“ Tu pourrais enregistrer un corpus pour moi ? ”

Pour une charte de qualité des corpus[1]

Alexis Michaud

ILPGA
19 rue des Bernardins, 75005 Paris
Mél: michaudalexis@yahoo.fr

ABSTRACT

The time-consuming task of archiving and disseminating data is not a priority with most phoneticians. As a result, finding a suitable ready-made corpus is no easy task; researchers often rely on corpora of questionable value. Looking back at a century of speech recording, the legacy is not as extensive—and nowhere as tidy—as the layman would think.

This paper calls for a “ Corpus quality standard ”. The argument (based on detailed examples) is that small-scale programs adhering to simple standards can actually go to build the databases we need. A quality standard would make data publication easier (thus fostering research) and allow for a smoother transition into the shelves of libraries, fulfilling the phoneticians’ key role in documenting the languages of the world.

INTRODUCTION

Les technologies numériques ouvrent la possibilité d’une conservation indéfinie de la documentation ; les potentiels pour la diffusion de l’information ne sont pas moins vertigineux. Mais ces progrès restent tout théoriques s’il n’existe pas un réseau de création, de diffusion et de conservation des corpus. Cette question est solidaire d’une réflexion sur les données que l’on souhaite recueillir. C’est la question de la *chaîne de la documentation parlée* dans son ensemble qui est donc posée ici, pour aboutir à une “ charte de qualité des corpus ”. Fondée sur un petit nombre de principes de bon sens, elle vise à ce que les multiples initiatives documentaires entrent dans une logique cumulative.

Le souci qui motive cet article déborde le champ de la phonétique, puisqu’il concerne la conservation du patrimoine linguistique. Le besoin est réel d’une collecte soigneuse des langues du monde ; ce travail intéresse diverses disciplines dans le champ des sciences humaines, et revêt également une grande importance pour la mémoire des peuples concernés. Mais l’existence de corpus fiables n’est pas moins nécessaire au travail sur les grandes langues, et à la recherche “ strictement linguistique ”, si on tient à l’opposer à l’“ ethno-linguistique ”.

Un coup d’oeil sur la situation actuelle (dans les laboratoires mais aussi dans les phonothèques) fournira l’occasion d’une réflexion sur la qualité des données. Un second volet, plus positif, formulera des propositions en vue d’une “ charte de qualité des corpus ”. Il se terminera par la description d’une initiative d’archivage, qui illustre de façon exemplaire les problèmes qui se posent.

1. SITUATION ACTUELLE

1.1 Les enregistrements sonores

En chambre sourde, on obtient aisément un enregistrement d’une très grande qualité. Cela encourage à constituer son propre corpus à mesure des besoins, plutôt que de chercher dans les fonds existants.

Mais il est illusoire de penser que l’on peut à tout moment créer le corpus dont on a besoin. Sans parler des problèmes techniques (dont on ne s’aperçoit bien souvent qu’après coup), il faut trouver les informateurs. Les usagers des laboratoires de phonétique sont régulièrement sollicités pour ces tâches de routine : tests de perception, enregistrements audio (et aussi séances de prises de données physiologiques, réservées aux informateurs de bonne composition). Les avantages sont clairs : les sujets ont l’expérience du travail en chambre sourde et des tâches demandées, tandis que les non-initiés peuvent être intimidés ou perplexes. Par ailleurs, étudiants et collègues rendent service bénévolement, tandis qu’il est nécessaire de rémunérer les personnes de l’extérieur.

Mais il n’est pas évident de s’assurer du sérieux d’informateurs non rétribués. Leur bonne volonté n’est pas infinie : on n’ose pas leur imposer de longs préliminaires, ni faire refaire les tâches. Le fait de recourir à un informateur linguiste, comme c’est très souvent le cas, pose aussi des problèmes

[1] *Nous tenons à remercier tout particulièrement M. Pascal Cordereix, Responsable des Collections sonores de la Bibliothèque Nationale, Mme Florence Gétreau, Conservateur des Archives sonores du Musée national des Arts et traditions populaires, et l’équipe du programme Archivage du LACITO.*

épistémologiques évidents. Sans entrer dans le détail de ce phénomène que la psycholinguistique connaît bien, retenons que le chercheur risque de travailler sur des données gravement faussées.

En outre, les habitués des laboratoires de phonétique sont souvent polyglottes, ce qui est un autre facteur indésirable. Un exemple de première main : les mots français que nous avons accepté d'enregistrer pour un cours de lecture de spectrogrammes se sont avérés " non canoniques " au point d'induire en erreur des déchiffreurs chevronnés. L'examen des spectrogrammes révèle des gestes articulatoires insolites en français, diagnostic qui confirme que le fait d'avoir une expérience linguistique très variée (en l'occurrence : pratique régulière de langues d'Asie) a une influence sensible sur la prononciation de sa langue maternelle. Autre exemple : un corpus de français enregistré aux Etats-Unis, et qui voudrait servir de norme, présente des voyelles diphtonguées, des /p/-/t/-/k/ aspirés et des /b/ dévoisés. Telle informatrice japonaise, en France depuis quelques mois, réalise des montées de continuation (contraires aux contours intonatifs du japonais) lorsqu'elle parle sa langue maternelle.

N'est-il pas dommage qu'un étudiant qui scrute pendant des années les courbes de F0 de l'anglais choisisse, faute de mieux, les *pubs* parisiens comme vivier d'informateurs? La large diffusion de corpus fondés sur une " charte de qualité " permettrait de lever cette hypothèque qui pèse sur la fiabilité des données.

1.2 Les prises de données physiologiques

Dans le domaine de l'étude physiologique de la parole, les difficultés techniques sont exacerbées. Il faut travailler avec un médecin, organiser le rendez-vous en hôpital, et mettre en place les équipements. Le dispositif est très vite complexe, puisqu'il est préférable de prendre simultanément diverses mesures : fibroscopie, acoustique et mesure de pression d'air nasal, par exemple. Il n'est pas rare que l'on s'aperçoive après coup d'un défaut d'enregistrement qui ôte toute valeur aux données, et oblige à tout recommencer. Si l'on évoque ici ces tâtonnements de la recherche, c'est pour souligner que ces expériences ne sont pas anodines, et que les données recueillies devraient profiter au plus grand nombre de chercheurs possible. Ce serait par ailleurs une forme de reconnaissance à l'égard des informateurs, importante même si elle ne figure pas parmi les sommations du Comité d'éthique !

2. RÔLE DU CHERCHEUR ET DES EQUIPES DE RECHERCHE

Si l'on souscrit à ces exigences de fiabilité, la constitution d'un corpus se révèle être une entreprise de taille. La documentation, travail qui peut occuper à plein temps, apparaît comme incompatible avec l'activité de recherche et avec l'exigence de publication à laquelle sont soumis les chercheurs. Mais certaines opérations essentielles, et qui demandent relativement peu de temps, ne peuvent guère être réalisées que par le chercheur. Le premier travail consiste à distinguer les documents importants du tout-venant (copies, documents de travail...). Les transcriptions et commentaires doivent être conservées avec les enregistrements (idéalement, dans le même boîtier). A défaut de ce tri élémentaire, la documentation risque de devenir inutilisable. Les collections sonores de la Bibliothèque nationale contiennent des centaines d' heures d' enregistrements en langues étrangères sans transcription et à peine identifiés. Ceux qui sont une trace de langues disparues, potentiellement les plus précieux, sont inutilisables. Si la langue est toujours vivante, une re-transcription demanderait un investissement considérable, pour un résultat sans doute médiocre. De ce fait, le dépôt non documenté est aujourd'hui refusé par les phonothèques, et d'abord par la Bibliothèque nationale, qui a vocation à être un lieu de conservation, mais aussi de consultation : le travail de mise en forme doit être réalisé par le chercheur et son équipe de rattachement, à qui il incombe ensuite de proposer le fonds en dépôt aux bibliothèques, qui ne font pas de prospection pour récupérer les données de ce type.

Le travail de mise en forme comporte également la numérisation des enregistrements. L'adoption du support numérique est nécessaire, non parce que les nouveaux supports (CD, DVD) sont plus durables, mais parce que la procédure de migration d'un support à l'autre ne s'accompagne d'aucune perte de qualité dès lors qu'il s'agit de données numériques, et peut être automatisée. Le passage d'une bande magnétique à un document numérique nécessite que l'on règle le niveau d'entrée (pour que le son ne sature à aucun moment, et, à l'inverse, ne soit pas trop faible).

Ce transfert est rendu nécessaire par le vieillissement des supports magnétiques. D'après l'expérience solidement documentée de la Bibliothèque nationale, mises à part les bandes BASF et les bandes tri-acétate des années 50, très robustes, les bandes magnétiques doivent être numérisées, de même que les petites cassettes et les cassettes DAT, qui ne sont pas les plus solides : quatre à cinq ans après l'enregistrement, elles présenteraient déjà des irrégularités. (Pour un exposé encyclopédique sur ces questions, 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, et les techniques sont faciles à maîtriser. Mais cela représente une charge de travail bien réelle. Il faut l'aide d'un technicien pour éviter qu'il n'y ait un maillon faible dans la chaîne. Pour prendre un exemple peu connu : si un disque compact est destiné à l'archivage, il ne faut pas écrire dessus avec les "feutres spécial CD", dont les solvants finissent par traverser le plastique et risquent de rendre le CD illisible... Il faut donc souhaiter que les centres de recherche fassent participer leur personnel technique aux tâches de création de bases de données. Il ne paraîtrait pas non plus absurde

d'associer à ces tâches des étudiants (en qualité de vacataires).

3. CHARTE DE QUALITÉ

Les propositions qui suivent constituent une synthèse fondée sur l'expérience documentaire de diverses institutions, et guidée par un souci de simplicité. Le lecteur est également renvoyé à l'ouvrage [BGP01].

3.1 L'indexation et la transcription

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

(1) la LANGUE, la REGION où elle est parlée, une brève PRESENTATION DU LOCUTEUR : nom, sexe, date et lieu de naissance, langue des parents, et langues pratiquées au quotidien à la date de l'enquête

(2) l'identité de l'ENQUÊTEUR, les LIEU et DATE de l'enregistrement, sa DUREE, ses CARACTERISTIQUES TECHNIQUES : type de micro et distance du micro à la source sonore, matériel, vitesse d'enregistrement ou fréquence d'échantillonnage, qualité générale.

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

La transcription doit indiquer en détail l'OBJECTIF de l'enregistrement et les CONSIGNES données à l'informateur. Si on lui demande des mots qui n'existent pas, ou des logatomes, il importe de savoir comment on les lui a présentés. Si le questionnaire est écrit, il n'est pas inutile de savoir si l'informateur utilise couramment cette orthographe. S'agissant d'un corpus " spontané ", prendre soin de donner dans la transcription les explications nécessaires pour comprendre ce qui se joue dans le texte enregistré ; le corpus ainsi documenté pourra intéresser des chercheurs de diverses spécialités. 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

(1) si un traitement est appliqué au signal (filtre passe-haut ou passe-bas, changement de la fréquence d'échantillonnage, redécoupage...), cela doit être indiqué, et le document retravaillé doit être **accompagné de l'original numérisé non découpé** (fréquence d'échantillonnage minimum : 16.000 Hz, 16-bit).

(2) la compression en format MP-3 ne satisfait pas aux normes d'archivage. Le document non compressé doit donc être disponible. L'argument de l'économie de place n'est nullement déterminant, étant donnée les capacités actuelles de stockage. L'usage du mini-disque, qui enregistre directement en format compressé, est exclu pour la même raison : les algorithmes de compression dénaturent le son.

3.3 Relation avec l'informateur

L'importance de la relation avec l'informateur est souvent négligée. En particulier, le fait de définir la tâche de l'informateur comme un travail (rémunéré) est essentiel au sérieux de l'entreprise. Inscrire cette exigence dans la "charte de qualité" vise à attirer l'attention vers la question de la relation à l'informateur.

4. VERS LES BASES DE DONNÉES " DE POINTE "

Si un corpus satisfait à la charte exposée ci-dessus, et dès lors que les données (sonores, physiologiques...) et la transcription sont conservés ensemble sur support numérique, l'essentiel est acquis. A partir de cette base solide, on peut souhaiter aller plus avant. La mise en forme dépend de la nature des données et de l'utilisation que l'on souhaite en faire. Parmi les programmes les plus ambitieux, le corpus de parole lue et de parole spontanée réalisé par l'Institut de Phonétique de l'Université de Kiel offre un modèle spécifiquement destiné aux phonéticiens, qui permet une annotation détaillée et offre de puissants moteurs de recherche (voir [KRS97]). Dans une perspective " généraliste ", le programme Archivage du LACITO [JLM01] s'engage dans une autre direction : plutôt que de créer un logiciel nouveau, ce qui pose des problèmes de compatibilité, le programme Archivage recourt à des outils standard, autour du langage de balisage de texte XML, dont on peut résumer l'ambition ainsi : "[XML] will let us build great libraries simply by building our own sites (...), simply by tagging everything properly so it fits into the larger schema of the Web" [Sie97 p. 21]. L'investissement de temps que représente l'apprentissage des rudiments de XML et XSL 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 de codage de caractères paraît très prometteuse pour le codage des caractères (voir [Jac99]). L'ensemble du projet est présenté dans [JLM01]. Pour notre propos, retenons qu'un champ très intéressant s'ouvrel'énorme potentiel qui existe dans le domaine des bases de données

5. UN EXEMPLE : LE FONDS OUBYKH

En dernier lieu, nous souhaiterions exposer un exemple qui montre tout à la fois l'importance et la fragilité des données produites par les phonéticiens. 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]. Ils sont irremplaçables à plusieurs titres : tout d'abord, les données cinéradiographiques sont plus précises que ce que l'on obtient aujourd'hui par les techniques qui remplacent les rayons X (trop dangereux). En outre, 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. Le débat sur l'interprétation de certaines d'entre elles n'a pas encore été tranché, et 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 sans espoir, en l'absence de transcriptions) dans le cadre du programme Archivage du LACITO.

Les films aux rayons X sont en cours de numérisation avec le concours du Service du Film de Recherche Scientifique. La numérisation des bandes magnétiques a eu lieu au LACITO, où elle est une opération routinière. L'ensemble des dix bobines numérisées selon le standard du son CD (16 bit et 44.100 Herz, ce qui laisse une marge considérable, puisqu'on considère généralement que 16.000 Herz suffisent pour la parole) 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 l'article 3, 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 d'un réseau de " phonothèques phonétiques " qui assurerait une petite clientèle. Si l'on propose, à défaut de publication, une diffusion gratuite et informelle, les détenteurs des droits auront lieu de s'y opposer, puisqu'il n'y aurait pas de diffusion auprès d'institutions pérennes (Bibliothèque nationale, par le dépôt légal, et bibliothèques universitaires), de sorte que le problème de la conservation ne recevrait pas de réponse satisfaisante.

Cet exemple veut montrer que les initiatives documentaires ont besoin d'être soutenues et relayées institutionnellement, pour que les corpus les plus intéressants 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 lorsqu'on s'adresse à un public de chercheurs. Conscient de la fragilité de cette position, l'auteur souhaitait néanmoins affirmer cette conviction : qu'avec le problème de la qualité des données, c'est la question de l'objet même de la linguistique qui est en jeu.

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.

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

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

http://lacito.vjf.cnrs.fr/ARCHIVAG

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

[KRS97] Kohler K., Rettstadt T., Simpson A.P. (1997) *The Kiel Corpus of Read/Spontaneous Speech*, AIPUK vol. 32, Institut de Phonétique de l'Université de Kiel.

http://www.ipds.uni-kiel.de/publikationen

[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

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