Bonjour à tous,

Voici le compte-rendu de la réunion du projet innovant du 30 septembre 2004 (idem en fichier attaché)

Etaient présents Maria Candéa, Serge Fleury, Cédric Gendrot, Michel Jacobson, Luiggi Sansonetti, Pollet Samvelian, Patrick Renaud, Sonia Branca et Thierry Pagnier.

Présentation des 2 nouveaux-venus :
Maria Candea : Maitre de Conf à Paris 3 spécialisée dans l'analyse du discours, **Centre de recherches sur le français contemporain, EA 1483** , (Marie Annick Morel),
Luiggi Sansonetti : vacataire au LACITO et à l'ILPGA en thèse deTAL.

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

Bien amicalement,
cedric

Cette réunion fut l'occasion de faire notre tout premier bilan sur le recueil des données…. De quoi dispose-t-on dès aujourd'hui ? Il s'agit de se servir d'un ensemble varié de corpus pour effectuer notre premier travail : fournir un fichier de Méta Données qui décrira le corpus mis à disposition sur le serveur, et ce bien sûr le plus précisément possible.

Voici les points principaux qui ont été abordés :

1- <u>Résumé du projet innovant (rappel pour les nouveaux venus)</u>

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.

<u>2- Le point corpus multimédia – quels types de données ? ? ?</u>

D'après le compte rendu de la précédente réunion :
« 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 ... »

Il semble que nous sommes tombés d'accord sur ce point. Les attentes d'un corpus écrit sont très différentes de celles d'un corpus oral, et donc le travail à fournir est également très différent… Mieux vaut éviter de se disperser dès le départ. La question se pose tout de même concernant les transcriptions de corpus oraux dont l'oral a disparu. Nous déciderons au cas par cas …

<u>3- Résultat de l'appel à données :</u>

Nous disposons des corpus apportés par :

- Luiggi Sansonetti, plutôt orientés « acquisition du langage », récoltés pour une bonne part (dites moi si je me trompe) par des étudiants pour leurs travaux.

- Sonia Branca et Thierry Pagnier, corpus de sociolinguistique, également récoltés et transcrits par des étudiants pour leurs travaux.
- Patrick Renaud, je n'ai pas encore de description
- Pollet Samvelian, le corpus dont j'ai connaissance a été réalisé avec un groupe de chercheurs de syntaxe et sémantique (Paris 3 et Paris 7). Il contient un certain nombre de phrases contrôlées et peut-être un peu de parole spontanée.
- Maria Candéa, qui nous offre son corpus de thèse, et nous proposera certainement des corpus enregistrés au sein du laboratoire Marie Annick Morel.

Plus les miens, et surtout ceux que j'ai pu récolter au laboratoire de Phonétique, en voici une liste plus complète, mais non exhaustive

- un corpus de 4 locuteurs (2h + 2f) lisant des phrases très contrôlées (phonèmes du français, syllabes dans des positions prosodiques ciblées), 3 textes français couramment utilisés en français (normal, lent, rapide, hypo et hyperarticulé), 39 phrases ambiguës dont la prosodie aide à la désambiguïsation. Phrases découpées mais non étiquetées
- Un corpus français de 2 locuteurs (1h + 1f) lisant une 50aine de phrases contrôlées (syllabes dans des positions prosodiques ciblées). Complètement étiqueté phonémiquement avec Praat
- Un corpus français de parole spontanée - 4 locuteurs (étiqueté phonémiquement en partie)
- Un corpus allemand de phrases lues par 2 locuteurs. Etiqueté phonémiquement par Kiel … très complet
- Un corpus de simulations émotionnelles (une 30 aine de phrases – 3 locuteurs - phrases classées non étiquetées)
- Un corpus français de 40 locutrices lisant un texte (2 répétitions). Etiqueté en phonèmes et syllabes
- Un corpus du LIMSI – parole radiophonique, anglais, allemand et français – étiqueté phonémiquement.

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

Nous avions abordé ce point à la dernière réunion :
« 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 ... »

Patrick Renaud nous a parlé de CLAPI (corpus de langue parlée en interaction ; http://gric.univ-lyon2.fr/projets/nomex-clapi/presentation/presentation.html), projet qui porte sur les droits que doivent avoir les locuteurs ainsi que les auteurs des corpus. Il a également été fait mention de ICAR (laboratoire à LYON). Nous nous appuierons sur les réflexions existantes les plus complètes afin de nous éviter tout problème. Cependant, une grande quantité des données que nous possédons n'ont pas pris soin de respecter ces principes, et nous devrons penser à un accès modulaire aux données que nous placerons sur le serveur, en fonction de ce que nous aurons décidé au cas par cas. L'accès se fera de toute façon au moyen d'un « Log-in ».
N'oublions pas que l'un de nos objectifs est de pouvoir continuer à placer de nouveaux corpus sur cette base de données, et il faudra que les futurs auteurs de corpus y attachent plus d'importance.

4 – Exclusivité

Nous avons brièvement parlé d'exclusivité. Quelqu'un qui nous soumet un corpus peut-il le donner à quelqu'un d'autre ? La réponse est oui !

5 – Mise en place du serveur

Les dernières signatures ont été obtenues ce jeudi…. Je crois qu'on peut dire que c'est gagné, et que ça n'a pas été de main morte.
Nous pourrons ainsi placer quelques extraits afin de simuler l'installation d'une base de données. Ceci devrait nous permettre de mettre en évidence la diversité des données, et ainsi de mieux réfléchir aux solutions à apporter…. Suite dans quelques jours .

5 – Achats - budget

Deux disques durs de 500 G ont été achetés; ils permettront de stocker temporairement les données avant de les placer sur le serveur. Je vais passer à Serge les données déjà existantes sur un de ces disques afin qu'il puisse les mettre en place en accès privé.
Le prochain achat sera donc du matériel d'acquisition. Nous en avions déjà discuté : si nous souhaitons que les futurs corpus soient de bonne qualité, il est nécessaire de se pourvoir d'un matériel solide et fiable. La décision s'est portée sur des petits enregistreurs dont je vous envoie la référence dans mon mail suivant (comptez 1000 euros l'unité)
Nous achèterons également pour Thierry qui numérise actuellement quantités de données, une platine Mini-Disc, ainsi qu'une carte d'acquisition. Les références seront fournies par Luiggi.

6 – Programme pour la suite

- D'après le formulaire fourni par Michel, ainsi que d'après les corpus que nous avons à notre disposition, nous pourrons réfléchir à compléter le formulaire
- L'étape suivante sera la prise en compte des annotations des corpus, leur récupération et leur disposition dans la base de données.