Annotation collaborative de corpus

9 mars 2012

Salle Internationale – MSH-Lorraine
91 avenue de la Libération
54000 Nancy

9h30 -9h50	*Accueil et présentation de la journée*
9h50 -10h30	*Annotation manuelle de corpus : mais de quoi parle-t-on ?* **Karën Fort - Inist-CNRS et LIPN UMR 7030 CNRS /Université Paris Nord**
10h30 – 10h50	*Pause*
10h50 – 11h30	*Exploitation éditoriale du balisage scientifique* **Pierre-Yves Buard – GREYC UMR 6072, équipe Document numérique Langue Usages /Pôle Document numérique, MRSH Caen**
11h30 – 12h10	*Projet ProDescartes (à confirmer)* **Vincent Carraud - GREYC UMR 6072 CNRS/Université de Caen/Ecole nationale supérieure d'ingénieur de Caen**
12h10 – 13h30	*Pause déjeuner*
13h30 – 13h50	**Chrestien de Lihus dans l'hypertexte : retour d'expérience sur une expérimentation d'annotation de ressources textuelles dans le réseau Wicri** **Thierry Daunois - Direction des Partenariats/Université de Lorraine**
13h50 – 14h10	*Plateforme d'annotation terminologique semi-collaborative* **Evelyne Jacquey – Atilf UMR 7118 CNRS/Université de Lorraine**
14h10 – 14h30	*Corpus littéraire : critique de la cité de Dieu* **Béatrice Stumpf – Atilf UMR 7118 CNRS/Université de Lorraine**
14h30 – 14h50	*Présentation d'une expérience d'annotation en parties du discours de corpus oraux : évaluation qualitative* **Christophe Benzitoun - Atilf UMR 7118 CNRS/Université de Lorraine**
14h50 - 15h10	*Les encyclopédies médiévales et* **les digital humanities** *: l'évolution du programme* **Sourcencyme** **Eduard Frunzeanu - Centre de Médiévistique Jean-Schneider - ERL 7229 Université de Lorraine** **Philippe Pons - Atilf UMR 7118 CNRS Université de Lorraine**
15h10 - 15h30	*Pause*
15h30 - 16h00	*Questions relatives aux interventions de l'après-midi*
16h00 – 17h00	*Discussion*

Résumés des interventions

9h50 - 10h30 *Annotation manuelle de corpus : mais de quoi parle-t-on ?*

Karën Fort - Inist-CNRS et LIPN UMR 7030 CNRS /Université Paris Nord

On assiste depuis les années 90 à un regain d'intérêt pour les corpus, en particulier les corpus annotés, et ce dans de nombreux domaines de recherche. Ces corpus annotés doivent offrir la meilleure qualité d'annotation possible, et nécessitent donc de faire intervenir des experts humains dans le processus d'annotation, que ce soit pour annoter directement le corpus ou pour corriger une annotation réalisée automatiquement. Or, cette phase manuelle est extrêmement fastidieuse et nécessite un travail de longue haleine et de qualité si possible constante. Il n'existe cependant à ce jour aucune grille permettant d'évaluer précisément la complexité de l'annotation envisagée. Les difficultés que présente l'étiquetage morpho-syntaxique du Hindi ne sont pas les mêmes que pour de l'annotation de renommages de gènes ou de structure de textes médiévaux, mais on peine à les définir, donc à les réduire.

10h50 – 11h30 *Exploitation éditoriale du balisage scientifique*

Pierre-Yves Buard - GREYC UMR 6072, équipe Document numérique Langue Usages /Pôle Document numérique, MRSH Caen

Présentation des méthodes et outils utilisés pour les projets d'édition multisupport de sources aux Pôle Document Numérique de la MRSH et aux Presses universitaires de Caen. Cette chaîne, organisée autour des recommandations de la TEI comme format d'échange entre chercheurs et éditeurs matériels, propose une organisation du travail et un ensemble d'outils permettant d'attribuer des formes adaptées aux différents supports de diffusion, d'une part, et aux catégories de textes identifiées et manipulées par les chercheurs, d'autre part.

11h30 – 12h10 *Projet ProDescartes*

Vincent Carraud - GREYC UMR 6072 CNRS/Université de Caen/Ecole nationale supérieure d'ingénieur de Caen

(résumé à venir)

13h30 – 13h50 Chrestien de Lihus dans l'hypertexte : retour d'expérience sur une expérimentation d'annotation de ressources textuelles dans le réseau Wicri

Thierry Daunois - Direction des Partenariats/Université de Lorraine

A partir d'une demande initiale de mise en ligne de ressources textuelles, il a rapidement semblé intéressant de dépasser ce cadre, pour proposer différents niveaux d'annotation. Aussi la version en ligne actuelle comporte dorénavant des annotations "techniques" (liens vers des ressources externes, commentaires de notes), et un dispositif permettant d'ajouter des commentaires disciplinaires (agronomie, histoire, géographie...). Cette expérimentation démontre la faisabilité (technique et humaine), avec une infrastructure légère, de ce type d'opération et permet d'envisager d'effectuer ce travail pour d'autres textes libres de droits. Mais elle ouvre également des perspectives en matière de recherche, par l'ajout de nouvelles fonctionnalités (balisage en TEI, curation...), qui n'étaient pas utiles dans ce contexte.

13h50 – 14h10 Plateforme d'annotation terminologique semi-collaborative

Evelyne Jacquey – Atilf UMR 7118 CNRS/Université de Lorraine

Projet ASTTIC (Annotation Sémantique et Terminologique de Textes pour leur Indexation et leur Catégorisation) – Axe 2 : Langues, textes et documents – MSH-Lorraine

L'objet de l'exposé est de présenter l'utilisation de la plateforme d'annotation terminologique qui a été développée pour assister les annotateurs dans la tâche de validation de candidats-termes en Sciences du Langage. Il fera aussi une petite incursion dans les objectifs plus larges du projet ASTTIC. L'annotation terminologique consiste à repérer et à mettre en valeur des *termes* en texte intégral. Les *termes* appartiennent au lexique scientifique des Sciences du Langage selon deux ressources terminologiques principales : Thesaulangue et le vocabulaire de la linguistique de de la base Francis accessibles via le portail Termsciences. Les textes annotés sont des articles scientifiques en Sciences du Langage mis à disposition par le LIDILEM dans le cadre de l'ANR Scientext. Les termes sont d'abord repérés automatiquement par des extracteurs de termes (Acabit et TermoStat) basés sur l'utilisation de règles statistiques et linguistiques. Les extracteurs de termes fournissent pour chaque texte un ensemble de candidats-termes que les annotateurs experts en SdL doivent ensuite valider ou rejeter. La tâche de validation étant complexe, l'annotation proposée dans la plateforme a deux particularités. Premièrement, elle se déroule en couches successives afin de diviser la question de la validation ou de la non-validation d'un candidat-terme en répondant à plusieurs questions élémentaires. Deuxièmement, elle permet de réaliser une annotation semi-collaborative dans la mesure où tous les annotateurs travaillent sur un même fichier et sur une même couche de validation, mais ceci l'un après l'autre, participant ainsi à la construction d'une validation fondée sur le consensus plutôt qu'au plus fort taux d'accord inter-annotateur.

14h10 – 14h30 *Annotation et édition de texte. Edition critique de la première traduction française de la Cité de Dieu par Raoul de Presles* (1371-1375)

Béatrice Stumpf – Atilf UMR 7118 CNRS/Université de Lorraine

ERC Histoire du lexique politique français **-** *Starting Grant - Projet N°208986-1-HFPSL*

L'édition critique de cette imposante traduction qui repose sur l'œuvre de saint Augustin, composée plus de neuf siècles auparavant, pourrait n'être qu'une publication imprimée classique, comme la plupart des éditions jusqu'à présent. Les atouts majeurs que présentent les annotations raisonnées de l'édition électronique ont orienté l'équipe vers une transcription de texte au format universel TEI qui exploite un langage défini en même temps qu'il garantit à l'édition une harmonisation interne dans la manière de transcrire. Dans notre présentation, nous mettrons l'accent sur les différents niveaux d'annotations retenus en vue d'une édition imprimée, d'une part, et de recherches lexicologiques, d'autre part, notamment pour établir l' « Histoire du lexique politique français » qui est précisément l'objet du programme ERC.

14h30 – 14h50 *Présentation d'une expérience d'annotation en parties du discours de corpus oraux : évaluation qualitative*

Christophe Benzitoun - Atilf UMR 7118 CNRS/Université de Lorraine

Dans notre intervention, nous présenterons notre travail portant sur l'élaboration d'un corpus d'apprentissage en vue de construire un étiqueteur morpho-syntaxique pour le français parlé. Pour élaborer cette ressource, nous avons sollicité des étudiants que nous avons fait travailler par binôme. Nous dresserons un panorama de la méthodologie suivie et de l'importante réflexion autour du contrôle de qualité de la ressource.

14h50 - 15h10 *Les encyclopédies médiévales et* les digital humanities *: l'évolution du programme* Sourcencyme

Eduard Frunzeanu - Centre de Médiévistique Jean-Schneider - ERL 7229 Université de Lorraine Philippe Pons Atilf UMR 7118 CNRS Université de Lorraine

Projet Sourcencyme (La compilation scientifique et philosophique dans les encyclopédies latines médiévales: textes et sources) - Axe 2 : Langues, textes et documents – MSH-Lorraine

Issu de l'intérêt pour les compilations médiévales, pour la transmission du savoir, pour la réécriture en tant que pratique d'appropriation des textes, le programme *Sourcencyme* vise la constitution d'une base de données qui réunisse un vaste ensemble de textes rédigés entre le XIIIe et le XVe siècle, notamment dans le domaine de la philosophie naturelle. Bénéficiant d'un balisage XML qui respecte les recommandations de la TEI, le corpus propose plusieurs types d'apparat (identifications des sources, annotations scientifiques, fiches bio-bibliographiques) destinés à mettre en lumière le rôle et la hiérarchie des autorités dans l'écriture encyclopédique, l'assimilation des doctrines des théologiens et des philosophes, l'impact de l'état (plus ou moins corrompu) des manuscrits sur la compréhension des textes, les distorsions et les raccourcis qu'entraîne l'acte de récrire, les interventions propres à chaque compilateur, etc. Qu'il s'agisse de textes pseudépigraphes ou de textes dont les versions diffèrent par leur division ou leur organisation, les particularités des œuvres médiévales ont imposé de faire des choix informatiques adaptés au cas par cas. Dans notre présentation, nous aborderons les différents problèmes de structuration informatique du corpus de *Sourcencyme* et ferons le point sur les recherches menées en vue de son enrichissement.