Semaine de formation intensive NooJ
Max Silberztein, Université Marie et Louis Pasteur
INALCO, 2 rue de Lille, 75007 Paris
18-22 janvier 2027

NooJ est un environnement de développement utilisé pour décrire formellement les langues et effectuer des analyses linguistiques sur des textes à plusieurs niveaux : orthographe, 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. Le moteur de NooJ a été intégré à de nombreuses applications de TAL qui exigent fiabilité et précision : reconnaissance d'entités nommées, génération automatique de textes, informatique décisionnelle, interface en langue naturelle, moteur de recherche conceptuel, production automatique de paraphrases, annotation psychologique, 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 tourne en natif sur Linux, macOS et Windows et dispose d'une WEB API qui est utilisée en particulier pour la version WEB de NooJ https://webnooj.univ-fcomte.fr.

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 linguistiques du TAL. Les étudiants de Master qui ont besoin d'une note doivent me prévenir dès leur inscription. Le stage dure une semaine : les matins sont dédiés au cours ; pendant les après-midis, des chercheurs et doctorants présenteront diverses applications et organiseront des ateliers avec NooJ.

ATTENTION : les places sont limitées. Les inscriptions sont gratuites mais obligatoires : il faut s'inscrire en envoyant son nom, son statut (étudiant / doctorant / universitaire / autre) et le nom de son institution à max.silberztein@gmail.com.

Lundi
9h-12h30 : **Corpus** : Limites des approches empiriques (Deep Learning, LLM). Corpus, requêtes et concordances.
14h-15h : Agnieszka Kaliska (Poznan Univ., Pologne) Affinement du modèle de flexion verbale du polonais.
15h-16h : Morgane Bona-Pellissier (Univ. de Nanterre). Modéliser la morphologie du latin.
16h-17h : Olena Saint-Joanis (INALCO). Gestion des variantes non standard de l'ukrainien avec NooJ.

Mardi
9h-12h30 : **Lexique & morphologie** : Unités Linguistiques Atomiques, dictionnaires et morphologie.
14h-15h : Kristina Kocijan (Zagreb Univ., Croatie). How to turn Croatian texts into linguistic resources using NooJ.
15h-16h : Maria Piá di Buono (Univ. di Napoli, Italie). From Linguistic Resources to Medical Text Simplification.

Mercredi
9h-12h30 : **Syntaxe locale** : Requêtes syntaxiques, grammaires locales, annotation de textes, entités nommées.
14h-15h : Marzieh Rabiei (Univ. Marie et Louis Pasteur). Formalisation morphologique des noms persans dans NooJ
15h-16h : Fabio Meroni (Lattice, HéLiCéO, CNRS, ENS, PSL, USN). Etiqueter des prédicats complexes du mwotlap

Jeudi
9h-12h30 : **Syntaxe structurelle** : Grammaires hors-contexte, contraintes d'accord et arbres.
14h-15h : Masako Watabe (INALCO) : Désambiguïsation des phrases AN (adjectif-nom) en rromani.
15h-16h : Leandro Kintal, Sandra Mendizaba (UNR, Argentine). A corrector to help use Spanish inclusive language.
16h-17h : Ana Gigon, Lucía Soda (UNR, Argentine). A new pedagogical approach to youth literacy in Spanish.

Vendredi
9h-12h30 : **Sémantique** : Grammaires transformationnelles. WEB sémantique. Génération automatique de textes.
14h-14h40 : Maximiliano Duran (Univ. Marie et Louis Pasteur). Annotations structurées de phrases quechua
14h40-15h20 : Ekgoan Susanti Utami (Univ. Diponegoro, Indonésie). NooJ POS tagger for Javanese V.0.0.1
15h20-16h : Divna Petkivic (INALCO). Étude linguistique préliminaire de l'accentuation des langues slaves.