Voici le compte-rendu de la réunion du projet innovant du mardi 20 septembre 2005

Etaient présents Frédérique Bénard, Serge Fleury, Cédric Gendrot, Michel Jacobson, Thierry Pagnier, Sonia Branca, Alexis Michaud.

Cette réunion de rentrée nous a permis de faire le point sur quelques aspects du PI, avant de préparer l'échéancier pour cette dernière année. Je vous rappelle que le PI est supposé ne durer que 2 ans, et le rapport devra être envoyé pour la fin de l'année 2005.

Voici les 3 points abordés au cours de cette réunion :

- quelques questions diverses
- Publicité du Projet, bonnes pratiques …
- Centres de compétences
- Echéancier

1. Quelques questions diverses

- Sonia Branca nous a demandé quels types de transcriptions étaient acceptables pour que les corpus puissent être insérés …
    - Il n'y a pas de restriction. De même, deux transcriptions différentes peuvent être accolées à un même corpus : une fine et grossière par exemple. Il est également possible de faire cohabiter une transcription non alignée au signal sonore (faite sur Word par ex.) avec une transcription alignée (faite sur Praat, Transcriber, CLAN, ELAN …)

- Serait-il possible de procéder à des échanges avec d'autres laboratoires : par exemple Paul CAPO du GARS ?
    - Bien sûr, comme nous en avons discuté : le point essentiel de ce projet est la NORMALISATION pour que ces corpus soient échangeables. Nous avons également des contacts avec le Québec (avec François Daoust pour le projet SATO). Si Paul Capo (ou d'autres) veut prendre contact avec nous, ce sera donc avec plaisir (auprès responsable du projet plutôt : serge.fleury@univ-paris3.fr

- Il est dommage que la fiche de métadonnées soit un peu « pauvre » … Par exemple, aucune case n'est prévue pour mentionner l'âge du locuteur, sa catégorie socio-professionnelle, l'origine des parents.
    - Oui en effet, nous avons rencontré ce problème et avons du TRANCHER. D'une part, la fiche de métadonnées décrit le corpus principalement, plus que le locuteur. Ensuite, nous avons besoin de cette normalisation : rajouter une simple case avec l'âge du locuteur en plus de son nom nous retire de la normalisation fournie par Dublin Core/OLAC. Rien n'empêche les concepteurs du corpus d'insérer ces informations dans la case « DESCRIPTION », et il est ensuite possible de les retrouver grâce au moteur de recherche mis en place (par exemple en recherchant le mot « banlieue ». Voici un extrait du compte rendu de la réunion du 1er avril 2005 qui répondait à cette question.

```
##########-------------------------------------------------
-     L'élément DC Description :
o     une description du contenu de la ressource. Peut contenir un résumé,
une table des matières, une référence à une représentation graphique du
contenu ou un texte libre sur le contenu. En commentaire donc, la langue
utilisée pour écrire ce texte (le français ici).
Pour l'instant ? ILEST LE SEUL endroit pour indiquer par exemple l'âge et
l'origine des locuteurs ...
Thierry, en parfait sociolinguiste, a insisté sur l'importance de ces
paramètres dans sa discipline, qui ne sont actuellement pas prévus par
OLAC/DC.  Michel nous a mentionné l'existence d'enquêtes très précises (du
Max Planck Institute : http://www.mpi.nl/world/tg/lapp/lapp.html), .....
mais trop en fait : trop lourd pour tout le monde bien que ce soit
théoriquement la panacée.
L'autre problème est que rajouter des éléments nous éloigne définitivement
de la normalisation. Une solution intermédiaire consiste en un ajout de
```

quelques éléments qui nous semblent indispensables, qui peuvent être enregistrés malgré tout, mais à part afin de ne pas mettre en péril l'aspect normalisé/échangeable que nous mettons en place actuellement. Thierry nous proposera la semaine prochaine le fruit de ses réflexions, en cherchant parmi les corpus dont il est responsable, mais aussi en visitant les pages de PFC (Phonologie du Français Contemporain), voire celle du Max Planck Institute (http://www.mpi.nl/world/tg/lapp/lapp.html) pour s'inspirer de leurs formulaires. Si Patrick veut nous faire part (par mail ou lors de la prochaine réunion) de quelques catégories qu'il estime indispensables, nous sommes preneurs ! Il faudra malgré tout trouver un compromis, à suivre lors de la prochaine réunion !
##########----------

- Donc si Thierry peut proposer une façon de mettre en forme ces informations dans la partie « DESCRIPTION », nous sommes preneurs. De même, il est possible d'indiquer des liens vers des fiches plus complètes telles que celles proposées Musée des arts et traditions populaires ou la BNF (OPAL)

    - Je vous rappelle également la mise en place d'un groupe de réflexion CATCOD, dont je vous renvoie un petit résumé ci-dessou, qui permet de réfléchir sur ce dont les linguistes ont besoin pour coder un corpus, afin de pouvoir proposer le fruit de cette réflexion à la TEI.

    « Suite à l'action spécifique ASILA (Interaction Langagière et Apprentissage), le groupe "comment cataloguer, comment coder" (CATCOD) a démarré avec S Heyden, E. Schang et M. Jacobson. Le but de ce projet est de faciliter les échanges au sein de la communauté (catalogage/codage des corpus oraux) et éventuellement de pouvoir proposer une analyse a la TEI. Dans le but de normaliser notre pratique et pour mener à bien cette tâche, un correspondant pour chaque domaine identifié serait idéal. J'ai proposé mon aide pour le secteur Phonétique, Thierry pourrait également s'y insérer pour la socio-linguistique. Voir Michel pour plus d'informations. »

2. Publicité du Projet, bonnes pratiques …

La journée C-Oraux à la BNF a bien mis en évidence l'importance de récolter des corpus oraux et d'en assurer la pérennité. Je vous rappelle que le guide C-Oraux est téléchargeable à l'adresse suivante (version non définitive) : http://www.culture.gouv.fr/culture/dglf/corpus_oraux.htm

Sur ce guide très bien fait, il est possible de trouver quelques fiches qui rappellent les choses à savoir avant d'enregistrer un corpus …
Très pratique si l'on ne veut pas commettre d'erreurs irréparables (comme prendre un magnétophone à mini-K7 par exemple), ou oublier de poser des questions importantes (telles que l'origine du locuteur, de ses parents)

Dans le but d'assurer la pérennité de ce projet, voici les points qui sont prévus :

- Un nouveau cours sera intégré aux prochains savoir-faire (date provisoire 27 janvier 2006) : un point sur la base de données, comment en profiter, comment y contribuer ? + un cours sur Transcriber que nous conseillons !

- Un nouveau cours dans le cursus TAL « Linguistique de corpus » permettra de familiariser les étudiants avec l'archivage de corpus de manière générale ainsi que la base de données de l'ED bien sûr. Les divers façons reconnues d'annoter un signal en fonction de la discipline et/ou des objectifs de recherche. Les prochains cours de Phonétique/Phonologie devraient intégrer des TDs à rendre qui consistent en un petit corpus étiqueté ... à insérer dans la base de données !!

Nous comptons sur vous tous qui demandez aux étudiants de fournir des corpus pour vos cours :

- essayez d'insérer systématiquement ces corpus dans la base de données
- indiquez les synthèses/consignes existantes,
- indiquez les prochains savoir faire
- proposez matériel disponible

3. Centres de compétences

J'avais écrit dans mon précédent mail pour la réunion PI que nous parlerions d'un appel à centre de compétences (que je vous transmets) à remplir pour le 15 octobre.

Il s'avère que c'est un peu plus pressé que prévu, et beaucoup de choses ont déjà avancé. Serge, Michel, Alexis (Michaud) et moi-même sommes plus que partants pour proposer quelque chose et nous sommes en train de rédiger un texte. Le laboratoire porteur du projet sera le LACITO qui a une longue réputation dans le domaine des corpus (le laboratoire porteur doit être affilié au CNRS, ce qui est la seule restriction à priori !).

Le projet tel qu'il est conçu pour l'instant porte sur la "normalisation" des données. Il s'agit d'un projet bien distinct du projet innovant sur lequel nous travaillons, mais nous pourrons avoir l'occasion de mettre à profit le travail effectué (**i.e. TRANSMETTRE NOS COMPETENCES**)
Je vous laisse lire ci-dessous un petit résumé de cet appel (que je vous transmets également). Merci de nous contacter si vous êtes partants!!

----------résumé----------

Le but général de la mise en place des centres de compétences est de dynamiser la recherche en évitant les redondances et en facilitant l'accès aux services disponibles grâce à une centralisation des connaissances et des moyens. Il s'agit de structurer les communautés de recherche en regroupant et en partageant ressources et outils divers (ressources classiques mais aussi ressources avancées plus rares). Cette capitalisation des résultats de chacun doit permettre des expérimentations de nouvelles fonctionnalités plus faciles et des mises en œuvre plus rapides. De plus, cette structuration devra faciliter une meilleure transmission de l'information scientifique qui ne soit plus seulement fondée sur les publications.
Un centre de compétences accompagne les équipes de recherches dans leurs besoins de créer, gérer et diffuser des ressources numériques (données et outils pour les traiter) à destination de la communauté scientifique. Diverses actions existent déjà qui vont dans ce sens et nombre d'organisations fonctionnent comme ce que nous souhaitons mettre en place (par exemple en SHS, dans le domaine des sciences sociales et économiques ou en géopolitique). Ces actions, d'une part serviront d'exemples et, d'autre part, pourront être partie prenante, dans un second temps, au présent appel. La première étape envisagée concerne essentiellement les domaines où ce mode de fonctionnement est encore à développer fortement et où il sera d'un bénéfice certain. C'est la raison des choix de domaines que nous avons effectués. Cette première étape appelle également un bilan détaillé de l'existant (en particulier, les articulations possibles avec les CCT [centres de compétences techniques] et la mission aux archives scientifiques du Réseau national des MSH). On soulignera ici qu'il ne s'agit pas seulement, dans les centres de compétences tels que nous les concevons, d'archives de documents, mais de ressources en général (c'est-à-dire de données, de corpus et d'outils pour les produire, les gérer, les modifier). Les étapes suivantes de cette initiative généraliseront ensuite cette approche aux autres disciplines
----------résumé----------

4. Echéancier

a- les sous à dépenser ...

JE VOUS RAPPELLE LA CHOSE SUIVANTE (POUR THIERRY, LUIGGI ET MOI-MEME !)

Il nous reste à vérifier si les 3000 euros qui ont été versés l'année dernière ont été complétés par un 2ème virement de 3000 euros

Pour l'instant, les propositions d'achat sont les suivantes :
- 2 Marantz (à réserver pour les doctorants et +) (devis à fournir par Cédric)
- 4-5 Mini Discs (comme celui de Thierry) avec micro de rechange (devis à fournir par Thierry)

- Logiciels Cordial pour 13 machines (devis à fournir par Luiggi)

b- Programme pour la rentrée – suite du projet !!

Nous allons progressivement installer toutes les données sur le serveur.
Il reste donc un certain nombre de fiches de Métadonnées (à remplir avec MKM) qui ne nous ont pas encore été communiquées, et parfois les corpus qui n'ont pas été fournis dans leur intégralité.
Pourriez vous svp nous transmettre toutes ces informations/données ?
Je me tiens bien sûr à votre disposition … De même si vous rencontrez quelques problèmes, n'hésitez pas à les faire connaître en envoyant un mail à l'ensemble du groupe.

Je vous rappelle le lien vers MKM nouvelle version

http://pi-ed268.univ-paris3.fr/files/makeMETADATA-W32-1.09.zip