Bonjour à tous,

voici un petit compte rendu de la précédente réunion du 18 juin 2004 (idem en fichier attaché)

Etaient présents :
- Serge Fleury, Cedric Gendrot, Michel Jacobson et André Salem pour les membres déjà connus ...
- Sonia Branca et Thierry Pagnier qui sont venus après avoir répondu à l'appel qui avait été effectué quelques semaines auparavant. Thierry Pagnier est Allocataire-Moniteur en 1ère année de thèse de socio-linguistique, sous la direction de Sonia Branca, professeur de Sociolinguistique.
- Frédérique Bénard, qui est étudiante en Licence TAL, et qui fera une maîtrise entre le TAL et la Phonétique l'année prochaine, son sujet n'est pas encore parfaitement défini à ma connaissance, mais son travail portera sur les corpus obtenus grâce à ce projet innovant.

Bien sûr, toutes vos réactions sont les bienvenues...

Bien amicalement,
cedric

Voici les points principaux qui ont été abordés :

1- Résumé du projet innovant

Le cœur du projet est l'élaboration **d'une base de stockage multimédia** (écrit / oral / vidéo, regroupant des données de langues de différentes natures) **ainsi que des propositions de normalisation pour l'encodage de corpus de langue**.
Le but de ce projet est notamment de rendre ces données visibles aux chercheurs de Paris 3 : ceux qui souhaitent intégrer leur corpus dans cette plate-forme ou tout simplement ceux qui souhaitent utiliser les données stockées dans la base. Ce projet sera conduit dans un cadre interdisciplinaire, et dont l'objectif évident est une contribution à la linguistique de corpus.
La mise en place d'un serveur constitue donc la première étape de ce projet. Dès que nous aurons récolté un maximum de corpus, l'étape suivante sera de tenter d'organiser toutes ces données en une base cohérente, qu'il sera possible de consulter librement, d'analyser, mais également de compléter grâce à des outils appropriés.

2- Résultat de l'appel à données :

Très peu de réponses. Malgré tout, cet appel nous a permis de prendre contact avec Sonia Branca et Thierry Pagnier qui se sont proposés de nous rejoindre en tant que membres actifs dans le projet ; leur collaboration motivée sera un atout précieux, d'autant qu'ils représentent la Sociolinguistique, et le dialogue multi-disciplinaire était l'un de nos objectifs. Pour tenter de résumer, Sonia possède une grande quantité de données audio (dont la qualité sonore peut varier), qui ont été collectées plusieurs années durant, et étiquetées par une variété d'étudiants. Thierry en numérise actuellement une grande quantité.

D'autres corpus pourront facilement être obtenus de la part du laboratoire de Phonétique (je me charge de cette collecte bien sûr !), mais également des corpus d'origines divers comme par exemple un corpus annoté de la part d'une étudiante en Maîtrise TAL. Les sections FLE peuvent-elles aussi nous fournir des corpus audio qu'ils sont habitués à obtenir ... je connais quelques personnes, je les contacterai.
En ce qui concerne le LACITO, nous devrons mieux définir quels sont les entrecoupements avec l'ED268. Certains corpus du Lacito sont déjà sur le site, d'autres en attente de l'être. Que faire .... ?

Il faudra sans nul doute refaire un appel à données, mais j'insiste à nouveau sur le fait que les contacts personnels de chacun, des « appels personnalisés » en quelque sorte en amélioreront l'efficacité.

3- Le droit des locuteurs, les droits des propriétaires des corpus

Voici un point que les sociolinguistes semblent beaucoup plus habitués à gérer ... en effet, en Phonétique, les locuteurs sont souvent amenés à lire des corpus de phrases très préparées, ou bien de la conversation dont le sens même est bien peu important .... les locuteurs sont en général peu soucieux de l'utilisation de leur voix .... quoique... ! Ceci est loin d'être le cas pour des corpus audio de Sociolinguistique (et à plus forte raison vidéo).
Pour l'instant, la question a été en quelque sorte contournée, en maintenant un accès individualisé à ces données ...

4- Le point corpus multimédia – quels types de données ? ? ?

Dans le projet innovant, il avait été fait mention de collecter les textes, sons et vidéos analysés par les différents membres de l'ED. Or, si l'utilité de collecter des corpus vidéo et audio annotés d'une manière ou d'autre paraît indéniable, il en va différemment des textes (corpus textuels) sur lesquels les syntacticiens, Taliens, ... ont travaillé en apportant des annotations (balisages), par exemple. La discussion reste apparemment ouverte ...

Un argument du oui :
les textes sur lesquels un travail a été fourni, soit par la création même d'un texte permettant l'étude de points de linguistique, ou bien par le balisage qui a été réalisé, sont des corpus sans nul doute .... Pour prendre l'exemple des laboratoires de Paris7 (comme LATTICE ou .....), de nombreux chercheurs en syntaxe, sémantique, pragmatique, vont chercher dans des base de corpus, des structures syntaxiques, ou des enchaînements lexicaux, afin de vérifier sur des textes réels, la pertinence de ce qu'ils étudient.
Un argument du non :
Les attentes d'un corpus écrit peuvent s'avérer différentes de celles d'un corpus oral. Il faudrait alors scinder dès le départ sur le serveur les corpus écrits des corpus audio/vidéo. La présence de textes purement écrits (qui ne seraient pas la transcription de données orales) pourrait ainsi nuire à la cohérence du projet innovant.

4 – Mise en place du serveur

Comme il est dit dans le résumé de ce projet innovant : « La mise en place d'un serveur constitue donc la première étape de ce projet. ». Nous allons en fait bénéficier d'un espace disque sur un serveur qui était réclamé par l'ILPGA depuis quelque temps déjà.... L'installation passe donc par le CRI de Paris3, et devrait être réalisée avant l'été (sic). Cette opération nous permet d'éviter l'achat d'un serveur et soulage considérablement notre budget... elle permet également un gain de temps appréciable.

5 – Achats - budget

Deux disques durs de 500 G ont déjà été commandés ; ils permettront de stocker temporairement les données avant de les placer sur le serveur.
Un des prochains achats sera certainement du matériel d'acquisition (un lecteur de Mini-Discs avec sortie numérique + carte son compatible), notamment pour Thierry qui numérise actuellement quantités de données, sans avoir le matériel adéquat ...
Une petite série d'appareils d'enregistrements pour les étudiants qui souhaitent faire quelques enregistrements, mais qui faute de matériel, se rabattent sur le dictaphone bon marché.

Le choix se porte principalement entre :
- MiniDisc (numérique) : à la fois économique et solide ... le son est de bonne qualité .... malgré, ce n'est pas la panacée, puisque les données sont compressées dès le départ.
- DAT (numérique), fragile et cher, malgré tout, il fournit le meilleur enregistrement possible.
- Enregistreur à cassettes classique (analogique) .... il peut faire sourire, mais s'il est de bonne qualité, il reste un outil robuste, fournissant un enregistrement de très bonne qualité .... par définition, étant un outil analogique, le son n'est pas dénaturé, n'oublions pas les vieilles recettes !

6 – suite ...

- Nouvel appel à données ... il est nécessaire de décider pour les textes annotés/balisés ..... chacun peut préciser ses préférences par mail.

- Il faudra penser à obtenir des propriétaires de corpus des informations détaillées.... Cf, ce que Michel a déjà fait pour le site du Lacito.
- Thierry continuera donc de numériser ses données ... Le mieux serait certainement de s'occuper du matériel d'acquisition au plus vite !
- Le serveur devrait bientôt être opérationnel... l'installation de premières données, accessibles à nous seuls pour l'instant, nous permettra de mieux réfléchir à l'organisation des données.