Bonjour à tous,

Voici le compte-rendu de la réunion du projet innovant du 16 novembre 2004 (idem en fichier attaché)

Etaient présents Frédérique Bénard, Maria Candéa, Serge Fleury, Cédric Gendrot, Michel Jacobson, Luiggi Sansonetti, Patrick Renaud.

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

Bien amicalement,
cedric

Voici les points principaux qui ont été abordés :

1- Etat du recueil de données et mise en ligne
Nous avons recueilli un certain nombre de corpus, que nous avons regroupés sous 10 intitulés.
Ces regroupements ont principalement été faits sur la base de leur source, c'est à dire en fonction des personnes qui ont fourni ces corpus.

BDDED268-1 - corpus Aurelf_upelf
corpus lu de phrases françaises phonologiquement équilibrées, logatomes, et textes lus à différents débits. En partie étiqueté phonétiquement (étiquetage récupérable en format texte)

BDDED268-2 - contrat PLM – dret
corpus français, turc et vietnamien de textes lus à différents débits + parole spontanée. En partie étiqueté phonétiquement (étiquetage récupérable en format texte)

BDDED268-3 - corpus simulations émotionnelles
36 phrases françaises de 8 syllabes lues de façon neutre + 4 simulations émtionnelles (joie, colère, tristesse, surprise) ; phrases classées mais non étiquetées

BDDED268-4 - corpus Kiel
phrases allemandes lues et intégralement étiquetées (étiquetage récupérable en format texte)

BDDED268-5 - corpus émissions journalistiques allemandes LDC
mélange d'interviews préparées et de flashs d'informations (en allemand). intégralement étiquetées automatiquement (étiquetage récupérable en format texte)

BDDED268-6 - corpus Nathalie DM

interviews (en français) préparées. étiqueté lexicalement en séquences sous word

BDDED268-7 - corpus Benguerel
phrases françaises lues et intégralement étiquetées (étiquetage récupérable en format texte)

BDDED268-8 - corpus Maria Candea
Conte raconté par une élève de 4ème ; étiquetage lexical en séquences peut-être récupérable en format texte)

BDDED268-9 - corpus Thierry Pagnier
ensemble de corpus de sociolinguistque. Etiquetage varié (souvent sous Word), pas toujours de son numérisé.

BDDED268-10 - corpus Luiggi Sansonetti
ensemble de corpus d'acquisition du langage. Etiquetage varié, pas toujours de son numérisé.

Ces données ont été placées sur les deux disques durs .... Voici le code permettant d'accéder à ces données en ligne sur le site suivant
http://pi-ed268.univ-paris3.fr/ffiles/index.html (public)
http://pi-ed268.univ-paris3.fr/files/index.html (privé avec log in)

Dans la majorité des cas, les données ne sont pas complètes, puisque nous avions demandé quelques extraits seulement afin de créer les fiches de Méta données ... donc pas d'urgence pour récupérer l'intégralité de ces corpus, ce ne sera pas l'étape la plus longue de toute façon.
Nous avons décidé de partir de cet ensemble de données pour réaliser notre travail ... Un nouvel appel à données pourra être fait plus tard, mais nous avons suffisamment de données pour commencer à travailler. Bien sûr, il n'est pas trop tard pour intégrer les données proposées par Pollet Samvelian et Patrick Renaud.

2- Fiche de méta données
Nous avions fait circuler quelques jours avant la réunion une petite fiche à remplir (pour parler en termes techniques : une fiche simplifiée de Dublin Core).
Nous avions commencé à en parler lors de notre dernière réunion ... Cette fiche permet de décrire chaque corpus aussi précisément que possible pour les catégoriser. Évidemment le découpage en 10 intitulés, tel qu'il est indiqué ci-dessus devra être affiné, puisqu'ils contiennent parfois des sous-corpus extrêmement diversifiés.

La catégorisation de ces données est une étape indispensable mais également délicate, il s'agit de ne pas classifier trop rapidement un corpus pour qu'il ne puisse plus livrer ses secrets. Mais certains champs de description ne seront pas nécessairement remplis, ou ils peuvent être doublés, triplés ... lorsqu'il y a ambiguïté (au moyen des 'raffinements' notamment) De même certains champs ne sont pas du tout applicables aux données sur lesquelles nous travaillons.

Nous allons reprendre des feuilles de DUBLIN CORE complètes pour améliorer cette petite fiche avec un maximum de champs applicables à la parole. Nous pourrons ensuite ajouter le travail fourni par Frédérique Bénard.
Frédérique travaille à une traduction des méta données proposées par OLAC. Ces méta données permettront de fournir plus de champs et ainsi compléter les descriptions possibles d'un corpus.
Luiggi pourrait également nous aider grâce aux descriptions fournies par CHAT (Codes for the Human Analysis of Transcripts) : principes systématiques de transcription et de codage implémentés dans CLAN.

3- Méthodes d'annotation des corpus
D'après les corpus que nous avons recueillis, la première observation fut l'abondance des corpus annotés grâce à un éditeur de texte tel que Word. L'énorme inconvénient est l'absence de marquage temporel sur le signal

acoustique.... ce sui est donc peu pratique pour la personne qui doit retrouver le son à partir du texte et vice-versa. C'est surtout peu pratique pour nous ... puisqu'il faudra refaire le travail pour retrouver sur le signal les « marqueurs » temporels. Nous sommes tous conscients du fait que les étudiants ,et autres qui font des corpus, se moquent bien de ce qui nous plairait le plus. Leur premier argument est la simplicité d'utilisation pour une efficacité suffisante.
Or il existe des outils prévus à cet effet, et qui permettraient de faciliter la vie des étudiants (des concepteurs de corpus en général) et de nous faciliter la tâche par la suite. Le seul inconvénient est qu'il n'existe pas d'outil parfait pour répondre aux besoins de tous. L'un est meilleur pour un étiquetage phonétique fin (Praat), l'autre pour l'annotation de grands corpus interactionnels (Transcriber, Clan)...
Michel avait déjà au LACITO un panorama d'outils d'annotations qui avait été réalisé il y a quelques temps. Luiggi pourra ainsi reprendre, avec l'aide de Michel, cet état des outils disponibles et l'actualiser pour nous le présenter à la prochaine réunion. Maria et moi-même sommes en contact (pour un autre projet) avec un des programmeurs de Transcriber, et une version récente a été entamée avec la visualisation de spectrogrammes. Transcriber pourrait alors permettre un étiquetage beaucoup plus fin .... affaire à suivre !

Le but est de pouvoir dès la prochaine rentrée, par exemple à l'AG de l'ED268, inciter les nouveaux venus (mais également les anciens) à utiliser ces outils pour leurs recherches, et ainsi participer à la dynamique de cette base de données. Nous pourrions très facilement ajouter ces logiciels à la liste déjà longue des savoir faire pour la rentrée prochaine .... Je fais une présentation sur Praat depuis 4 ans, et il y a toujours un public nombreux. Mais j'en ferai un (ou 2) autre(s) très volontiers ! !

4- Achats pour février 2005
Les projets d'achats s'étaient arrêtés sur 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. Les mini-Discs n'apportent pas que des avantages ;
Nous allons donc essayer des « petits » enregistreurs ... Le labo de Phonétique en a commandé un exemplaire que nous essaierons avant de le commander sur le budget du Projet Innovant. Cet achat ne pourra, quoi qu'il en soit, être effectué avant février, date d'ouverture des budgets ...

6 – Programme pour la prochaine réunion

Nous reprendrons les fiches d'annotations les plus complètes possibles afin de les remplir pour décrire nos corpus. Le but est d'arriver le plus vite possible à une fiche (quasi) définitive qui nous permettra de décrire les corpus sur le site, mais également d'intégrer au mieux les nouveaux corpus que nous pourrions recueillir.