Stage intensif NooJ à l'INALCO, 1-5 février 2010

INALCO, 49 bis avenue de la Belle Gabrielle,
75012 Paris (RER Nogent-Sur-Marne)

NooJ est un environnement de développement linguistique qui propose des méthodologies et des outils pour formaliser les langues en construisant des ressources linguistiques, tester ces ressources linguistiques en les appliquant à des textes de taille importante, et gérer, accumuler et combiner un grand nombre de ressources.

NooJ permet de formaliser cinq niveaux de phénomènes linguistiques : orthographe, morphologie, lexique, syntaxe et sémantique. Pour chacun de ces niveaux, NooJ propose une méthodologie, un ou plusieurs formalismes adaptés, des outils-logiciels de développement et un ou plusieurs analyseurs automatiques de textes. Par exemple, au niveau morphologique, NooJ fournit deux formalismes pour décrire la flexion et la dérivation, un formalisme pour décrire la morphologie lexicale (par ex. pour représenter les familles de mots) et un formalisme pour entrer des règles de morphologie productive (par ex. pour formaliser la création de néologismes).

Les outils et formalismes de NooJ sont tous compatibles entre eux de façon ascendante, et sont graduellement plus puissants au fur et à mesure qu'on monte dans la hiérarchie linguistique. Par exemple, le niveau orthographique utilise des machines à états finis ; le niveau syntaxique utilise des grammaires hors contexte ; le niveau sémantique utilise des réseaux de transition augmentés (Augmented Transition Networks ou ATN) dont la puissance est équivalente à celle d'une machine de Turing. Cette approche « multiple » apporte de nombreux avantages pour les travaux de description linguistique car les linguistes disposent d'outils de développement et d'analyse adaptés à chaque niveau de formalisation. Par ailleurs, des phénomènes très spécifiques à des langues très différentes, comme par exemple la variation orthographique (massive) en chinois, le traitement des voyelles absentes en arabe, la morphologie massive en hongrois etc. sont traités avec des outils spécifiques. NooJ fournit un environnement unifié à l'intérieur duquel ces outils spécialisés communiquent entre eux grâce à une structure d'annotations (« Text Annotation Structure » ou TAS). La TAS permet de formaliser des phénomènes à cheval sur plusieurs niveaux linguistiques.

INSCRIPTIONS

Le stage est gratuit, mais les inscriptions sont obligatoires et les places sont limitées. Chaque participant doit venir avec son ordinateur portable sur lequel NooJ a déjà été installé.

Pour s'inscrire, envoyez un message à : max.silberztein@univ-fcomte.fr. en spécifiant votre nom, votre statut (étudiant / doctorant / M1 / M2, chercheur, enseignant, etc.), votre institution (laboratoire, université, etc) ainsi que votre domaine d'intérêt.

PROGRAMME

Deux séances de cours chaque matin ; deux présentations d'utilisateurs l'après-midi.

Lundi 1er février

Cours 9H-12H30 : Traitement de corpus : ouvrir un texte, gérer des corpus, la norme XML, lancer des requêtes et construire des concordances

Présentation 14H-14H45 : Sandrine Traïdia, SEDYL : **Grammaire des dates en kurde**. *Je présenterai des grammaires syntaxiques de dates réalisées sous forme de graphes. Ces grammaires permettront d'annoter automatiquement les textes pour rechercher des compléments circonstanciels de temps.*

Présentation 14H45-16H : Huei-Chi Lin, Université de Franche-Comté : **Traitement des homonymes et des variantes graphiques en chinois**. *En chinois contemporain, il existe plusieurs types d'homonymie et de variantes graphiques : les variantes monosyllabiques, les variantes de mots simples polysyllabiques et les variantes des mots composés. Dans cette présentation, nous présenterons les cas de correspondance entre les sons, les morphèmes et les caractères graphiques. Ensuite, nous décrirons leur formalisation lors du développement du module chinois dans NooJ. En conclusion, nous montrerons le résultat d'analyse.*

Mardi 2 février

Cours 9H-12H30 : Morphologie flexionnelle (ex. conjugaison des verbes), morphologie dérivationnelle (ex. nominalisation d'un verbe), morphologie productive (ex. néologismes).

Présentation 14H-14H45 : Mathieu Roy, INALCO : **Reconnaissance automatique des chaînes d'accords nominaux en kiswahili**. *Le kiswahili est une langue bantu appartenant au groupe G62 selon la classification de Malcom Guthrie. Cette langue se caractérise principalement par le fait que chacun de ses nominaux appartient à une classe. Les classes se repèrent par la chaine d'accords réguliers qu'un nominal appartenant à une classe déterminée pilote sous la forme de préfixes sur les différentes parties du discours en relation avec lui (verbes, adjectifs). La détection des chaines d'accords nominaux permet un premier niveau d'analyse grammaticale et de reconnaissance des diverses catégories d'un énoncé. Cette identification pourrait par la suite être le support de différentes transformations, comme le passage à titre d'exercice de cours du singulier au pluriel. L'analyse sera appliqué à un roman écrit en kiswahili standard, standard qui s'appuie sur les variétés tanzaniennes du kiswahili et qui est soutenu par l'Etat tanzanien.*

Présentation 14H45-15H30 : Yaakov Bentolila, Univ. Ben Gurion : **morphologie de l'Hébreu**. *Le mot hébraïque se constitue par l'association d'une racine de trois ou quatre lettres et d'une configuration morphologique. À l'instar d'autres langues sémitiques, l'hébreu présente une morphologie riche car une grande partie de l'information est véhiculée par des morphèmes liés, préfixés ou suffixés à une base ou à un radical: des prépositions, des conjonctions, des relatifs, la détermination, etc. Les verbes se conjuguent en sept configurations différentes, et dans chacune d'elles les informations de personne, de nombre, de genre, de temps, et même de complément se manifestent en affixes. Pour les noms on compte plus de cent configurations. Le nom aussi se décline, non seulement pour rendre le nombre et le genre, mais aussi pour signifier la possession, l'appartenance ou sa relation avec un autre nom. Nous présenterons les grammaires morphologiques du module hébreu de NooJ. Nous comptons des grammaires pour la déclinaison de noms ou la conjugaison de verbes, ainsi que pour la tokénisation. En général, on applique les grammaires de déclinaison pour les suffixes et les grammaires de tokénisation pour les préfixes.*

Mercredi 3 février

Cours 9H-12H30 : Dictionnaires NooJ : mots simples, mots composés et expressions figées

Présentation & tutoriel 14H-15H15 : Odile Piton, Université Paris 1 : **Morphologie avec NooJ, application à l'albanais**. *Nous montrons comment reconnaître les mots albanais, qui est une langue à déclinaison. Nous donnerons des exemples de déclinaison des noms, de la conjugaison des verbes et de familles de mots dérivés. La morphologie productive est particulièrement intéressante pour reconnaître des mots albanais car il existe des "listes ouvertes" qui ne peuvent être représentées dans un dictionnaire. Nous verrons comment NooJ nous permet d'effectuer de telles reconnaissances, et nous apprendrons à combiner dictionnaires et graphes.*

Présentation & tutoriel 15H15-16H30 : Slim Mesfar, Institut Supérieur d'Informatique, Tunisie : **Gestion et analyse de corpus**. *Un corpus est un ensemble de fichiers textes qui partagent les mêmes langue, format et codage. NooJ peut importer des fichiers XML. Dans un premier temps, nous utilisons diverses fonctionnalités pour analyser un corpus, construire des concordances, l'annoter et finalement l'exporter. Ensuite, nous nous concentrons sur la manipulation de textes au format XML. Après l'importation de ces textes, nous ajoutons des annotations au corpus construit via l'application d'expressions rationnelles et de grammaires locales pour, enfin, l'exporter dans son format d'origine.*

Jeudi 4 février

Cours 9H-12H30 : Syntaxe : grammaires locales, groupe nominal, analyse structurelle

Présentation & tutoriel 14H-15H30 : Anaïd Donabédian, INALCO : **L'élaboration des ressources linguistiques pour le module arménien de NooJ : défis et méthode**. *Nous décrirons le développement des ressources utilisées pour lemmatiser le corpus arménien occidental avec NooJ, qui répond aux spécificités linguistiques de l'arménien : alphabet propre, signes de ponctuation au fonctionnement spécifique, morphologie nominale agglutinante, morphologie verbale flexionnelle, nominalisations étendues, dérivation proliférante. Une grande partie du travail a été réalisée avec peu de moyens humains et matériels, et cette présentation pourra donner des points de repère à ceux qui souhaitent entamer la réalisation d'un module dans une langue non encore traitée avec NooJ.*

Vendredi 5 février

Cours 9H-12H30 : Sémantique : entités nommées, construction automatique de paraphrases

Présentation 14H-14H45 : Denis Le Pesant, Université Paris 10 : **Analyse syntaxique d'une classe de verbes de communication avec NooJ**. *Après avoir présenté trois très grands dictionnaires électroniques de Jean Dubois, nous faisons une analyse linguistique détaillée d'une classe de verbes de communication. Puis nous présentons le dictionnaire NooJ des verbes français, ainsi que les grammaires NooJ associées. Nous présentons la tâche d'annotation syntaxique et sémantique de grands corpus, et analysons les résultats.*

Présentation & tutoriel 14H45-16H : Mei Wu, Université de Franche-Comté : Traduction automatique. **Traduction français-chinois : Un traducteur automatique français-chinois pour les groupes nominaux simples**. *Nous avons procédé à l'analyse et la traduction de 612 groupes nominaux français. Nous discuterons des problèmes linguistiques et techniques rencontrés : la traduction polysémique ; la construction de classes sémantiques ; l'ordre des mots ; les contraintes lexicales.* **Tutoriel :** *Créer un dictionnaire bilingue et les fichiers de propriétés. Transformer une phrase active en phrase passive (EN). Traduire les expressions de dates (EN-FR). Traduire une phrase simple française en anglais (FR-EN).*