Propositions de Normalisation pour une Base de Corpus Multimédia à l'ED268.

*Cédric Gendrot, Frédérique Bénard **

Université Paris 3, Sorbonne Nouvelle
Laboratoire de Phonétique et de Phonologie
19, rue des Bernardins - 75005 Paris, FRANCE
Tél. : ++33 (0)1 43 26 37 80 - Fax : ++33 (0)1 44 32 05 78
Courriel : fred-benard@freesurf.fr, cgendrot@univ-paris3.fr
* Les noms des auteurs sont indiqués par ordre alphabétique.

Mots-clés : base de données, linguistique de corpus, normalisation, pluridisciplinaire, XML

ABSTRACT - RESUME

This paper aims at promoting the creation of a pluridisciplinary database for the members of the Ecole Doctorale 268, which gives the possibility to archive and share oral and video corpora. This work lies within the scope of an innovating project financed by the Scientific Council of the University Paris 3 – Sorbonne Nouvelle. This project is realized within a pluridisciplinary framework, and brings together researchers with complementary scientific concerns, comparing diverse experimentations and emerging needs between varied disciplines using language corpora.

Cet article vise à informer les membres de l'Ecole Doctorale 268, de la création d'une base de données pluridisciplinaire par des membres de l'ED, pour des membres de l'ED, dans laquelle il sera possible de déposer et de partager des corpus oraux et vidéos. Ce travail s'inscrit dans le cadre d'un projet innovant financé par le Conseil Scientifique de l'Université Paris 3 - Sorbonne Nouvelle (responsable du projet : Serge Fleury).

1. INTRODUCTION

Avec l'essor de la linguistique de corpus (voir notamment [Hab97], principalement depuis 1990, il est devenu nécessaire d'archiver des ressources informatisées volumineuses. Dans cette optique, l'Ecole Doctorale 268 élabore un prototype de plate-forme gérant une base de données multimédia dans un format libre, universel, polyvalent et qui assurera la pérennité des données. Ce travail, réalisé dans un cadre pluridisciplinaire (les différents domaines de la linguistique et le traitement automatique des langues), réunit des chercheurs aux préoccupations scientifiques complémentaires, en confrontant les expérimentations diverses et les besoins émergents entre les disciplines utilisant des corpus linguistiques. Dans ce cadre, nous travaillons en étroite collaboration avec le LACITO (en la personne de Michel Jacobson), qui a déjà mis sur pied une base de données aux objectifs similaires aux notres[1]. L'objectif de notre projet est de proposer une réflexion pour une démarche de normalisation, lors de l'élaboration d'une base de données de ressources linguistiques (orales et vidéos), regroupant des données de langues et de natures différentes. Des propositions de normalisation pour l'encodage de corpus de langue, et des méthodes de constitution d'un corpus par l'utilisation d'outils performants (enregistrement, annotation / transcription, analyse, etc.) ont été faites sur la base de critères tels que la disponibilité des outils, leur généricité, leurs formats d'entrées/sorties, etc.

2. UTILITES D'UNE BASE DE CORPUS

2.1. Archivage de langue du LACITO

Une base de corpus permet notamment, de conserver des enregistrements de langues rares ou disparues, comme c'est le cas au LACITO, qui archive des enregistrements de langues rares disparues avec le décès du dernier locuteur, puis du linguiste ayant analysé ce corpus. L'intérêt de l'archivage est de pouvoir assurer une certaine pérennité des données en possession des chercheurs, en transférant les données sur des supports de meilleure qualité en fonction de l'avancée technologique. Une fois les données protégées des aléas du temps, il faut également pouvoir les retrouver. En effet, un bon archivage doit permettre de connaître facilement ce qui est mis à disposition, et d'accéder aux données qui intéressent l'utilisateur.

2.2. Analyse / synthèse de la parole

La disponibilité de grands corpus audio et d'outils automatiques d'alignement de plus en plus performants apporte de nouvelles perspectives : en possédant une base de donnée de taille conséquente (au minimum 20k mots), il devient possible de confirmer ou d'infirmer les analyses effectuées sur de la parole de laboratoire, i.e. des corpus restreints et soigneusement calibrés.

Dans le domaine de la synthèse de la parole, la synthèse par concaténation de diphones, la plus commune à l'heure actuelle, utilise non plus une série limitée de diphones reprenant l'ensemble de possibilités autorisées par la langue, mais choisit parmi un grand corpus pré-étiqueté les meilleurs diphones (« candidats ») possibles pour une synthèse la plus naturelle possible.

2.3. Favoriser les collaborations inter et pluri-disciplinaires :

[1] http://lacito.vjf.cnrs.fr/archivage/

La mise en commun de corpus recoltés parmi des disciplines diverses est indispensable pour tenir compte des besoins de chacun en fonction des spécialités de chaque discipline. Ce point a naturellement favorisé une collaboration interdisciplinaire au sein de l'équipe pour ne pas orienter de manière irrémédiable cette base de corpus vers une discipline unique. Ce projet vise à rapprocher dans un futur proche différentes disciplines de la linguistique afin de les amener à travailler ensemble sur un même corpus. Des expériences ont déjà été tentées dans ce sens, notamment par le biais de conférences appelant différentes communautés à travailler sur un corpus unique.

2.4. Aider les jeunes chercheurs pour la constitution de leurs corpus

Le premier travail du linguiste consiste souvent en la constitution d'un corpus. Ce dernier est le carrefour d'aspects théoriques essentiels pour la vérification de ses hypothèses. Mis à part ces aspects théoriques, un corpus doit également répondre à des exigences méthodologiques. L'optique de pouvoir intégrer des corpus dans une base de données, comme celle proposée par l'ED268, permet aux jeunes chercheurs désireux de vouloir y participer, d'être attentifs aux informations et données qu'il leur sera impératif de fournir sous peine de fournir un corpus qui ne serait qu'un banal enregistrement anaonyme dont on ne connaitrait pas le nom des locuteurs, le type exact de données reccueillies, les noms et rôles des différents contributeurs, le format des données, etc.
Les soins à apporter à l'enregistrement, bien que cela ne fasse pas partie de notre premier objectif, mérite malgré tout quelques lignes pour que l'on abandonne au plus vite le magnétocassette bas de gamme par exemple. De nombreuses fiches pourront être consultées sur le site de la DGLFLF qui propose un guide[2] des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux, paru à l'occasion de la journée C-Oraux à la BNF le 17 mai 2005.

3. LE CŒUR DU PROJET : OBJECTIFS

- Dans un premier temps, nous proposons une base de corpus essentiellement orale, du fait que les exigences des données purement textuelles diffèrent considérablement de celles des données orales. En fait, la normalisation de corpus oraux pose davantage de contraintes, puisque l'on est confronté à des enregistrements sonores, accompagnés de données textuelles pour les transcriptions et d'annotations, rédigées au moyen d'outils différents, pas nécessairement compatibles entre eux.

- Une réflexion sur la normalisation de l'encodage et de la description de corpus de langues devient nécessaire dans la mesure où l'on entre dans une démarche d'exploitation, de conservation et de diffusion des données. En effet, il est important d'utiliser des normes de description utilisées par le plus grand nombre de communautés, ainsi qu'un format de représentation universel.

4. EXPLOITATION, CONSERVATION, DIFFUSION

4.1. Pourquoi est-il nécessaire de normaliser ?

Dans le but de chercher ou retrouver des corpus, il est indispensable d'utiliser les mêmes termes et le même schéma pour décrire les données. Cela permet d'effectuer une recherche automatique, avec un moteur de recherche de type « google », efficace, pour retrouver ou chercher un corpus pertinent, un extrait de corpus, voire même une simple expression, sans avoir à lire la totalité de la ressource. De plus, l'utilisation des mêmes normes de descriptions et d'encodage que celles déjà proposées par la communauté des linguistes permet de faciliter le partage des données, et surtout, leur diffusion (voir également [Hab98]).

4.2. Comment normalise-t-on ?

La normalisation s'opère grâce à un format de représentation universel XML (eXtensible Markup Language), et des organismes de normalisation et de standardisation comme Dublin Core et OLAC (Open Language Archive Community), et un concept d'interopérabilité comme celui de l'OAI (Open Arhcive Initiative).

4.2.1. XML (voir notamment [Har01]) est un langage à balises qui permet d'annoter et de structurer une ressource, c'est-à-dire qui structure des données textuelles avec des balises ouvrantes et des balises fermantes. Fonctionne comme du HTML. C'est un langage qui a l'avantage d'être libre de droit, multi-plateforme, et échangeable.

<balise attribut="valeur">donnée</balise>

<titre lang="fr">Propositions de Normalisation pour une Base de Données.</titre>

4.2.2. Le Dublin Core[3] est une norme de métadonnées (« données qui décrivent d'autres données »), qui définit une quinzaine d'éléments de base simples, mais efficaces pour décrire tout type de données électroniques :

•Title, (creator), subject, description, publisher, contributor, date, type, format, identifier, language, relation, coverage, rights, source.

4.2.3. L'OAI [4] est un concept d'interopérabilité, autrement dit d'échanges, qui permet d'effectuer une recherche sur les métadonnées. L'OAI permet de parcourir les métadonnées des ressources mises en ligne, sur le web ou les bases de données, afin de retrouver le chemin d'accès physique des archives, sans avoir à les télécharger pour en vérifier le contenu. Il s'agit donc d'un indicateur qui fonctionne comme un catalogue. De plus, le concept de l'OAI implique que chaque ressource

[2] http://www.culture.gouv.fr/culture/dglf/corpus_oraux.htm

[3] www.dublincore.org
[4] www.openarchives.org

mentionnée est « open source », c'est à dire accessible à tous.

OLAC[5] est une communauté de linguistes, qui à partir du Dublin Core généraliste, et du concept de l'OAI, essaye de répondre au mieux aux besoins spécifiques des linguistes (voir [Bir00] et [Bir04]). Pour ce faire, OLAC propose cinq extensions au Dublin Core, rattachées à la linguistique:

•Discourse Type : (drama, narrative, language play,…)

•Language Identification : (code ISO: fr, en,…)

•Linguistic Field : (sociolinguistique, phonétique,…)

•Linguistic Data Types : (lexicon, primary-text, language-description, ...)

•Participant Roles : (annotator, author, speaker,…)

Afin de respecter la norme du Dublin Core, OLAC ne remplace pas les métadonnées du Dublin Core, mais les spécifie par rapport aux attentes de la communauté linguistique.

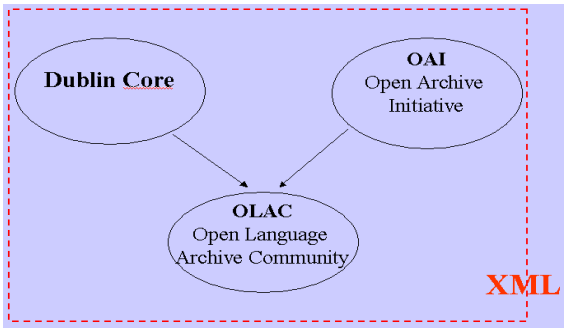

Figure 1 : Récapitulatif de la normalisation des corpus oraux.

5. METHODOLOGIE

5.1 Les types de corpus

Afin de commencer notre travail, nous avons au préalable réuni 13 corpus intégrant une annotation linguistique, provenant de formats différents (audio et vidéo) et de disciplines variées: acquisition du langage, syntaxe et sémantique, sociolinguistique, phonétique et phonologie. Précisons qu'il est nécessaire de prendre en compte trois types de données : les données brutes des ressources (fichiers audio et/ou vidéo), les annotations linguistiques (qui accompagnent et décrivent les données brutes), et les métadonnées que nous devons créer. Ces dernières permettent de décrire ces différents fichiers de manière précise, sur le même principe qu'une fiche de bibliothèque qui décrit les caractéristiques d'un ouvrage.

5.2 Les annotations

Les travaux d'annotation fournis avec nos 13 corpus ont été effectués avec des outils différents: outils d'annotation linguistique (par ex. Transcriber), outils de traitement du son (par ex. Praat), outils de traitement de l'image (par ex. Anvil / Clan), éditeurs de textes (par ex. Word). Ces outils ont souvent des formats propriétaires qui ne sont pas toujours compatibles entre eux. Notre objectif étant de mettre à disposition des ressources accessibles et réutilisables par tous, il nous était donc indispensable de trouver un format de normalisation au niveau de la structure de ces annotations. XML est un métalangage (langage de description utilisant des balises) libre de droit, qui permet l'insertion d'informations très complètes dans les documents décrits. Des passerelles peuvent ainsi être facilement construites (certaines sont déjà disponibles sur Internet, notamment Praat2XML pour Praat) pour passer des formats propriétaires générés par les logiciels mentionnés, au format XML.

5.3 MakeMETADATA

En ce qui concerne la constitution des métadonnées, nous avons décidé d'utiliser quatorze éléments de la norme du Dublin Core (DC), complétés par le standard OLAC (Open Language Archive Community). Le DC est une « fiche informatique » standardisée qui permet de décrire précisément et simplement tout type de données. OLAC permet quant à lui de spécifier les éléments du DC, en fonction des besoins propres à la communauté linguistique. L'étape la plus récente de notre travail a abouti au développement d'un outil permettant l'encodage au format XML des métadonnées de manière conviviale, comme le montre la figure 2.

Figure 2 : Copie d'écran de MakeMETADATA.

MakeMETADATA offre à l'utilisateur la totalité des champs DC et OLAC sélectionnés, afin de générer les métadonnées correspondantes à son corpus dans un format normalisé. Pour en faciliter la saisie, il suffit de cocher les paramètres qui correspondent au format du corpus et le code est généré de façon automatique.

6. LES OUTILS

L'encodage des caractères au sein de l'annotation peut également apporter son lot de problèmes. L'utilisation du codage Unicode, qui attribue un même indice pour chaque caractère graphique, indépendamment de la plate-forme informatique, du logiciel ou de la langue, permet d'éviter ces difficultés s'il est utilisé dans les formats de sortie des outils. Pour approfondir ce dernier point, une réflexion

[5] http://www.language-archives.org

sur la TEI (Text Encoding Initiative (voir par ex. [Ide 96a & b]), qui permet l'échange d'informations stockées sous forme électronique, notamment pour les sciences humaines), s'avère utile pour fournir des outils intégrant cette normalisation de l'annotation. Nous allons donc indiquer quelques outils qui tiennent compte de la TEI, génèrent (ou peuvent générer) une sortie XML, et qui sont libre de droits et gratuits.

6.1 Praat (*http://www.fon.hum.uva.nl/praat*)

Praat est un outil qui permet une analyse acoustique du signal en plus de l'annotation linguistique. Il permet une transcription fine, et huit niveaux d'analyse possibles (niveaux indépendants avec un ancrage temporel nécessaire). Praat est convertible dans une bonne mesure en un format XML.

Figure 3 : Copie d'écran du logiciel Praat : fenêtre d'édition su signal affichant de haut en bas le signal sonore, le spectrogramme, la fréquence fondamentale mesurée à partir du signal, et l'annotation.

6.2 Transcriber (*http://www.etca.fr/CTA/gip/Projets/Transcriber/*)

Transcriber est un logiciel d'annotation linguistique spécifiquement conçu pour la transcription lexicale des dialogues. Il propose un seul niveau d'analyse, et génère un format XML avec une grammaire (DTD) spécifique.

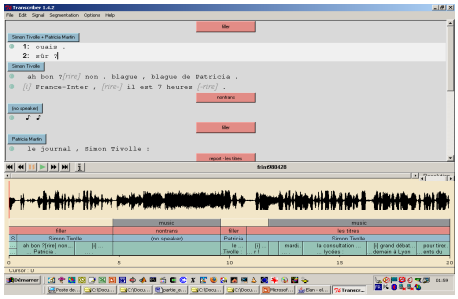

Figure 4 : Copie d'écran du logiciel Transcriber.

6.3 ELAN (http://www.mpi.nl/tools/elan.html)

Elan est un logiciel d'annotation linguistique très complet: Il autorise notamment l'annotation de la vidéo, plusieurs niveaux d'analyse possibles (dépendants / indépendants, nécessité de spécifier l'ancrage temporel pour chaque niveau) et une sortie XML (DTD spécifique).

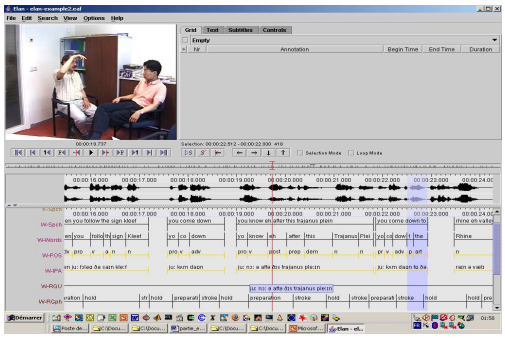

Figure 5 : Copie d'écran du logiciel ELAN.

CONCLUSION

Cette base de données devrait être opérationnelle à la fin de l'année 2005. Le travail mené autour de ce projet est visible tout au long de sa progression sur son site Web (http://pi-ed268.univ-paris3.fr).

BIBLIOGRAPHIE

[Bir00] Bird Steven, M.L. (2000), "A formal Framework for Linguistic annotation", *Speech Communication* 33((1,2)):23-60.

[Bir04] Bird Steven and Gary Simons (2004), "Building on Open Language Archives Community on the DC Foundation", in Hillman and Westbrooks (editors), *Metadata in Practice: AWork in Progress*, ALA Editions.

[Hab98] Habert Benoît, F. C. e. I. F. (1998). *De l'écrit au numérique : constituer, normaliser, exploiter les corpus électroniques*, Paris, InterÉditions/Masson.

[Hab97] Habert Benoît, N. Adeline et Salem André. (1997), *Les linguistiques de corpus*, Paris, Armand Colin/Masson.

[Har01] Harold E.R., W. S. Means (2001), *XML in a nutshell*, O'REILLY.

[Ide96a] Ide Nancy, Jean Véronis (1996). "Une application de la TEI aux industries de la langue : le Corpus Encoding Standard", *Cahiers GUTenberg* **24**.

[Ide96b] Ide Nancy, Véronis Jean (1996). "Présentation de la TEI : Text Encoding Initiative." *Cahier Gutenberg* **24**: 4-10.

[Ver00] Véronis Jean. (2000). "Annotation automatique de corpus : panorama et état de la technique", *Ingénierie de langues*, J. M. Pierrel. Paris, Hermès.

Remerciements : Les auteurs tiennent à remercier tous les membres du projet innovant non mentionnés parmi les auteurs de cet article : Sonia Branca, Maria Candea, Serge Fleury, Michel Jacobson, Thierry Pagnier, Patrick Renaud, Luiggi Sansonetti, André Salem, Pollet Samvelian, Jacqueline Vaissière.