

Conservation des langues menacées et partage des ressources : exemples et conseils pratiques

Alexis Michaud* ** et Michel Jacobson**

*Laboratoire Phonétique et Phonologie CNRS/ Sorbonne Nouvelle (UMR 7018)
ILPGA, 19 rue des Bernardins, 75005 Paris

**Langues et Civilisations à Tradition Orale (LACITO) CNRS
<http://lacito.vjf.cnrs.fr/>
alexis.michaud@vjf.cnrs.fr, jacobson@idf.ext.jussieu.fr

Ce document présente l'historique d'un projet de conservation de documents linguistiques sur des langues en danger : le programme *Archivage* du LACITO. Pour des références plus complètes nous renvoyons au site web du projet, qui contient des explications sur la partie technique de ce que nous présenterons : les outils, les formalismes et les méthodes de travail :

<http://lacito.vjf.cnrs.fr/archivage/>

Comme il n'existe pas à notre connaissance de manuels à l'usage des linguistes sur les outils, les méthodes et les formalismes existants pour la gestion de corpus oraux, nous proposerons ici un certain nombre de conseils issus de notre pratique, concernant la constitution, la conservation à long terme et la diffusion de corpus oraux sur des langues rares.

Notre intervention dans cette école vise à aider ceux qui ont des données de langues rares (ou comptent en recueillir) et souhaitent assurer efficacement leur conservation et leur diffusion. La perspective étant celle d'un "témoignage", nous nous concentrerons sur :

- I. Quelques principes pour s'orienter dans le travail d'archivage
- II. Présentation d'un cas: le corpus oubykh. Aperçu des problèmes à résoudre.
- III. Présentation des outils du programme Archivage du LACITO.

Enjeux de la documentation des langues

En ce qui concerne le phénomène général de disparition des langues, l'urgence de la documentation, le profit qu'on peut escompter de la conservation des langues pour la recherche, pour les communautés concernées, et des actions en cours pour la conservation de données sur ces langues, on trouve sur internet des informations à jour. Les projets existants ont une image internet

soignée, et donnent souvent beaucoup d'informations sur les principes de collecte et les méthodes:

- Endangered Languages Fund¹,
- Gesellschaft für bedrohte Sprachen²,
- Foundation for Endangered Languages³,
- projet financé par la fondation Volkswagen⁴,
- projet de la School of Oriental and African Studies de Londres⁵

Il existe des livres généraux, mais les plaidoyers pour la survie des langues ne sont pas vraiment utilisables comme guides pratiques de l'archivage.

Dans l'optique de la mise en place de banques de données bien organisées, concernant les questions pratiques/ documentaires, nous conseillons un livre généraliste plein d'expérience:

Bonnemason, B.; Ginouvès, V.; Pérennou, V.: Guide d'analyse documentaire du son inédit, pour la mise en place de banques de données (Modal - AFAS, Parthenay, France 2001)⁶.

Situation actuelle

Les bases de données sonores abritées par les centres de recherches en phonétique et linguistique sont relativement peu développées. Les centres de recherche assurent rarement le suivi des documents enregistrés par leurs chercheurs.

¹ <http://www.ling.yale.edu/~elf/>

² <http://www.uni-koeln.de/gbs/>

³ <http://www.unizh.ch/spw/aspw/dang/>

⁴ <http://www.mpi.nl/DOBES/>

⁵ <http://www.eldp.soas.ac.uk/>

⁶ <http://www.famdt.com/>

I. Organiser la collecte des données dans un souci de partage des données

Les corpus dont il est question ici sont des « données de chercheurs », constituées par des personnes comme vous, dont la spécialité n'est pas la constitution de bases de données mais qui se servent de celles-ci comme outils. Un chercheur, en approfondissant l'étude d'une langue menacée, ou de quelques-unes, est à même de rassembler des données d'une finesse qu'un coup de filet documentaire mené par un non spécialiste peut difficilement égaler. Notre « témoignage » porte sur ces fonds individuels qui disparaissent souvent avec le chercheur, pour indiquer comment ils peuvent rejoindre un fonds documentaire partagé. Ces données demandent en fait un traitement documentaire exigeant pour être partageables avec d'autres.

Objectif: que les données soient aussi utilisables par d'autres personnes que par le seul chercheur qui les a recueillies.

D'après notre expérience, les chercheurs et étudiants ont tendance à constituer leur propre corpus à mesure des besoins de leur recherche, plutôt que de raisonner en termes de patrimoine documentaire partagé. Si vous participez à cette école d'été, c'est que vous êtes sensible aux limites de cette logique :

- il est illusoire de penser qu'on peut à tout moment créer le corpus dont on a besoin.
- le partage de ressources sur Internet demande plus que le simple dépôt d'un fichier sur une machine.

En général, les publications des chercheurs concernent les analyses de leurs données. Il est beaucoup plus exceptionnel que les données brutes d'enregistrement soient elles aussi publiées. L'accès aux enregistrements est pourtant essentiel pour évaluer l'adéquation des analyses qui en sont données, pouvoir remettre en cause ces analyses et en proposer de nouvelles, bref, de prolonger la recherche de façon cumulative. En fait, l'un des principaux obstacles au développement de bases de données de bonne qualité, de corpus publiables, est que le travail de documentation n'est pas vraiment reconnu pour l'avancement du chercheur (aspirant ou confirmé!), et apparaît donc souvent comme incompatible/en conflit avec l'activité de recherche, et avec l'exigence de publication à laquelle sont

soumis les chercheurs. A chacun de savoir déjouer cette logique, en attendant que ça change.

Collecter des données c'est à la fois stocker les résultats d'une enquête (enregistrements audio, vidéo, physiologiques, etc.), les analyses qui en sont faites (transcriptions, traductions, analyses syntaxiques...), ainsi que la description de l'enquête (les *métadonnées*). Le manque d'un de ces trois aspects risque de compromettre l'exploitation des données.

En général, c'est le chercheur qui possède des données inédites qui est le plus à même de préparer l'archivage de son fonds. C'est aussi lui qui doit décider quels sont les documents à conserver et à quel niveau de description il souhaite le diffuser. Ce travail requiert quelques compétences *techniques*, notamment pour effectuer la numérisation des enregistrements, mais aussi *scientifiques* pour les analyses linguistiques (transcriptions, etc.) ainsi qu'en matière de *gestion documentaire* (constitution des métadonnées).

II. L'exemple du fonds oubykh

Cet exemple, qui montre la fragilité des « documents de chercheur », concerne la langue oubykh, langue du Caucase du Nord-Ouest, étudiée de façon suivie par G. Dumézil et G. Charachidzé, ainsi que C. Paris, Ch. Leroy et R. Gsell. Des enregistrements minutieux ont été réalisés, ainsi que des films cinéradiographiques. L'historique de ces documents est présenté dans une « Etude articulatoire de quelques sons de l'oubykh d'après film aux rayons X » (Leroy Ch., Paris C. 1974, Bulletin de la Société de Linguistique de Paris, tome LXIX, fasc. 1.). La langue sur laquelle ils nous renseignent, et qui est aujourd'hui éteinte, était l'une des langues les plus riches en consonnes jamais observées. Les publications auxquelles le corpus a déjà donné lieu n'épuisent pas son intérêt.

Les documents, qui datent de la fin des années 1960, se sont trouvés répartis entre les divers chercheurs concernés. Lorsque l'Institut de Linguistique et Phonétique générales et Appliquées (ILPGA, Université Paris 3) s'est vu confier la donation René Gsell, nous avons souhaité que les collections sonores fassent partie de la donation, et en avons entrepris l'inventaire. Grâce au concours de Mme Agnès Gsell-Noy, les transcriptions, les films et plusieurs bandes magnétiques originales d'oubykh ont pu rejoindre le fonds dépareillé que

Mme Dabjen-Bailly s'efforçait de son côté de mettre en ordre (tâche qui était sans espoir, en l'absence de transcriptions) dans le cadre du programme Archivage du LACITO.

Les films aux rayons X ont été numérisés avec le concours du Service du Film de Recherche Scientifique. La numérisation des bandes magnétiques a été effectuée au LACITO, où elle est une opération routinière. L'ensemble des dix bobines numérisées selon le standard du son CD tient sur deux CD de données. Le corpus reconstitué contient de nombreux mots rangés par paires minimales, des phrases et des récits. Ce fonds, qui a failli disparaître, est d'un usage relativement aisé parce que

1) il s'inscrit dans le cadre de recherches menées sur le long terme par des linguistes qui ont publié sur cette langue une documentation assez abondante : dictionnaires, récits, grammaires

2) il comporte des transcriptions ; les quelques parties non transcrites risquent de rester inutilisables. Certaines des transcriptions ont visiblement été faites par le chercheur sans souci d'éventuels lecteurs autres que lui-même ; surchargées de corrections, peu lisibles, elles risquent d'être « perdues pour la science ».

Les principes à retenir de cet exemple sont :

- un bon corpus repose sur une analyse approfondie de la langue : analyse phonémique/phonétique et lexicale pour les paires minimales et les listes de mots ; analyse syntaxique pour les énoncés et les textes suivis. A défaut, un corpus de langue menacée (dont on ne peut pas facilement, ou plus du tout, trouver de locuteurs) est peu ou pas exploitable.
- le chercheur doit réaliser lui-même une annotation détaillée. Au jour d'aujourd'hui, le plus avisé paraît être de faire plusieurs sauvegardes sur CD-ROM, contenant sur un même support les fichiers (son, vidéo...) et leur annotation (dans le format qui paraît le plus commode au chercheur : fichier Word ou .pdf, ou autre format, ou même manuscrit scanné).

Supposons maintenant que le corpus a été bien annoté, numérisé, et qu'il en existe plusieurs copies de sauvegarde, chez le chercheur, à son centre de recherche/dans une institution aussi pérenne que possible. En fonction du temps qu'il est encore prêt

à consacrer au travail documentaire, le chercheur peut

- signaler l'existence du corpus, en l'état, et le diffuser tel quel
- envisager une mise en forme plus sophistiquée, mais aussi plus facile à diffuser. C'est ce que propose le programme Archivage du LACITO.

III. Le programme Archivage

Le programme Archivage du LACITO⁷ a débuté il y a une dizaine d'années avec pour objectifs la conservation et la diffusion des matériaux récoltés lors des enquêtes linguistiques menées sur le terrain par les membres du laboratoire depuis sa création il y a une trentaine d'années. Ces matériaux consistent principalement en récits, listes de mots, cérémonies, chants, accompagnés de leurs transcriptions, traduction, etc. Aujourd'hui ce programme d'archivage n'est plus propre au LACITO dans la mesure où nous traitons déjà des données de plus de cinq laboratoires de linguistique de terrain.

Les enregistrements font l'objet d'une numérisation et sont stockés sur des CD-ROM dans des fichiers au format wav (44.1KHz 16bits, mono ou stéréo, sans compression). (Sur le site web assurant la diffusion des archives, seules des versions compressées au format mp3 -- bitrate : 128 -- sont actuellement diffusées, mais l'archive proprement dite est en format non compressé.)

Les annotations comportent des transcriptions, des traductions, des gloses, des indications scénographiques, des analyses en phrases, en mots, en morphèmes, quelques informations typologiques et un ancrage temporel des parties d'analyses. Les chercheurs sont libres, en fonction de ce qui leur paraît le plus adéquat aux documents qu'ils traitent et au temps dont ils disposent, de fournir un plus ou moins grand nombre de niveau d'annotation : certains documents ne comportent pas le niveau d'analyse morphème à morphème, par exemple. Cette relative souplesse de codage tient au langage choisi pour le codage de ces annotations : le langage de balisage de texte XML. Ce langage nous permet d'explicitement la structure de nos annotations, de les coder avec n'importe quel alphabet d'Unicode (actuellement nous utilisons le

⁷ <http://lacito.vjf.cnrs.fr/archivage/>

Latin, le Cyrillique, le Devanagari ainsi que l'Alphabet Phonétique International). Ce langage nous assure aussi une indépendance vis à vis des plate-formes (Unix-Linux, Windows, Mac-OS), une bonne intégration au web, une sécurité pour l'avenir dans la mesure de sa standardisation (il s'agit d'une recommandation du « World Wide Web Consortium », le W3C ; et l'ancêtre de XML, le langage SGML, a été normalisé au sein de l'ISO) : à mesure de l'évolution des langages et des supports, les problèmes de « migration des données » qui ne manqueront pas de se poser seront les mêmes que pour un très grand ensemble d'autres bases de données, de sorte que le programme Archivage n'aura pas besoin de développer des outils spécifiques compliqués pour se maintenir aux normes : il pourra « suivre le courant » des évolutions technologiques, au prix d'une « veille technologique » limitée. XML est un outil qui est au centre d'un ensemble de technologies, et est employé par une communauté très importante dépassant largement le cadre de la recherche linguistique. La définition de nos annotations est explicitée dans une DTD (Document Type Definition) inspirée de celle de la TEI (Text Encoding Initiative). Les métadonnées sont elles aussi encodées en XML en suivant les recommandations de *Open Language Archives Community*⁸ (OLAC).

Pour faciliter le travail de structuration de document, le programme Archivage du LACITO a mis au point deux outils :

SoundIndex : logiciel couplant un éditeur de texte XML et un éditeur de son. Il permet d'assister le linguiste pour l'ancrage temporel des unités (texte, mots, phrases, morphèmes).

ITE : logiciel permettant de saisir des transcriptions ainsi que de les gloser. Il présente les documents sous forme interlinéaire.

Enfin, pour faciliter le partage des données préparées selon nos normes, le LACITO a mis en place un serveur web qui héberge et diffuse les données, et offre une interface de consultation pratique pour les utilisateurs. Nos efforts n'ont pas porté sur le côté convivial/ludique de l'interface ; mais les visiteurs qui, comme vous, ont le projet de participer à une entreprise de ce type apprécieront la fonctionnalité du site. Sur la base solide et fiable

de l'architecture XML du corpus, le passage à une base de données grand public conviviale ne pose pas de difficulté, comme le montre l'exemple de la base de données de langues de Nouvelle Calédonie réalisée par le LACITO en collaboration avec le Centre culturel Jean-Marie Tjibaou de Nouméa.

Le programme Archivage étant réalisé au sein d'un laboratoire du CNRS, il n'est donc pas, par nature, pérenne. Ce programme propose aux chercheurs une aide en termes d'équipement audio (pour la numérisation) et informatique, ainsi qu'un certain nombre d'outils de gestion de corpus oraux (création, maintenance, interrogation et diffusion), de conseils et d'assistance pour qu'ils effectuent leur propre archivage ; la diffusion de leur corpus, notamment par internet, n'a lieu que si les auteurs le souhaitent, mais la préparation des données rend du moins envisageable une telle diffusion.

<http://lacito.vjf.cnrs.fr/archivage/>

alexis.michaud@vjf.cnrs.fr

jacobson@idf.ext.jussieu.fr

⁸ <http://www.language-archives.org/>