Bonjour à tous,
Voici le compte-rendu de la réunion du projet innovant du 14 février 2005 (idem en fichier attaché)

Etaient présents Frédérique Bénard, Serge Fleury, Cédric Gendrot, Michel Jacobson, Thierry Pagnier, Luiggi Sansonetti.

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

Bien amicalement,
cedric

##
DATE DE PROCHAINE REUNION….
Le vendredi semblait être la seule journée qui pourrait convenir à tout le monde...
Compte tenu du petit travail qui vous sera demandé, nous avons décidé de la fixer à la fin mars.
Je vous propose le vendredi 25 mars 2005.
(ou bien comme seconde proposition : le vendredi 1er avril 2005)
J'espère que cette date conviendra également à ceux qui n'étaient pas présents. N'hésitez pas à me prévenir si ce n'est pas le cas.
##

Voici les points principaux qui ont été abordés :

1- Description du corpus BDD1 : contrat AUPELF UREF

Fred et moi-même avions en notre possession deux documents pour réaliser et présenter notre travail :

- la feuille de description DC/OLAC (Méta-Données) que nous avons mise en place au cours de nos toutes dernières réunions, c'est à dire les 14 éléments de Dublin Core, avec leurs raffinements éventuels, ainsi que les attributs OLAC. Nous avons ajouté à chaque élément une description très sommaire de ce qui est attendu.
- une grille reprenant tous ces éléments de la feuille de description DC/OLAC avec de simples cases à cocher.

Ces documents ont été fournis à chacun (merci Serge !) pour pouvoir débattre plus librement sur les choix que nous avions faits. Vous recevrez tous dans mon prochain mail ces deux documents, avec les améliorations que nous avons apportées, pour que nous puissions réaliser ce travail sur les 11 autres corpus.

Je vais détailler ci-dessous les 14 éléments de description que nous avons passés en revue, et principalement ceux qui ont pu poser problème. Je reprends ici la description sommaire proposée pour chacun de ces éléments. Lorsqu'il y a eu discussion et que celle-ci a abouti à une clarification de la description, je me suis permis de l'insérer dans notre feuille de description.

Rappelons que ces feuilles de méta-données sont théoriquement écrites en anglais ... Puisque nous remplirons (pour l'instant) toutes ces descripteurs en français (pour les cas où il faut insérer du texte libre, comme par exemple pour description), il sera nécessaire de le préciser en tant que commentaire dans le document XML.

Précisons également que cette fiche de Méta-Données sera doublée : une pour le fichier son/video, et une seconde pour l'annotation. Ces fiches seront par définition très similaires et nous ne nous occupons pour l'instant que de celle décrivant le son.

1. **Title :** nom donné à la ressource, (celui par lequel elle est connue officiellement). Ce titre est unique même si un corpus a été divisé en petits groupes.

2. **Subject :** sujet du contenu de la ressource, (pour gloser, le sujet/but pour lequel pour le corpus a été construit) décrit par un ensemble de mots clés, de phrases ou d'un code de classification. Utilisé avec *Subject* pour identifier une ressource en tant que genre particulier.

   ⇒ Peut-être complété par les extensions OLAC suivantes :

- **discourse-type :** fournit un vocabulaire contrôlé pour identifier environs dix types de discours différents. (interactive_discourse ; language_play ; narrative ; unintelligible_speech ; ...). Cette extension n'est pas vraiment adaptée pour **Subject.** C'est à dire qu'il faudrait que ce soit un corpus construit POUR étudier la narration, le discours interactif. *Or, il est logique de penser que l'on utilise plus souvent la narration, le discours interactif pour y étudier un point précis*

- **language :** fournit des codes pour identifier toutes les langues connues, à la fois vivantes et disparues. Nous utiliserons le code ISO 639-1 à 2 lettres : le + simple, mais qui point vers le SIL's Ethnologue à trois lettres. (préfixe: x-sil-) lorsqu'une langue n'y est pas référencée. Pour remplir simplement cette extension, il sera nécessaire de proposer un menu déroulant puisqu'on ne peut bien évidemment pas demander à qui que ce soit de retenir ces codes ISO. Pour le moment, nous changerons manuellement pour inscrire le code correspondant (français --> fr)

- **linguistic-field :** ces codes décrivent le contenu d'une ressource comme relevant d'une sous-catégorie particulière des sciences du langage. (**discourse_analysis : phonetics : phonology : pragmatics, psycholinguistics, semantics, sociolinguistics, syntax, text and corpus linguistics).** Ce champ est simple à remplir si l'on considère qu'il vaut mieux qualifier très largement le corpus. C'est à dire s'il y a ambiguïté, il n'est pas dérangeant de remplir plusieurs champs même s'ils ne correspondent pas parfaitement... Ce raisonnement est choisi de façon pragmatique puisque l'on considère que ces méta-données servent avant tout à créer des intersections pour permettre de retrouver les corpus. Patrick Renaud avait mentionné certaines lacunes dans cette liste, on pourra en rediscuter à la prochaine réunion.

3. **Description :** 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).

4. **Publisher :** une entité responsable de la diffusion de la ressource, dans sa forme actuelle. Pour nous, ce sera toujours l'ED 268.

5. **Contributor :** une entité qui a contribué à la création du contenu de la ressource.

   ⇒ Peut être complété par les extensions OLAC suivantes : **role, annotator, author compiler consultant, data_inputter, depositor, developer , speaker, sponsor, transcriber, translator.**
   Par souci de normalisation, il est impératif d'inscrire les noms de la façon suivante : Nom, Prénom
   S'il s'agit d'un contributeur qui a tout fait lui-même, on peut par souci d'économie indiquer le nom sous « author » et indiquer en commentaire cette convention de notation.

6. **Date :** une date associée à un événement dans le cycle de vie de la ressource. norme iso et donc raisonnement identique à **Language** plus haut.

7. **Type :** la nature ou le genre du contenu de la ressource. Par opposition à Subject (puisque les extensions OLAC sont très similaires à **Subject**), il s'agit du type de données utilisées, par exemple l'extrait, le texte, ou la série de phrases prononcé par le locuteur ;
   ⇒ L'élément « type » peut-être complété par les extensions OLAC suivantes :

   - **discourse-type :** fournit un vocabulaire contrôlé pour identifier environs dix types de discours différents.
     - **interactive_discourse : language_play : narrative** : **unintelligible_speech :**
   - **language :** identique à Subject.
   - **linguistic-type :** fournit une classification de la nature de la forme de la ressource d'un point de vue linguistique.
     - **lexicon :** la ressource incluse une liste systématique des items lexicaux. uniquement pour des listes de mots
     - **primary_text :** matériel linguistique qui consiste en lui-même au sujet de l'étude. Presque systématiquement coché, même pour de la parole spontanée.

- **language_description :** la ressource décrit une langue ou quelques aspects d'une langue, à travers une documentation systématique des structures linguistiques. Par exemple un article de linguistique... !

8. **Format :**

Pour Format, la fiche de description sera à fortiori différente lorsqu'elle décrit le son ou l'annotation.

- **Medium :** la matérialisation physique ou digitale de la ressource. DC suggère d'utiliser une valeur de type MIME. Comme pour les normes ISO dans language, nous changerons manuellement pour inscrire le code correspondant ... pour le moment. A noter que les types nous sont peu familiers et ne comprennent pas le format WAV par exemple, fred s'en occupe
- **Durée :** en secondes, à faire de manière précise, je peux m'en occuper.

9. **Identifier :** référence non ambiguë à la ressource dans un contexte donné. Une URL.
10. **Source :** référence à une ressource à partir de laquelle la ressource actuelle a été dérivée, par exemple des DATs. En général, si cette ressource est introuvable, il est préférable de ne rien indiquer. Dans la grille d'exemple que je vous enverrai, nous avons décidé de laisser la mention « DATs introuvables » pour bien comprendre l'information requise.
11. **Language :** la langue du contenu intellectuel de la ressource. Convient donc à l'annotation plutôt qu'au son, contrairement à **Type**.

    ⇒ Peut être complété par l'extension OLAC suivante :
    - **language :** (voir ci-dessus)

12. **Relation :** référence à une autre ressource qui a un rapport avec cette ressource.

13. **Coverage :** la portée ou la couverture spatio-temporelle de la ressource.

Nous n'avons pas complètement compris ces extensions ...

- **spatial :**
- **temporal :**

14. **Rights :** formation sur les droits sur et au sujet de la ressource. Nous en avions déjà parlé, ce champ est optionnel, certains préfèrent insérer un Copyright sans que cela change grand chose et d'autres n'insèrent rien ce qui n'indique pas une absence de Copyright. Quelle que soit notre décision, il est possible d'indiquer « accès restreint » ce qui permet de visualiser la fiche de description, mais impose de contacter le « **Publisher** » pour pouvoir éventuellement obtenir plus d'informations.

2- Présentation de quelques logiciels utilisables pour l'annotation de corpus.

Nous avons passé beaucoup de temps sur le premier point qui était essentiel. Luiggi et Michel n'ont pas vraiment eu le temps de nous présenter le travail qu'ils avaient préparé. Pour la prochaine fois donc ... Rien d'urgent pour l'instant.

Je rappelle que le but de cette étape est de proposer aux membres de l'ED268 (étudiants, chercheurs) des outils adaptés pour l'annotation linguistique de corpus. La première communication sur ce sujet aura lieu le samedi 21 mai pour les rencontres de l'Ecole Doctorale à l'ILPGA puisque nous comptons faire une communication orale sur le projet. Les savoir-faire de la rentrée 2005 incluront, je l'espère, un ou deux logiciels supplémentaires ; ceux que nous aurons considéré comme particulièrement utiles, et qui seront donc mentionnés lors des rencontres de l'Ecole Doctorale.

3- Présentation de l'enregistreur Marantz

J'ai finalement présenté notre enregistreur numérique MARANTZ (acheté sur un contrat du labo de Phonétique) et j'ai fait une petite démonstration. Il est très simple d'utilisation et son principal intérêt réside dans la possibilité de transférer l'enregistrement, tel un simple copier/coller sur un ordinateur. Malgré tout, il est très

coûteux (1000€) et je ne suis pas sûr que l'on pourrait le prêter librement à tous les étudiants souhaitant faire du corpus. Quoi qu'il en soit, son prix nous limite également dans le nombre d'appareils que l'on pourra acheter. Une solution intermédiaire proposée par Thierry s'avère intéressante. Il utilise une nouvelle génération de Mini-Discs SONY (200€) qui permettent d'une part d'enregistrer en format non compressé, et d'autre part de transférer l'enregistrement exactement comme avec le Marantz. Le seul inconvénient est que le format de sortie du SONY est un format propriétaire et qu'ils ne proposent pas actuellement de sortie en format libre non compressé (WAV, AIFF, ...). De petites applications « artisanales » permettant cette conversion, peuvent être téléchargées sur internet, mais il paraît encore difficile de proposer publiquement cette solution sur le site web du projet ...

4 – Programme pour la prochaine réunion

- Nous avons décidé de laisser quelques semaines afin que chacun puisse compléter la feuille de description que je vous envoie dans mon prochain mail ... La présentation de plusieurs descriptions de corpus sera donc le point principal du programme de la prochaine réunion.

- Michel et Luiggi se seront concertés d'ici à la prochaine réunion et pourront nous proposer la suite de notre discussion sur les logiciels d'annotation linguistique à conseiller.