Bonjour à tous,
Voici le compte-rendu de la réunion du projet innovant du 1er avril 2005 (idem en fichier attaché)

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

Bien amicalement,
cedric

##########
DATE DE PROCHAINE REUNION….
Le vendredi semblait être à nouveau la seule journée qui pourrait convenir à tout le monde...
Je vous propose le vendredi de la rentrée des vacances du printemps, soit si je ne me trompe pas le vendredi 13 mai

J'espère que cette date ne fera peur à personne ; et plus sérieusement conviendra également à ceux qui n'étaient pas présents. N'hésitez pas à me prévenir si ce n'est pas le cas.
##########

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

- je vous envoie la contribution que nous (Fred et moi-même) avons envoyée aux prochaines rencontres de l'ED. Nous allons également d'essayer d'en faire une sorte de session plénière, permettant de sensibiliser le plus grand nombre au travail qui est proposé ici ... La linguistique de Corpus est un domaine très en vogue à l'heure actuelle et fera l'objet d'un cours spécifique en TAL : L6F01 (Linguistique de corpus ; oral et écrit) .
- Serge et André ont été contactés par François Daoust d'une université Canadienne en visite en France pour le projet SATO :
- *Le «Réseau d'échanges de ressources, de connaissances et de méthodologies en analyse de texte assistée par ordinateur» vise à développer les conditions pour une mise en commun de nos ressources et de nos méthodes à des fins d'enseignement et de recherche dans le domaine de l'analyse de corpus textuels. Le projet s'articule de façon prioritaire autour de trois volets de convergence technologique : un volet méthodes et expérimentation, un volet normalisation XML des formats de documents électroniques et un volet terminologie. (http://www.ling.uqam.ca/sato/index.html)*

Le projet que nous mettons en place les intéresse au plus haut point. Le format normalisé que nous avons choisi (XML) pour faciliter ces échanges montre ici toute son utilité.

- Le 1er mai à la BNF, présentation du guide des corpus oraux ; description de la juridiction sur les corpus oraux. N'hésitez pas à y aller (pour + d'infos, demandez à Michel, j'ai trouvé cette présentation sur internet où l'on retrouve Michel en bonne place :
  http://users.info.unicaen.fr/~anne/siteDialint/7baude.pdf
- Nous n'avons pas eu le temps d'aborder le dernier point sur les outils d'annotation du signal. Pour éviter ce problème, ce sera le premier point abordé lors de notre prochaine réunion .... et Michel nous transmettra par mail son tableau explicatif.

##########

Nous avions prévu de passer en revue quelques corpus dont celui de Maria (BDD8) qui m'avait laissé ses commentaires, et celui de Thierry (BDD9). Ensuite nous avions prévu de parler de MKM (MakeMetaData) et des problèmes rencontrés avec ce dernier.
En fait cette fois-ci, nous n'avons pas passé en revue tous les points de la feuille de Méta-Données (Dublin Core + OLAC), mais sommes directement passés aux problèmes, et avons mélangé les remarques avec l'utilisation de MKM. Mon compte rendu sera donc moins chronologique que les fois précédentes, et regroupera plutôt les thèmes les plus importants.

- 1. La fiche de Méta-Données Dublin-Core/OLAC (Subject ; Relation ; Format ; Contributor ; Source ; Title)

- 2. Description du fichier ANNOTATION :
- 3. MAKE METADATA (MKM) : Remarques et modifications possibles ! (Elément « description » du Dublin Core », Iteration des champs de réponses, ajout de catégories, nouvelle appellation de certains éléments)
- 4. Catalogage des Méta-Données
- 5. Programme pour la prochaine réunion

1. LA FICHE DE META-DONNEES (DUBLIN CORE / OLAC)

Nous avons à nouveau répondu à quelques points qui étaient sensibles lors de la dernière réunion sur **la fiche de Méta-Données (Dublin Core / OLAC) :**

- L'élément DC « **Subject** » : le sujet/but pour lequel pour le corpus a été construit.
    - Par exemple, si l'on se sert d'un dialogue pour valider les réalisations phonétiques qui y apparaissent, l'objet d'étude est bien la langue et sa réalisation phonétique et non le dialogue. Par conséquent dans l'élément « Subject » > extension « Discourse type », la case « interactive discourse » ne sera pas cochée.... Par contre, elle sera cochée dans l'élément DC « Type ».
    - Je rappelle l'explication de « Subject » fournie par Fred et proposée dans le compte-rendu précédent.
        - **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.
        - **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*

- L'élément DC **Relation** :
    - Si je le mentionne ici, c'est qu'il est tentant d'utiliser cet élément pour faire référence aux groupes pré-découpés d'un grand corpus (+ voir élément « Title ») ....
    - En fait, il est réellement utilisé dans le cas de versions différentes, comme par exemple, un fichier sonore en format .mp3 et un second en format WAV. Ou un étiquetage en format txt et l'autre en format XML. Un point qui permettra peut-être de clarifier la notion de RELATION ; n'oubliez pas que chaque fiche ainsi constituée avec MKM devra être reliée à une URI, par définition unique ...
    - Par contre n'oublions pas que l'annotation ayant sa propre fiche de méta-données, la relation entre annotation et fichier sonore devra être précisée.

- L'élément DC **Title :**
    - nom donné à la ressource, (celui par lequel elle est connue officiellement). Pour ceux dont le corpus a été divisé en petits groupes... Je propose une convention, qui n'est peut-être pas une très bonne idée , j'attends vos réactions ...

  Pourquoi ne pas indiquer dans le titre une forme de relation en utilisant une partie commune « Récit Boucle d'Or » puis indiquer entre parenthèses.... partie1, partie voyelles .... Qu'en pensez vous ?

- L'élément DC **Format** :
    - Prenons exemples de formats connus : WAV et MP3, (le MP3 pour des raisons de place peut-être mis à disposition sur un serveur, tout en gardant précieusement la version WAV bien sûr). Ces formats ne correspondent pas aux types MIME qui permettent de décrire le format du fichier sonore.... Ces informations peuvent être insérées dans Source par exemple (voir ci-dessous). Mais de toute façon, fournir des informations telles que la Fréquence d'échantillonnage ou le codage n'indique rien sur la qualité du fichier son ... puisque par exemple, il est possible de numériser à 44100Hz un fichier enregistré au préalable avec un magnétophone de très mauvaise qualité ! Les formats de type MIME seront insérés (fournis à l'utilisateur) dans la prochaine version de MKM.

- L'élément DC **Contributor :**
    - ⇒ Peut être complété par les extensions OLAC suivantes : **role, annotator, author compiler consultant, data_inputter, depositor, developer , speaker, sponsor, transcriber, translator.**
    - S'il s'agit d'un contributeur qui a tout fait lui-même, on pourrait par souci d'économie indiquer le nom sous « author » et indiquer en commentaire cette convention de notation., ce qui s'avère trompeur en fait ! Mieux vaut laisser le travail à l'utilisateur ... qui devra cocher toutes les cases nécessaires.
    - En ce qui concerne le nom des locuteurs .... l'anonymat est de rigueur pour certains enregistrements ... Mieux vaut malgré tout conserver dès le départ cette information pour la rendre anonyme ensuite plutôt que l'inverse !

- L'élément DC **Source :**
    - **Rappel :** 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, nous avons décidé de laisser la mention « DATs introuvables » pour bien comprendre l'information requise.
    - Il est important de préciser que pour les cas où vous devez écrire votre réponse, vous avez la possibilité de décrire ... même s'il n'y a qu'une case. Il est parfaitement possible d'y insérer un texte long donnant des précisions supplémentaires. Ne vous fiez donc pas à la petite taille de la case ...

2. Description du fichier ANNOTATION :

Je rappelle que chaque fiche de Méta-Données sera doublée : une pour le fichier son/video, et une seconde pour l'annotation. Ces fiches seront à priori très similaires .... Voici quelques détails supplémentaires.

Le même MKM sera utilisé pour le son et son annotation ; à l'utilisateur d'enregistrer ces deux fichiers sous 2 noms différents mais complémentaires. Quelques précisions seront à rajouter dans la documentation locale pour préciser que ce point n'est pas pertinent pour décrire une annotation (par exemple ... publisher,). ... et inversement !
Actuellement le manque de champs pertinents peut être compensé an ajoutant toute l'information que l'on (le responsable du corpus) souhaite dans l'élément « Description ». Dans le moteur de recherche qui sera mis en place pour faire une recherche préliminaire sur ces corpus ... il sera tout de même possible de retrouver toutes ces informations par mots-clés ... Mais puisque ces informations ne seront pas catégorisées, il ne sera pas possible d'aller plus loin que cette recherche par mots-clés. Généralement parlant, l'élément « Description » semble être souvent utilisé comme la « poubelle » des méta-Données :-)

Les formats de type MIME (donc de l'élément DC « Format ») par exemple seront sous-informatifs ... mais il en va de même pour le son : ils restent très vagues; les grilles de type Praat ou Clan seront considérées comme du « Plain text » . Leur format propriétaire exact devra être mentionné ailleurs si on le souhaite, il peut de toute façon être obtenu lorsqu'on accède aux données.

Nous avons eu une petite discussion concernant la normalisation de l'annotation. Comme je l'ai souvent mentionné dans les précédents compte-rendus. Un problème important sera l'énorme diversité des outils d'annotations utilisés par les responsables de corpus (en plus des méthodes d'annotation).
« La TEI (Text Encoding Initiative) est un projet international visant à mettre au point des directives pour l'élaboration et l'échange de documents électroniques à des fins de recherche érudite, et pour répondre aux besoins les plus variées des industries de la langue en général. Ce qui est le plus réutilisable du TEI, c'est plutôt sa façon générique de décomposer un document en une séquence de trois grandes parties ainsi qu'en définissant une série d'éléments pouvant être insérés dans n'importe laquelle de ces parties (préliminaires, corps, post-liminaires) »
source : http://www.autoroute.gouv.qc.ca/publica/normes/norme44.htm
La conversion vers une structure organisée telle que celle proposée par la TEI est aux marges de notre projet, mais la conversion de certains fichiers d'annotation en un format XML simple est déjà possible. La question suivante se pose : Doit-on mettre deux ou plusieurs versions (TextGrid et XML par exemple), ou doit-on n'en laisser qu'une, en mettant à disposition un outil qui permet la conversion de grilles on-line (comme sur la page de Xi-tools par exemple.)... la deuxième solution s'impose pour ne pas encombrer la BDD !

3. MAKE METADATA (MKM) : Remarques et modifications possibles !

Conseils en vrac

- Bien lire la notice !!! Il est évident que bon nombre de réponses aux questions qu'il est légitime de se poser se retrouvent dans la documentation. Peut-être serait il utile d'insérer une « Foire Aux Questions » reprenant les points les plus évidents de la documentation ...
- En ce qui concerne la taille du logiciel sur l'écran (version .exe), le problème peut-être résolu en augmentant la zone de l'écran dans les propriétés de l'affichage ... Je crois qu'il faut au moins 800 pixels ! Je rappelle que MKM est également disponible en version HTML, comme il est coutume de faire . Merci Serge !

Iteration de champs

Rappelons que ces feuilles de méta-données sont théoriquement écrites en anglais... Puisque nous remplirons (pour l'instant) toutes ces informations en français (pour les cas où il faut insérer du texte libre, comme par exemple pour description), il sera nécessaire d'améliorer ce point, et d'avoir :

a. la possibilité fournir l'information en plusieurs langues
b. en indication la langue utilisée pour fournir l'information

Certains champs pourront ainsi être complétés également en anglais pour plus de possibilités d'échange. Comme je l'avais précisé dans mes précédents compte-rendus, tous (ou presque) les éléments Dublin Core / OLAC doivent pouvoir être réitérés. Serge combinera ces deux points dans MKM pour donner à l'utilisateur la possibilité d'insérer plusieurs réponses, et ce en indiquant la langue qu'il utilise.
Serge, qui va beaucoup plus vite que moi a déjà commencé ses modifications, pour que bon nombre d'éléments puissent être multipliés (possibilité d'inscrire plusieurs réponses : principe d'itérativité) et/ou traduits en anglais.
...
> 2. Possibilité de construire plusieurs métadonnées du même type (par exemple, plusieurs éléments de type
>"title" dont la description serait écrite dans des langues différentes).
> Cette mise à jour est disponible uniquement sur l'élément title (dans cette version 1.05), mais elle le sera sous
>peu (version 1.06) sur tous les éléments pour lesquels la multiplication des métadonnées est possible.
> Pour réaliser la multi-description d'une métadonnée (title pour le moment), dans l'onglet contenant cet élément,
>il y a désormais un bouton "EDIT" donnant accès à un éditeur permettant (via un système d'onglet) de
>construire plusieurs descriptions pour une même métadonnée (avec possibilité pour chacune de coder la langue
>d'écriture)

Ajout de descripteurs

- L'élément DC 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).

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 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 !

Nouvelle appellation de certains éléments

Pour un grand nombre d'étiquettes Dublin Core (ou OLAC), nous avons décidé de les renommer (masquer) dans les 2 premières colonnes de MKM, puisqu'elles ne sont pas toujours intuitives (nécessitant alors de consulter la doc), voire même trompeuses (comme « Subject » par exemple).
Ces étiquettes seront en fait remplacées par des gloses en français, mais également en anglais.
Attention, je précises qu'elles seront masquées ... le fichier XML résultant contiendra ces étiquettes puisqu'il s'agit d'un format normalisé, elles sont donc indispensables !

4. CATALOGAGE :

Michel a évoqué 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 ...

Michel a insisté nécessité d'un identifiant (Identifier : référence non ambiguë à la ressource dans un contexte donné) selon les principes de l'OAI. Voici un petit résumé pris sur un site fourni par Serge :
http://www.figoblog.org/document566.php

« Dans cet article, les auteurs abordent la problématique de l'utilisation de l'OAI quand on veut non pas se contenter d'échanger des métadonnées, mais échanger les ressources elles-mêmes. Parmi les problèmes soulevés, il y en a un qui m'est cher en ce moment : la difficulté de faire correspondre les métadonnées et les identifiants avec la localisation réelle de la ressource. Enfin le propos est d'utiliser l'OAI pour échanger des formats de métadonnées complexes, comme METS et MPEG21, qui permettent à la fois de localiser précisément toutes les parties d'une ressource, et de connaître toutes les modifications qui l'affectent. Le protocole OAI rejoint alors le modèle OAIS, deux standards qui à part ça et malgré leur ressemblance phonétique n'ont rien à voir entre eux. »

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 ...)

5 – Programme pour la prochaine réunion

- Michel et Luiggi nous proposeront une discussion sur les logiciels d'annotation linguistique à conseiller. Nous commencerons par ce point.

- Nous avons décidé de laisser quelques semaines afin que chacun puisse utiliser librement MKM sur son ou ses corpus ... La présentation de plusieurs descriptions de corpus sera à nouveau au programme de la prochaine réunion.

- Thierry nous proposera la semaine 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 de IMTI 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 !

- ... suite de la discussion sur le catalogage ...