Voici le compte-rendu de la réunion du projet innovant du 14 décembre 2004

Etaient présents Frédérique Bénard, Maria Candéa, Serge Fleury, Cédric Gendrot, Michel Jacobson, Thierry Pagnier, Pollet Samvelian.

Voici les points principaux qui ont été abordés :

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

Nous avions discuté lors de la précédente réunion de l'utilité de montrer aux différents membres de l'ED (susceptibles de nous fournir des corpus) quels types d'outils ils pouvaient utiliser pour annoter leurs corpus, afin que ce soit un travail plus simple, rapide et efficace pour eux (et pour nous également).
L'objectif principal étant que plus aucun étudiant n'annote son corpus sous Word, et en perdant donc toute l'information temporelle qui nous est si chère.
Michel nous a mentionné l'existence d'une page web intitulée XI-tools qui fait un inventaire non exhaustif de quelques outils d'annotation du signal. Ce travail avait été réalisé lors d'un précédent contrat dans lequel Michel est impliqué. Michel et Luiggi se proposent donc de nous en faire un petit compte rendu dès la prochaine réunion. Il serait également possible de le compléter … Ce sera donc au programme de la prochaine réunion, enfin si nous disposons d'assez de temps, puisque comme vous pourrez le constater, le programme sera chargé !
En attendant, Serge a mis un lien vers cette page web sur notre site « projet innovant » dont je vous rappelle l'url … Je vous invite à jeter un œil, il commence à contenir beaucoup d'informations …

http://pi-ed268.univ-paris3.fr/ffiles/index.html (public)
http://pi-ed268.univ-paris3.fr/files/index.html (privé avec log in)

2- Réception de nos deux derniers corpus

Nous avons stoppé la réception des corpus pour commencer notre travail de description et de normalisation de ces corpus.
Cependant je dois vous préciser l'arrivée de 2 nouveaux corpus promis par 2 de nos membres, qui ont fini par céder à mes pressions amicales, mais néanmoins fermes ! :-)

- corpus persan, fourni par Pollet Samvelian, il comprend un certain nombre de phrases lues par 2 locutrices … et étiquetées sous Word … Son intérêt réside dans les gloses morphologiques qui accompagnent l'annotation, et que nous n'avions pas dans nos précédents corpus.
- Corpus vidéo fourni par Patrick Renaud. Il comprend un interview audio-vidéo annoté sous CLAN. Son intérêt est double puisqu'il réunit ces deux caractéristiques que nous n'avions pas jusqu'ici : la vidéo et l'utilisation de CLAN pour l'annotation.

Merci donc à Pollet et à Patrick de nous avoir fourni des corpus qui apporteront un peu plus de diversité dans les données que nous possédions jusqu'ici …

3- Description de nos corpus …

Voici le point principal de notre dernière réunion ; il s'agit également d'un des points d'intérêt principaux de notre projet …
Il s'agit donc de décrire au mieux chacune des données que nous avons reçues, et à partir desquelles nous avons décidé de réaliser notre travail… Nous avons passé en revue les questions suivantes : Quels sont les éléments descripteurs que nous pouvons utiliser pour « cataloguer » nos données ? Jusqu'à quel niveau de précision doit-on aller ?

Nous avions décidé d'utiliser une feuille de DUBLIN CORE (DC), qui permet de décrire des ressources de manière structurée grâce au format XML. Je vous rappelle qu'un de nos objectifs est qu'un utilisateur puisse visualiser les données qui correspondent à tous ces critères descriptifs, qui agissent ensuite comme un simple moteur de recherche.

Les **éléments** descripteurs de Dublin Core ne sont pas excessivement nombreux (13), mais peuvent apporter une description plus détaillée au moyen de **raffinements**. Afin d'affiner plus encore nos descriptions, nous avons décidé d'utiliser OLAC. …
pour reprendre la traduction de la présentation d'OLAC réalisée par Frédérique, en voici une brève présentation :

« OLAC (Open Language Archive Community) est un partenariat d'institutions et d'individus qui crée des librairies virtuelles de ressources linguistiques au niveau mondial. Les métadonnées de l'OLAC sont une spécialisation des métadonnées du Dublin Core, et cet article décrit un cadre de travail d'interopérabilité qui valide, distribue et combine ces métadonnées. Les préoccupations de la communauté de l'OLAC sont la création et le maintien de standards et de pratiques dans trois domaines : les métadonnées, l'interopérabilité et les procédures. OLAC espère montrer comment une communauté spécialisée peut répondre à ses besoins de recherche à partir du Dublin Core. »

OLAC ne vise pas à ajouter des éléments à la liste fournie par DC, mais plutôt à la spécialiser. Sous un certain nombre d'éléments seront donc proposés des **attributs** (qui diffèrent donc des **raffinements**) et permettent quand c'est nécessaire de compléter une description trop réduite pour nos types de données.

Notre travail de la précédente réunion consistait à passer en revue une liste classique d'éléments du DUBLIN CORE, de vérifier leur utilité dans notre cas, de s'assurer de la non ambiguïté de ces termes, et d'ajouter au besoin les attributs proposés par OLAC. Voici un par un les éléments que nous avons passés en revue et les commentaires qui en ont été faits …..
(EGALEMENT COMPLETE PAR FREDERIQUE qui je vous le rappelle fera son mémoire de maitrise sur ce projet). Ces points seront développés par Frédérique et Michel puisque leurs versions d'OLAC étaient discordantes, et nous n'avons pu statuer définitivement sur certains points ! Ils nous présenteront leurs choix à la prochaine réunion …

1/TITRE :
il s'agit du titre simple (fourni par les contributeurs du corpus … on ne parle plus de « créateur », voir ci-dessous). Il est possible d'ajouter comme raffinement un « titre alternatif » dans une autre langue. Le nombre de raffinements doit pouvoir être récursif. Je dois préciser dès à présent, et ceci est valable pour tous les éléments, que la langue par défaut est l'anglais, et qu'un certain nombre d'éléments et de raffinements doivent pouvoir être complétés dans plusieurs langues (en français notamment).

2/ CREATOR
Dublin Core conseille désormais de le supprimer … Plusieurs contributeurs, avec des rôles différents, sont souvent à l'origine d'un corpus et permettent de détailler de manière plus précise ce point.

3/ SUBJECT
Quelques mots-clés …. DC conseille de choisir parmi une liste fixe . Nous n'en aurons pas à priori, et quelques mots clés seront choisis par l'auteur de la description (ci-dessous) … comme cela se fait pour un mémoire, une thèse, etc…
OLAC propose de rajouter l'attribut « subject language » : la langue que décrit la feuille de DC. Cet aspect sera complété par l'élément « Language » plus loin.

4/ DESCRIPTION
Il s'agit d'un résumé et/ou d'une table des matières.

5/ PUBLISHER
Nécessairement l'ED 268, même si ce corpus est distribué ailleurs …

6/CONTRIBUTOR
Nous utiliserons ici une liste contrôlée, définie par OLAC, avec les différents rôles des différents contributeurs potentiels.

7/ DATE
Il existe un nombre important de raffinements pour définir de quelle date il s'agit. Nous les laisserons tels quels tout en sachant qu'une grande partie ne sera pas complétée. Cet élément me permet de signaler qu'il n'est pas toujours aisé de compléter une feuille de DC, et comme c'est le cas pour le site du LACITO, les contributeurs ne peuvent vraisemblablement pas remplir les différents éléments sans l'aide d'une personne « compétente ».

8/ TYPE
Nature ou genre du contenu (vidéo, audio … pour les plus grandes catégories). OLAC propose un certain nombre d'attributs qui visent à préciser la nature du contenu … plus principalement axé sur la linguistique. Michel et Fred s'en occupent !

9/ FORMAT
Là encore OLAC propose un certain nombre d'attributs qui visent à préciser la nature du contenu … Michel et Fred s'en occupent !

10/ IDENTIFIER URL
Lieu sur le serveur de chaque fichier décrit ….

11/ SOURCE
Cet élément ne peut-être complété que si le support physique qui a « permis l'enregistrement » est disponible et bien référencé. Par exemple, un mini disc qui contient l'enregistrement original est la source d'un fichier décrit, si bien sûr ce mini disc a été conservé, et qu'on sait où le retrouver.

12/ LANGUAGE
Encore une fois, OLAC propose un certain nombre d'attributs qui visent à préciser la nature du contenu …. Michel et Fred s'en occupent !

13/ RELATION
Ce dernier point a mené à une longue longue discussion … Pour reprendre depuis le départ, je dirais que les corpus qui nous ont été communiqués sont au nombre de 13. Cependant, la quantité de ressources contenues dans un seul et même corpus peut être excessivement variable et diversifiée. Considérer un corpus comme un élément unique et indissociable parce qu'il a été fourni par une seule personne est donc purement arbitraire…. Et ceci doit être reconsidéré! Jusqu'où peut-on aller dans la description de ces corpus …. ? Réponse, chaque fichier doit être repéré comme distinct (étiquetage/annotation et fichier audio/vidéo sont également considérés comme distincts). Il y aura donc autant de fichiers que de descriptions. Lorsqu'il y aura une recherche effectuée, pourront alors apparaître seuls quelques fichiers d'un ou de plusieurs corpus, ou bien tous les fichiers d'un seul corpus, en fonction de la requête.

Je vais prendre quelques exemples pour détailler ce dernier point :
- Le corpus fourni par Patrick Renaud contient un long fichier video avec deux fichiers d'annotation ; ce qui fait donc trois fichiers à décrire. Il s'agit d'une interview que je n'ai pas écoutée … mais si Patrick Renaud considère qu'elle forme une ressource indissociable, nous la conserverons telle quelle, et une recherche ne pourra aboutir qu'à l'intégralité de ce fichier .
- Le corpus fourni par Thierry Pagnier contient un long fichier audio avec un fichier d'annotation ; ce qui fait donc deux fichiers à décrire. Il s'agit d'une interview qui comprend plusieurs parties bien distinctes qui requièrent des descriptions différentes concernant certains éléments comme la « description » ou le « type ». Il est possible de découper physiquement ce fichier en fonction des thèmes, mais il est également possible d'attribuer des indices temporels à chacune de ces parties (sans avoir à découper ce fichier donc). Il est alors possible d'extraire automatiquement le morceau concerné. Dans tous les cas, c'est au contributeur principal du corpus de faire ce choix.
- Un certain nombre de corpus qui m'ont été fournis (ou que j'ai fournis) sont des corpus typiquement phonétiques/ phonologiques (evidemment ! !) … et nous avons souvent l' (a mauvaise) habitude de découper tous nos fichiers en phrases pour les analyser finement ensuite. Cela fournit parfois un très (très) grand nombre de phrases … Pour ces corpus, je me charge de les regrouper en un ensemble plus raisonnable de phrases. Un fichier d'annotation viendra alors mentionner le début et la fin de chaque phrase et re-créera ce découpage … en effet les phrases auparavant découpées seront alors collées (concaténées) les une aux autres. (je fais cela automatiquement, je ferai au passage le corpus de Pollet qui correspond tout à fait à ce que je viens de décrire, si elle n'y voit pas d'inconvénient bien sûr)

Pour conclure sur l'élément relation, les différents fichiers appartenant au même corpus seront décrits quant à la relation qu'ils ont entre eux …. Michel et Fred nous détailleront ce point …

6 – Programme pour la prochaine réunion

- Présentation par Michel et Frédérique du fruit de leur travail … comment OLAC peut-il nous aider à décrire nos données ?

- Il nous sera rapidement indispensable d'utiliser ou de créer un outil qui permette d'insérer tous ces éléments descripteurs simplement (en complétant des boites de dialogue par exemple) et ainsi écrire le fichier XML correspondant …. En effet, les éléments de description sont pour l'instant insérés dans des tableaux et ce format n'est ni ergonomique, ni économique. Serge a repéré outil qui pourrait correspondre à ce que nous recherchons i.e. KEPLER. Il nous en fera une description lors de la prochaine réunion.
- Je vous montrerai, s'il est arrivé, l'enregistreur numérique Marantz que nous avons commandé.
- Michel et Luiggi nous présenteront quelques outils d'annotation du signal (si assez de temps)

QUESTIONS SUBSIDIAIRES pour la prochaine fois :
- Pourra-t-on se limiter à OLAC pour compléter notre description …. Que faire/penser de la TEI (text encoding initiative) par exemple ?