Workshop NooJ 26-30 janvier 2009-01-06
INALCO, 49 bis avenue de la Belle Gabrielle, 75012 Paris (RER Nogent-Sur-Marne)

Cours le matin : 9H-10H30 et 11H-12H30 ; présentations l'après-midi : 14h-15H30 et 16H-17h30.

lundi 26 janvier matin

Gestion des textes et corpus. Requêtes et concordances. Expressions rationnelles. Grammaires.

lundi 26 janvier après-midi

Gestion et analyse de corpus, Slim Mesfar, Institut Supérieur d'Informatique, Tunis

Un corpus est un ensemble de fichiers textes qui partagent certaines propriétés telles que la langue, le format et le codage. NooJ emploie son propre format de fichier pour le traitement de corpus. En l'occurrence, il utilise les fichiers de type ".noc" où il stocke tous les textes appartenant au corpus, leurs informations structurelles (par exemple, les unités de texte), différents indices linguistiques ainsi que la structure d'annotation interne de chacun d'entre-eux.

Toutes les fonctionnalités disponibles pour le traitement de textes sont aussi valables au niveau des corpus. En effet, nous pouvons procéder à un ensemble de mesures statistiques (les caractères, les tokens, les digrammes, les annotations, etc.). En outre, nous pouvons construire des concordances simples et complexes sur l'ensemble de tous les textes appartenant au corpus. Chaque ligne de la concordance obtenue exhibe, successivement, le nom du texte source, le contexte avant, la séquence retenue et le contexte après.

Nous notons que, dans NooJ, nous disposons aussi de la possibilité d'exporter, sous format XML, les textes et les corpus annotés à l'issue d'une phase d'analyse linguistique. Nous pouvons, également, importer des textes préalablement annotés, à l'aide d'outils extérieurs, pour construire des corpus au format NooJ.

Les variations en chinois, Huei-Chi Lin, Université de Franche-Comté.

Dans cette présentation, nous tentons de déterminer la formalisation des variations en chinois, dans l'objectif du traitement automatique des langues. De cette perspective, nous sommes amenés à engager les démarches suivantes :

-- Standardiser les variantes uni-graphiques et multi-graphiques ;
-- Mettre en œuvre les variations concernant le vocabulaire chinois, par exemple, les formes répétitives des adjectifs, des adverbes, des verbes, etc.

Pour mieux réaliser ces deux démarches, nous introduirons tout d'abord les systèmes de codage des caractères, et puis les variations uni-graphiques et multi-graphiques ainsi que leur standardisation. Par la suite, nous monterons les variétés dans le vocabulaire chinois, et comment nous les traitons en servant les formalismes proposés dans NooJ.

mardi 27 janvier matin

Décrire une langue avec NooJ. Les Unités Linguistiques Atomiques. Dictionnaires. Morphologie flexionnelle et dérivationnelle.

mardi 27 janvier après-midi

Grammaire des dates en chinois, espagnol, japonais, russe, Adriana Amaya, Murielle Fabre, Claire Olivier, Estelle Delavennat, INALCO

Nous présentons un ensemble de grammaires locales qui représentent des compléments de date et qui peuvent être utilisées pour annoter automatiquement les textes.

Morphologie de l'albanais, Odile Piton, Université Paris 1.

Nous allons montrer comment NooJ permet de reconnaître les mots albanais. Précisons que l'albanais est une langue à déclinaison.

-- La morphologie flexionnelle sera traitée par des exemples concernant la déclinaison des noms et la conjugaison.
-- Nous utiliserons la morphologie dérivationnelle pour reconnaître des familles de mots.
-- La morphologie productive est particulièrement intéressante pour reconnaître des mots albanais car il existe des "listes ouvertes", qui ne peuvent être listées intégralement dans un dictionnaire, et qui nécessitent donc des outils permettant leur reconnaissance dynamique.

Nous verrons comment les graphes de Nooj nous permettent d'effectuer de telles reconnaissances.

mercredi 28 janvier matin

Editeur graphique de grammaires. Graphes imbriqués. Contrat et débogueur.

mercredi 28 janvier après-midi

Entités nommées en arabe, Slim Mesfar, Institut Supérieur de l'Informatique, Tunis

Les entités nommées incluent les expressions de noms propres (noms de personnes, lieux, organisations, etc.), les expressions temporelles (dates et heures) et les expressions numériques (expressions monétaires et pourcentages). La reconnaissance des entités nommées en arabe posent un ensemble de problèmes tels que l'absence d'une majuscule à la tête des noms propres, l'absence de voyelles dans les écrits courants, les problèmes de délimitation et de polysémie, etc. Nous préconisons une approche de reconnaissance à base de règles écrites manuellement. Ces règles sont fondées sur des preuves internes et externes pour l'identification et la catégorisation des entités nommées où :

*-- **Les preuves internes** sont fournies par les constituants de l'entité nommée. Elles peuvent être contenues dans des listes de mots déclencheurs ou de noms propres.*
*-- **Les preuves externes** sont fournies par le contexte dans lequel une entité nommée apparaît. Elles se basent sur les relations syntaxiques au sein d'une phrase pour attribuer la catégorie de l'entité*

retenue. Cette catégorisation utilise les informations morpho-syntaxiques fournies suite à une phase d'analyse morphologique.

L'ensemble des règles est représenté à l'aide de grammaires syntaxiques NooJ. Nous présenterons un système pour le repérage d'entités nommées ainsi que le typage de celles-ci. Un ordre de passage est attribué à l'ensemble des grammaires ainsi construites afin de pouvoir utiliser les entités reconnues.

Analysis of the psychological content of narrative texts in Hungarian, Orsolya Vincze, Hungarian Academy of Science.

Narrative psychological content analysis connects psychological contents to linguistic-structural features of narratives. One important condition for a research based on narrative psychological content analysis is to develop computer programs, which can relatively safely identify narrative categories carrying psychological content in large databases. Collaboration between the Linguistics and the Psychology Institutes of the Hungarian Academy of Sciences have resulted in a set of linguistic categorization algorithms (e.g. activity-passivity, emotions, cognitive states, intentionality) using Nooj. In my presentation I would like to introduce some of these algorithms: the process of their developing, application and some unsolved problems for further discussion.

jeudi 29 janvier matin

Transducteurs et annotations. Structure d'annotations du texte (TAS). Exportation XML.

jeudi 29 janvier après-midi

Arbre syntaxique. Transformations syntaxiques. Analyse sémantique. Traduction automatique.

vendredi 30 janvier matin

Analyse morphologique de l'hébreu, Yaakov Bentolila, Univ. du Negev.

Tutoriel : nous analysons un petit texte hébraïque, transcrit en lettres latines. Pour ce faire nous préparerons un dictionnaire, des grammaires flexionnelles simples visant à générer les différentes formes que prennent deux substantifs, un verbe au présent et un adjectif. Nous allons préparer une grammaire morphologique pour la tokénisation d'un substantif préfixé par une préposition.

Présentation : nous considérerons différents problèmes caractéristiques du traitement morphologique d'un texte hébraïque, notamment la présence de digrammes, (p.ex. dans la lettre SHIN) ou par des conditions phonologiques ou morphologiques (présence du DAGUESH). Nous verrons comment certains opérateurs NooJ ont été adaptés ou créés pour résoudre ces cas. Ensuite nous examinerons la grammaire flexionnelle du verbe hébraïque et en discuterons les particularités.

Analyse syntaxique du hongrois, Kata Gabor, Académie des Sciences de Budapest

En plus des requêtes qui s'appuient sur l'analyse lexicale et morphologique, le moteur de NooJ est également exploitable pour l'analyse et l'annotation syntaxiques. La robustesse ainsi que la conformité XML font du NooJ un outil favorable pour l'analyse de corpus à tous les niveaux de la description

linguistique. Cette présentation se concentrera sur les divers aspects de la construction d'un analyseur syntaxique avec NooJ, y compris l'analyse de surface et l'analyse de dépendances.

La première partie se réalisera sous forme d'un tutoriel "Comment construire un analyseur syntaxique avec NooJ?", et s'étendra sur les problématiques suivantes: Comment créer des grammaires syntaxiques? Construction de graphes, exploitation de traits lexicaux, teste de grammaires : concordances, debugging, Modélisation de phénomènes locaux par des grammaires locales, Comment exploiter la vitesse et l'efficacité de la technologie NooJ? Fonctionnalités avancées: variables et contraintes lexicales pour l'annotation des dépendances à distance.

Le tutoriel sera suivi de la présentation du module syntaxique hongrois de NooJ: du dictionnaire syntaxique à la cascade des grammaires locales et au-delà : l'analyse des dépendances verbales.

vendredi 30 janvier après-midi

Syntaxe : requêtes syntaxiques et grammaires locales, Christine Fay-Varnier, LORIA.

L'équipe TALARIS du LORIA a participé à la campagne d'évaluation PASSAGE consistant à fournir pour chacune des phrases provenant d'un corpus de référence, d'une part l'ensemble des constituants et, d'autre part, l'ensemble des relations. Nous présentons dans cet article l'approche que nous avons suivie pour la recherche des constituants en s'appuyant sur des automates et plus particulièrement sur des FST. Afin de limiter les erreurs d'analyse, nous nous sommes efforcés, d'une part, de réduire les ambiguïtés au plus tôt en affinant certaines catégories utilisées, et d'autre part, de prendre en compte de façon plus significative le contexte dans l'application des règles. Les résultats obtenus sur le corpus de référence sont encourageants et permettent de valider la méthodologie mise en œuvre.

Tutoriel : nous présentons la mise en œuvre d'une analyse syntaxique. Sur un exemple réduit, nous montrons comment créer une grammaire, élargir cette grammaire en prenant en compte le contexte, comment analyser un texte à partir de cette grammaire et récupérer les résultats de l'analyse.

Traduction automatique, Mei Wu, Université de Franche-Comté.

Tutoriel : Créer les dictionnaires bilingues. Les fichiers de propriétés. Grammaires morphologiques et syntaxiques? Traduire les groupes nominaux anglais et l'expression de dates. Traduction d'une phrase simple (Fr-En).

Présentation : traduction français-chinois (40'): Nous décrivons un traducteur automatique qui "paraphrase" les groupes nominaux français en produisant leur traduction en chinois. Basé sur le corpus de «La mare au diable (1846) » de George Sand, nous avons procédé à l'analyse et la traduction de 612 groupes nominaux simples français. Nous discuterons des problèmes linguistiques et techniques que nous avons rencontrés : La traduction polysémique : comment construire une lexicographie conventionnelle pour traiter les unités polysémiques ? La construction de classes sémantiques : comment donner les différentes classes sémantiques aux entrées lexicales ? L'ordre des mots : comment ranger l'ordre des mots dans la traduction ? Les contraintes lexicales : comment contrôler l'application de dictionnaires avec les contraintes lexicales ? Comment ajouter les éléments linguistiques supplémentaires dans la langue-cible ?