

Semaine de formation intensive NooJ
Max Silberztein, Université de Franche-Comté
Auditorium Dumézil
INALCO, 2 rue de Lille, 75007 Paris
8-12 janvier 2024

NooJ est un environnement de développement utilisé pour analyser des textes à l'aide de formalisations linguistiques : orthographe et typographie, morphologie flexionnelle, dérivationnelle et agglutinative, syntaxe locale, structurelle (en constituants) et de dépendance, grammaire transformationnelle et analyse sémantique. NooJ offre tous les outils formels de la hiérarchie de Chomsky-Schützenberger pour formaliser chaque type de phénomène : grammaires régulières (comme XFST), hors-contexte (comme GPSG), contextuelles (comme LFG) et non restreintes (comme HPSG), ainsi que des outils de développement pour créer, tester et partager des ressources linguistiques à large couverture.

NooJ est utilisé comme outil de linguistique de corpus et dans les humanités numériques pour effectuer des analyses statistiques sur des corpus historiques, littéraires, psychologiques ou socio-linguistiques en analyse de discours, cf. <http://atishs.univ-fcomte.fr>. Le moteur de NooJ a été intégré à de nombreuses applications de TAL : reconnaissance d'entités nommées, génération automatique de textes, informatique décisionnelle, interface en langue naturelle, moteur de recherche conceptuel, production automatique de paraphrases, web sémantique, traduction automatique, cf. plus de 500 références à la page nooj.univ-fcomte.fr/references.html.

NooJ est soutenu depuis 2013 par le programme Metashare de la Communauté européenne, est gratuit et open source et fonctionne sous Windows (.NET), Mac OSX, LINUX et Unix (JAVA). Il existe des modules de ressources linguistiques au format NooJ pour plus de 30 langues, cf. : nooj.univ-fcomte.fr/resources.html. Son nouveau moteur linguistique RA est en chantier sur GitLab, tourne en natif sur Linux, MacOS et Windows, dispose d'une WEB API, et sera utilisé pour la version WEB de NooJ en développement.

Ce stage s'adresse particulièrement aux étudiants de Master, doctorants et collègues universitaires intéressés par la formalisation linguistique, les outils d'exploration de corpus et les applications du TAL. Les étudiants en Master qui peuvent et désirent valider le stage auprès de leur université doivent me prévenir dès leur inscription. Le stage dure une semaine : les matins sont dédiés au cours ; pendant les après-midis, des chercheurs présenteront diverses applications et tutoriels de NooJ.

ATTENTION : le stage est gratuit mais les places sont limitées. Les inscriptions sont obligatoires : il faut s'inscrire en envoyant son nom, son statut (étudiant / doctorant / universitaire / autre) et le nom de son institution à max.silberztein@univ-fcomte.fr.

Lundi 8

9H-12H30 : **Corpus** : Limites des approches empiriques. Corpus, requêtes et concordances.

14H-15H : Prihantoro (Universitas Diponegoro, Indonésie) : SANTI-morf: A Morpheme Tagger for Indonesian.

15H-16H : Rafik Kassmi (Université de Kenitra, Maroc) : Implementation d'un correcteur orthographique de l'arabe.

Mardi 9

9H-12H30 : **Lexique & morphologie** : Unités Linguistiques Atomiques, dictionnaires et morphologie.

14H-15H : Masako Watabe (Université de Franche-Comté) : Un dictionnaire polylectal pour le rromani.

15H-16H : Asmaa Amzali (Université de Kenitra, Maroc) : Implémenter un lexique-grammaire des verbes psychologiques arabes.

Mercredi 10

9H-12H30 : **Syntaxe locale** : Requêtes syntaxiques, grammaires locales, annotation de textes, entités nommées.

14H-15H : Olena Saint-Joanis (Université de Franche-Comté) : Grammaires pour la levée d'ambiguïté de l'ukrainien.

15H-16H : Xavier Blanco Escoda (Universitat Autònoma de Barcelona), Rafael Garcia Pérez (Universidad Carlos III de Madrid) : L'expression de l'intensité en ancien français et en castillan ancien : lexiques et corpus.

Jeudi 11

9H-12H30 : **Syntaxe structurelle** : Grammaires hors-contexte, contraintes d'accord et arbres.

14H-15H : Mario Monteleone (Université de Salerne, Italie) : Analyse transformationnelle de l'italien.

15H-16H : Héla Fehri (MIRACL, Tunisie) : Un générateur de paraphrases pour la langue arabe.

Vendredi 12

9H-12H30 : **Sémantique** : Grammaires transformationnelles. WEB sémantique. Génération automatique de textes.

14H-15H : Maximiliano Duran (CRIT, Besançon) : Analyse transformationnelle du quechua.

15H-16H : Anabela Barreiro (INESC Lisbonne) : eSPERTo: an automatic paraphraser for Portuguese