Bonjour à tous,
Voici le compte-rendu de la réunion du projet innovant du vendredi 13 mai 2005

Etaient présents Frédérique Bénard, Serge Fleury, Cédric Gendrot, Michel Jacobson, Thierry Pagnier, Luiggi Samsonetti.
Bien sûr, toutes vos réactions sont les bienvenues...

Bien amicalement,
cedric

##
DATE DE PROCHAINE REUNION….
Je propose le vendredi 24 juin à 9h30, j'attends vos réponses.
!##

############################ POINTS PRELIMINAIRES ###################################

- Je vous rappelle en premier lieu la tenue d'un colloque le 17 mai 2005 (demain !) sur les corpus oraux ; cela a lieu à la BNF.

  Pour info : http://www.bnf.fr/pages/cultpubl/colloque_388.htm

- N'oubliez pas que le 21 mai 2005 auront lieu les VIIIèmes rencontres de l'ED. Fred et moi-même y présenterons le travail de notre groupe. Notre but est tout d'abord de faire la publicité de ce projet, mais également de recueillir les réactions et attentes des membres de l'ED268. Venez nombreux ! (présentation à 17h).

- 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.

- Dans la même optique (spécialisation acquisition), Luiggi fait partie d'un projet ATILF composé notamment de l'équipe CRAPEL (Emmanuelle Canut, Jeanne-Marie Debai Sieng, ...). Voir Luiggi pour plus d'informations.

##

Voici les 3 points abordés au cours de cette réunion :

- 1. La fiche de Méta-Données Dublin-Core/OLAC : est-elle complète ?
- 2. Les logiciels d'annotation linguistique a conseiller
- 3. Catalogage des fiches de Méta-Données

1. LA FICHE DE META-DONNEES (DUBLIN CORE / OLAC) : est-elle complète ??

Lors de la réunion précédente, Thierry nous avait fait part de sa déception concernant l'absence de catégories sur l'âge, l'origine, la profession (etc...) du locuteur. Les socio-linguistes plus particulièrement, mais pourquoi pas les phonéticiens, ont recours à ce type d'informations dans leurs études. Ces informations peuvent être insérées mais ne sont pas normalisées dans notre fiche de Méta-Données, dans le sens où il est certes possible de les y insérer dans l'élément « Description », mais en tant que texte seul, et il ne sera donc pas possible d'effectuer des requêtes du type :

« je cherche tous les corpus obtenus sur des locuteurs de 10 à 12 ans. »

Ajouter un quelconque élément à cette fiche de Méta-Données reviendrait à se couper de la normalisation qui nous est si chère ; en effet, la normalisation des méta-données est effectuée sur les ressources et non sur les personnes/locuteurs. En tout cas, rien n'est prévu à ce propos. L'ajout d'un fichier attaché contenant ces informations serait possible mais nécessiterait une nouvelle table ronde pour laquelle il serait nécessaire de décider quels critères sont indispensables ou non ....
Une solution acceptable consisterait en une structure organisée des informations sur le locuteur dans l'élément « Description ». Par exemple,
locuteur.age = 12
enquete.lieu = Nantes

Thierry pourrait se mettre d'accord avec d'autres (Luiggi, Maria, ...) pour structurer cette information dans le texte. Il serait alors nécessaire d'indiquer le choix de cette structure à ceux qui souhaitent opérer des requêtes de ce type !

2. LES LOGICIELS D'ANNOTATION LINGUISTIQUE A CONSEILLER

Luiggi et Michel nous ont fait un passage en revue des outils d'annotation du signal les plus utilisés. Je vous rappelle qu'un des objectifs de ce projet consistera en une série de recommandations pour les membres de l'ED qui réalisent des corpus.
Il est indispensable de leur proposer des outils pratiques et efficaces (qui seront par ailleurs compatibles avec notre base de données, mais cela fait partie du côté « efficace »).
Luiggi nous a envoyés vendredi un fichier .pdf qui résume sa présentation. Je vous rappelle que Serge l'a déjà mis en ligne à l'adresse suivante :

http://pi-ed268.univ-paris3.fr/files/documents/synthese-outils-annotation.pdf
log-in : habituel
psswd : idem

Pour résumer... et puisque nous allons mentionner (de façon très sommaire) ces outils lors de notre présentation aux VIIIèmes rencontres de l'ED268 ...
Aucun outil d'annotation du signal n'est évidemment parfait, veuillez excuser les conclusions parfois simplistes que je présente ci-dessous. Tous les logiciels sont libres de droit sauf « WinPitch ». Ce critère est primordial pour nous, et si nous pouvons le mentionner, nous ne pouvons pas véritablement le conseiller. Les plus intéressants et complémentaires semblent être ELAN, Transcriber et Praat (peut-être Sound Index pour certains cas précis).

CLAN : complet, mais très peu pratique, et nécessiterait une mise à jour, utilise XML mais aucune sortie XML possible.
http://childes.psy.cmu.edu/clan/

WinPitch : complet mais non libre de droit ...
http://www.winpitch.com

AGTK tabletrans : sortie XML, pas de transcription fine possible. Un seul niveau d'analyse possible.
http://sourceforge.net/projects/agtk

AGTK multitrans : sortie XML, pas de transcription fine possible. Plusieurs niveaux d'analyse possibles (niveaux indépendants, nécessité de spécifier l'ancrage temporel pour chaque niveau)
http://sourceforge.net/projects/agtk

Praat : permet une transcription fine mais peu pratique pour les dialogues. Sortie texte convertible dans une bonne mesure en un format XML. Très utilisé en phonétique (mon chouchou, mais vous le savez déjà !) . Autant de niveaux d'analyse que possible (niveaux indépendants, nécessité de spécifier l'ancrage temporel pour chaque niveau)
http://www.fon.hum.uva.nl/praat

Elan : sortie XML (dtd spécifique), très complet. Plusieurs niveaux d'analyse possibles (niveaux dépendants/indépendants, nécessité de spécifier l'ancrage temporel pour chaque niveau).

http://www.loria.fr/equipes/protheo/SOFTWARES/ELAN/index.html

Transcriber : XML (dtd spécifique), très pratique pour les dialogues, pas de transcription fine, tours de parole, un seul niveau d'analyse possible.
http://www.etca.fr/CTA/gip/Projets/Transcriber/IndexFr.html

Sound Index : écrit par Michel. Sortie XML (pas de dtd spécifique), Plusieurs niveaux d'analyse possibles (niveaux dépendants/ indépendants, possibilité de spécifier l'ancrage temporel pour chaque niveau mais pas indispensable)... Mais peu ergonomique.
http://michel.jacobson.free.fr/

3. Catalogage des fiches de Méta-Données

Nous avions évoqué dans notre précédente réunion la mise en place imminente d'un catalogue qui sera la première partie visible accessible de notre BDD. Serge en a déjà fait circuler un exemplaire ...

Cette récolte de Méta-Données permettra également de proposer le moteur de recherche principal dont nous parlons depuis quelques réunions maintenant. Il sera possible ensuite de hiérarchiser nous-mêmes les corpus, pour définir nos propres regroupements (sur le français par exemple ...)

4 – Programme pour la prochaine réunion

- Eventuel dé-briefing de la journée à la BNF (17 mai) et des VIIIèmes rencontres de l'ED (21 mai).
- **Etape Suivante :** je communiquerai à Serge un ou deux corpus (si nécessaire) réorganisés en fonction des regroupements dont nous avions parlés, accompagnés de la fiche de Méta-Données pour le son, et de celle pour l'annotation. Nous les placerons ainsi sur le serveur (et complèterons progressivement avec les autres corpus, je vous rappelle que nous en avons 12 au total). Ceci nous permettra de servir de nouveau point de départ pour le projet :
    - la constitution d'un fichier de catalogage des fiches de Méta-Données de chaque corpus. Ce fichier de catalogage nous permettra d'opérer des requêtes sur la base de données.
    - A nouveau, il sera probablement nécessaire de créer un petit logiciel (peut-être proche de MKM) capable de produire ces requêtes de façon conviviale.