Université Paris 3 - Sorbonne Nouvelle

Projet innovant 2005-2006 :

Propositions de normalisation pour

une base de corpus multimédia à l'ED 268

Rapport scientifique et financier

Equipe initiale du projet :

Serge Fleury
- Maître de Conférences à Paris 3/I.L.P.G.A. (Institut de Linguistique et de Phonétique Générale Appliquées). serge.fleury@univ-paris3.fr
- Equipe de recherche : Centre de Lexicométrie et d'Analyse Automatique des Textes (CLA^{2}T) / SYLED

André Salem
- Professeur à Paris 3/I.L.P.G.A. salem@msh-paris.fr
- Equipe de recherche : Centre de Lexicométrie et d'Analyse Automatique des Textes (CLA^{2}T) / SYLED

Michel Jacobson
- Ingénieur d'études au LACITO (laboratoire LAngues et Civilisations à Traditions Orales, – Equipe de recherche : CNRS UMR 7107) jacobson@idf.ext.jussieu.fr

Patrick Renaud
- Professeur à Paris 3/I.L.P.G.A. prenaud@univ-paris3.fr
- Equipe de recherche : LACITO - EA 1483 Observatoire du français contemporain

Pollet Samvelian
Maître de Conférences à Paris 3/I.L.P.G.A. pollet.samvelian@univ-paris3.fr
- Equipe de recherche : UMR 7528 Monde Iranien

Cédric Gendrot
A.T.E.R. à Paris 3/I.L.P.G.A. Doctorant : cgendrot@univ-paris3.fr
- Equipe de recherche : Laboratoire de Phonétique et de Phonologie – CNRS UMR 7018

Membres associés au projet :

Maria Candéa : Maître de Conférences à Paris 3 (Centre de recherches sur le français contemporain, EA 1483) Maria.Candea@Univ-Paris3.Fr
Sonia Branca : Professeur à Paris 3. (UPRES SYLED, EA 2290) branca@msh-paris.fr
Luiggi Sansonetti : Doctorant à l'I.L.P.G.A. (CLA^{2}T / SYLED / EA2290), associé à l'EA 1324 – CALIPSO) luiggi.sansonetti@noos.fr
Thierry Pagnier : Allocataire-moniteur à Paris 3. (UPRES SYLED, EA 2290) Thierry.Pagnier@Univ-Paris3.Fr
Frédérique Bénard : étudiante en maîtrise (M1) de TAL fred-benard@freesurf.fr

Mots-clés :

- **documentation (ou signalétique) :** informations générales sur le corpus. Le terme **métadonnées** est utilisé ici, notamment parce qu'elles sont représentées sous une forme bien définie, quantifiée et normalisée.

- **transcription :** La parole de chaque locuteur doit aussi être transcrite orthographiquement.

- **annotation :** annoter (ou enrichir) consiste à ajouter ou retrancher des informations à des données (Habert 2005). Dans notre projet, les annotations recouvrent des informations sur le contenu thématique, les locuteurs et la qualité acoustique (terme plus générique)

- **métadonnée :** une donnée décrivant une autre donnée (ici des corpus oraux).

- **normalisation :** dans ce projet, il est fait mention de normalisation à propos de l'utilisation de normes existantes pour la représentation des informations manipulées (XML pour le codage des documents électroniques par exemple)

Description du projet :

Le cœur du projet était l'élaboration d'une plateforme gérant une base de données multimédia (regroupant des données de langue de différentes natures), ainsi qu'une réflexion sur la normalisation pour l'encodage de corpus de langue (oral/vidéo). L'objectif était non seulement de pouvoir regrouper des données normalisées pour assurer la diffusion, l'échange et la pérennité de ces données, mais également de pouvoir les rendre visible aux membres de l'ED268 : ceux qui souhaitent intégrer leur corpus dans cette plateforme ou tout simplement ceux qui souhaitent utiliser les données stockées dans la base.

Sommaire :

I Introduction et résumé du projet de candidature

L'Ecole Doctorale 268[1] regroupe différentes disciplines des sciences humaines qui utilisent en permanence des corpus divers (textes, sons, vidéos) leur servant de base de travail pour leurs recherches.

L'utilisation de bases de données s'est répandue depuis quelques années puisque ces dernières permettent de regrouper les données de manière cohérente. Ce regroupement vise ainsi à récupérer les données stockées plus facilement, mais également à pouvoir les travailler plus facilement, et ce grâce à l'organisation qui est un des principes fondamentaux d'une base de données. L'innovation proposée par ce projet réside dans sa tentative de regrouper dans une approche ergonomique des données diverses afin d'une part (i) de laisser une possibilité aux chercheurs de mettre en commun leurs travaux et de favoriser ainsi les collaborations par un accès simplifié et efficace : **corpus interactionnels** (ii) de permettre aux étudiants de mieux consulter les applications et implications de leur(s) discipline(s) : **e-learning.** iii) de privilégier un travail en collaboration entre les différentes disciplines de l'ED268.

Les chercheurs réunis dans le cadre de ce projet partagent une expérience certaine dans la constitution et l'archivage de corpus de différentes natures. Un des objectifs du projet était de confronter des expérimentations diverses et de besoins émergents communs pour conduire à proposer des méthodes et des outils réutilisables par des chercheurs de disciplines diverses dont un des points communs est de travailler sur des corpus de langue. Pourvus d'un sens de l'adaptation important tant du point de vue humain que professionnel, nous souhaitions organiser une équipe de travail regroupant des chercheurs ayant des préoccupations scientifiques complémentaires dans le cadre de ce projet pluridisciplinaire.

Les participants au Projet innovant mesurent l'importance qu'a eu l'acceptation de ce Projet dans le déroulement de leurs travaux et sont reconnaissants du soutien apporté par le conseil scientifique de l'université. Le présent rapport est l'occasion de revenir sur les réalisations que le projet a permises au cours de l'année écoulée. Le travail mené autour de ce projet fut visible tout au long de sa progression sur un site web régulièrement mis à jour (http://pi-ed268.univ-paris3.fr). Un des objectifs de ce projet était de construire des ressources linguistiques normalisées, ces données étant elles aussi consultables sur le site associé au projet.

II Phase préliminaire : principes de mise en place de la base de corpus

II.1 Etat du recueil de données et mise en ligne sur un serveur

La mise en place d'un serveur constituait la première étape de ce projet. Il a été rapidement convenu que nous bénéficierions d'un espace disque sur un serveur. L'installation de cet espace a été réalisée par le CRI de Paris3. Cette opération nous a permis d'éviter l'achat d'un serveur et a soulagé considérablement notre budget, qui pouvait alors s'orienter sur d'autres

[1] http://ed268.univ-paris3.fr

aspects matériels à résoudre. L'utilisation du serveur de Paris 3 a également permis de disposer d'un gain de temps appréciable dans la conduite de ce projet.

Dans la candidature pour le projet innovant, il avait été fait mention de collecter des textes, sons et vidéos analysés par les différents membres de l'ED. Les attentes liées à l'utilisation d'un corpus écrit peuvent s'avérer différentes de celles d'un corpus oral. Il aurait alors été nécessaire de distinguer dès le départ sur le serveur les corpus écrits des corpus audio/vidéo. La présence de textes « purement » écrits (qui ne seraient pas la transcription de données orales) pourrait ainsi nuire à la cohérence du projet innovant, et nous avons décidé de les laisser de côté.

Après avoir procédé à un premier « appel à données », nous avons recueilli douze corpus, que nous avons regroupés sous 12 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 DRET
corpus lu de phrases françaises phonologiquement équilibrées dans un nombre non précisé de langues

BDDED268-3 - corpus simulations émotionnelles
36 phrases françaises de 8 syllabes lues de façon neutre + 4 simulations émotionnelles (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. Etiqueté 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.

BDDED268-11 - corpus persan, fourni par Pollet Samvelian,
phrases lues par 4 locutrices étiquetées sous Word, gloses morphologiques qui accompagnent l'annotation.

BDDED268-12 - Corpus vidéo FNAC (Céline Charles-Fontaine)
Il comprend un interview audio-vidéo annoté sous CLAN.

Ces différents corpus ont l'avantage de présenter pour la plupart des spécificités complémentaires. Par exemple, les gloses morphologiques qui accompagnent l'annotation du corpus **BDDED268-11**, et que nous n'avions pas dans nos précédents corpus. Quant au corpus **BDDED268-12**, son intérêt est double puisqu'il réunit deux caractéristiques que nous n'avions pas jusqu'ici : la vidéo et l'utilisation de *CLAN* (cf infra) pour l'annotation.

II.2 Achats - budget

Le laboratoire de Phonétique et de Phonologie (abrité dans les locaux de l'Ecole Doctorale 268) ayant récemment fait l'acquisition d'un enregistreur numérique MARANTZ, une démonstration de son utilisation a été effectuée au cours d'une des réunions de travail au sein de cette équipe. Il s'agit d'un matériel semi professionnel, 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 semble être, à cet égard, réservé à des utilisateurs avertis. Quoi qu'il en soit, son prix nous a également limités dans le nombre d'appareils achetés. Une solution intermédiaire proposée par l'un de nos membres s'est avérée intéressante. Ce dernier utilise une nouvelle génération de mini-discs SONY (environ 200 € HT l'unité) qui permettent d'une part d'enregistrer dans un 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 mini-disc SONY est un format propriétaire[2] et qu'ils ne proposent pas actuellement de sortie en format ouvert non compressé (WAV, AIFF, ...). De petites applications « libres », permettant cette conversion, peuvent à ce jour être téléchargées sur internet.

II.3 Le droit des locuteurs, les droits des propriétaires des corpus

Voici un point que les sociolinguistes de notre groupe de travail semblaient beaucoup plus habitués à gérer. En effet, en phonétique/phonologie par exemple, les locuteurs sont souvent amenés à lire des corpus de phrases très préparées, ou bien de la conversation dont le sens même est bien peu important... Les locuteurs peuvent ainsi être moins soucieux de l'utilisation de leur voix, ce qui est loin d'être le cas pour des corpus audio (et à plus forte raison vidéo) de Sociolinguistique. Cette question est parfaitement d'actualité comme il en sera fait état dans la partie IV.2, et puisqu'elle ne connaît pas de réponse juridique et éthique satisfaisante, nous avons décidé de choisir au cas par cas, les corpus que nous pouvions mettre en libre accès et ceux pour lesquels nous réservions un accès sur requête auprès du

[2] Un format est dit propriétaire s'il a été élaboré par une entreprise, dans un but essentiellement commercial.

« fournisseur » du corpus, c'est-à-dire un accès modulaire aux données que nous avons placé sur le serveur. L'accès à une partie du serveur du projet a été restreint au moyen d'une connexion par identification.

Dans cette optique, il a été fait mention de *CLAPI* (corpus de langue parlée en interaction ; http://gric.univ-lyon2.fr/projets/nomex-clapi/presentation/presentation.html), projet qui porte sur les droits que doivent avoir les locuteurs ainsi que les auteurs des corpus. Il a également été fait mention de *ICAR*[3] (Unité Mixte de Recherche ICAR (UMR 5191)). A l'avenir, nous nous appuierons sur les réflexions existantes les plus complètes. Cependant, pour une grande quantité des données que nous possédons, les auteurs des corpus n'ont pas pris soin de respecter les principes nécessaires, comme par exemple de faire signer aux locuteurs un formulaire stipulant par exemple qu'ils abandonnent leurs droits après enregistrement.

N'oublions pas que l'un de nos objectifs est de pouvoir continuer à placer de nouveaux corpus sur cette base de données, et il faudra que les futurs auteurs de corpus y attachent plus d'importance. Nous avons également abordé le thème de l'exclusivité au cours de nos réunions. Quelqu'un qui nous soumet un corpus est tout à fait libre de le fournir à quelqu'un d'autre, un autre organisme visant à collecter des corpus dans une optique de libre échange.

III Réflexion et activité à partir d'une base de corpus ainsi constituée

III.1 Feuille de métadonnées (fiche de documentation des corpus)

> « Sans une documentation jointe, un corpus est d'emblée moribond. L'un des dangers de la facilité actuelle de rassembler des données est précisément que les objectifs de regroupement ainsi que ceux des annotations effectuées ne soient pas enregistrées : le corpus cesse d'être utilisable dès que se perd la mémoire de ces choix. » (Habert, 1997b)

Il est nécessaire de documenter un corpus, selon l'éternel principe d'une fiche de bibliothèque qui permet de décrire chaque corpus aussi précisément que possible pour les catégoriser. 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 délivrer ses secrets. La description complète de la fiche de documentation est fournie en annexes.

Les différents éléments qui décrivent ces corpus sont appelés des « métadonnées ». Nous avons utilisé le **Dublin Core**[4] qui est une norme de métadonnées composée d'éléments simples mais efficaces pour décrire une grande variété de ressources. Elle comprend quatorze éléments (`Title, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights`) dont la sémantique a été établie par un consensus international de professionnels. Chaque élément est optionnel, peut être répété, et possède un ensemble limité de qualificatifs et d'attributs utilisés pour raffiner (et non pas étendre) la signification de l'élément. Certains champs ne sont pas nécessairement applicables aux données sur lesquelles

[3] http://gric.univ-lyon2.fr/
[4] http://dublincore.org/

nous travaillons et ils ne seront pas nécessairement instanciés dans notre travail de description des données manipulées pour ce projet.

Il était également important d'améliorer cette petite « fiche » en y intégrant des descripteurs pertinents pour le domaine de la linguistique ce qui fut une partie du travail de Maîtrise[5] en Sciences du langage effectué par Frédérique Bénard.

OLAC[6] (Open Language Archive Community) est une autre norme de métadonnées : elle ajoute des attributs aux éléments de la norme **Dublin Core** afin de les compléter pour une meilleure description des corpus linguistiques. Le principal intérêt d'utiliser cette norme de description est lié au fait que son utilisation est déjà largement développée par des organismes internationaux réputés tels que le Linguistic Data Consortium[7], le SIL[8] International et la Linguistic List[9]. En choisissant ces deux types de documentation normalisée, nous nous intégrons dans une démarche mettant en avant un format international de libre échange.

La version d'OLAC 1.0 propose les 5 attributs suivants :

- **discourse-type :** qui s'applique à **type** et **subject** du DC.
- **language (pour language identification)** :qui s'applique à **language** et subject
- **linguistic-field** : qui s'applique au **subject**
- **linguistic-type** : pour **type**
- **role** : pour **contributor** (et **creator**)

Une nouvelle fois, la description complète de cette fiche de documentation normalisée (**Dublin Core** et **OLAC**) est fournie en annexes.

Certains membres de notre groupe nous ont fait part de leur déception concernant l'absence de catégories sur l'âge, l'origine, la profession (etc.) du locuteur. Les sociolinguistes plus particulièrement (mais également 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 métadonnées, 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 métadonnées revenait à se couper de la normalisation qui nous était si chère ; en effet, la normalisation des métadonnées 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 un nouveau groupe de travail qui devrait décider de définir quels critères sont indispensables ou non ....
Une solution acceptable consisterait à définir une structuration de ces informations associées au locuteur dans l'élément `description`. On aurait par exemple,

[5] http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/sitespp/maitrise-2005/memoire-fred-2005.pdf
[6] http://www.language-archives.org/
[7] http://www.ldc.upenn.edu/
[8] http://www.sil.org/
[9] http://www.linguistlist.org/

```
locuteur.age = 12
enquete.lieu = Nantes
```

III.2 Réorganisation des données

Du type de corpus dépend un certain nombre d'habitudes comme précisé ci-dessus. En fonction des disciplines, certains utilisateurs travaillent avec des fichiers longs qui ne peuvent être découpés ; d'autres utilisent des séries d'énoncés découpés en d'innombrables petits extraits.

Un regroupement des fichiers doit être effectué pour le cas où une série d'extraits ne nécessiterait pas de feuilles de description bien distinctes. Ce regroupement peut être fonction de l'objectif linguistique (ce qui est pertinent dans le corpus), mais peut dépendre de choix plus subjectifs. Il s'agit à prime abord d'un compromis, puisqu'un regroupement trop grossier ferait perdre de l'information, alors qu'un regroupement trop minutieux « noierait » l'utilisateur dans un surplus d'information.

III.3 Remplissage/Ecriture des fiches de métadonnées : MakeMetaData

Une fois la feuille de documentation finalisée et que les principes (conventions) pour remplir cette dernière ont été définis, notre objectif consistait à proposer un petit formulaire/outil qui nous permettrait de remplir les différents champs de façon ergonomique. En effet, la fiche de documentation normalisée proposée par **Dublin Core** et **OLAC** est écrite en XML[10] (eXtensible Markup Language), un langage qui permet de structurer le contenu dans les documents électroniques facilitant leur échange et la récupération des informations représentées. Malgré tout, l'écriture en XML est « très laborieuse » pour le non-initié et nous devons reléguer l'écriture à un niveau sous-jacent, invisible pour l'utilisateur.

III.3.1 Présentation et utilité de *MaKeMetadata* (MKM)

Serge Fleury a commencé par tester un outil disponible en ligne (*KEPLER*[11]), celui-ci permettait d'insérer les éléments de **Dublin Core** et d'**OLAC** dans une feuille de description, au moyen de différentes boîtes de dialogue. Ce logiciel n'était cependant pas utilisable tel quel dans le cadre de notre projet. Serge Fleury s'est proposé de développer un logiciel dévolu à cette tâche : *MaKeMetadata* (MKM)[12].

Rappelons qu'une première fiche de métadonnées est constituée pour le fichier son (ou video) et une seconde pour l'annotation. Ces fiches sont à priori très similaires bien que différentes ; à l'utilisateur d'enregistrer ces deux fichiers sous 2 noms différents mais complémentaires. De même certains éléments ne pas pertinents pour décrire une annotation (par exemple le locuteur) et inversement. Comme précisé ci-dessus, le manque de champs pertinents peut être compensé en ajoutant toute l'information que l'on souhaite dans l'élément « `description` ». Dans le moteur de recherche qui a été mis en place ultérieurement pour

[10] http://www.w3.org/XML/
[11] http://kepler.cs.odu.edu/
[12] http://pi-ed268.univ-paris3.fr/MKM-doc/mkMETADATA.pdf

faire une recherche préliminaire sur ces corpus, il est possible de retrouver toutes ces informations par mots-clés. Mais puisque ces informations ne sont pas catégorisées, il n'est pas possible d'aller plus loin que cette recherche par mots-clés.

De par l'aspect international de la normalisation, ces fiches de métadonnées sont théoriquement écrites en anglais... Puisque nous avons rempli (pour l'instant) toutes ces informations en français (pour les cas où il faut insérer du texte libre, comme par exemple pour le champ `description`), nous avons choisi (i) de laisser par défaut le code « français » pour rédiger l'information ; (ii) de laisser à l'utilisateur la possibilité de fournir l'information en plusieurs langues en indiquant la langue utilisée pour fournir l'information. Certains champs pourront ainsi être complétés également en anglais pour faciliter les échanges.

Pour un petit 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 documentation), voire même trompeuses (comme « `subject` » par exemple). Ces étiquettes seront en fait remplacées par des gloses en français, mais également en anglais. Le fichier produit (au format XML) restitue bien entendu ces étiquettes puisqu'il s'agit d'un format normalisé, elles sont donc indispensables !

III.4 Méthodes d'annotation des corpus

III.4.1 Présentation du problème

Les corpus obtenus lors de notre premier recensement étaient très variables, tant par leurs conditions d'enregistrement, que par leur annotation. En effet les phonéticiens par exemple étudient souvent un élément très précis en fonction de leurs hypothèses (dans le meilleur des cas, les mots, phonèmes, mais aussi certains phonèmes, quelques phénomènes bien précis tels que le relâchement des occlusives). D'autres s'intéresseront de façon prépondérante à la transcription de l'oral (en étudiant la syntaxe de l'oral par exemple) et procèdent généralement à une transcription orthographique, la précision et l'ancrage temporel chers aux phonéticiens ne sont plus alors des aspects privilégiés. Cette variabilité de l'annotation s'avère capitale dès lors que l'on considère le point suivant.

D'après les retours obtenus lors de nos différentes présentations officielles, les utilisateurs potentiels d'une base de corpus espèrent généralement pouvoir extraire/afficher des occurrences prononcées par des locuteurs, par exemple avoir la possibilité d'afficher tous les extraits dans lesquels des locuteurs ont prononcé le mot « table ». Pour pouvoir laisser la possibilité aux utilisateurs d'opérer un certain nombre de requêtes, il faudrait que l'annotation (ou plus précisément ici, la transcription) puisse être « intégrée » dans un langage adapté pour ce genre de requête (XML par exemple) sur le même principe que celui utilisé pour documenter les corpus via les métadonnées[13].

Pour les corpus dont l'annotation est la plus complète, il sera utile (pour l'exemple) de faire une conversion à titre d'exemple. En effet, pour les logiciels dont l'annotation peut-être

[13] Cette démarche peut notamment être observée sur le site du LACITO (http://lacito.vjf.cnrs.fr/archivage/). Le LACITO (LAboratoire des CIvilisations et Traditions Orientales) est réputé pour ses compétences en matière de gestion et de constitution de corpus. Notre précieuse collaboration avec le LACITO s'est manifestée en la personne de Michel Jacobson, membre du projet.

récupérée au format texte brut, un *script* permettrait la conversion au format XML de manière automatique. Le problème reste en fait que les types d'annotation autorisés/proposés par divers logiciels sont parfois très différents avec leurs avantages et leurs lacunes.

> *Transcriber* par exemple est particulièrement pratique pour noter des dialogues avec tours de parole, nombre de locuteurs variés etc., mais n'est pas vraiment idéal pour l'étiquetage fin de phonèmes ou même de syllabes). Un grand nombre de logiciels permettent cet étiquetage fin sont a contrario très peu pratiques pour étiqueter des dialogues.

L'utilisation de est très pratique puisque par définition, elle permet une certaine architecture et donc d'organiser les différents types d'étiquetage. Cependant, deux nouvelles questions se posent :

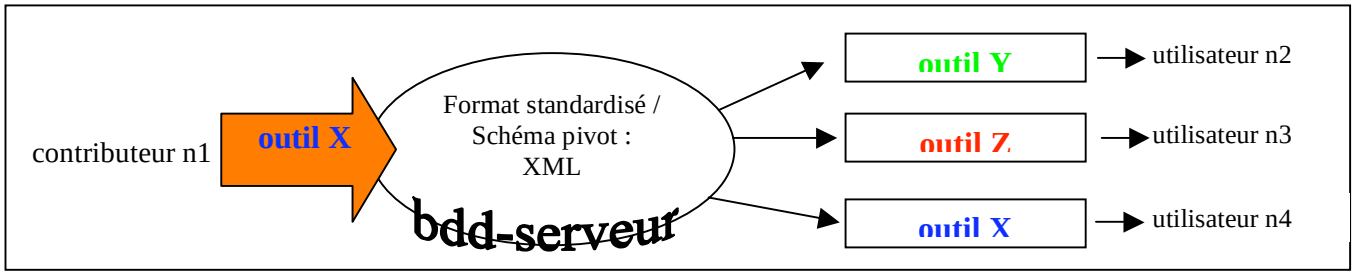
1- Est-il possible d'imposer un logiciel d'étiquetage, qui puisse ensuite produire des « segmentations » parfaitement intégrables à l'architecture toute prête pour l'accueillir ?
2- Est-il possible d'imposer une convention de notation comme il est parfois fait dès lors qu'on a pour but de créer une base de données (Cf. Base de données de Kiel, par Klaus Kohler) ?

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 MS-Office Word™. L'énorme inconvénient est l'absence de marquage temporel sur le signal acoustique, ce qui est donc peu pratique pour la personne qui doit retrouver le son à partir du texte et vice-versa.

Nous sommes tous conscients du fait que les étudiants/chercheurs qui constituent des corpus, ont pour premier argument la simplicité d'utilisation pour une efficacité suffisante. Or il existe des outils prévus à cet effet, et qui permettraient de leur faciliter la tâche et la notre par la suite. Le seul inconvénient est qu'il n'existe pas « d'outil idéal » pour répondre aux besoins de tous. Comme précisé ci-dessus, l'un est adapté pour l'annotation de grands corpus interactionnels (*Transcriber*, *Clan*), l'autre pour un étiquetage phonétique fin (*Praat*)… Ces logiciels seront plus longuement détaillés dans la partie III.4.3.

III.4.2 Solutions possibles

Considérons par exemple que les annotations de corpus puissent être stockées dans un format standardisé (non propriétaire), format qui serait un schéma pivot comme le montre la figure suivante



Puisque les métadonnées ont été insérées au format XML, dans un souci de normalisation, il serait possible de suivre ce même principe pour l'annotation : l'outil utilisé pour l'annotation de corpus devrait donc être compatible avec XML, il suffit pour cela que cet outil offre une sortie au format texte brut. Des passerelles peuvent être construites pour passer du format fourni par certains logiciels au format standardisé XML. Le problème principal est que chaque discipline utilise un petit nombre d'outils très pratiques pour certains aspects précis, donc très utilisés au sein d'une discipline, mais qui ne sont pas forcément très compatibles entre eux. Ce n'est malheureusement pas le seul problème. Par exemple nous devions faire en sorte que les caractères utilisés pour tel corpus puissent être lus sur tous types de plateformes. Ce genre de problème a été géré par le site du LACITO qui met directement à disposition les outils nécessaires, ou bien propose une visualisation en ligne. UNICODE[14] a été créé pour résoudre ces incompatibilités entre systèmes d'exploitation par exemple, mais tous les outils n'intègrent pas UNICODE (Cf. *Praat*).

Cette partie se trouve aux frontières de notre projet puisque ne nous comptons pas, dans le cadre du projet, proposer une manière standardisée d'annoter les corpus. Cependant, il serait préférable de proposer des outils qui puissent intégrer à l'avenir ces annotations standardisées. Ceci pose problème à la fois en amont et en aval de la base de données :

- il ne faut pas que les utilisateurs potentiels soient contraints par un outil qui leur semble peu pratique. Il est important de prendre en compte que la principale motivation des étudiants est de faire un corpus simplement pour faire leur travail, et répondre à leurs hypothèses ...
- il faut que les corpus stockés sur le serveur puissent être analysés par n'importe quel utilisateur, avec des (ses ?) outils d'analyse. On en revient donc au problème mentionné ci-dessus : les outils d'analyse utilisés par chacun au sein d'une discipline ne sont pas nécessairement compatibles avec le format standardisé que nous utiliserons.

Nous avons ainsi pu proposer aux membres de l'ED268 des outils adaptés pour l'annotation linguistique de corpus. La première présentation sur ce sujet a eu lieu le samedi 21 mai 2005 pour les rencontres de l'Ecole Doctorale à l'ILPGA lors d'une communication orale. Il est en effet indispensable de proposer des outils pratiques et efficaces (qui seront par ailleurs compatibles avec notre base de données, mais cela fait partie du côté « efficace »).

Dans cette lignée, un cours de linguistique de corpus a été créé dès la rentrée 2005-2006 afin de sensibiliser les étudiants de Licence 3 en Sciences du Langage à la constitution et la transcription de corpus.

III.4.3 Présentation de quelques logiciels utilisables pour l'annotation de corpus.

Ces outils ont été abordés et conseillés lors de notre présentation aux VIIIèmes rencontres de l'ED268. Aucun outil d'annotation du signal n'est évidemment parfait, et il est possible que les conclusions présentées ci-dessous soient parfois simplistes. Un passage en revue plus complet des outils d'annotation du signal les plus utilisés a été rédigé par Luiggi Sansonetti et Michel Jacobson, membres du projet innovant et est consultable sur le site du projet

[14] http://www.unicode.org/

innovant[15]. Les plus intéressants et complémentaires semblent être *ELAN*, *Transcriber* et *Praat* (peut-être *Sound Index* pour certains cas précis).

CLAN : complet, mais peu ergonomique, et nécessiterait une mise à jour, utilise XML bien qu'aucune sortie XML ne soit 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. 8 niveaux d'analyse possibles (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 Jacobson. 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/

III.5 Catalogage des métadonnées : pourquoi faire ?

La dernière étape de notre projet innovant fut la mise en place d'un catalogue de métadonnées i.e. la première partie visible accessible de notre Base de corpus. Puisque la constitution de métadonnées a permis de constituer un faisceau de mots-clés, l'objectif de ce catalogue était de pouvoir retrouver les différents corpus grâce aux mots-clés qui les caractérisent dans les fiches de métadonnées. Nous avons ainsi pu développer le moteur de recherche principal de notre base de corpus. Celui-ci est disponible sur la page :

[15] htpp<sckjjjjhkdskwdskhk

pi-ed268.univ-paris3.fr/catalogue/moteur-catalogue.html

L'objectif d'un moteur de recherche peut être, comme nous l'avons précisé infra, de pouvoir retrouver des occurrences bien précises (par exemple les réalisations du mot « table »). Actuellement, si l'intégralité de l'annotation est insérée dans la partie `description` de la fiche de métadonnées du fichier de transcription, il est alors possible d'y accéder de manière un peu détournée. Cependant, si ce moteur de recherche nous permet d'accéder au corpus dans son intégralité sur la base de ces mots-clés, il ne nous permet pas d'accéder directement à un extrait précis d'un long fichier sonore et de son annotation.

La prochaine étape au-delà du terme de ce projet innovant sera de mettre en place un moteur de recherche qui opèrera sur les transcriptions des différents corpus et non plus sur les métadonnées. Cette étape sera techniquement facile car des moteurs de recherche de ce type existent déjà, mais la diversité des formats d'annotations/transcriptions relève des difficultés plus longues à résoudre, comme précisé infra.

IV Evènements parallèles – Contacts – Suite du Projet

La mise en place de ce projet sur une période de 18 mois nous a permis de nouer des contacts et de nous impliquer dans des initiatives proches de la nôtre. Voici en quelques lignes les aspects les plus intéressants.

IV.1 Contacts

Serge Fleury et André Salem 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.

Suite à l'action spécifique ASILA (Interaction Langagière et Apprentissage), le groupe "comment cataloguer, comment coder" (CATCOD) a démarré avec Serge Heyden, Emmanuel Schang et Michel Jacobson qui ont été rejoints par certains membres de notre groupe. 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 proposition conforme à la TEI.

Dans la même optique (en spécialisation acquisition), Luiggi Sansonetti a intégré un projet ATILF composé notamment de l'équipe CRAPEL (Jeanne-Marie Debaisieux et Evelyne Jacquey.
.).

IV.2 La journée C-Oraux à la BNF

Le 13 mai à la BNF (Bibliothèque Nationale François Mitterrand) a eu lieu un séminaire C(orpus)-Oraux qui avait également pour but de produire un guide des corpus oraux téléchargeables à l'adresse suivante

http://www.culture.gouv.fr/culture/dglf/corpus_oraux.htm

Ce séminaire a permis de mettre en évidence l'importance de récolter des corpus oraux et d'en assurer la pérennité, tout en insistant sur certains aspects peu développés à l'heure actuelle tels que la juridiction sur les corpus oraux.

- l'utilisation d'une petite fiche à faire signer aux locuteurs est suggérée. Ce n'est pas la panacée mais constitue pour l'instant la meilleure solution. (Ce procédé est utilisée notamment par des organismes (Cf. ELDA) dont l'objectif principal est de recueillir des corpus oraux).
- le guide contient quelques termes juridiques à utiliser permettant ainsi une meilleure compréhension entre juristes et linguistes.

Restent les questions éthiques qui méritent également réflexion. Notons qu'il peut être utile de faire la distinction entre l'archivage de corpus d'étudiants (ou de chercheurs), comme nous le faisons, qui est utile dans une optique universitaire; et l'archivage de langues rares (ce qui est fait au LACITO par exemple) qui à des visées linguistiques plus ambitieuses et qui est peut-être plus précisément concernée par les questions éthiques.

IV.3 Centres de compétences

La collaboration effectuée avec le LACITO lors de ce projet innovant a été particulièrement couronnée de succès puisqu'elle nous a permis, en l'associant à un plus grand nombre de laboratoires de recherche (LACITO, LPP, LACAN, SYLED et MODYCO), de mettre en œuvre une proposition pour les centres de compétences dans le domaine de la gestion de corpus oraux en linguistique. Cette proposition ayant été acceptée, le travail fourni au cours du projet innovant pourra ainsi être poursuivi. Il a d'ailleurs été reconnu l'importance de la mise en place de d'un projet innovant soutenu par le conseil scientifique dans l'acceptation de cette proposition.

Le projet tel qu'il est conçu pour l'instant porte sur la "normalisation" des données. Il s'agit d'un projet bien distinct du projet innovant, mais nous pourrons avoir l'occasion de mettre à profit le travail effectué et ainsi « transmettre nos compétences ».

> Le but général de la mise en place des centres de compétences est de dynamiser la recherche en évitant les redondances et en facilitant l'accès aux services disponibles grâce à une centralisation des connaissances et des moyens. Il s'agit de structurer les communautés de recherche en regroupant et en partageant ressources et outils divers (ressources classiques mais aussi ressources avancées plus rares). Cette capitalisation des résultats de chacun doit permettre des expérimentations de nouvelles fonctionnalités plus faciles et des mises en œuvre plus rapides. De plus, cette structuration devra faciliter une meilleure transmission de l'information scientifique qui ne soit plus seulement fondée sur les publications. Un centre de compétences accompagne les équipes de recherches dans leurs besoins de créer, gérer et diffuser des ressources numériques (données et outils pour les traiter) à destination de la communauté scientifique. Diverses actions existent déjà qui vont dans ce sens et nombre d'organisations fonctionnent comme ce que nous souhaitons mettre en place (par exemple en SHS, dans le domaine des sciences sociales et économiques ou en géopolitique). Ces actions, d'une part serviront d'exemples et, d'autre part, pourront être partie prenante, dans un second temps, au présent appel. La première étape envisagée concerne essentiellement les domaines où ce mode de fonctionnement est encore à développer fortement et où il sera d'un bénéfice certain. C'est la raison des choix de domaines que nous avons effectués. Cette première étape appelle également un bilan détaillé de l'existant (en particulier, les articulations possibles avec les CCT [centres de compétences techniques] et la mission aux archives scientifiques du Réseau national des MSH). On soulignera ici qu'il ne s'agit pas seulement, dans les centres de compétences tels que nous les concevons, d'archives de documents, mais de ressources en général (c'est-à-dire de données, de corpus et d'outils pour les produire, les gérer, les modifier).

V Résumé des résultats

La déroulement du projet innovant a permis d'aboutir aux points suivants qui nous permettent de suggérer que tous les objectifs que nous nous étions fixés ont été atteints.

+ Création d'un Générateur ergonomique de métadonnées : MaKeMetadata

+ Mise en place d'un Site Web : http://pi-ed268.univ-paris3.fr

+ Mise en place d'un Moteur de recherche : http://pi-ed268.univ-paris3.fr/catalogue/moteur-catalogue.html

+ Présentation aux Assises de la recherche de la recherche 2006: http://pi-ed268.univ-paris3.fr/publis/assises_de_la_recherche_fleury_gendrot_jacobson_projet_innovant.pdf

+ présentation du projet aux VIII RED : http://pi-ed268.univ-paris3.fr/publis/RJC-ED_fb_cg_050605.pdf

VI Bilan financier et remerciements

La quasi-intégralité du budget (1840 euros) a été dépensée en équipements informatiques : 2 disques durs de sauvegarde, 1 enregistreur numérique, 2 lecteurs-enregistreurs Mini-Disc.

Les membres du projet remercient le Service financier de l'Université pour les informations précises fournies, de l'attribution des lignes budgétaires jusqu'à la préparation du dossier comptable.

Notre gratitude à l'égard du Conseil Scientifique ressort, nous l'espérons, de l'ensemble de ce rapport, mais nous avons plaisir à redire nos remerciements en conclusion !

VII Références bibliographiques

- Bird Steven and Gary Simons (2004), "Building an Open Language Archives Community on the DC Foundation", in Hillmann and Westbrooks (editors), Metadata in Practice : A Work in Progress, ALA Editions
- Bird Steven, M. L. (2000). "A formal Framework for Linguistic annotation." Speech Communication 33 ((1,2)): 23-60.
- Habert Benoît (2005) Instruments et ressources électroniques pour le français, Collection L'essentiel français, Ophrys, Gap/Paris
- Véronis, J. (2000). Annotation automatique de corpus : panorama et état de la technique. Ingénierie de langues. J. M. Pierrel. Paris, Hermès.

Liens :

Norme de métadonnées pour le codage des ressources informatiques :
http://dublincore.org/
Standard de codage de métadonnées utilisé par les bibliothèques :
http://lcweb.loc.gov/marc/
Organisation d'archivistes de donnes linguistiques et proposition de codage de métadonnées pour cette communauté :
http://www.language-archives.org/
IMDI (ISLE Meta Data Initiative) autre proposition pour la communauté linguistique :
http://www.mpi.nl/IMDI/
Standard de diffusion de métadonnées :
http://www.openarchives.org/

VIII Annexes

Descriptif des éléments utilisés pour décrire les ressources constituant la base de corpus

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.
    - 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 ...

Il est possible d'indiquer dans le titre une forme de relation en utilisant une partie commune « Récit Boucle d'Or » puis indiquer entre parenthèses.... partie1, partie voyelles

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.

- 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 ».

⇒ 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 pourrait par souci d'économie être tenté 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 !

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.

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.

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.

    - 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.

13. **Coverage :** la portée ou la couverture spatio-temporelle de la ressource.
Nous n'avons pas complètement compris ces extensions ... à voir ...
    - **spatial :** éventuellement le lieu de l'enregistrement du corpus, bien que cela puisse être trompeur.
    - **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.