Archivage de patrimoine linguistique: langues menacées, tradition orale et normes numériques

Boyd Michailovsky, Michel Jacobson

Communication au colloque « Société de l'Information », ENS Lyon, 19-21 mai 2005

Archivage de patrimoine linguistique
Langues menacées, tradition orale et normes numériques

Nom du porteur : Boyd Michailovsky, Michel Jacobson
Laboratoire de rattachement : CNRS/LACITO
Thématique de recherche : Linguistique
Noms des partenaires et Laboratoires de rattachements des partenaires : : L.-J. Boë, ICP Grenoble

Introduction

Le laboratoire LACITO (LAngues et CIvilisations à Tradition Orale, UMR7107) du CNRS réunit des chercheurs en linguistique et en ethnologie qui travaillent sur le terrain, principalement en Afrique, en Asie et en Océanie. Au cours de leurs recherches, ils recueillent des enregistrements de parole, généralement dans des langues sans tradition d'écriture, souvent des langues en danger de disparition. Le programme « Archivage » a été conçu au Lacito au début des années 1990 pour étudier les moyens de pérennisation, de diffusion et d'exploitation pour la recherche de ce fonds documentaire. Le développement de l'archive a été poursuivi dans le cadre de l'appel d'offres « Société de l'Information » (2003), renouvelant une collaboration entre le Lacito et le laboratoire ICP de Grenoble datant du programme « Ingénierie des langues » (1997-1999).

Le caractère patrimonial de la documentation linguistique

Les données à archiver ont été collectées pour une variété de besoins scientifiques centrés sur la linguistique et l'étude de la tradition orale ; elles ont servi, et servent (ou peuvent servir) encore de base aux recherches linguistiques et ethnologiques, mais sont restés pour la plupart inédites et inaccessibles. Si certaines transcriptions ont été publiées, les documents sonores en particulier n'ont pu ni faire l'objet de publication, ni même être correctement conservés. Or, ces matériaux revêtent un caractère patrimonial à la fois culturel et scientifique.

Sur le plan scientifique, ces matériaux représentent la base observationelle de l'étude de la diversité des langues, science en grande partie fondée sur l'observation. Face à l'impossibilité de refaire les enquêtes dont ils sont issus (la « répliquabilité » prisée dans d'autres sciences), leur conservation et leur communication sont les conditions nécessaires à la validation des résultats acquis et au cumul de connaissances qui contribuera aux avancées futures.

La numérisation

La numérisation est fondamentale aux deux fonctions principales de l'archive, qui sont la pérennisation et la diffusion des documents dans un format adapté aux méthodes modernes de la recherche. Elle permet de conserver enregistrements et annotations sur un même support.

La copie des documents numérisés est facilement automatisable, et cela sans perte d'information, chose impossible avec des documents analogiques.

Le son archivé pour la conservation est numérisé en codage PCM à la norme de 44,1Khz/16bits. (Un format compressé est utilisé actuellement pour la diffusion sur Internet.) Les anciens supports analogiques (bandes, cassettes) sont bien sûr gardés, mais ils se détériorent inexorablement, d'autant plus que nous n'avons pas les moyens d'en assurer la conservation dans les règles de l'art (conditions climatiques controlées, déroulement régulier des bandes, transfert régulier sur nouveaux supports).

Pour les ressources textuelles nous avons adopté le format XML (Extensible Markup Language), avec le codage de caractères Unicode, dès 1997, suivant en cela un large consensus interdisciplinaire. Quant à la structure précise des documents à l'intérieur de ces normes, nous nous sommes inspirés des recommandations de la Text Encoding Initiative, mais en les adaptant à nos besoins.

La conservation : l'aspect institutionnel

La numérisation des données facilite leur entretien à long terme, et la normalisation du format et du codage des données et des métadonnées permettra de garantir leur interprétation correcte et leur transcodage ou reformatage lorsque l'évolution des techniques et des normes le rendra nécessaire. Ce sont des contraintes techniques qui seront imposées par tout conservateur éventuel. Nous ne discuterons pas ici l'aspect institutionnel de la conservation, sauf pour remarquer qu'un laboratoire de recherche n'a ni la pérennité ni la compétence nécessaire pour jouer le rôle de conservateur de patrimoine, fut-il scientifique. Nous sommes entrés en relation avec des institutions comme la BnF et des bibliothèques numériques universitaires pour assurer cette fonction.

Contenu et diffusion de l'archive

L'archive en ligne contient trois types de ressources :

- Les métadonnées, ou informations de catalogage.
- Les données, principalement des enregistrements sonores et les annotations XML correspondantes, en grande majorité des histoires et des récits autobiographiques, avec quelques listes de mots.
- Une interface de consultation des données et des métadonnées. L'interface consiste en un jeu de feuilles de style en XSL-T (Extended Stylesheet Language -- Transformations), langage de manipulation de données XML, et un processeur qui applique ces feuilles aux fichiers de données selon les requêtes de l'utilisateur ; les résultats, formatés en HTML (Hypertext Markup Language), sont transmis au *browser* du client.

Nous présentons ci-dessous ces ressources avec les moyens d'y accéder.

Les métadonnées

Pour chaque ressource archivée, l'archive contient des *métadonnées*, l'équivalent d'une fiche de catalogue comportant le titre de la ressource, sa description, la date et le lieu de sa création, la date de la dernière révision, les noms des participants à sa création, son URL, etc., dans un format normalisé.

Nos métadonnées sont conformes à la norme de l'« Open Language Archives Community » (OLAC), basée sur la norme de description de ressources multimédia « Dublin Core » ; elles répondent au protocol d'interrogation de l'« Open Archives Initiative ». Elles sont ainsi

consultables non seulement sur le site Archivage mais aussi via les portails de la LinguistList et du Linguistic Data Consortium, portails très fréquentés par la communauté scientifique des linguistes. Ces sites interrogent régulièrement nos métadonnées pour mettre à jour leurs liens, ce qui permet à un utilisateur de trouver nos ressources sans connaissance préalable du Lacito.

Sur le site Archivage on peut accéder au catalogue directement soit par une opération de « recherche dans les métadonnées », soit via un lien correspondant à une requête pré-établie (par exemple, « accès au corpus Népal »). La réponse est une liste des « fiches » de métadonnées disponibles, présentée comme ceci (réponse à la requête « documents sur la langue oubykh ») :

title	format	language	metadata	file	
Eating fish makes you clever	text/xml	Oubykh	about		browse
Eating fish makes you clever	audio/mp3	Oubykh	about		
Eating fish makes you clever	text/pdf	Oubykh	about		browse
Eating fish makes you clever	text/pdf	Oubykh	about		browse

Sur chaque ligne, le lien « about » affiche les métadonnées elles-mêmes ; le lien « browse » donne (pour les annotations XML) l'accès synchronisé à une annotation et à l'enregistrment correspondant via l'interface du programme Archivage ; les icônes sont des liens directs sur les ressources, permettant de les télécharger ou les ouvrir en utilisant les outils du client, sans passer par l'interface du programme Archivage. (Les deux dernières entrées de la liste correspondent aux images de deux transcriptions manuscrites, dont une de la main de G. Dumézil, d'un enregistrement en oubykh, langue morte, que nous avons archivé en 2004.)

Les données et leur consultation

Les données de l'archive sont en général des paires de documents, un enregistrement sonore et une annotation XML. L'enregistrement est linéaire, mais l'annotation a une structure hiérarchique (voir l' exemple ci-dessous) correspondant aux unités texte, phrase (<S>), mot (<W>), et morphème (<M>). Le lien entre enregistrement et annotation est assuré par des ancres temporelles (<AUDIO>) à au moins un de ces niveaux, souvent la phrase. Lorsqu'on accède à l'annotation d'une phrase, celle-ci comporte des indices temporels permettant de trouver la plage correspondant de l' enregistrement. L'exemple ci-dessous est extrait de l'annotation XML d'une phrase :

```
<S xml:lang="x-sil-LIF" id="SOGHAs136">
      <AUDIO start="612.1013" end="613.1817"/>
      <TRANSL xml:lang="en">The dead man spoke.</TRANSL>
      <FORM kindOf="phono">kɛsibʲɛn kɔ paɪrɛ.</FORM>
      <TRANSL xml:lang="fr">Le mort parlait.</TRANSL>
      <W>
            <M class="prefix">
            <FORM kindOf="phono">kɛ</FORM>
            <TRANSL xml:lang="fr">ACT</TRANSL>
            </M>
            <M class="v" sclass="prstem">
                  <FORM kindOf="phono">si</FORM>
                  <TRANSL xml:lang="fr">mourir.S1</TRANSL>
            </M>
            ...
      </W>
      ...
</S>
```

L'interface fournie sur le site de l'archive propose un choix de « vues » sur ces données. Pour parcourir le texte, on peut choisir de voir afficher les transcriptions au niveau de la phrase et/ou du mot ou du morphème, les traductions en français et/ou en anglais, etc., dans la limite

de l'annotation disponible. Le son est accessible, phrase par phrase ou en continu, en cliquant l'icône du haut-parleur.

D'autres « vues » sur les données sont également proposées : on peut demander de voir toutes les occurrences d'un morphème ou d'une glose en le choisissant dans le « search list » qu'on peut faire afficher (colonne de gauche) : les phrases contenant le morphème sont affichées. On peut demander une concordance du texte (dans le champ « View: »). Chaque fois qu'une phrase est affichée en réponse à une requête, le son correspondant est immédiatement disponible.

Pour utiliser un exemple tiré de l'archive dans un article ou une commmunication, on peut le copier à partir de l'affichage interlinéaire et le coller dans un fichier de traitement de texte (comme l'exemple ci-dessous) ou sur une diapositive, éventuellement avec le lien au son (l'icône haut-parleur). L'interface du site sert à formater les données et à fournir le lien vers la ressource sonore.

(1) **"lamsae *ləu* khɛnɛ kɛniŋwa tasuŋ piɛ *ləu*** (langue : limbu ; source : sogha193)

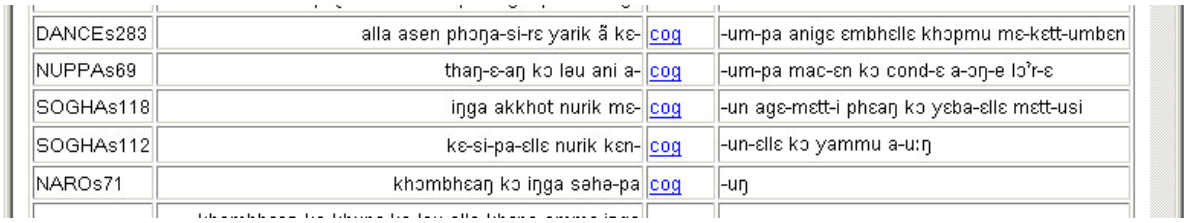
lamsa-e	***ləu***	**khɛnɛ**	**kɛ-niŋwa**	**tas-uŋ**
neveu-VOC	alors!	2SG	2SG-esprit	faire.arriver.S2-1SG→3
pi-nɛ		***ləu***		
donner.S1-1SG→2SG		EXCL		

— D'accord, neveu! Je vais te donner satisfaction.

Accès indépendant aux ressources de l'archive

L'utilisateur qui souhaite interroger les données d'une façon non prévue par l'interface peut accéder directement aux ressources. Ainsi il pourrait souhaiter faire une concordance non pas d'un seul texte mais de tout le corpus d'une langue, avec des critères spécifiques de sélection et de tri. (Le logiciel d'annotation « Interlinear Text Editor », disponible sur le site de Michel Jacobson, peut répondre à la première de ces exigeances, la concordance multi-texte.) L'image ci-dessous est extraite d'une concordance (en HTML) des seuls thèmes verbaux (identification codée dans les données) des 10 textes limbu -- ici quelques occurrences du thème *cog-* 'faire' -- triée en ordre alphabétique phonétique avec un tri secondaire sur le contexte de droite. La référence au texte pour chaque occurrence est indiquée dans la colonne de gauche ; cliquer sur le mot-clé (troisième colonne) donne accès à l'enregistrement de la phrase. Une concordance « parlante » de ce type, exploitant les données de l'archive, peut être

produite et utilisée localement (ou même rendue disponible sur le web) par tout chercheur disposant d'un accès Internet et de quelques connaissances en XML. Il est possible mais souvent inutile de télécharger les fichiers son et de les garder localement, car les liens peuvent être construits de façon à accéder directement à la ressource son sur le site Archivage.



DANCEs283	alla asen phɔŋa-si-rɛ yarik ã kɛ-	cog	-um-pa anigɛ ɛmbhɛllɛ khɔpmu mɛ-kɛtt-umbɛn
NUPPAs69	thaŋ-ɛ-aŋ kɔ lau ani a-	cog	-um-pa mac-ɛn kɔ cond-ɛ a-ɔŋ-e lɔ'r-ɛ
SOGHAs118	iŋga akkhot nurik mɛ-	cog	-un agɛ-mɛtt-i phɛaŋ kɔ yɛba-ɛllɛ mɛtt-usi
SOGHAs112	kɛ-si-pa-ɛllɛ nurik kɛn-	cog	-un-ɛllɛ kɔ yammu a-u:ŋ
NAROs71	khɔmbhɛaŋ kɔ iŋga sahə-pa	cog	-uŋ

Un utilisateur pourrait également proposer, non pas une vue nouvelle sur les ressources de l'archive, mais une annotation concurrente ou complémentaire.

Prospective

L'Archive contient actuellement 127 documents son/texte en 26 langues, préparés par une vingtaine de chercheurs, avec des corpus importants en langues de la Nouvelle-Calédonie, du Népal, et du Caucase du Nord-Ouest. Son contenu et son interface ont été développés en fonction des besoins d'un petit nombre d'utilisateurs connus, pour la plupart des membres du Lacito qui apprécient l'accès amélioré à leurs propres données. Nous prévoyons de continuer à développer l'archive dans un cadre élargi, en collaboration avec une communauté plus large d'utilisateurs et avec des institutions qui seraient d'accord pour jouer le rôle de conservateur à long terme.

Références

Jacobson, M., B. Michailovsky, J. B. Lowe. 2001. Linguistic documents synchronizing sound and text. Numéro spécial « Speech Annotation and Corpus Tools », *Speech Communication*, n°33 (2001).

Jacobson, Michel. 2004. Les archives sonores au LACITO. *Bulletin de liaison des adhérents de l'AFAS* 26. p. 2-8. (http://afas.mmsh.univ-aix.fr/Bulletin/Bulletin_AFAS_26.pdf)

Jacobson, Michel and Boyd Michailovsky. 2002. Linking linguistic resources: time-aligned corpus and computerized dictionary. International workshop on resources and tools in field linguistics. Third international conference on language resources and evaluation.. Las Palmas, Espagne. 26-27/05/2002. (http://www.mpi.nl/lrec/papers/lrec-pap-27-JACMICv2.pdf)

Sites Internet

Lacito, Programme Archivage : http://lacito.vjf.cnrs.fr/archivage

Outils de préparation de données (Interlinear Text Editor, SoundIndex), etc. : http://michel.jacobson.free.fr/informatique.htm

Open Archives Initiative : http://www.openarchives.org

Open Language Archives Community : http://www.language-archives.org

Text Encoding Initiative : http://www.tei-c.org

Unicode Consortium : http://www.unicode.org

XML : http://www.w3c.org/XML

XSLT : http://www.w3c.org/TR/xslt

Informations bibliographiques

Michailovsky, Boyd et Michel Jacobson

Archivage de patrimoine linguistique: langues menacées, tradition orale et normes numériques.

Communication parue dans :
Jean-Louis Lebrave, ed.,
La Société de l'information et ses enjeux : Actes du colloque de bilan du programme interdisciplinaire « Société de l'information » 2001-2005. ENS-LSH-Lyon, 19-20-21 mai 2005. CNRS. p. 237-240.

Ce PDF est disponible à l'adresse : http://halshs.ccsd.cnrs.fr/aut/michailovsky/