concernent plusieurs étapes du cycle de vie d'un corpus. Elles consisteront principalement : (1) à entretenir les bases constituées (maintenance, amélioration, diffusion, valorisation), (2) à veiller à leur intégration dans les outils utilisés par l'équipe, (3) à accompagner l'enrichissement (apport de textes nouveaux) de la base Presto. Il conviendra en outre de développer, en collaboration avec les membres de l'équipe, (4) des outils pour l'annotation automatique de ces données textuelles, généralement plus « bruitées » que l'écrit standard (écrits relevant d'états de langue anciens, oral transcrit, écrit médié par la machine), et (5) des outils pour l'accès et l'exploitation de ces données en ligne.

Dans le cas de la base Presto, l'ingénieur pourra s'appuyer sur les outils de traitement de corpus déjà développés dans le cadre du projet et qui doivent être enrichis sur plusieurs plans : couverture de nouveaux types de textes, amélioration de l'analyse en POS/lemmes existante, élaboration d'un nouveau niveau d'analyse en dépendances syntaxiques

Qualités personnelles : l'ingénieur devra faire preuve d'aptitudes relationnelles pour le travail en équipe, de qualités de rigueur scientifique et d'autonomie, d'esprit d'initiative

Compétences requises (**A**pplication, **M**aîtrise, **E**xpert)

	Compétences	A	M	E
Informatique	Développement Unix (Shell, logithèque GNU...)		×	
	Développement open-source (SVN, GIT, Javadoc, Doxygen...)	×		
	Programmation (Perl, Python, Java...) et conception logicielle (architectures n-*tiers* et MVC)			×
	Expressions régulières		×	
	Gestion des sites et développement Web (client et serveur) : PHP, JavaScript, CSS, Apache		×	
	Traitement XML: DOM, SAX, XPath, XSL		×	
	Langage SQL, SGBD MariaDB, ou à défaut MySQL		×	
	Développement d'IHM, ergonomie des interfaces	×		
TALN	Statistiques textuelles et R		×	
	Création de ressources et utilisation d'étiqueteurs pour l'analyse morpho-syntaxiques et en dépendances			×
	Traitement de l'écrit, en particulier « bruité »			×
	Outils d'exploration de corpus : CQP/CWB, TXM, Primestat, ScienQuest		×	
	XML-TEI		×	
	Gestion de projet en TALN			×
Langues	Français			×
	Anglais		×	