Conservation des langues et partage des ressources : le rôle des chercheurs dans la mise en place de banques de données

Alexis Michaud

Conservation des langues et partage des ressources : le rôle des chercheurs dans la mise en place de banques de données

Alexis Michaud

Laboratoire Phonétique et Phonologie CNRS/ Sorbonne Nouvelle (UMR 7018)
ILPGA, 19 rue des Bernardins, 75005 Paris
Alexis.Michaud@univ-paris3.fr

RÉSUMÉ

La réflexion part d'un constat paradoxal: les bases de données sonores abritées par les centres de recherches en phonétique sont relativement peu développées. Les centres de recherche assurent rarement le suivi des documents enregistrés par leurs chercheurs. Le présent article, qui se place principalement du point de vue de la conservation des langues en danger, présente une réflexion sur le rôle que pourraient jouer des « phonothèques universitaires », centres de diffusion mais aussi de création de bases de données.

ABSTRACT

Looking back at a century of speech recording, the legacy is not as extensive—and nowhere as tidy—as the layman would think. Research centres seldom keep track of the recordings made by their researchers. This paper, focusing primarily on endangered languages data, argues that a network of sound libraries associated with university libraries and research centres should be set up to build and disseminate corpora, following certain quality standards. Researchers could then have access to databases that would reflect the variety of research purposes as well as the variety of the world's languages.

INTRODUCTION

Les possibilités offertes par les nouvelles technologies pour la création de bases de données multi-média sont vertigineuses. Les corpus déjà réalisés atteignent des volumes impressionnants, comme en témoigne par exemple le catalogue du site ELRA (http://www.icp.grent.fr/ELRA/home.htm).

Mais ces entreprises documentaires sont destinées à tel ou tel usage précis, généralement dans le domaine du Traitement Automatique des Langues, et sont d'abord tournées vers les grandes langues de communication. Elles ne rendent donc nullement dérisoires les efforts pour constituer des bases de données des langues en danger, tâche non rentable économiquement qui est principalement le propre de centres de recherche universitaires (voir les travaux menés dans le cadre de : *Endangered Languages Fund,* http://www.ling.yale.edu/~elf, *Gesellschaft für bedrohte Sprachen*, http://www.uni-koeln.de/gbs, *Foundation for Endangered Languages* (http://www.unizh.ch/spw/aspw/dang).

Dans le domaine de ces bases de données de langues rares, il devient nécessaire, pour mettre à profit les nouvelles technologies, de mettre en place au sein des centres de recherche un réseau de création, de diffusion et de conservation de corpus. Le besoin est réel d'une collecte soigneuse des langues du monde. Ce travail, qui intéresse plusieurs disciplines dans le champ des sciences humaines et revêt également une grande importance pour la mémoire des peuples concernés, nécessite un archivage des *données de chercheur*, qui demandent un traitement documentaire aussi exigeant que les *bases de données ingénieur*.

La réflexion menée ici au sujet de la *chaîne de la documentation parlée* pourra également intéresser les chercheurs qui travaillent sur les grandes langues. Alors que se multiplient les corpus modelés sur les besoins de la recherche en Traitement Automatique des Langues, les linguistes doivent affirmer leurs propres critères de fiabilité, pour disposer de corpus qui reflètent la diversité des questionnements auxquels on peut soumettre les langues.

1. MÉTHODES DE TRAVAIL DES CHERCHEURS

Les bases de données sonores abritées par les centres de recherches en phonétique sont paradoxalement assez peu fréquentées et peu structurées, si on les compare, par exemple, avec les bibliothèques universitaires. 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. Les fonds d'archives sont peu connus, les grands corpus distribués sur Internet dépassent souvent les budgets du chercheur individuel, tandis que l'on peut enregistrer soi-même un corpus d'une qualité technique satisfaisante.

On voudrait souligner ici les limites de cette logique: il est illusoire de penser que l'on peut à tout moment créer le corpus dont on a besoin. Dans le cas des langues en danger, la mise en commun des données existantes est particulièrement nécessaire. Mais dans l'étude des grandes langues, le travail documentaire ne demande pas moins de sérieux. D'après notre expérience, ce sont souvent des usagers des laboratoires de phonétique qui sont sollicités comme informateurs pour ces « corpus personnels ». Les sujets ont l'expérience des tâches demandées, tandis que les non-initiés peuvent être intimidés ou perplexes; par ailleurs, étudiants et collègues rendent service bénévolement. Mais la bonne volonté d'informateurs non rétribués n'est pas infinie, ce qui encourage à abréger la préparation de la session, pourtant cruciale.

Le fait de recourir à un informateur linguiste, et souvent polyglotte, pose aussi des problèmes épistémologiques évidents. Des mots français enregistrés par un locuteur natif pour un cours de lecture de spectrogrammes se sont avérés « non canoniques » au point d'induire en erreur des déchiffreurs chevronnés ; c'est très vraisemblablement la conséquence d'une expérience linguistique très variée. Un corpus de français enregistré aux Etats-Unis présente des /p/ /t/ /k/ aspirés et des /b/ dévoisés, déformations quasi inévitables pour qui réside en pays anglophone. Telle informatrice japonaise, en France depuis quelques mois, réalise des montées de continuation à la française (contraires aux contours intonatifs du japonais) lorsqu'elle parle sa langue maternelle. G. Boulakia, T.D. Dô et T.H. Trân rapportent un exemple similaire pour le vietnamien [BTD98]. Cela remet en cause certaines données publiées dans les revues internationales, fournies par des locuteurs natifs mais résidents depuis fort longtemps dans un pays étranger. Outre leur prononciation, ce séjour perturbe également leur façon de percevoir (des vérifications, sur des Français « résidents longue durée » en France, vont être entreprises à l'ILPGA).

Il paraît donc important qu'étudiants et chercheurs disposent, au sein de leur laboratoire, d'une phonothèque qui s'enrichisse par des acquisitions parmi les corpus qui existent. Les chercheurs auraient ainsi l'occasion de se faire une idée précise de l'état des lieux; si le chercheur décide de réaliser lui-même un corpus pour combler une lacune de la documentation existante, le travail pourra alors se faire dans un souci de qualité, avec l'idée qu'il sera utilisable par d'autres. Dans le domaine des mesures physiologiques, l'utilité d'un partage des données est particulièrement claire, étant donné la complexité du dispositif de mesure.

2. LE RÔLE DU CHERCHEUR ET DES ÉQUIPES DE RECHERCHE

Mais la gestion d'une base de données, et la constitution d'un corpus publiable, sont des entreprises de taille. Le travail de documentation apparaît souvent comme incompatible avec l'activité de recherche, et avec l'exigence de publication à laquelle sont soumis les chercheurs, aspirants ou confirmés. De fait, la constitution de bases de données occupe des professionnels à plein temps. La seule question des normes de codage motive des programmes de grande ampleur tels que TEI [SMB94] ou MATE [DBO98] (pour d'autres références, voir [JLM01]). Est-ce alors à des projets spécialisés qu'incombe le travail de documentation? Le projet DOBES (Dokumentation Bedrohter Sprachen), financé par la fondation Volkswagen, par exemple, est spécifiquement dédié à la collecte de langues rares sur quelques années. On voudrait souligner ici l'intérêt d'un travail *à plus petite échelle mais à plus long terme*, et qui soit intégré aux centres de recherche. En effet, un chercheur, en approfondissant l'étude d'une langue, est à même de rassembler des données d'une finesse qu'un *coup de filet documentaire* peut difficilement égaler. Les remarques qui suivent concernent ces fonds individuels qui disparaissent souvent avec le chercheur, pour réfléchir à la façon dont ils pourraient rejoindre un fonds documentaire partagé.

Le chercheur qui possède des données inédites est seul à même de préparer l'archivage de son fonds. Il doit décider quels sont les documents à conserver, et en établir une transcription. A défaut, la documentation risque de devenir inutilisable. Les collections sonores de la Bibliothèque Nationale contiennent quantité de documents linguistiques, sans transcription (avec de très rares exceptions, par exemple une mission de Brunot dans les Ardennes [Cor01]). Parmi ces collections, des centaines d'heures d'enregistrements en langues étrangères, à peine identifiés. L'exploitation d'un document de ce type demande un investissement considérable, pour un résultat très hypothétique. Le dépôt non documenté est aujourd'hui refusé par les phonothèques, qui sont des lieux de conservation, mais aussi de consultation : le travail de transcription et de mise en forme doit être réalisé par le chercheur et son équipe de rattachement.

Outre la transcription, le travail de mise en forme comporte également la numérisation. (Pour un exposé encyclopédique sur les questions des supports, voir [CF96]). Pour cette deuxième étape du travail de documentation, c'est également le chercheur qui est le mieux à même de faire le travail. Mais cela représente une charge de travail. Il faut donc souhaiter que les centres de recherche organisent des programmes de création de bases de données auxquels participeraient leurs techniciens et ingénieurs, qui pourraient être assistés d'étudiants (en qualité de vacataires). Ces équipes assureraient également le suivi des fonds anciens. Leur présence dans les centres de recherche amènerait certains linguistes à prêter attention aux tâches documentaires urgentes qui concernent notre discipline (par exemple la conservation des collections sonores du Musée de l'Homme, patrimoine linguistique très fragile dans la gestion duquel, à notre connaissance, les linguistes ne jouent actuellement guère de rôle).

La présence de ces équipes serait également précieuse pour ceux (étudiants et enseignants étrangers) qui souhaitent réaliser un corpus de leur propre langue : à l'heure actuelle, faute de procédures simples à suivre, les bonnes volontés se découragent.

3. CHARTE DE QUALITÉ PROPOSÉE

Les propositions qui suivent, guidées par un souci de simplicité, résument très brièvement les critères essentiels pour la création de données diffusables. Le lecteur est également renvoyé à [BGP01].

3.1 L'indexation et la transcription

Un inventaire (qui peut dans un premier temps se faire avec un simple logiciel de traitement de texte) doit indiquer :

(1) la LANGUE, la REGION où elle est parlée, une brève PRESENTATION DU LOCUTEUR

(2) l'identité de l'ENQUÊTEUR, les LIEU et DATE de l'enregistrement, sa DUREE, ses CARACTERISTIQUES TECHNIQUES

(3) les DOCUMENTS CORRESPONDANTS : transcriptions, photographies, vidéos, publications.

La transcription doit expliquer l'OBJECTIF de l'enregistrement et les CONSIGNES données à l'informateur. Il importe également de donner des précisions sur les DROITS D'AUTEUR et les éventuelles RESTRICTIONS POUR LA DIFFUSION.

3.2 La qualité de l'enregistrement audio

Si un traitement est appliqué au signal, cela doit être indiqué. Le document retravaillé doit être accompagné de l'original numérisé non retouché. (La suppression de certaines parties pour des raisons de confidentialité constitue bien sûr une exception.) La compression en format MP-3 ne satisfait pas aux normes d'archivage. Le document original doit donc être disponible. L'argument de l'économie de place n'est nullement déterminant, étant donné les capacités actuelles de stockage. L'usage du mini-disque est fortement déconseillé, pour la même raison.

3.3 Rémunération de l'informateur

Le fait de définir la tâche de l'informateur comme un travail, rémunéré comme tel, est essentiel au sérieux de l'entreprise. Inscrire cette exigence dans la « charte de qualité » vise à attirer l'attention vers la question essentielle de la relation à l'informateur.

4. DES PERSPECTIVES NOUVELLES : VERS LES BASES DE DONNÉES « DE POINTE »

Dès lors que les données (sonores, physiologiques, visuelles...) et les « métadonnées » sont conservées ensemble sur support numérique, l'essentiel est acquis. Par la suite, la mise en forme dépend de l'utilisation que l'on souhaite en faire. Le choix de normes de codage est l'objet de discussions très actives ([SB00], [LSB01]) ; soulignons que les données de langues menacées, récoltées « artisanalement », se prêtent une mise en forme aussi rigoureuse que les grands corpus de parole lue ou de parole téléphonique, comme le montre l'exemple du programme Archivage du LACITO [JLM01], http://lacito.vjf.cnrs.fr/archivage/index.html.fr. Il recourt à des outils standard, autour du langage de balisage de texte XML, dont l'ambition est résumée par [Sie97 p. 21]: « [XML] will let us build great libraries (...) simply by tagging everything properly so it fits into the larger schema of the Web ». L'investissement de temps que représente l'apprentissage des rudiments de ces instruments a pour contrepartie la compatibilité avec de nombreux outils, ce qui est utile autant en « synchronie » (pour échanger des données d'un système à l'autre et d'une plate-forme à l'autre) qu'en « diachronie », pour que les données restent lisibles lorsque logiciels et systèmes d'exploitation évoluent. En particulier, la norme UNICODE pour le codage des caractères est très prometteuse (voir [Jac99]).

Un domaine extrêmement prometteur s'ouvre donc pour qui a des données intéressantes à mettre en valeur ; mais il faut souligner le rôle des chercheurs dans ce travail. Le programme Archivage est réalisé au sein d'un laboratoire du CNRS, et n'est donc pas *pérenne*. Ce programme propose des outils aux chercheurs pour qu'ils effectuent leur propre archivage; les fichiers sont mis sur Internet si les auteurs le souhaitent, mais le programme (si paradoxal que cela semble) n'a pas vocation à être dépositaire des documents. Les chercheurs restent responsables du sort des documents à long terme. Ce serait donc un contresens sur ce programme que d'en attendre qu'il « se charge de l'archivage des données sur les langues en danger », d'autant plus qu'il n'y a pas de personnel pour effectuer le travail de documentation (tenue d'un catalogue, conservation de copies intégrales des documents). Pour atteindre toute son ampleur, un programme comme celui du LACITO aurait besoin d'être relayé par l'effort documentaire de « phonothèques phonétiques » pérennes accueillant les corpus les plus divers.

5. UN EXEMPLE : LE FONDS OUBYKH

En dernier lieu, nous souhaiterions exposer un exemple d'initiative d'archivage qui montre la fragilité des « documents de chercheur ». Cet exemple 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 » [LP74]. La langue sur laquelle ils nous renseignent, et qui est aujourd'hui éteinte, était l'une des deux 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'ILPGA 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, qui contient de nombreux mots rangés par paires minimales, des phrases et des récits, remplit les conditions 1 et 2 de la « charte de qualité » esquissée dans le présent article. Quant à la question de la relation avec l'informateur, qu'il a paru important d'intégrer dans la « charte de qualité », précisons simplement que G. Dumézil appelait Tevfik Esenç, son unique informateur, « mon maître et ami Tevfik Esenç ».

Ce fonds de référence, qui est d'un usage très aisé, aurait sa place dans tout centre de recherche en phonétique. Mais pour en arriver là, les procédures restent à inventer. Un éditeur acceptera-t-il de se charger d'une publication de ce type ? Le marché n'existe pas, faute de « phonothèques phonétiques » qui assureraient une petite clientèle. Si l'on propose, à défaut de publication, une diffusion gratuite et informelle, la qualité technique sera nécessairement moins bonne, et les détenteurs des droits auront lieu de s'y opposer, puisqu'il n'y aurait pas de diffusion auprès d'institutions pérennes (en particulier la Bibliothèque Nationale), de sorte que le problème de la conservation ne recevrait pas de réponse satisfaisante.

CONCLUSION

Cet exemple veut montrer que les initiatives documentaires ont besoin d'être relayées institutionnellement, pour que les données les plus intéressantes parviennent aux chercheurs qui sauront en tirer profit. L'auteur s'exprimait ici en qualité d'informateur et de documentaliste amateur, fonctions l'une et l'autre marginales, et dont il n'est pas habituel de se prévaloir devant un public de chercheurs. Les questions soulevées dépassent le champ de la recherche individuelle ; elles touchent à des problèmes institutionnels délicats, par exemple celui des structures d'archivage au sein du CNRS. Si hypothétiques que soient les perspectives évoquées ici, on espère qu'elles encourageront certains linguistes à donner de leur temps pour contribuer à la mise en place d'un réseau de « phonothèques universitaires » de qualité. En effet, si la linguistique s'efforce sans cesse de *définir* son objet, il paraîtrait naturel que les chercheurs eux-mêmes consacrent une partie de leurs efforts à *matérialiser* celui-ci sous forme de corpus qui puissent être partagés, et à *préserver* celles de ses manifestations qui disparaissent actuellement.

BIBLIOGRAPHIE

[BGP01] Bonnemason B., Ginouves V., Perennou V. (2001), *Guide d'analyse documentaire du son inédit, pour la mise en place de banques de données*, éditions MODAL.

[BDT98] Boulakia G., Dô T.D., Trân T.H. (1998) « Intonation in Vietnamese », *in* Hirst et Di Cristo (1998), *Intonation Systems : A Survey of Twenty Languages*, Cambridge University Press, pp. 395-416.

[CF96] Calas M.-F, Fontaine J.M. (1996) *La Conservation des documents sonore*, éd. du CNRS.

[Cor01] Cordereix P. (2001) "Ferdinand Brunot, le phonographe et les 'patois' ", *Le Monde alpin et rhodanien* 1er- 3e trimestre 2001, pp. 39-54.

[DBO98] Dybkjaer L., Bernsen N, Ole, Dybkjaer H., McKelvie D., Mengel A. (1998) *The MATE Markup Framework.*

http://mate.nis.sdu.dk.information/d12/

[JLM01] Jacobson M., Michailovsky B., Lowe J.B. (2001) "Linguistic documents synchronizing sound and text", *Speech Communication* 33, 79-96

[Jac99] Jacobson M. "Les normes de codage de caractères", webzine *Prograzine* n°3. Consultable sur le site : http://michel.jacobson.free.fr

[LP74] Leroy Ch., Paris C. (1974) "Etude articulatoire de quelques sons de l'oubykh d'après film aux rayons X", Bulletin de la Société de Linguistique de Paris, tome LXIX, fasc. 1.

[LSB01] "The Digitization of Language Data: The Need for Standards", colloque organisé à Santa Barbara, Californie, 21-24 juin 2001. http://linguistlist.org/~workshop/reading.html

[SB00] Simons G. & Bird S. "Requirements on the infrastructure for digital language documentation". Site internet : voir [LSB01].

[Sie97] Siegel D. (1997) "The Web is ruined—and I ruined it", *in* Connolly D. (ed.) *XML: Principles, Tools, and Techniques*, WWW Journal n°2:4, Sebastopol, CA: O'Reilly & Associates, 1997.

[SMB94] Sperberg-McQueen, C.M., Burnard, L. (ed) (1994) *Guidelines for Electronic Text Encoding and Interchange* (TEI P3). Electronic Book Technologies, Providence, RI. http://www.ltg.ed.ac.uk/software