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

Bénard Frédérique, Maîtrise TAL, Directeur de recherche : S. Fleury – email : fred-benard@freesurf.fr.
Gendrot Cédric, LPP/CNRS UMR7018, Doctorat Phonétique 4ème année, Directeur de recherche : J. Vaissière.
email : cgendrot@univ-paris3.fr
Université Paris 3, Sorbonne nouvelle, 19 rue des Bernardins, 75005 Paris.
* Les noms des auteurs sont indiqués par ordre alphabétique.

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

Cet article vise à informer les membres de l'Ecole Doctorale 268, de la création d'une **base de données pluridisciplinaire**, dans laquelle il sera possible d'archiver et de partager ses 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).

Avec l'essor de la **linguistique de corpus**, 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 multimédia dans un format libre, universel, polyvalent et qui assurera la pérennité des données. Ce travail, réalisé dans un cadre pluridisciplinaire, réunit des chercheurs aux préoccupations scientifiques complémentaires, en confrontant les expérimentations diverses et les besoins émergents entre diverses disciplines utilisant des corpus de langue. Dans ce cadre, nous travaillons en étroite collaboration avec le LACITO (en la personne de Michel Jacobson), qui entretient déjà une base de données similaire à nos objectifs (http://lacito.vjf.cnrs.fr/archivage/). L'objectif de ce projet est de proposer une réflexion et 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. Cette base de données permettra aux membres de l'Ecole Doctorale 268, d'y déposer leurs travaux, de les partager, mais également d'effectuer de nouvelles investigations sur les corpus rendus disponibles. Cette base de données sera donc dans un futur proche (fin 2005) un lieu de stockage et d'archivage, ainsi qu'un lieu de recherche. Des propositions de normalisation pour l'encodage de corpus de langue (oral/vidéo), et des méthodes de constitution d'un corpus par la formation à des outils performants (enregistrement, annotation, analyse,…) seront également émises à l'issue de ce projet.

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 complètent et accompagnent les données brutes), et les méta-données que nous devons créer (ces dernières permettent de décrire les différents fichiers de manière précise, sur le même principe qu'une simple fiche de bibliothèque).

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) ou de l'image (Anvil / Clan), voir un simple éditeur de texte (Word)] attribuant parfois leur marque/format propriétaire, et par conséquent non compatibles les unes avec les autres. L'objectif étant de mettre à disposition des ressources accessibles et réutilisables par tous, il nous était indispensable de trouver un format de **normalisation** au niveau de la structure de cette annotation. **XML** est un métalangage (langage à balises) libre de droit, qui permet l'insertion d'informations très complètes sur la structure d'un document. Des

passerelles peuvent ainsi être facilement construites (certaines sont déjà disponibles sur internet) pour passer du format propriétaire, fourni par ces logiciels, au format standard et universel XML (Extensible Markup Language). L'encodage des caractères au sein de l'annotation peut également apporter son lot de problèmes ; ceux-ci peuvent être évités en utilisant la police de caractères Unicode qui attribue un numéro unique pour chaque caractère, indépendamment de la plate-forme, du logiciel ou de la langue. Pour approfondir ce dernier point, une réflexion sur la TEI (Text Encoding Initiative, qui permet l'échange d'informations stockées sous forme électronique, notamment pour les sciences humaines), s'avère nécessaire pour fournir des outils intégrant cette standardisation de l'annotation. Lors de notre présentation, nous proposerons, aux futurs utilisateurs de cette base de données, des logiciels gratuits, performants et pratiques d'utilisation afin d'annoter leurs corpus. Une formation complète à ces logiciels sera intégrée aux prochains savoir-faire (http://www.cavi.univ-paris3.fr/ilpga/ED/savoirFaireED268.htm) qui sont organisés chaque année pour tous les membres de l'ED268.

En ce qui concerne la constitution des méta-données, 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 méta-données de manière conviviale. Cet outil sera présenté, ainsi que les avancées réalisées, d'ici aux prochaines Rencontres de l'ED268.

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

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

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.

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

Habert Benoît, N. A. e. S. A. (1997). Les linguistiques de corpus Paris, Armand Colin/Masson.

Harold E.R., W. S. M. (2001). XML in a nutshell, O'REILLY.

IDE, N., VÉRONIS, J. (1994). MULTEXT (Multilingual Tools and Corpora). Proceedings of the 14th International Conference on Computational Linguistics, COLING'94, Kyoto, Japan.

IDE Nancy, J. V. (1996). "Une application de la TEI aux industries de la langue : le Corpus Encoding Standard." Cahiers GUTenberg **24**.

IDE Nancy, V. J. (1996). "Présentation de la TEI : Text Encoding Initiative." Cahier Gutenberg **24**: 4-10.

Véronis, J. (2000). Annotation automatique de corpus : panorama et état de la technique. Ingénierie de langues. J. M. Pierrel. Paris, Hermès.

Remerciements : Lea 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.