PEPS INS2I – INSHS 2011

A déposer au plus tard le 4 avril 2011 à minuit
Le nom du document attaché (**PDF uniquement**) devra impérativement se présenter sous la forme " **PEPS_NOM_prénom.pdf** "
(Le nom et le prénom à indiquer sont ceux du coordinateur du projet)

Titre long du Projet (maximum 150 caractères) : Modélisation linguistique et computationnelle des chaînes de coréférence dans des textes en français médiéval et contemporain

Titre court du Projet : MC4 (Modélisation Contrastive et Computationnelle des Chaînes de Coréférence)

Mots clés (choix libre) : référence, détermination, coréférence, morphosyntaxe, sémantique, grammaticalisation, repérage automatique

Résumé du Projet (maximum 10 lignes) :

Ce projet porte sur la coréférence dans des textes narratifs en français médiéval et en français contemporain, avec comme objectif une modélisation à la fois linguistique et computationnelle, destinée à améliorer les systèmes actuels de traitement automatique du langage, notamment en extraction d'information. L'approche met en avant, outre les expressions référentielles déjà étudiées par ailleurs, les indices morphosyntaxiques qui, sans référer, participent à la coréférence : appositions, constructions pronominales, sujets zéro des infinitifs et des participiales, etc. L'aspect contrastif prend ainsi tout son sens : on analysera en priorité les éléments qui fonctionnent comme expression référentielle en français médiéval et comme indice coréférentiel en français contemporain, et inversement. Expressions et indices constituant deux types de maillons de chaînes de coréférence, avec deux niveaux de contribution, l'approche computationnelle tiendra compte de cet aspect et ouvrira la voie du repérage et de la construction automatique de chaînes multi-niveau.

Porteur de projet (civilité, prénom et nom)	Code et Nom du laboratoire (en toutes lettres)	Courriel	Code postal et ville
Mr. Frédéric Landragin	UMR 8094, LaTTICe	frederic.landragin@ens.fr	92120 Montrouge
Equipes participantes (civilité, prénom et nom)	Code et Nom du/des laboratoire(s) (en toutes lettres)	Courriel	Code postal et ville
Mr. Benjamin Fagard	UMR 8094, LaTTICe	benjamin.fagard@ens.fr	92120 Montrouge
Mr. Thierry Poibeau	UMR 8094, LaTTICe	thierry.poibeau@ens.fr	92120 Montrouge
Mr. Bernard Victorri	UMR 8094, LaTTICe	bernard.victorri@ens.fr	92120 Montrouge
Mme. Sophie Prévost	UMR 8094, LaTTICe	sophie.prevost@ens.fr	92120 Montrouge
Mme. Céline Guillot	UMR 5191, ICAR	celine.guillot@ens-lyon.fr	69342 Lyon
Mme. Amalia Todirascu	EA 1339, LiLPa	todiras@umb.u-strasbg.fr	67084 Strasbourg
Mme. Cath. Schnedecker	EA 1339, LiLPa	cschnede@unistra.fr	67084 Strasbourg
Mme. Hélène Manuélian	UMR 7187, LDI	manuelia@u-cergy.fr	95011 Cergy

Moyens demandés en Euros HT :

Durée du projet :	Equipement :	Fonctionnement :	**Total :**
12 mois ☐ 24 mois ☒	2000 €	13000 €	15000 €

Avis du directeur de laboratoire :
Très favorable, le projet soumis fait suite à un groupe de travail pluridisciplinaire très dynamique, animé par le porteur. Les travaux déjà engagés sont parfaitement en adéquation avec le quadriennal de notre équipe. Les résultats obtenus sont déjà robustes et très encourageants pour le bon déroulement du projet soumis. – Michel Charolles, directeur de l'UMR 8094 LaTTICe.

PEPS INS2I – INSHS 2011

LE PROJET

Exposé scientifique (2 pages maximum) : mettez en valeur le caractère exploratoire du projet, l'originalité des approches et la prise de risque.

Suite aux activités d'un groupe de travail actif depuis deux ans, et dans l'optique de préparer une soumission de projet à plus grande échelle (projet ANR ou projet européen, soumission envisagée pour 2012 ou 2013), ce projet PEPS (thème « traitement automatique du langage ») s'intéresse à la référence et à la coréférence dans des textes écrits, en français médiéval et en français contemporain, avec des objectifs à la fois théoriques et pratiques qui regroupent divers chercheurs en linguistique et en informatique.

1. D'un point de vue linguistique, il s'agit de modéliser les éléments d'une chaîne de coréférence, en tenant compte non seulement des expressions référentielles (noms propres, syntagmes nominaux, syntagmes sans nom, pronoms personnels, démonstratifs, adverbiaux, possessifs, etc.) dont le rôle est de porter l'attention du destinataire sur un référent identifiable, mais aussi des expressions et indices qui, sans référer, rappellent ou évoquent dans l'esprit du destinataire l'existence d'un référent (appositions, constructions pronominales, constructions attributives, etc.). En complément des travaux déjà effectués sur la coréférence (Corblin 1995 ; Schnedecker 1997 ; Charolles 2002), ce projet a pour objectif d'étudier et de modéliser la contribution de ces expressions et indices aux chaînes de coréférence, en partant du principe que tous les maillons d'une chaîne de coréférence n'ont pas la même importance. C'est dans ce principe « multi-niveau » et dans l'étude des indices coréférentiels que repose le caractère exploratoire du projet pour son versant linguistique.

Dans un même ordre d'idée, le projet s'intéressera aux maillons qui n'ont pas de trace linguistique marquée (les sujets zéro des infinitifs et des participiales, notamment), du fait de phénomènes d'ellipse ou de grammaticalisation. Pour ce faire, l'approche contrastive, qui met en rapport français médiéval et français contemporain, prend tout son sens : elle permettra de mieux appréhender ces phénomènes et de mieux les prendre en compte dans les modélisations.

A partir de ces éléments, une annotation de corpus (sur un ensemble de textes narratifs courts) sera mise en œuvre : elle consistera à repérer manuellement les expressions et indices des chaînes de coréférence correspondant aux différents personnages en présence. Cette phase d'annotation a plusieurs buts. Il s'agit premièrement de tester la procédure de détection des éléments de coréférence, ce qui passe par l'écriture d'un guide d'annotation incluant un ensemble de tests linguistiques. Il s'agit deuxièmement de construire et annoter les chaînes de coréférence de manière à étudier leurs propriétés : longueur, nature du premier maillon, typologie des maillons, structure de la chaîne, part des anaphores pronominales, déterminations successives des expressions, etc. Il s'agit troisièmement de mettre en perspective les analyses ainsi effectuées par rapport aux théories linguistiques classiques. On pensera par exemple aux rôles du nom propre, aux indices de reprise d'une chaîne de coréférence (Schnedecker 1997), ou encore aux différences entre déterminant défini et déterminant démonstratif : en analysant les apparitions de ces deux types de déterminant sur l'ensemble des textes annotés, on testera des hypothèses linguistiques portant notamment sur le caractère classifiant du défini (qualifiant pour le démonstratif) et la fonction de re-dénomination du défini (re-prédication pour le démonstratif).

Le corpus obtenu sera scindé en deux parties. La partie en français médiéval viendra enrichir les corpus électroniques de manuscrits anciens existants (BFM, CoRPTeF, SRCMF), en ajoutant un niveau d'annotation qui n'existe pas pour l'instant : l'annotation des références (niveau sémantique). La partie en français contemporain servira de corpus de test et d'entraînement pour les aspects TAL (Traitement Automatique du Langage).

2. D'un point de vue informatique, il s'agit d'une part d'adapter les outils d'annotation existants à la gestion des chaînes de coréférence, et d'autre part d'explorer la voie de la détection automatique de chaînes de coréférence. Pour le premier point, des outils tels que MMAX 2 (Müller & Strube 2006), GLOZZ 0.9.9 (Widlöcher & Mathet 2009) ou ANALEC 0.6 (Victorri 2010) sont dotés de fonctionnalités d'annotation et de visualisation, mais ne comportent pas les fonctionnalités de représentation et d'analyse que l'on attendrait d'eux pour ce qui concerne les chaînes de coréférence : visualisation des différentes chaînes d'un texte sous la forme de graphes ; mise en relief de spécificités morphosyntaxiques ou sémantiques d'une chaîne ; identification automatique de la structure d'une chaîne (en utilisant par exemple un système de motifs, cf. Mellet & Longrée 2009) ; calcul d'indices numériques à partir des éléments annotés (calcul de saillance, notamment). Ce projet vise à satisfaire ces

besoins via le développement d'une nouvelle version du logiciel ANALEC, qui joue le rôle de plateforme d'analyse de textes écrits, plateforme intégrant divers modules, non seulement de visualisation, mais aussi d'analyse.

Pour le second point, il existe déjà des systèmes d'identification de chaînes de coréférence (Longo & Todirascu 2010), et par là même beaucoup de méthodes d'annotation de la coréférence, notamment des méthodes adaptées en vue d'un traitement automatique ultérieur (Van Deempter & Kibble 2000). Néanmoins, ces méthodes et systèmes se restreignent aux expressions référentielles, voire à certaines expressions référentielles (celles qui sont détectables facilement). Plus que cela, ce projet vise à explorer la voie du traitement automatique non seulement pour les expressions référentielles mais aussi pour les indices qui font l'objet d'une étude dans le versant linguistique. C'est là que se trouvent le caractère exploratoire du projet pour son versant informatique et la prise de risque. L'objectif est double : premièrement, et c'est l'objectif à long terme, il s'agit de détecter automatiquement les coréférences à l'aide de tous les indices qui pourront être formalisés (lexique, morphosyntaxe, sémantique), en explorant les possibilités de mise en œuvre de modèles à base d'apprentissage ; deuxièmement, et c'est un objectif moins risqué, il s'agit d'envisager le pré-repérage des références et des coréférences pour faciliter la tâche d'un annotateur humain : pré-délimitation des syntagmes concernés, pré-remplissage de certains champs, tout ceci sans nécessité de justesse absolue. Des méthodes d'apprentissage automatique seront également testées sur ce point, de manière à ouvrir la voie de l'annotation automatique de la coréférence sur de grands corpus. Les développements réalisés constitueront une maquette qui sera intégrée à la plateforme ANALEC.

Le projet fournira ainsi : une modélisation linguistique et computationnelle de la coréférence ; un ensemble d'études linguistiques (études contrastives, études sur la détermination) ; un corpus et son manuel d'annotation ; une maquette TAL ; une plateforme logicielle intégrant les modules de visualisation, d'analyse et de TAL autour de la coréférence, cette plateforme étant destinée à la fois aux linguistes qui s'intéressent aux phénomènes de référence, et aux informaticiens qui développent des algorithmes de résolution d'anaphores et de coréférences. Cette plateforme aura ainsi le rôle de nouveau logiciel à l'interface des sciences informatiques et des sciences humaines et sociales.

Utilisation principale des moyens demandés :

- frais de mission dans le but de rechercher des experts français (Strasbourg, Toulouse, Lyon, Nancy) et de fédérer une communauté de chercheurs en linguistique et en TAL autour de cette thématique de coréférence, à la fois pour la réalisation de ce projet et pour sa pérennisation via la soumission dans un deuxième temps d'un projet ANR ;
- frais d'organisation d'un colloque sur la coréférence ;
- frais de fonctionnement et d'équipement.

Références bibliographiques :

Charolles M. (2002). *La référence et les expressions référentielles en français*. Paris : Ophrys.

Corblin F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Rennes : Presses Universitaires de Rennes.

Landragin F. (2004). « Saillance physique et saillance cognitive », *Cognition, Représentation, Langage (CORELA)* 2(2) : http://edel.univ-poitiers.fr/corela.

Longo L., Todirascu A. (2010). RefGen : un module d'identification des chaînes de référence dépendant du genre textuel, *Actes de la 17e Conférence sur le Traitement Automatique des Langues Naturelles*.

Mellet S., Longrée D. (2009). « Syntactical 'Motifs' and Textual Structures », *Belgian Journal of Linguistics* 23: 161-173.

Müller C., Strube M. (2006). Multi-Level Annotation of Linguistic Data with MMAX2. In: Braun S., Kohn K., Mukherjee J. (Eds.) *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Frankfurt: Peter Lang.

Schnedecker C. (1997). *Nom propre et chaîne de référence*. Paris : Klincksieck.

Van Deempter K., Kibble R. (2000). « On Coreferring: Coreference in MUC and related annotation schemes », *Computational Linguistics*, vol. 26, n° 4 : 629-637.

Victorri B. (2010). « Analec : logiciel d'annotation et d'analyse de corpus écrits », logiciel téléchargeable sur : http://www.lattice.cnrs.fr/-Analec-.

Widlöcher A., Mathet Y. (2009). « La plate-forme Glozz : environnement d'annotation et d'exploration de corpus », 16e Conférence sur le Traitement Automatique des Langues Naturelles, Senlis.