

Les archives sonores au LACITO

Une présentation en ligne de ce projet a été exposée lors de la dernière journée d'étude de l'AFAS, à Paris, à la Bibliothèque nationale de France le 2 octobre 2003. Pour plus d'information, vous pouvez consulter le site Web du LACITO <http://lacito.vjf.cnrs.fr> ou celui de Michel Jacobson : <http://michel.jacobson.free.fr>

Les chercheurs du laboratoire LACITO (Langues et Civilisations à Tradition Orale) du CNRS travaillent depuis plus de 30 ans à l'étude et à la description de langues minoritaires, souvent sans écriture et parfois en danger de disparition. Les données recueillies sur le terrain (un peu partout dans le monde : Océanie, Népal, Afrique...) comportent des enregistrements audio et des annotations de ces derniers (transcriptions phonétiques, analyses en mots et morphèmes, lexiques, etc.). Le volume actuel de ces données représente plusieurs centaines d'heures d'enregistrement.

Ces enregistrements ont été stockés suivant les époques sur des supports magnétiques analogiques de type bandes ou mini-cassettes, puis plus récemment en numérique sur des cassettes DAT ou des miniDisc.

Les analyses sont, elles, pour partie publiées, pour partie manuscrites (cahiers de terrain). Pour certaines, il existe un support électronique informatique. Les formats de fichier utilisés sont divers et vont de ceux proposés par les outils de traitement de texte à ceux proposés par des outils de structuration plus spécifiquement linguistique comme Lexware ou Shoebox.

Ces données d'enquête ont tendance à disparaître avec le temps. En particulier, les supports magnétiques de types bandes ou cassettes se dégradent physiquement (oxydation, etc.) et se démagnétisent. Dans de bonnes conditions de conservation (pièce à température et hygrométrie constante), le vieillissement est ralenti mais non pas arrêté. Pour ne pas détériorer ces matériaux originaux, il faut éviter de le manipuler. Mais ne pas les manipuler du tout risque aussi de les abîmer. De plus, la duplication de ces supports se faisant avec une perte de qualité, il faut toujours repartir de l'original afin d'obtenir la meilleure copie possible, original que l'on a beaucoup de réticences à manipuler pour les raisons invoquées plus haut. Un autre inconvénient est que les appareils de lecture (Revox, Nagra, Uher, etc.) de ce type de supports commencent parfois à manquer ou deviennent plus difficilement accessibles par ces mêmes linguistes qui ont fait ces enregistrements quelque années auparavant. D'un point de vue pratique, les bandes et cassettes sont d'un maniement lourd car, outre qu'elles requièrent du matériel spécifique et différent de celui utilisé pour son annotation, elles sont à accès séquentiel c'est-à-dire qu'il faut les dérouler jusqu'au moment que l'on souhaite écouter. Un autre facteur de disparition des données est leur non publication. La quasi-totalité des enregistrements et la plus grosse partie des analyses n'ont pas fait l'objet de publications. Comme nous n'avons pas dans le laboratoire de recensement ni de catalogue de nos matériaux d'enquêtes, nous risquons, à terme, de n'avoir plus conscience de leur existence.

La séparation physique des enregistrements et de leurs analyses est aussi un facteur de disparition de la connaissance. On peut aboutir à des situations où l'on possède des bandes dont on ne sait rien (ni de quelle langue il s'agit, ni de quand date l'enregistrement, ni où a été faite l'enquête, etc.), ou bien des transcriptions dont on ignore où se trouve l'enregistrement. Ce manque de lien entre les ressources peut être

très dommageable, dans la mesure où les transcriptions étant des analyses secondaires et non primaires, le recours aux enregistrements est indispensable en vue d'une éventuelle falsification.

Toutes ces données (enregistrements et analyses) sont aujourd'hui propriété du linguiste. Elles sont de plus physiquement dispersées puisqu'il n'y a pas de gestion centralisée de ce type de matériaux dans notre laboratoire. Les documents sont donc difficilement accessibles à la fois pour le linguiste pour les raisons que l'on a vues plus haut (manque de matériel d'écoute, fragilité des supports originaux), mais aussi pour les autres personnes intéressées car il faut se déplacer et venir consulter les documents sur place ou bien demander au propriétaire d'en faire une copie. Ces données matérielles mais aussi un certain nombre de connaissances sur celles-ci sont fortement liées au linguiste qui les possède. Or ces derniers partent à la retraite, changent de pôle d'intérêt, etc. Il faut donc se poser le problème de la transmission de ces données. Certaines de nos enquêtes ne pourraient pas être reconduites car les langues changent et meurent. Un état de langue est une observation à un moment donné, refaire la même enquête reviendrait à décrire un autre état de langue à un autre instant. Reproduire ces enquêtes peut aussi être difficile à d'autres égards, car les locuteurs se déplacent, vieillissent, meurent, que certains terrains deviennent inaccessibles pour des raisons politiques... Tout autant de raisons pour se préoccuper dès aujourd'hui de la transmission de nos données et connaissances qu'elles datent d'hier ou d'aujourd'hui. La conservation, la publication et le catalogage sont trois tâches indispensables pour garantir cette transmission.

Le programme archivage

Depuis plusieurs années, notre laboratoire a lancé un vaste projet de gestion de ces données brutes d'enquêtes. Trois principaux buts ont été définis :

- 1) Le sauvetage en vue de la pérennisation des données
- 2) Leur diffusion
- 3) La facilitation d'accès à ces données.

La pérennisation des données audio

Ce qui a semblé le plus urgent était de trouver un support de conservation des données audio. Dans un premier temps, nous avons pensé tout naturellement au CompactDisc, qui représente encore aujourd'hui un bon compromis : le prix des supports est très faible, le prix du matériel pour graver et pour écouter n'a fait que baisser, la quantité de données que l'on peut stocker (74 minutes en qualité CD-audio, soit 44100Hz, 16bits, stéréo) est adéquate aux enregistrements jusqu'alors pratiqués. Du format CD-audio envisagé au départ, nous avons ensuite opté pour celui du CD-rom, qui permet de gérer nos données avec toutes les facilités que nous procure l'informatique. Mais la pérennité des données passe plus par le format que par le support physique. Les CD-roms seront sans doute remplacés à terme par d'autres supports plus durables, plus fiables et de plus grande capacité de stockage (ce qui arrive déjà en partie avec l'arrivée des DVD). En revanche un fichier informatique restera un fichier informatique quel que soit le support sur lequel il est stocké. L'avantage d'un fichier informatique est qu'il est duplicable à l'infini (les notions d'original et de copies se confondent), que leur accès est direct (plus de déroulement jusqu'au moment voulu) et qu'il peut être véhiculé à travers de vastes réseaux d'informations comme Internet. Le plus important devient donc de

bien choisir le format de fichier, c'est-à-dire de choisir comment seront organisées les données dans le fichier informatique. Nous avons choisi le codage PCM¹, c'est-à-dire " non compressé ", que nous utilisons conjointement avec un format de fichier RIFF/wav connu pour sa simplicité, sa popularité et sa bonne documentation. Ce que nous avons principalement voulu éviter est l'utilisation de formats propriétaires qui risquent de nous lier à un éditeur logiciel particulier et de ne pas être maintenus dans l'avenir. Nous avons aussi été conduits à choisir les caractéristiques de qualité que nous jugerions acceptables pour la gestion de notre programme. Notre choix s'est inspiré de la norme CD-audio. Nous numérisons donc toutes nos données audio en 44100 Hz, 16 bits, mono ou stéréo suivant la nature de l'original, même si cette qualité peut parfois sembler trop élevée pour certains de nos supports.

La pérennisation des annotations

Beaucoup d'annotations d'enregistrements ont été faites sur des cahiers de terrain. Une partie de ces derniers a été saisie dans des fichiers informatiques et une partie encore plus faible a fait l'objet de publications. Le reste demeure, tout comme les enregistrements, la propriété des linguistes. Là encore, la pérennisation de ces analyses doit passer par leur numérisation et par le choix d'un format de codage de cette analyse.

Historiquement notre choix s'est porté sur le formalisme SGML². En fait XML³ étant arrivé en 1998 au moment même où a débuté le projet, nous l'avons adopté immédiatement. Avec XML nous bénéficions d'un système générique de balisage de texte aussi puissant mais plus simple et plus récent que SGML. Ce formalisme nous permet d'utiliser tous les caractères d'Unicode dont nous avons besoin principalement pour l'alphabet phonétique international (API), mais aussi pour des écritures comme le cyrillique, le dévanagari, etc. Nous pouvons expliciter en XML l'analyse linguistique (en mots, morphèmes, gloses, etc.) et formaliser cette analyse dans une syntaxe formelle (DTD⁴). Le choix de la définition d'une DTD particulière plutôt que celle d'une DTD existante comme celle(s) de la TEI⁵ a été guidé par la volonté de restreindre notre analyse à un domaine plus étroit que ne le proposaient les DTD existantes (trop complexes). Nous avons tout de même utilisé une terminologie semblable à celle de la TEI afin de rendre le passage de notre codage à celui de la TEI simple, voir trivial.

Notre DTD définit quatre niveaux d'analyses : le texte, la phrase, le mot et le morphème. Chacun de ces niveaux peut contenir des transcriptions, des traductions et un ancrage temporel. Les transcriptions peuvent être de différents types (phonologique, phonétique, orthographique, translittérée), de plusieurs transcrip-teurs et de différentes époques. Une traduction peut, elle aussi, être de plusieurs transcrip-teurs et de différentes époques et bien sûr en différentes langues cibles. Il est possible d'incorporer des notes à tous les niveaux, ainsi que des indications scénographiques. Cette DTD relativement simpliste reste proche de l'analyse morpho-phonologique de documents bruts tels que nous les trouvons dans nos documents.

L'explicitation de cette structure nous a permis de faire la part entre ce qui relève de la structure logique de l'analyse et ce qui relève de règles d'interprétation typographique pour représenter cette structure. Ces dernières règles sont codées à part, dans des

-
- 1 Pulse Code Modulation
 - 2 Standard Generalized Markup Language
 - 3 Extensible Markup Language
 - 4 Document Type Definition
 - 5 Text Encoding Initiative

feuilles de styles XSL-T⁶. L'indépendance entre structure logique et les règles de présentation nous permet de présenter les mêmes données de plusieurs manières possibles (texte interlinéaire, concordance, liste de mots, etc.). Elle nous permet aussi de définir des présentations différentes suivant le support utilisé pour le consulter (papier, écran d'ordinateur, outil multimédia, etc.)

La pérennisation de l'archive

Le troisième élément constitutif de notre archive, après les enregistrements et les annotations, est le catalogue des métadonnées. Lui aussi est codé dans un formalisme XML mais en respectant une syntaxe propre à son domaine d'application. Nous utilisons en fait OLAC⁷ qui lui même est basé sur Dublin-Core⁸ étendu et respécifié pour le domaine d'application des métadonnées aux archives de ressources linguistiques. Ces métadonnées nous permettent de documenter le contenu de toutes les ressources disponibles dans les archives (enregistrements et analyses) en précisant les participants (auteurs, locuteurs, transcripateurs,...), les lieux et dates des enquêtes, les durées, les formats, etc. C'est aussi elles qui explicitent le lien entre les ressources (enregistrement et annotation d'un enregistrement). L'ensemble de toutes nos métadonnées forme ce que l'on pourrait appeler le catalogue de notre archive.

La diffusion

Pour le moment la diffusion des données que nous avons mis en place repose sur une architecture web (client-serveur). Nous avons créé deux sites web, accessibles à partir du site principal du laboratoire, qui hébergent et diffusent l'un sur Internet l'archive publique, l'autre en Intranet avec un accès protégé le reste des archives numérisées (travaux en cours, et données confidentielles).

Sur le site web public, nous donnons accès à l'heure actuelle à une centaine de contes dans une vingtaine de langues soit environ 11 heures d'enregistrements. Les fichiers d'enregistrements (dégradés au format mp3 pour diminuer leur taille et donc le temps de téléchargement) et ceux d'annotation peuvent être simplement téléchargés, ou bien consultés en ligne avec une interface proposant un certain nombre de choix de présentation (texte interlinéaire, liste de mots, ...), des outils de recherche (occurrences d'un mot ou morphème, entrées du dictionnaire, ...) , et bien sûr une consultation multimédia puisqu'à tout moment le lien entre une annotation et son enregistrement est maintenu.

La consultation multimédia repose sur les outils du web (navigateur classique Internet Explorer, Netscape, Mozilla, etc.) auxquels on ajoute un outil (plug-in, application externe ou applet Java) pour la gestion des fichiers audio. Cette consultation peut se faire dans les deux sens : à partir de l'annotation vers l'enregistrement (en cliquant sur une partie de l'annotation, on peut écouter la partie correspondante dans l'enregistrement), à partir de l'enregistrement vers l'annotation (en choisissant un moment dans l'enregistrement par un slider, on peut accéder aux annotations de celui-ci). Enfin le déroulement dans le temps de l'enregistrement est associé, par un mécanisme de gestion d'événements d'activation et d'inactivation, à la mise en valeur des parties d'annotation associées. Il est possible aussi d'accéder aux ressources de nos archives, voire à l'interface de consultation de ces archives au moyen de moteurs

6 XML Stylesheet Language - Transformations

7 Site web de Open Language Archives Community (<http://www.language-archives.org/>)

8 Site web de Dublin-Core Metadata Initiative (<http://dublincore.org/>)

de recherche spécialisés sur le web (comme celui de la LinguistList⁹) qui recherchent les ressources disponibles en fonction d'un certain nombre de critères portant sur les métadonnées dans tout un ensemble d'archives linguistiques (celles qui sont enregistrées dans OLAC) et non plus uniquement dans celles du LACITO. Cette fonction est assurée par l'adoption d'un protocole de dialogue entre les détenteurs d'archives et les moteurs de recherche appelé OAI (Open Archive Initiative).

La constitution d'un document d'archive

Une grande partie de nos enregistrements a été réalisée avec du matériel analogique de type bande magnétique ou mini cassette. Plus récemment, nous avons utilisé des cassettes DAT et des miniDisc. Afin de numériser et/ou d'informatiser toutes ces données, nous nous sommes équipés de matériel de restitution audio (lecteurs de bandes, de cassette, de MiniDisc, de DAT), d'une carte de conversion analogique/digitale, d'un ordinateur, d'un logiciel d'édition de son et d'un graveur de CD. Avec cet équipement en libre service, les chercheurs peuvent venir avec leurs enregistrements de terrain, les numériser eux-mêmes et repartir avec un instantané de ces derniers dans un format numérique et informatique.

Pour la constitution ou la récupération des analyses, le travail est un peu plus compliqué. La DTD que nous avons mis au point ne peut nous servir que pour la création d'une nouvelle annotation, ou bien pour contrôler qu'une annotation particulière est bien formée. La plupart du temps, les annotations préexistent et il faut juste les reformater pour qu'elles respectent à la fois le formalisme XML et notre DTD. Suivant le formatage initial (MS-Word, Lexware, Shoebox, etc.) nous avons pu ou non construire des scripts de reformatage. Par exemple le passage de Shoebox vers XML est une opération facilement automatisable que nous avons implémentée dans un script PERL. A contrario, le passage d'un format de MS-Word vers XML est en général une opération trop incertaine et nous sommes la plupart du temps obligés de le faire, au moins partiellement, à la main. A cela il faut ajouter une conversion du codage des caractères pour les nombreux linguistes qui ont utilisé des codages propriétaires (fontes SIL...). Nous avons par ailleurs écrit deux outils pour aider le linguiste à faire son annotation. 1) '*SoundIndex*¹⁰': un logiciel d'aide à l'ancrage temporel. Ce logiciel allie un éditeur de son et un éditeur de texte XML et permet de déterminer en sélectionnant dans la forme d'onde les moments que l'on veut faire correspondre à des parties d'annotation. 2) '*Interlinear Text Editor*¹¹': un outil d'aide à la détermination des gloses de mots ou de morphème. Il s'agit là d'un logiciel permettant de saisir les gloses des mots ou morphèmes tout en enrichissant au fur et à mesure un lexique de toutes les gloses utilisées. Ainsi le linguiste peut plus facilement être cohérent dans son propre système de glose et gloser systématiquement la même unité de la même manière.

Perspectives

Le caractère patrimonial de nos données d'archives, nous pousse à trouver une institution plus adéquate que le CNRS pour conserver et diffuser ce type de ressources. En effet, un laboratoire CNRS est constitué par un contrat de 4 ans. Ce contrat peut être renouvelé mais on ne peut pas considérer une période de 4 ans comme du long terme. Une autre raison motivant cette recherche de partenaires est que le cœur du métier du LACITO, reste avant tout la constitution et l'enrichissement de l'analyse linguistique de

9 Site web de la LinguistList (<http://www.linguistlist.org/>)

10 Téléchargeable sur le site web (<http://michel.jacobson.free.fr>)

11 Idem

ces ressources et que nous ne sommes spécialisés ni en conservation de supports d'information, ni en gestion documentaire.

L'architecture que nous visons pour la gestion, l'enrichissement et la diffusion de nos données, pourrait s'organiser autour de trois rôles principaux que sont l'auteur, le conservateur, et l'éditeur. Le rôle de l'auteur, c'est-à-dire celui qui conçoit et est responsable d'un document d'archives, doit rester dans le champ de la linguistique, puisque notre archive est composée de documents d'enquêtes linguistiques. En revanche, l'auteur n'est pas forcément le seul contributeur, il y aura la plupart du temps des locuteurs, des informateurs, éventuellement des transpositeurs, des traducteurs, etc. Le rôle de conservateur pourrait être confié à un professionnel du domaine (BnF, Archives Nationales, INA, etc.). Le problème du statut de ce type d'archives reste à définir, car il n'existe à ce jour ni dépôt légal, ni institution dont la mission est la conservation de ce type de document.

Enfin la mission de publication, c'est-à-dire de rendre les données accessibles par le public, doit en partie être confiée au conservateur qui posséderait les 'originaux' numériques. Ce serait sans doute aussi à lui d'appliquer des éventuelles restrictions d'accès (pour des raisons de confidentialité, de droits, etc.). Aux côtés du conservateur/éditeur, il existe une large place pour d'autres éditeurs de contenus, spécialisés en ressources linguistiques. Ces ressources pourraient aussi servir à d'autres corps de métier que celui des linguistes (par exemple des éditeurs d'applications pédagogiques ou ludo-éducatives pourraient aussi s'en servir). Une autre forme d'édition est celle que proposent les 'consolidateurs' en offrant des accès transparents non pas à une archive mais un ensemble d'archives. Actuellement, cette fonction est assurée sur le web par ce qu'on appelle des moteurs de recherche. Par exemple, la LinguistList offre un portail d'accès à toutes les archives enregistrées auprès de OLAC.