# Création, diffusion et archivage de bases de données des langues rares : enjeux scientifiques et méthodes

Alexis Michaud

UNIVERSITE PARIS 3 – SORBONNE NOUVELLE
DIPLOME D'ETUDES APPROFONDIES

# Création, diffusion et archivage de bases de données des langues rares : enjeux scientifiques et méthodes

Mémoire présenté par Alexis MICHAUD

discipline : Phonétique

Directrice de recherche : Mme Jacqueline VAISSIÈRE

soutenue le 21 juin 2002 devant le jury composé de :

Mme Martine MAZAUDON, Présidente
M. Boyd MICHAILOVSKY
Mme Annie RIALLAND
Mme Jacqueline VAISSIÈRE

# Table des matières

# *Introduction :*

*Enjeu de la documentation des langues rares pour la science phonétique*

De septembre 1999 à juillet 2001, un séjour à Hanoi nous a fourni l'occasion de découvrir la richesse linguistique du Vietnam. Au début de l'année 2000, nous avons pris contact avec des linguistes qui travaillent dans le domaine des langues d'Asie du Sud-Est : Michel Ferlus et Barbara Niederer en France, Trần Trí Dõi et Nguyễn Văn Lợi au Vietnam. La charge de travail qui était la nôtre ne permettait pas d'enquêtes de terrain suivies. En revanche, tous reconnaissant l'indigence des fonds documentaires existant sur les « petites langues » du Vietnam, nous avons souhaité œuvrer à la mise en place d'un fonds documentaire à composante sonore, qui irait de pair avec un programme de collecte des langues rares du pays. (Le projet est reproduit en Annexe.)

Ce projet, qui s'est élaboré entre mai et novembre 2000, devait être réalisé à l'Université Nationale de Hanoi, en collaboration avec le Musée d'ethnologie du Vietnam, et en partenariat avec deux laboratoires du CNRS : le CRLAO (Centre de Recherches Linguistiques sur l'Asie Orientale) et le LACITO (laboratoire Langues et Civilisations à Tradition Orale). M. Trần Trí Dõi, professeur au Département de linguistique de l'Université Nationale de Hanoi, et Mme Barbara Niederer, chercheur au CRLAO, avaient accepté le principe d'une co-direction du projet, dont le Musée d'ethnologie était également partenaire. Le budget pour la première phase était ciblé par rapport au programme de collecte. Le projet a attiré à l'automne 2000 l'attention de M. le Premier conseiller de l'Ambassade de France, et de M. l'Ambassadeur, qui a sollicité un financement sur « amendements parlementaires » (accordés par le Sénat, avec un délai de mise en paiement rapide). Le projet a été accepté ; un financement à 100% a été accordé en décembre 2000. Mais cette somme a été redirigée vers un autre poste budgétaire par M. le Conseiller Culturel, ce qui a empêché la mise en route du projet.

De retour en France, accueilli à la Sorbonne Nouvelle et au LACITO (programme Archivage), nous souhaitions poursuivre la réflexion sur l'apport que nous pourrions faire dans le domaine de la documentation linguistique, de façon plus modeste et plus suivie.

Le point d'orgue du travail documentaire mené cette année dans le cadre du programme Archivage du LACITO a été la participation à la reconstitution d'un corpus de la langue oubykh (langue caucasienne disparue) qui s'était trouvé dispersé. Mais le présent mémoire décrit également d'autres travaux réalisés et en cours, dont l'un est directement lié au travail de recherche projeté pour une thèse de doctorat. Le souci de la documentation rejoint ainsi un travail personnel de recherche sur une langue rare.

## *1. Mais est-ce bien de la phonétique ?*

Un mémoire au sujet des corpus, présenté pour solliciter l'attribution d'un Diplôme d'Etudes Approfondies en phonétique ? Le rapport avec la recherche en phonétique paraît bien ténu. N'y a-t-il pas eu erreur d'aiguillage ? Ne s'agit-il pas plutôt d'un travail de bibliothécaire-documentaliste, ou de traitement automatique des langues ?

Le présent mémoire ne prétend pas innover au plan technique : écrit par un débutant à mesure de ses apprentissages (par exemple en ce qui concerne les langages XML et XSL), il reste rudimentaire au plan des techniques informatiques présentées. Le fait d'inscrire ce travail comme mémoire de DEA de phonétique est une façon d'exprimer une conviction concernant la nature du travail de conservation : il est important que des linguistes prennent en charge le travail de description des langues rares, pour produire des documents de grande qualité, qui aient des chances d'être conservés lorsque la majorité des langues existant actuellement seront éteintes.

En d'autres termes, le point de départ du présent travail est l'idée selon laquelle il appartient aux linguistes d'aujourd'hui de consacrer une partie de leurs efforts à la conservation du patrimoine linguistique mondial, la théorie linguistique ayant tout à y gagner, à court terme comme à long terme. La tradition de description de langues et de constitution méticuleuse de corpus par des linguistes existe depuis longtemps ; elle est en fait intimement liée au travail philologique de conservation du patrimoine culturel. A l'ère électronique, il est utile de faire le point des réalisations et des besoins, pour que puisse se poursuivre les recherches de *linguistique des langues*. L'accent mis sur la pluralité des langues ne doit pas être imputé à une quête d'« exotisme » : la prise en compte de cette pluralité permet seule d'éviter que les théories ne soient « glottocentriques ». *Le développement des recherches et les progrès dans la compréhension des phénomènes nécessitent la mise en place d'archives sonores par des linguistes.* L'époque actuelle représente un moment charnière, entre la découverte des moyens d'enregistrement modernes (audio et vidéo) et l'uniformisation linguistique (disparition des langues et civilisations à tradition orale, diffusion des langues nationales, influence universelle de la langue anglaise). Le phonéticien, soucieux, comme tous les

linguistes, de *définir son objet d'étude*, ne doit-il pas s'efforcer de *matérialiser son objet d'étude* sous forme d'enregistrements aisés à partager ? Or cela n'est pas encore réellement entré dans les habitudes : les Actes de la conférence *Speech Prosody 2002* (Aix-en-Provence, 8-11 avril 2002), publiés sur CD-ROM, ne contiennent presque pas d'illustrations sonores, qui paraîtraient pourtant particulièrement nécessaires dans le domaine de la prosodie. Le CD contient moins de 4 Mo d'illustrations sonores : seuls sept articles sont illustrés, par des fichiers de quelques secondes. Il ne s'agit pas ici de problème de capacités de stockage, puisque le CD contient en tout 38 Mo de données : il aurait donc été possible d'y adjoindre plus de 600 Mo de documents sonores. L'absence d'habitude documentaire chez un certain nombre de chercheurs en linguistique fait qu'ils ne sont guère sensibles aux questions des formats de données : dans les illustrations des Actes de *Speech Prosody* 2002 dont il vient d'être question, on trouve des documents numérisés à 11.025 Hz. Ces échantillons paraissent bien trop courts pour qu'ils permettent de répondre aux questions nouvelles que l'on veut poser, pour tester des hypothèses nouvelles. Ne serait-ce pas une attitude plus scientifique de transmettre des données plus étendues ? L'expérience du chercheur confirme l'intuition de simple bon sens selon lequel un va-et-vient entre données et modélisation est nécessaire[1]. Le développement de publications électroniques qui permettent de fournir des illustrations sonores d'articles marque certes un tournant dans le domaine des publications en phonétique, mais il ne paraît pas réaliste d'espérer que cette pratique apporte une solution au problème de la documentation sonore, puisque les illustrations de ce type sont nécessairement parcellaires au regard du système linguistique dont elles sont extraites.

## *2. Pourquoi chercher à conserver les « petites langues » ?*

Au seuil de ce travail, il paraît nécessaire de se demander pourquoi les réalisations dans le domaine de la documentation des langues rares restent relativement modestes, surtout au vu de la taille actuelle de la communauté internationale des linguistes. Une raison déterminante paraît être qu'aux yeux de beaucoup de chercheurs, l'enjeu de la recherche n'est pas à l'heure actuelle de collecter des données, mais de procéder à leur analyse. L'entreprise de conservation des langues menacées peut apparaître vaine : qu'espère-t-on au juste en emmagasinant fébrilement des données sur de nombreuses langues en voie de disparition ?

[1] Cette idée est par exemple formulée par Culioli : «...il faudra analyser la langue de plus en plus précisément et pour cela on peut faire plusieurs retour aux données empiriques et ce faisant réaménager son système d'analyse » (Culioli 1976, p. 9).

N’est-il pas plus éclairant de tisser des liens entre les connaissances existantes, pour réunir les descriptions fragmentaires et proposer des modèles universels ?

Cette attitude sceptique n’est pas nouvelle. On y retrouve la critique formulée jadis par La Bruyère :

> Quelques-uns par une intempérance de savoir, et par ne pouvoir se résoudre à renoncer à aucune sorte de connaissance, les embrassent toutes et n'en possèdent aucune : ils aiment mieux savoir beaucoup que de savoir bien, et être faibles et superficiels dans diverses sciences que d'être sûrs et profonds dans une seule. Il trouvent en toutes rencontres celui qui est leur maître et qui les redresse ; ils sont les dupes de leur curiosité, et ne peuvent au plus, par de longs et pénibles efforts, que se tirer d'une ignorance crasse. D'autres ont la clef des sciences, où ils n'entrent jamais : ils passent leur vie à déchiffrer les langues orientales et les langues du nord, celles des deux Indes, celles des deux pôles, et celle qui se parle dans la lune. Les idiomes les plus inutiles, avec les caractères les plus bizarres et les plus magiques, sont précisément ce qui réveille leur passion et qui excite leur travail ; ils plaignent ceux qui se bornent ingénument à savoir leur langue, ou tout au plus la grecque et la latine. Ces gens lisent toutes les histoires et ignorent l'histoire ; ils parcourent tous les livres, et ne profitent d'aucun ; c'est en eux une stérilité de faits et de principes, mais à la vérité la meilleure récolte et la richesse la plus abondante de mots et de paroles qui puisse s'imaginer : ils plient sous le faix ; leur mémoire en est accablée, pendant que leur esprit demeure vide. *(Les Caractères ou les Mœurs de ce siècle)*

Le linguiste « de terrain » serait de ces personnes qui « trouvent en toutes rencontres celui qui est leur maître et qui les redresse » : un spécialiste de phonétique trouvera à redire à sa description phonétique (trop peu expérimentale, ou sans statistiques ni même données chiffrées, par exemple), un spécialiste de syntaxe le jugera peu stable sur ses bases théoriques, les psycholinguistes auront également du mal à tirer parti de ses données, et ainsi de suite. Au vu du temps, de l’énergie et de l’investissement financier que demandent les enquêtes de terrain, et au vu des multiples obstacles matériels et psychologiques que rencontre le chercheur, le résultat peut paraître mince. Si l’on croit faire des découvertes, c’est parce qu’égaré par la fascination de l’exotique on « découvre » dans une langue rare ce qui existe aussi bien dans d’autres langues, bien décrites, dont on n’a pas pris le temps de prendre connaissance.

Face à ce raisonnement, il est nécessaire d’affirmer que la linguistique, comme toute discipline scientifique, se nourrit d’observations, et d’ouverture sur les faits de langues. Certes, l’enjeu est la compréhension « du langage » ; régulièrement, le linguiste veut croire à la promesse que contient le titre d’un article, ou d’un nouveau livre, qui promet de tirer au

clair *la structure de l'information linguistique*, ou les « maximes de la conversation », ou promet d'accéder au coeur d'une « grammaire universelle ». Cet espoir est ensuite déçu, avec tous les déchirements d'une rupture, puis relancé par une nouvelle tentative, dans un *mouvement sans mémoire*[2] des théories linguistiques. Pour prendre du recul face aux entreprises sans lendemain qui promettent l'universel sans passer par les langues, il peut être utile de citer les « pères fondateurs » comme Saussure : « ce qui nous est donné, ce sont les langues. Le linguiste est obligé d'en connaître le plus grand nombre possible, pour tirer de leur observation et de leur comparaison ce qu'il y a d'universel en elles. »[3] Il ne paraît pas nécessaire de multiplier les citations à l'appui de cette idée : il est vain d'opposer la « linguistique de terrain » à la recherche « strictement linguistique ». Le « terrain », c'est les faits de langues dont on veut rendre compte ; en ce sens chaque linguiste choisit son « terrain », mais certains sont plus riches que d'autres, ouverts sur un plus grand nombre de réalités. Il n'est pas fortuit que l'ouverture sur les langues du monde aille volontiers de pair, chez les linguistes, avec le soin dans le choix des données.

> On sait que la plupart des phrases de linguistes ne sont pas des textes réels, pas même une simulation dont on essaie de faire qu'elle soit de plus en plus adéquate à des énoncés véritables. Ce sont souvent des énoncés constitués de manière à se prêter à une analyse qui elle-même est fondée essentiellement sur une langue en tant que ramenée à une pratique écrite. (Culioli 1976, p. 10)

Pour envisager « un travail de type cumulatif sur les systèmes de représentation correspondant à des langues diverses » (ibid., Introduction), il faut des données en contexte, permettant l'étude de « tous les problèmes d'ajustements qui sont le propre de l'activité signifiante » (ibid., p. 5). « On ne doit pas considérer qu'il y a une partie de l'activité (ou du domaine) qui est un noyau de règles programmées et que là-dessus viennent s'ajouter les fioritures rhétoriques. » (ibid., p. 20.)

Par ailleurs, les linguistes ne doivent pas méconnaître en eux-mêmes le goût des « faits de langues », qui souvent motive leur vocation, et soutient l'intérêt qu'ils portent à leur discipline : ainsi de Saussure, comme en témoigne la lettre souvent citée qu'il adresse à Meillet (4 janvier 1894) :

> ... je vois de plus en plus (...) l'assez grande vanité de tout ce qu'on peut faire finalement en linguistique. C'est en dernière analyse seulement le côté pittoresque d'une langue, celui qui fait qu'elle diffère de toutes autres comme appartenant à un certain peuple ayant certaines origines, c'est ce côté presque ethnographique, qui conserve pour moi un intérêt (...).

---

[2] Le constat, et la formule, appartiennent à Antoine Culioli.

[3] *Cours de linguistique générale*, éd. Payot, 1972, p. 44 (= p. 92 de l'édition originale).

S'il était permis de calquer la formule de Pascal selon laquelle « Se moquer de la philosophie, c'est vraiment philosopher », on aimerait défendre l'idée selon laquelle se moquer de la théorie linguistique, et étudier les langues, c'est vraiment faire de la linguistique. Pour reprendre l'exemple de Saussure, c'est son expérience des langues dans ce qu'elles ont de plus vivant, expérience « presque ethnographique », qui entretient chez lui l'insatisfaction à l'égard des théories existantes, et motive en fin de compte le travail qui aboutira au *Cours de linguistique générale* :

> Sans cesse l'ineptie absolue de la terminologie courante, la nécessité de la réformer, et de montrer pour cela quelle espèce d'objet est la langue en général, vient gâter mon plaisir historique, quoique je n'aie pas de plus cher voeu que de n'avoir pas à m'occuper de la langue en général. Cela finira malgré moi par un livre où, sans enthousiasme ni passion, j'expliquerai pourquoi il n'y a pas un seul terme employé en linguistique auquel j'accorde un sens quelconque.

N'est-il pas dommage que cet appel à explorer les langues dans leur réalité, et à se défier des théories, trouve si peu d'écho parmi les professionnels de la langue ? En revanche, le détail des propositions théoriques de Saussure donne lieu à une exégèse méticuleuse, alors même que Saussure envisage sa contribution théorique comme purement négative : montrer les faiblesses des théories linguistiques.

Les découvertes « ethnolinguistiques » qu'évoque Saussure prennent une autre dimension lorsqu'il s'agit de langues vivantes. Outre la découverte de systèmes phonologiques et syntaxiques qui réservent encore des surprises au chercheur, l'étude de langues rares offre à l'enquêteur l'occasion d'un cheminement entre cultures, dépaysement dont il ne faut pas méconnaître la profondeur : il correspond sans doute (pour adopter le vocabulaire des jeunes sciences cognitives) à des modifications de « structures cognitives » profondément ancrées. Pour que les chercheurs de demain disposent de documents qui les renseignent de façon aussi complète que possible sur ces langues et les cultures qu'elles véhiculent, le plaisir du dépaysement linguistique doit, chez les chercheurs, se doubler du souci de faire parvenir des documents fiables à la postérité.

Nous avons voulu évoquer au seuil de ce travail la mine sévère des détracteurs de la linguistique de terrain, et rappeler que la linguistique de terrain fait parfois figure de « petite soeur » de disciplines jugées plus sérieuses. Pour lever ce préjugé, il est nécessaire que ceux qui pratiquent la « linguistique des langues » témoignent de ce que le but des enquêtes linguistiques n'est pas de jouer aux aventuriers (cela peut être une motivation au départ, mais ne doit pas éclipser le propos scientifique), mais de ramener des données de qualité. Cette idée est un fil conducteur du présent travail. Le problème du manque d'intérêt pour les « petites langues » tient pour partie à l'absence de données de qualité facilement accessibles.

Ainsi, A. Culioli souligne la nécessité de données fiables (*Ibid.*, p. 27) : « on ne peut généraliser qu'à partir d'études très sérieuses sur des langues dont la relation avec le français ou l'anglais est assez ténue : le japonais, le chinois, le malgache, les langues africaines... ». Il faut qu'existe une base de documents sonores transcrits de qualité, permettant une vérification de détail, faute de quoi les réflexions théoriques sont coupées de leurs racines. Beaucoup de chercheurs utilisent des travaux tels que *Patterns of Sounds*, de Ian Maddieson (Maddieson 1984), collection commentée des systèmes phonémiques de plus de 300 langues génétiquement diverses. A une étudiante qui soulevait le problème du son précis auquel faisait référence une notation extraite de cet ouvrage, le professeur Nick Clements a répondu qu'il n'était pas possible de le savoir exactement. Il concluait : « c'est le meilleur outil dont on dispose à l'heure actuelle » ; il incombe à la statistique de dégager des faits significatifs de cet ensemble non homogène, les approximations et partis pris des enquêteurs devant se compenser à l'échelle de plusieurs centaines de langues. Mais le chercheur qui procède ainsi s'interdit de prendre au sérieux les formes « uniques ». Or les formes encore jamais observées, dont on peut argumenter qu'elles sont de nature à faire progresser la réflexion en lui offrant un objet nouveau, n'ont aucune chance de faire surface dans les statistiques.

### *3. Urgence du travail de documentation*

Le constat d'ensemble est clair :

> Of the 6,000 languages listed in Ethnologue (Grimes 1992) for which there are population figures,
>
> - 52% are spoken by less than 10,000 people;
> - 28% by less than 1,000; and
> - 83% are restricted to single countries, and so are particularly exposed to the policies of a single government.
> - 10% are spoken by less than 100 speakers
>
> At the other end of the scale, 10 major languages, each spoken by over 109 million people, are the mother tongues of almost half (49%) of the world's population.
>
> There is agreement among linguists who have considered the situation that over half of the world's languages are moribund, i.e. not effectively being passed on to the next generation. We and our children, then, are living at the point in human history where, within perhaps two generations, most languages in the world will die out. (Source : "Manifesto of the Foundation for Endangered Languages"; Iatiku #2, p.2.)

L'ampleur des bouleversements linguistiques qu'introduisent aujourd'hui les échanges « mondialisés » apparaît clairement si l'on observe à quel point les systèmes phonologiques des langues du monde présentent des particularités de nature aréale : ainsi, les langues du Caucase sont riches en groupes de consonnes ; les langues d'Asie du Sud-Est présentent couramment des phénomènes de glottalisation au niveau lexical (tons glottalisés, phonèmes glottiques, registres phonatoires) ; les systèmes de tons ponctuels de nombreuses langues d'Afrique subsaharienne ont des similitudes entre eux. Lorsqu'une langue présente une situation typologique extrême, elle n'est généralement guère différente de langues voisines : ainsi, le système consonantique très complexe de l'oubykh a pu apparaître et perdurer au sein des langues caucasiennes du Nord-Ouest, au voisinage de langues comme le tcherkesse, l'abzakh..., elles aussi très riches en consonnes. En d'autres termes, outre les évidentes différences entre langues proches, entre dialectes, entre idiolectes, et même entre registres d'expression, il existe des différences qui n'ont pu apparaître que par un phénomène de relative isolation géographique, qui définissait un certain nombre d'ensembles linguistiques dont chacun pouvait présenter des caractéristiques typologiques affirmées, que les contacts continus entre langues ne modifiaient pas profondément, du fait que les langues (pour simplifier) étaient principalement en contact avec leurs voisines. La suprématie des langues nationales et l'influence universelle de l'anglais créent des « court-circuits » qui font disparaître de nombreux systèmes linguistiques et suppriment les conditions qui avaient permis leur apparition[4].

[4] Outre le domaine phonétique, cette évolution touche les composantes les plus diverses des langues. Soit l'exemple des systèmes de numération. Une comparaison fondées sur quelques grandes langues (anglais, allemand, français, chinois, hébreu, espagnol, vietnamien, japonais...) fait ressortir l'emploi généralisé du système décimal, ainsi que la très large diffusion des chiffres latins. Sur ce fond d'uniformité, il reste certes des différences entre langues : en chinois, 十三 (« dix-trois ») signifie 'treize', et 三十 (« trois-dix ») signifie 'trente', tandis qu'en hébreu on ne change pas le nombre auquel on fait référence en permutant les deux chiffres : les deux combinaisons signifient pareillement *treize*. S'agissant des unités, le chinois a un mot pour 10.000, et un autre pour ce nombre élevé au carré (100.000.000), à la différence du système « mille-million-milliard ». Par ailleurs, quelques sous-ensembles de nombres se singularisent par leur morphologie : les *–teen* de l'anglais ; les *quatre-vingt-dix* du français, opposés au *nonante* du dialecte de Suisse romande. Mais ces quelques spécificités sont bien loin de refléter la diversité des systèmes de numération qui ont existé dans les langues humaines. Martine Mazaudon (communication au sujet de la numération dans les langues tibéto-birmanes, prononcée en janvier 2002, à paraître dans le *Bulletin de la Société de linguistique de Paris*) présente des exemples de systèmes de numération à base quatre *(un, deux, trois, quatre, un-un, un-deux, un-trois, deux-un*, etc.) et de systèmes de numération duodécimaux *(un, deux, trois, ..., vingt, vingt et un, vingt-deux, ...vingt-dix, vingt-onze, vingt-douze, vingt-treize,* etc.) dans les langues tibéto-birmanes. Dans les systèmes duodécimaux, un mot particulier désignait le carré du chiffre le plus élevé, soit 400, qui correspond dans son principe au 100 du système que nous employons. Ma. Mazaudon a établi que la disparition de ces systèmes est liée à l'intervention de missionnaires occidentaux, qui ont voulu soulager les indigènes des complications de

La démarche adoptée dans le présent mémoire consiste à présenter certains fonds documentaires existants, qui donnent une idée de l'état actuel de la documentation, et les modèles d'archivage possibles, ainsi que leur finalité. Entre autres paradoxes, la tâche de conservation des langues en train de disparaître, et des cultures qui s'effacent avec elles, passe par un apprentissage des toutes dernières technologies, les plus hautement « mondialisées », définies pour le réseau planétaire par quelques groupes d'experts qui travaillent bien sûr en anglais. De fait, cela représente un effort, pour la même personne, de se tenir à jour, par le biais de l'anglais, sur les formats d'archivage informatique (et bien sûr sur les recherches linguistiques en cours) tout en étudiant une langue rare, dans laquelle se disent des choses que l'on espère restées assez éloignées des préoccupations « occidentales postmodernes »[5]. Mais c'est là la seule solution réaliste : il n'est pas imaginable de contrecarrer sur le terrain la « mondialisation linguistique », seulement de garder trace des langues qui aujourd'hui ne sont plus transmises d'une génération de locuteurs à l'autre.

En conclusion, le présent travail repose sur la conviction selon laquelle le développement des recherches et les progrès dans la compréhension des phénomènes langagiers nécessitent un travail de mise en place d'archives sonores par des linguistes.

leur système, qui paraissait contre nature aux yeux des missionnaires. Cette influence étant relayée par d'autres (en particulier : monnaies nationales décimales, scolarisation), le système décimal s'est répandu, au point qu'il ne reste plus que des vestiges de l'ancien système. Les changements linguistiques peuvent aller très vite. Il ne paraît pas vraisemblable que des systèmes autres que le système décimal resurgissent à l'avenir, pour la même raison de compatibilité dans un contexte de contact permanent entre langues.

En un sens, le système à base 4 et le système duodécimal ne disparaissent pas avec les langues qui les employaient : il s'agit de modes de fonctionnement que savent décrire les mathématiques modernes, qui connaissent l'ensemble du paradigme des systèmes de numération. L'esprit curieux a même aujourd'hui à sa disposition des objets d'étude que ne lui proposaient pas les langues naturelles : système hexadécimal, système binaire... qui trouvent une implémentation en informatique. La linguistique, discipline mal équipée pour traiter de ces questions, peut se décharger de la tâche de spéculer sur les nombres, confiant à des spécialistes ce domaine de l'invention humaine. Mais si l'on franchit ce pallier d'abstraction, il ne s'agit plus alors de linguistique, seulement de spéculation mathématique : étude du fonctionnement de systèmes mathématiques, non du langage humain.

[5] Le choix de s'attacher à ce qu'une langue a d'ancien plutôt qu'à ses efforts pour s'adapter aux réalités nouvelles est répandu parmi les linguistes qui mènent des enquêtes de terrain, bien que ce choix ne soit en général pas explicité. Ainsi, B. Michailovsky, étudiant des langues du Népal (hayu, limbu) dont tous les locuteurs parlent également la langue nationale, le népali (à des degrés divers), fait le choix, dans ses travaux

# *Première partie :*
# *Un état des lieux décevant*

## *1. Des phonothèques peu fréquentées par les phonéticiens*

Il peut être utile de présenter un état des lieux institutionnel de la documentation sonore, dont nous n'avons pour notre part acquis quelque idée qu'à l'occasion de la présente recherche, les programmes et collections en question n'étant guère fréquentés par la majorité des phonéticiens. Pour une présentation qui vise à l'exhaustivité, on pourra consulter *Les archives sonores en France* (éditions MODAL), brochure qui contient le rapport d'étude présenté en 1998 à la Mission du Patrimoine Ethnologique (Ministère de la Culture) et le compte-rendu des journées de réflexion sur les archives sonores organisées à Cordes en novembre 1999. Nous évoquerons d'abord ici les collections sonores des grandes institutions que sont la Bibliothèque Nationale et le Musée national des Arts et traditions populaires.

### *Les collections sonores de la Bibliothèque nationale de France*

A l'origine des collections sonores de la Bibliothèque nationale se trouvent les Archives de la Parole fondées par Ferdinand Brunot, dont P. Cordereix propose une analyse historique approfondie dans « Ferdinand Brunot, le phonographe et les 'patois' » (Cordereix 2001). Il peut être intéressant de rappeler cette entreprise visionnaire d'un « phonéticien » (même si le terme n'avait pas encore cours à l'époque), le lien qu'elle entretient avec la fondation de l'Institut de Phonétique de Paris (dont l'ILPGA est le descendant direct), et d'interroger ses motivations. Cette réflexion pourrait conduire un plus grand nombre d'universitaires à exploiter les collections gérées par la Bibliothèque nationale, dans un souci de profondeur historique de la recherche, et de collaboration avec les conservateurs dans le travail de mise en valeur des collections. Beaucoup de phonéticiens d'aujourd'hui se reconnaîtraient sans doute dans les préoccupations qui ont donné naissance aux Archives de la Parole : « la philologie et la phonétique expérimentale, qui toutes deux témoignent d'un intérêt, nouveau pour l'époque, porté à la langue parlée », au moment de « la naissance d'une phonétique qui ne va plus être une phonétique des signes, mais une phonétique des sons, prenant pour base, suivant l'expression de l'abbé Pierre-Jean Rousselot, 'non des textes morts mais l'homme

scientifiques, d'écarter systématiquement les emprunts népalis de l'analyse phonologique (communication personnelle), choix qui reflète un intérêt plus vif porté aux états anciens de la langue.

vivant et parlant' » (*Ibid.*, p. 40). La réflexion sur ces Archives de la parole (devenues Musée de la parole puis Phonothèque nationale) peut permettre de préciser certains des enjeux du travail de recherche, par exemple en mesurant l'écart entre le projet initial qu'avait Brunot d'un Atlas linguistique phonographique et les enregistrements effectivement réalisés par lui et Charles Bruneau. Il paraîtrait raisonnable de définir un projet de recherche au vu des fonds existants : pour les compléter et les exploiter, ou au contraire pour critiquer leurs fondements et s'engager dans une toute autre direction, mais sans les ignorer, car alors aucun enseignement n'est tiré d'un siècle d'entreprises et d'expériences d'enregistrement. C'est l'efficacité de tout un pan du travail de recherche qui est en jeu[6].

Il faut souligner ici le rôle des équipes de recherche. Le Département de l'audiovisuel de la Bibliothèque nationale n'a pas vocation à endosser la responsabilité de poursuivre l'entreprise des pères fondateurs : inventaire et transcription doivent impérativement être réalisés avant le dépôt en phonothèque. Les praticiens de l'enquête linguistique « de plein terrain » sur des langues jusque-là pratiquement non décrites ont formulé ces principes avec une grande clarté :

> « on doit toujours penser que les cahiers d'enquête doivent être rédigés comme s'ils étaient destinés à être repris par un autre chercheur. On doit rédiger ses enquêtes en envisageant que quelqu'un d'autre pourra les utiliser. C'est une exigence essentielle pour le travail en équipe : le travail individuel doit être accessible et pouvoir servir à tous ; d'autre part, si pour quelque raison il se trouve interrompu, il doit pouvoir être repris et continué par un collègue. » (Bouquiaux et Thomas 1971, p. 34)

En pratique, les dépôts réalisés au fil du vingtième siècle n'ont pas toujours satisfait à ces exigences fondamentales. Les collections sonores de la Bibliothèque Nationale contiennent quantité de documents linguistiques sans transcription. Parmi ces collections, des centaines d'heures d'enregistrements en langues étrangères, à peine identifiés. Les enregistrements de langues disparues, qui sont en un sens les plus irremplaçables, risquent de ne jamais être utiles à la recherche faute du catalogage qui permettrait leur repérage par les spécialistes qui sauraient en faire usage. 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.

[6] Ainsi, des chercheurs proches de la retraite nous ont dit avoir prêté main-forte à des collègues lors du dépouillement de corpus contenant des patois de France dont ils étaient familiers. La somme des efforts individuels déployés de Brunot à nos jours n'aurait-elle pas suffi à mener à bien le projet d'Atlas linguistique phonographique, si chaque contributeur avait orienté son travail au vu d'inventaires de la documentation existante et des tâches restant à réaliser ?

Il paraît donc important qu'étudiants et chercheurs disposent, au sein de leur laboratoire, d'une phonothèque contenant copie de nombreux corpus existants (fonds anciens et offres commerciales actuelles). 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.

Cet aperçu (bien trop rapide) de l'histoire et du rôle institutionnel des collections sonores de la Bibliothèque nationale souhaitait faire ressortir la nécessité de « phonothèques universitaires », centres de diffusion mais aussi de création de bases de données, qui aient, dans la « chaîne de la documentation parlée », un rôle intermédiaire entre les chercheurs (échelon individuel) et la Bibliothèque nationale.

## Le Musée national des Arts et traditions populaires

Les collections sonores du Musée national des Arts et traditions populaires correspondent à ce qui était auparavant la section France du Musée de l'homme. Elles contiennent des documents d'une grande richesse, qui renseignent sur de nombreux aspects des parlers et des musiques des « pays de France ». Les efforts déployés pour assurer la numérisation des fonds et pour répondre de façon adéquate aux « urgences patrimoniales », et la perspective d'une redéfinition des missions de l'institution dans son ensemble lors de sa délocalisation prochaine en région méditerranéenne, sont des sujets passionnants pour qui s'intéresse au patrimoine sonore. Pour le propos qui est le nôtre, on relèvera que le problème d'absence de transcriptions complètes n'est pas évité ici non plus : le volume des transcriptions est loin d'atteindre celui des enregistrements. Une indexation très précise des documents permet, dans la plupart des cas, d'avoir une idée précise du contenu ; en outre, s'agissant de dialectes français, on peut espérer, au plan linguistique, que beaucoup de documents soient quasi-transparents au chercheur d'aujourd'hui. Néanmoins, l'expérience de certaines équipes pourtant familières des patois concernées confirme le fait qu'un enregistrement dont il n'a pas été établi de transcription complète est d'un emploi difficile et hasardeux. Au risque de lasser par la répétition, tirons ici aussi la conclusion qui s'impose : si l'auteur d'un l'enregistrement le juge digne de devenir un document d'archive, il lui appartient d'en établir une transcription complète.

Par ailleurs, une importante règle de fonctionnement des grandes phonothèques est qu'il n'appartient pas à ces institutions de « prospecter » et de solliciter des donations auprès des

détenteurs d'archives sonores. C'est à ces derniers qu'il incombe de préparer leur fond pour le proposer ensuite en dépôt.

## Autres institutions

S'agissant de la phonothèque du Musée de l'Homme, qui contient également des fonds linguistiques qui intéressent au premier chef les linguistes, un courrier adressé « au Responsable de la Phonothèque du Musée de l'Homme » étant resté sans réponse, il n'a pas paru utile d'insister pour nouer un contact. Il faut espérer que des membres de la communauté des linguistes (et si possible des membres à l'assise institutionnelle plus solide qu'un étudiant) puissent être associés, d'une manière ou d'une autre, à la gestion de ce très riche fonds, et à sa bonne conservation, laquelle constitue une tâche de grande ampleur.

Parmi les institutions qui possèdent de grands fonds linguistiques se trouve également la Société des Missions Etrangères de Paris ; à l'heure actuelle, nous ne nous sommes pas encore renseigné sur leurs collections sonores.

D'autres institutions gèrent des collections sonores : les Archives nationales, et l'Institut National de l'Audiovisuel, INA. Les Archives nationales étant dédiées principalement à la mémoire institutionnelle, elles paraissent relativement éloignées des préoccupations des linguistes, de même que l'Institut National de l'Audiovisuel, dont les missions (pour simplifier) sont plutôt « synchroniques » que « diachroniques », et plutôt « utilitaires » que « scientifiques » : les documents archivés par l'INA peuvent bien sûr intéresser les linguistes, mais il paraîtrait saugrenu, par exemple, de vouloir adosser à l'INA une phonothèque universitaire de « langues rares ».

## Le programme Archivage du LACITO

L'objectif du programme est « la pérennisation, l'exploitation et la diffusion de documents linguistiques intégrant texte et son, en particulier les enregistrements faits et transcrits sur le terrain par les chercheurs du laboratoire » (Jacobson, Michailovsky et Lowe 2001)[7]. Il faut néanmoins souligner que le LACITO n'entre pas dans le cadre du chapitre « Institutions

[7] La version anglaise du même texte mentionne exclusivement les enregistrements réalisés par les chercheurs du LACITO (p. 80) : « The purpose of the LACITO Linguistic Archive project is the archiving of hundreds of hours of taped linguistic and ethnographic speech data, mainly in little-known and endangered languages, collected over the years by members of the LACITO research group of the French Centre National de la Recherche Scientifique (CNRS) ».

d'archivage », dans la mesure où le LACITO n'est pas une institution pérenne détenant des fonds d'archive, mais un laboratoire du CNRS, dont l'existence est suspendue à la décision de ses tutelles de le prolonger ou non, au vu du programme de recherche proposé sur une durée de quelques années. Pour atteindre pleinement son objectif de pérennisation des données linguistiques dont il assure la mise en forme et la diffusion, le programme Archivage du LACITO devra trouver à s'adosser à une institution pérenne.

### Les collections de l'Institut de Phonétique à Paris

L'Institut de Phonétique, qui est devenu ILPGA (Institut de Linguistique et Phonétique Générales et Appliquées), ne possède pas à l'heure actuelle de phonothèque. Cette institution n'a pas conservé ses fonds anciens. Quant aux travaux actuellement en cours, les solutions choisies pour la réalisation et l'annotation de corpus n'ont pas encore pu être définies de façon consensuelle. Les logiciels employés pour la visualisation et l'annotation de données ne sont pas conçus pour la création de corpus : les logiciels d'analyse du signal SNOORI et PRAAT proposent des outils d'annotation, qui permettent d'établir divers types de transcriptions, mais ne sont pas conçus pour de grands volumes d'annotations. Ainsi, l'annotation créée par SNOORI est enregistrée directement dans le document sonore, ce qui peut l'endommager ; l'annotation par PRAAT est plus complète, mais est loin de régler tous les problèmes de codage de caractères.

Au bilan, l'ILPGA n'a pas encore mis en place de base de données destinées aux étudiants et aux enseignants-chercheurs. Une réflexion est en cours pour trouver une solution à ce problème. Dans l'immédiat, ce sont les chercheurs qui ont l'entière responsabilité de leurs corpus.

## *2. La grande fragilité des fonds individuels*

Le devenir des fonds individuels lorsqu'un chercheur disparaît constitue une question importante, même si elle peut paraître de très mauvais goût. On aimerait employer une tournure moins brutale que l'évocation de la disparition du chercheur, mais la périphrase « arrive au terme de sa carrière » évoquerait plutôt l'aboutissement d'une vie de travail ; « parvient à la retraite » ne convient pas, les chercheurs n'ayant pas coutume d'interrompre leur activité à l'âge de la retraite. Sans s'apesantir sur ces réflexions, il faut reconnaître que les chercheurs repoussent généralement à plus tard le travail de mise au propre de leur fonds. Cela est compréhensible : cette mise au propre revient à préparer les données personnelles accumulées au fil des ans, avec l'idée que d'autres puissent s'en servir ; cela

revient en réalité à admettre que l'on va disparaître, et que les recherches vont être reprises par d'autres. Il faut pour cela recul et abnégation ; les rivalités qui peuvent exister entre savants n'encouragent pas les habitudes de partage des précieuses données. En outre, les chercheurs n'ont pas l'obligation de déposer leurs enregistrements et transcriptions auprès de l'institution qui a financé leur mission (comme cela a été le cas jusqu'à une certaine époque pour les chercheurs financés par le Musée des Arts et Traditions Populaires). Le dépôt progressif des données n'ayant plus lieu, à aucun moment la question de la transmission du fonds individuel n'est directement posée. Le résultat de cette situation est une perte documentaire considérable, particulièrement regrettable s'agissant de documents sur les langues en danger, littéralement irremplaçables. Dans ce dernier domaine en particulier, il y a lieu de penser que c'est dans les fonds individuels que se trouvent potentiellement les données les plus précieuses. 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 très difficilement égaler. (Un exemple concret, celui du fonds de langue oubykh, est décrit dans la troisième partie.)

### *3. Pléthore de formats, peu de données*

Une recherche sur Internet est maintenant une étape obligée du travail de documentation. S'agissant des langues rares, il existe un certain nombre de programmes internationaux bien visibles sur Internet. Globalement, un examen approfondi oblige à constater que le rendement documentaire des programmes n'est pas très élevé. Les formats proposés (indépendamment de la langue concernée) sont d'une grande complexité. Ils visent à proposer des modèles unifiés pour l'annotation des documents linguistiques. En revanche, la quantité de matériau linguistique et sa qualité ne sont pas très élevées.

Commençons par l'anecdotique. Une fondation d'inspiration chrétienne se propose de constituer une archive de mille langues, et de graver ces données sur une « pierre de Rosette » en nickel, pour que les ethnologues d'après l'apocalypse aient les moyens de redécouvrir notre civilisation, comme les savants du XIXe siècle se sont fondés sur la pierre de Rosette pour déchiffrer les hiéroglyphes égyptiens.

> The Rosetta Project is a global collaboration of language specialists and native speakers working to develop a contemporary version of the historic Rosetta Stone. In this updated iteration, our goal is a meaningful survey and near permanent archive of 1,000 languages. Our intention is to create a unique platform for comparative linguistic research and education as well as a functional linguistic tool that might help in the recovery of lost languages in unknown futures. We are creating

> this broad language archive through an open contribution, open review process and we invite you to participate. The resulting archive will be publicly available in three different media: a micro-etched nickel disk with 2,000 year life expectancy; a single volume monumental reference book; and through this growing online archive. We offer a growing collection of descriptions, texts, analytic materials and audio files for 1,000 languages. Contribution and review of translations, glossed vernacular texts, orthographies, core word lists, inventories of phonemes and audio files for languages in which you have expertise.

A cet effet, une liste de mille langues a été établie. L'inventaire en liste est impressionnant. Parmi ces langues se trouve la langue naxi, que nous sommes en train d'apprendre en vue d'étudier sa prosodie (travail de thèse de doctorat). Le site promet entre autres des données audio, qui nous intéressent grandement. Mais cliquer sur le nom de cette langue fait apparaître un écran décevant :

***Naxi***

*Ethnologue code: NBF*
*Alternate names: "Mo-su", "Moso", "Mosso", Lomi, Mu, Nahsi, Nakhi, Nasi*
*Family: Sino-Tibetan*
*Countries where spoken: China*

***Browse - Review - Contribute***
*Naxi texts are available in the categories below. The numbers in parenthesis indicate how many versions of each text type are currently in the archive. All texts need verification and review comments.*

*Detailed descriptions (1)*
*Genesis translations (0)*
*Glossed vernacular texts (0)*
*Orthographies (0)*
*Swadesh word lists (0)*
*Inventories of phonemes (0)*
*Audio files (0)*
*Basic color terms (0)*
*Miscellaneous texts (0)*

Les priorités documentaires pour remplir les blancs (qui forment la majeure partie de ce programme démesuré) sont clairement affichées : d'abord traduire le chapitre 1 de la Genèse.

Ce programme à la valeur scientifique très contestable illustre ce que peuvent être des programmes documentaires coupés des groupes de recherches scientifiques. Très visibles sur Internet, ces programmes retiennent une partie de l'attention du grand public et des mécènes. Il est important que des programmes sérieux leur fassent concurrence de façon constructive.

Une liste de ces programmes sérieux est fournie par Randy LaPolla, disponible à l'adresse http://ctspc05.cphk/lapolla/el.rtf. Présentons rapidement certaines des institutions qui travaillent dans ce domaine, à commencer par The Endangered Language Fund, qui offre un certain nombre de propositions pour contribuer au travail de conservation.

La nécessité de normes partagées, jugée essentielle pour que les travaux de documentation entrent dans une logique cumulative, a notamment donné lieu à un colloque récent à Santa Barbara (« Workshop on The Digitization of Language Data: The Need for Standards »). Le « Linguistic Data Consortium » de l'Université de Pennsylvanie est particulièrement actif dans ce domaine (Steven Bird, Mark Liberman). Un document rédigé par Steven Bird et Gary Simons (« Requirements on the infrastructure for digital language documentation and description », version du 14 novembre 2000) envisage les besoins des utilisateurs, des concepteurs, des techniciens, des archivistes, et des financeurs. Mais il n'est pas sûr que ces ambitieuses tentatives permettent de parvenir à un modèle complet, qui satisfasse les besoins et desiderata les plus divers. L'effort gigantesque qui est nécessaire pour concevoir des architectures telles que MATE (Dybkjær et Bernsen 2000) ou TEI (Sperberg-McQueen et Burnard 1994; Burnard et Sperberg-McQueen 1995) est certes admirable, mais il ne paraît pas indispensable d'attendre une (improbable) standardisation des formats et pratiques pour organiser nos bases de données sur les langues rares.

Le primat du format sur le contenu est un problème particulièrement sensible dans le projet DOBES (Dokumentation Bedrohter Sprachen), financé par la fondation Volkswagen. L'enjeu du programme est résumé simplement en quelques lignes :

> Approximately 6500 languages are currently spoken world-wide. It can be assumed that around two-thirds of these languages will become extinct in the 21st century. All languages are intimately interlinked with the culture of their speakers, and all languages and cultures represent specific expressions of human thought and social organisation. Therefore, with every language which becomes extinct priceless intellectual values will be lost forever. The project DOBES will contribute to the conservation of this cultural heritage. The MPI for Psycholinguistics in Nijmegen (NL) will house the data archive which will cover sound material, video recordings, photos, and various textual annotations.

Le programme s'appuie sur un partenariat scientifique prestigieux, fonctionnant par « appel d'offres » venant d'équipes de spécialistes des langues concernées, les données étant centralisées au célèbre Max-Planck-Institut. La partie « technique » du site est extrêmement détaillée. Un effort tout particulier est apporté à la définition claire des objectifs, du statut des équipes, des formats d'archivage, des questions de copyrights et de diffusion. En revanche, si l'on entre dans le détail des tâches de collecte en cours, l'état actuel du travail

amène à penser que ce projet qui bénéficie d'un important financement ne permettra pas de rassembler une documentation de qualité sur un nombre élevé de langues. La rigidité du cadre imposé est sans doute pour quelque chose dans le retard pris par le programme, qui indique par exemple (page Web consultée le 29 avril 2002) :

> # Page in Work
> # DOBES Archive
>
> This will be the entry page to the DOBES archive of resources about Endangered Languages. The metadata description of the resources will be open to anyone wishing to obtain information about the type of resources in the archive. The resources themselves will in many cases only be available by special request.
>
> The metadata will be browsable and searchable with the help of the IMDI tools. Assuming that the user has appropriate access privileges he may use the EUDICO tool set or view the files with another tool that works on the standard formats as adopted within the DOBES project, such as WAVE, MPEG, or XML. Information concerning these standard tools is available on the tools page.
>
> We expect a first version of the archive to become available during the fall 2001.
>
> Last updated: July 16, 2001

Soulignons en outre un principe du projet DOBES qui est clairement formulé dans les textes définissant l'enjeu du programme : la prise en charge à terme de la documentation (et donc sa pérennisation, qui représente l'enjeu premier du programme) n'est pas assurée par le projet. Ainsi, dans la page Web de la fondation Volkswagen :

> ...responsibility for the general accessibility of the documentation and continued data maintenance will lie with the applicants.

Enfin, le projet DOBES fixe une norme à appliquer par les équipes de chercheurs. DOBES apporte un modèle technique et une aide financière pour la constitution d'archives, mais ne se charge pas de l'intégralité du traitement des données. En d'autres termes, ce sont les équipes de recherche qui doivent réaliser elles-mêmes l'effort documentaire ; dans le projet DOBES, comme dans le quotidien de la recherche au LACITO, par exemple, c'est aux équipes de chercheurs qu'il incombe de réaliser, en pratique, le travail de documentation.

## *4. Nécessité d'une collaboration et d'un dialogue lors de la création de corpus : réflexions au sujet d'un corpus réalisé à l'ILPGA*

L'objet de ce paragraphe est de mettre le doigt sur certaines difficultés qui existent actuellement du fait de la faible diffusion des corpus, et du fait d'un faible degré de coordination entre tâches documentaires et tâches de recherche. L'exemple choisi est celui d'un contrat passé entre l'ILPGA et l'ancienne AUPELF-UREF (désormais AUF : Agence Universitaire de la Francophonie) : contrat S8-940/98.SD.1, « Constitution d'une base de données de français lu et spontané », sous la direction de Mme le Professeur Jacqueline Vaissière. Les auteurs nous ont aimablement communiqué un exemplaire du corpus et du document d'accompagnement. Les noms des destinataires sont indiqués sur la page de garde : A. Bonneau et Y. Laprie du CRIN (Nancy), et Khalef Boulkroune à l'AUPELF-UREF. Les auteurs n'ont apparemment pas reçu de nouvelles de l'utilisation qui a été faite du corpus après sa réalisation.

Le document d'accompagnement ne comporte que six lignes d'informations au sujet des locuteurs :

> Locuteur 1 : Femme, 49 ans, française (région lyonnaise), niveau d'étude supérieur.
>
> Locuteur 2 : Femme, 25 ans, française (région parisienne), niveau d'étude supérieur.
>
> Locuteur 3 : Homme, 37 ans, français (région lyonnaise), niveau d'étude supérieur.
>
> Locuteur 4 : Homme, 25 ans, français (région nantaise), niveau d'étude supérieur.
>
> Les quatre sujets ne présentent pas de troubles connus de la parole ni de l'audition. Ils s'expriment dans un français « standard », sans accent régional.

Cette présentation laconique a de quoi surprendre. Si les locuteurs s'expriment « dans un français 'standard', sans accent régional », quelle est la finalité des indications entre parenthèses, (région nantaise) (région lyonnaise) (région parisienne) ? Est-ce que la seule mention « Femme, 49 ans, française », « Homme, 37 ans, français » aurait paru trop simpliste, de sorte que les auteurs du corpus aurait ressenti le besoin d'ajouter une précision ?

À la consultation du corpus, un habitué des laboratoires de phonétique reconnaîtra certains grands classiques, par exemple l'histoire *La bise et le soleil*, mais sans explication sur le choix de ce texte, les références de la version choisie, les enregistrements déjà existants, et le projet de documentation phonétique auquel ce texte est associé : celui de l'Association Phonétique Internationale. Il peut paraître dommage de s'en tenir, comme c'est le cas, au seul commentaire « Texte de type *fable* : la bise et le soleil. » S'agissant des énoncés « spontanés », il n'est pas dit à l'observation de quels phénomènes ils peuvent servir, mais le

corpus n'est pas « tous usages » pour autant. Si l'on souhaite interroger ce corpus de parole spontanée pour une étude de l'intonation du français, les paramètres font défaut : faute de contexte, l'architecture pragmatique du texte est absente, ce qui complique (et même hypothèque d'emblée) l'interprétation de la prosodie de ces énoncés. Si, suivant le parcours implicitement proposé par les auteurs, on souhaite comparer ces données avec la parole spontanée relue, on doit postuler qu'il s'agit bien du même énoncé, motivé par la même intention de communication, avec cette seule différence que l'intention n'est plus réelle mais simulée. Le registre « Parole spontanée relue » serait donc une « lecture sentie », en mettant l'intonation appropriée, plutôt qu'une relecture « mise au propre ». Mais la transcription, qui adopte certaines traditions orthographiques de transcription de la parole orale, contient des tournures comme *je m'promenais dans la campagne* que le locuteur *lit* telles quelles, hyper-articulant le m qui devient une géminée [ʒəm#prɔm] (document F1L0014), alors qu'on entend, dans le document original, une longue pause : « parce que j'me... promenais dans la campagne ».

L'impression d'ensemble qui ressort de la consultation de ce corpus est le manque de clarté des enjeux : public visé, besoin auquel le corpus doit répondre, en un mot le cadre. Cette critique n'est bien sûr pas une attaque *ad hominem.* En formulant un point de vue critique sur ce corpus, le propos est de mettre au jour les exigences que peuvent avoir les utilisateurs, et souligner que l'architecture d'un corpus ne va pas de soi. Dès lors que l'on souscrit à certaines exigences de qualité, la collecte et la mise en forme de données devient une entreprise de taille. L'ampleur de la tâche suggère que la constitution de corpus doit être un travail d'équipe.

### *Bilan : Le rôle des équipes de recherche : Pour une charte de qualité des bases de données*

Les bases de données sonores abritées par les centres de recherches en phonétique sont assez peu fréquentées et peu structurées, si on les compare, par exemple, avec les bibliothèques universitaires. Les centres de recherche assurent rarement le suivi des documents enregistrés par leurs chercheurs. En se plaçant principalement (mais pas exclusivement) du point de vue de la conservation des langues en danger, essayons d'ébaucher une réflexion sur le fonctionnement qui pourrait être celui de « phonothèques universitaires », centres de diffusion mais aussi de création de bases de données.

Les possibilités offertes par les nouvelles technologies pour la création de bases de données multimédia sont vertigineuses. Les corpus déjà réalisés atteignent des volumes impressionnants, comme en témoigne par exemple le catalogue du site ELRA. 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. Dans le domaine de ces bases de données de langues rares, il paraîtrait opportun de mettre en place au sein des centres de recherche un réseau de création, de diffusion et de conservation de corpus. La collecte soigneuse des langues du monde intéresse plusieurs disciplines dans le champ des sciences humaines et revêt également une grande importance pour la mémoire des peuples concernés. Cette collecte nécessite un archivage des *données de chercheur*, qui demandent un traitement documentaire aussi exigeant que les *bases de données ingénieur*. 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.

A l'heure actuelle, 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 (Dô Thê Dung, Trân Thien Huong et Boulakia 1998). 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).

Une logique patrimoniale dans la gestion des corpus paraît donc nécessaire. Or le chercheur qui possède des données inédites est seul à même de préparer l'archivage de son fonds. Outre la transcription, le travail de mise en forme comporte également la numérisation. (Pour un exposé encyclopédique sur les questions des supports, voir Calas et Fontaine 1996). 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 plus d'attention aux tâches documentaires urgentes qui concernent notre discipline.

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.

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 Bonnemason, Ginouvès et Pérennou (2001).

### *L'indexation et la transcription*

Dans l'analyse documentaire du son inédit, on peut bénéficier d'indications et conseils rassemblés dans le guide de Bénédicte BONNEMASON, Véronique GINOUVES, Véronique PERENNOU (2001), *Guide d'analyse documentaire du son inédit, pour la mise en place de banques de données*, éditions MODAL[8]. Ce guide contient notamment des modèles de fiche de métadonnées ; le Musée des Arts et Traditions Populaires, la Bibliothèque nationale, l'Institut de Phonétique de Grenoble possèdent également des

[8] Il en existe une version plus ancienne, moins fournie : GENDRE Claude, FONTAINE Jean-Marc, GIOUX Andrée (1984), *L'oral en boîte : guide pratique pour la collecte et la conservation des enregistrements sonores*, Paris : Association française d'archives sonores.

formats de métadonnées bien adaptés. Pour retenir l'essentiel : un inventaire doit indiquer pour chaque document :

TITRE
LANGUE, REGION
HISTOIRE PERSONNELLE DU LOCUTEUR Sauf dans les cas où les convenances rendent malséante une « enquête sociologique » sur l'informateur, il est bon de disposer d'une brève HISTOIRE PERSONNELLE DU LOCUTEUR : nom, sexe, date et lieu de naissance, langue des parents, bilinguisme éventuel et langues pratiquées au quotidien. La langue qui fait l'objet de l'enquête est-elle sa langue d'usage ?
LIEU ET DATE DE L'ENREGISTREMENT
COLLECTEUR/ENQUÊTEUR
DUREE
GENRE/OBJECTIF DETAILLE DE L'ENREGISTREMENT accompagné des consignes détaillées données à l'informateur. Si on lui demande des mots qui n'existent pas (logatomes, par exemple), indiquer comment on les lui a présentés : « c'est un mot nouveau de ta langue, comment ça se dit ? », etc. Si le questionnaire est écrit, préciser : s'agissant de langues sans tradition écrite, le texte « lu » ne signifie pas la même chose que si l'informateur utilise couramment cette orthographe (exemple : anglais, français, chinois, japonais…). Si c'est un dialogue ou un corpus « spontané », donner dans la transcription les explications nécessaires à la bonne intelligence de ce qui se joue dans le texte enregistré.
DOCUMENTS CORRESPONDANTS : annotations, photographies, vidéo, et travaux non publiés et publications auxquelles le document a donné lieu. Doit se trouver sur le même support.
CARACTERISTIQUES TECHNIQUES :
Lieu d'enregistrement : chambre sourde, lieu d'habitation…
Type de micro et distance du micro à la source sonore
S'il s'agit d'une bande magnétique : préciser la vitesse d'enregistrement et l'état du support au moment de la numérisation.
qualité générale : excellente, bonne, médiocre. Préciser les problèmes : rapport signal/bruit médiocre, écho de la pièce…
COPYRIGHTS et DROITS D'AUTEUR... si, pour le document en question, des éléments de réponse particuliers peuvent être donnés à cette question délicate.
Il peut également être utile (pour l'archivage, même si le problème ne se pose plus vraiment après numérisation) d'indiquer où se trouve l'original, et qui en est le détenteur.

*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 déconseillé pour la même raison.

*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. Dans les enquêtes linguistiques de terrain, le principe est le suivant : qu'il s'agisse d'un salaire, de cadeaux, « un point est certain, de l'une ou l'autre façon, l'informateur doit être rémunéré » (Bouquiaux et Thomas 1971, p. 73). Inscrire cette exigence dans une « charte de qualité » qui s'appliquerait également aux enquêtes « en laboratoire » attirerait l'attention vers la question essentielle de la relation à l'informateur.

La « charte de qualité » ainsi formulée peut paraître contraignante, et lourde à mettre en place. En réalité, elle présente des avantages au plan de la flexibilité.

## Une application flexible

Le « cahier des charges » indiqué ci-dessus n'exige pas de contrôle par un organisme certificateur : c'est le laboratoire de recherche qui a la responsabilité de son interprétation et de son application. Les procédures de diffusion sont pareillement ouvertes. Corpus payants ou gratuits, l'essentiel est qu'ils soient élaborés dans les règles de l'art par le laboratoire qui les a conçus, et adossés à une institution pérenne. Etant donné la rapidité de l'évolution actuelle des supports, il ne paraît pas essentiel de fixer des procédures ; ce qui paraît en revanche urgent est l'implication des laboratoires de recherche.

Si l'on envisage les choses de façon optimiste, ces travaux soigneusement préparés pourront aboutir dans une phonothèque, lieu de conservation et de consultation. L'activité déployée par les linguistes pourra ainsi aboutir à un accroissement cumulatif de données sur les langues du monde. Si l'on envisage les choses de façon plus réaliste, et à plus court terme, des efforts pour la mise en place de phonothèques-bases de données au sein des centres de recherche permettront un état des lieu des données actuellement disponibles, qui convaincra rapidement les chercheurs du manque de données fiables et systématiques. Pour prendre un

exemple (qui ne recouvre nullement une critique adressée aux auteurs des travaux existants), si les petites collections de données diffusées par le *Journal of the International Phonetic Association* étaient en libre accès aux chercheurs et aux étudiants dans les universités, les limites de ces travaux pionniers apparaîtraient rapidement, telles que : enregistrements de qualité inégale, absence d'enregistrements systématiques de *l'ensemble* des oppositions phonématiques existant dans une langue donnée, emploi systématique de la traduction anglaise du mot (et non de sa forme phonématique en API) comme portail d'accès au fichier son correspondant.

# *Deuxième partie :*

*guide pratique d'archivage et présentation du programme Archivage du LACITO et de la base de données phonétique de Kiel*

Il a paru nécessaire de consacrer l'intégralité de la première partie à une réflexion générale sur « l'état des lieux », et à l'exposition d'une « charte de qualité des corpus » simple et indépendante du choix des outils (en particulier informatiques) à mettre en oeuvre. Cette démarche a permis d'offrir une rapide vue d'ensemble, qui donne maintenant tout son sens à l'exposition de la réponse que le programme Archivage du LACITO (www.lacito.archivage.vjf.cnrs.fr) apporte aux besoins constatés.

Le choix du programme Archivage tient au fait qu'il est particulièrement destiné à la conservation des langues rares ; de façon indépendante, un modèle de base de données permettant d'annoter un corpus de façon spécifiquement phonétique sera présenté : la base de données mise au point par l'Institut de Phonétique de l'Université de Kiel, qui prend pour point de départ une transcription « canonique », la transcription phonétique se faisant ensuite sur la base d'un examen des spectrogrammes. Ce programme, créé pour l'allemand, se prête à l'archivage et l'étude d'autres langues (dans le domaine des langues européennes, il a d'ores et déjà été utilisé pour le français). D'autres solutions techniques pour la création d'un corpus seront évoquées au cours de la discussion.

La présentation de ces programmes sera précédée d'un bref *guide pratique de l'archivage*[9] récapitulant les étapes préalables à la mise en forme des données selon le format proposé par le LACITO.

[9] Des expériences personnelles semblent montrer qu'une synthèse de ce type peut avoir une utilité : en 1998-99, souhaitant construire un corpus de qualité de la langue émérillon (langue tupi-guarani parlée en Guyane), nous nous sommes heurté à des difficultés, tant au plan de la technique que de la méthode. Après de nombreuses heures passées à ce travail mal engagé, le corpus n'était pas encore mis en forme.

## *1. Premières étapes de la création des documents*

### a. L'enregistrement

Qu'il soit audio ou vidéo, c'est la phase cruciale pour la qualité des données. La chaîne ne doit comporter aucun maillon faible : acoustique du local où on enregistre, type de micro, niveau des piles de l'appareil le cas échéant... Il vaut la peine de prendre le temps d'effectuer des tests. La séance doit être préparée ; par la suite, les documents doivent être étiquetés sans tarder.

### b. La réécoute et la transcription

C'est l'étape la plus gourmande en temps. Au plan technique, si l'on a fait le choix d'un enregistrement sur cassette standard ou cassette DAT, c'est une étape délicate puisqu'on soumet à rude épreuve le support qui contient l'enregistrement original. Une solution (lorsque c'est possible) consiste à recopier le document sur cassette standard, ou sur une autre cassette DAT, pour pouvoir le réécouter à loisir sans craindre d'endommager l'original. La réécoute avec un appareil à cassettes normales bon marché type Philips AQ 6350 ou Philips AQ 640 est commode, car ces appareils permettent de réécouter au ralenti.

### c. Numérisation et découpage

La numérisation des documents analogiques s'impose du fait du vieillissement progressif des supports analogiques. Au fil des ans, l'enregistrement perd de sa qualité, à commencer par les aigus, et les différentes couches superposées sur la bande magnétique s'impriment les unes sur les autres, créant un effet d'écho. D'après l'expérience solidement documentée de la Phonothèque nationale, les bandes magnétiques les plus solides sont les bandes BASF (quelle que soit l'année de fabrication), et les bandes tri-acétate des années 50. Les autres marques sont de moins bonne qualité, les bandes ne se conservent pas bien. Doivent être numérisées en priorité les bandes AMPEX et SCOTCH, et les bandes de toutes marques datant des années 1970, dont la qualité est plus mauvaise que jusque dans les années 1960. (Pour baisser les coûts de production, et produire en masse, les normes de qualité ont été revues à la baisse dans les années 1970.) Mais la perte de qualité ne concerne pas seulement les « vieilles » bandes magnétiques : les petites cassettes (produites en grandes quantités depuis les années 80), moins larges et dont la bande magnétique est moins épaisse, connaissent les mêmes phénomènes de vieillissement et d'usure. Enfin, les cassettes DAT (Digital Audio Tape), enregistrées en format digital, n'en sont pas moins des supports magnétiques, à la merci des champs magnétiques (proximité d'un téléphone ou d'un autre appareil contenant un aimant, etc.). Quelques années après leur enregistrement (trois à quatre ans, là encore selon l'expérience de techniciens de la Bibliothèque nationale), elles présenteraient des irrégularités. Une « migration des données » est donc indispensable.

Les supports qui accueilleront ensuite les données ne sont en fait pas plus durables que les anciens. Les CD-R et les DVD-R créés avec un graveur intégré à un micro-ordinateur reposent sur des techniques différentes de celles employées pour les CD de musique « du commerce », imprimés par les maisons de disques, qui sont gravés sur la galette en plastique du CD et sont de ce fait d'une très grande solidité, tandis que les CD-R et les DVD-R emploient une technique de *colorants* : la surface de ces supports est couverte d'une pellicule de pigments ; le graveur « brûle » une partie de ce colorant pour créer la séquence de vides et de pleins (les 0 et 1 du code binaire). Ces supports risquent donc de se détériorer lorsqu'ils sont exposés à la lumière. Une grande quantité d'information est stockée sur la faible surface du support, ce qui fait peser une hypothèque sur sa durée de vie. L'avantage des supports numériques réside dans le fait que le processus de « migration » d'un support à l'autre peut être automatisée dans le cas de supports numériques, et ne s'accompagne d'aucune perte d'information : une machine peut copier toute une pile de CD-R sur de nouveaux supports (par exemple des disques durs de grande capacité) sans qu'il y ait besoin d'une manipulation pour chaque disque, tandis que la copie de support analogique sur support analogique nécessite de répéter les mêmes gestes pour chaque document, et s'accompagne d'une perte de qualité des données.

En théorie, la numérisation est une simple tâche de routine (« a routine task » selon Jacobson, Lowe et Michailovsky 2001). En pratique c'est une opération qui demande un matériel relativement coûteux : appareil lisant le support original (cassettes, bandes UHER, cassettes DAT), éventuellement platine DAT qui effectue la numérisation, ordinateur possédant une bonne carte son, logiciel d'acquisition tel que SoundForge. Les linguistes peuvent sans difficulté avoir accès à ces matériels coûteux, qui sont des équipements de base des laboratoires de phonétique. Mais il est bon d'avoir 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 est donc préférable d'écrire un minimum sur le CD lui-même. Par ailleurs, il est important de choisir un matériel robuste, pour minimiser les risques de perte de données. Il existe des marques de CD-R plus spécifiquement destinées à l'archivage de données ; la Phonothèque Nationale (Bibliothèque de France) utilise la marque Mitsui. Dans le choix d'un graveur de CD-R, c'est apparemment la marque Yamaha qui donne les meilleurs résultats. Enfin, il est apparemment plus sûr de graver à petite vitesse, non à la vitesse maximale (proposée par défaut par les logiciels de création de CD-R).

Outre la formation minimale qu'il faut acquérir pour bien réaliser la numérisation, l'opération prend un certain temps et demande une certaine attention. Il serait concevable que les chercheurs demandent ce service à des étudiants qui travaillent dans le même

domaine de recherche. C'est aux débutants d'aujourd'hui que profiteront plus tard ces fonds ; il ne devrait donc pas être impossible de les intéresser à participer à leur constitution. Ces tâches pourraient faire l'objet d'un TD obligatoire, et figurer parmi les *savoir-faire* que les étudiants pourraient faire valoir au terme de leur cursus.

Lors de la numérisation des documents (ou plus précisément de leur re-numérisation, s'agissant de support déjà numériques tels que le DAT), la fréquence d'échantillonnage utilisée pour l'archivage au LACITO est 44.100 Hz, 16 bit : le standard du son CD. C'est beaucoup plus que ce qu'utilisent actuellement les phonéticiens dans la plupart des corpus ; le « standard » serait plutôt 22.050 Hz ou même 16.000 Hz, parce que ce choix suffit pour la lecture de spectrogrammes. Le logiciel Snoori, qui ajuste l'échelle d'affichage des fréquences à la fréquence d'échantillonnage du fichier sonore, oblige de fait ses usagers à retailler les fichiers à la taille de 16.000 Hz, faute de quoi les spectrogrammes ne sont pas lisibles. Il faut souligner que ce nivellement par le bas de la qualité des fichiers son n'a pas lieu d'être étant donné l'augmentation rapide des capacités de stockage et de la rapidité de travail des processeurs : un ordinateur personnel « milieu de gamme » (autour de 1.000 euros) peut aujourd'hui de gérer des fichiers son de plus de 300 Mo, soit environ une heure d'enregistrement Mono à la fréquence de 44.100 Hz ; le logiciel SoundForge permet très commodément de les retailler pour en extraire le passage que l'on souhaite visualiser en détail, et de rééchantillonner cet extrait pour le ramener à 16.000 Hz. Il est donc important de disposer, à des fins d'archivage, de fichiers numérisés à 44.100 Hz. Cette fréquence d'échantillonnage permet par ailleurs de graver directement des CD audio, ce qui peut être utile dans certains cas. Ainsi, le programme Archivage du LACITO diffuse les documents audio sur Internet en format compressé (MP-3), mais conserve par ailleurs les « originaux » numériques en « format étendu » (44.100 Hz), qui pourront être mis à disposition lorsque les « autoroutes de l'information » seront en place, et dans tous les cas demeurent l'indispensable point de référence.

L'étape suivante consiste à graver ces données sur CD-R : les fichiers .wav et leur transcription.

Cela représente une grosse somme de travail, mais 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. Si le linguiste ne souhaite pas entreprendre la mise en forme de ses données selon un format plus élaboré, il aura du moins permis la conservation des données, ce qui permet d'envisager leur transmission. 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.

## *2. L'analyse documentaire des fonds*

Une phonothèque, si petite soit-elle, a besoin d'un fichier qui présente l'ensemble du fonds de manière à contenir toutes les informations nécessaires pour retrouver un document. L'analyse documentaire permet de mettre en relation documents sonores originaux (typiquement sur support magnétique), documents manuscrits (cahiers de transcription-annotation), et supports numériques qui conservent les documents sonores originaux.

Soulignons d'abord que l'analyse documentaire des documents numérisés et préparés par le chercheur est beaucoup plus facile que celle de documents magnétiques originaux. Une bande enregistrée sur le terrain peut contenir des choses disparates : quand le stock de bandes est épuisé, on commence à utiliser les espaces restants sur les bandes magnétiques déjà utilisées : deuxième piste ou deuxième face, ou espace libre à la fin d'une piste déjà enregistrée. Cela crée rapidement un puzzle : un document peut ou non correspondre à une bande magnétique entière. Il peut n'en occuper qu'une partie, et partager l'espace avec d'autres documents. A l'inverse, un même récit peut occuper plusieurs bandes magnétiques : qu'il se trouve sur les pistes A et B ou sur deux bobines différentes, il y a une coupure en deux morceaux. Pareillement, les cahiers de transcriptions ont leur histoire : ils ont été mis au propre (avec des modifications), enrichis avec les informateurs, réagencés selon une logique donnée qui peut être différente pour chaque cahier : projet éditorial, classement par informateur ou classement par type de document. A l'aide des techniques très commodes de retravail du son numérique (logiciel SoundForge, par exemple), un fonds cohérent peut être reconstitué ; il paraît plus avisé de parvenir ainsi à une architecture simplifiée et améliorée, que reflétera le fichier d'inventaire. L'architecture du fonds ressort alors clairement, et l'analyse documentaire est grandement simplifiée. Si, en revanche, le travail de numérisation et de regroupement des documents sous forme numérisée n'a pas pu être fait en collaboration avec le chercheur, et que l'on choisisse de conserver sous forme numérisée l'« architecture par défaut » qui était celle du fonds sous forme magnétique, cette architecture non seulement n'est pas plus claire que celle des bandes magnétiques, mais elle est nécessairement un peu plus complexe, puisque les supports numériques sur lesquels sont conservés les fichiers sonores numérisés représentent une autre forme de découpage, qui s'adapte à la taille du format. On peut faire tenir environ 60 minutes d'enregistrement en 44.100 Hz stéréo sur un CD-R, ou 120 minutes en mono ; ne graver sur un CD qu'un document correspondant à une bande magnétique peut amener un gaspillage considérable de place ; en revanche, si l'on calcule en fonction de l'espace disponible sur le support numérique, on va aboutir à un découpage différent de celui des bandes magnétiques (par exemple, 4 CD pour 12 bandes), ce qui obligera à compliquer l'index : il faudra préciser, à

propos de tel document référencé comme « bande magnétique 3 », sa position sur tel ou tel CD. Il est donc préférable de revoir l'architecture du fonds à l'occasion de sa numérisation.

De même pour les documents papiers originaux, « notes de terrain » et transcriptions : si le chercheur peut effectuer lui-même la synthèse des notes, et produire un document (informatisé si possible, ou à défaut un manuscrit) qui représente l'état le plus avancé de son annotation, les versions précédentes, toutes moins complètes, n'ont plus besoin d'être référencées en détail, sauf éventuellement par une indication du genre « Des papiers préparatoires sont conservés dans mon bureau dans tel carton ». Si tel ou tel état du travail présente un intérêt particulier (ex.: notation phonétique, remplacée dans les derniers états du travail par une notation plus phonologique et moins proche des réalisations de surface), le mieux est de le saisir aussi sur informatique, et de le mettre sur le CD d'archivage. A défaut, on peut indiquer où le retrouver. Mais il vaut infiniment mieux avoir quelque chose d'abouti. Cette conviction va à l'encontre du respect « philologique » du document dans tout son détail, voire dans sa matérialité : quiconque est rompu à l'exercice de l'*explication de texte* peut argumenter solidement que chaque détail d'un manuscrit a son intérêt et sa portée, que les différentes versions d'un texte ont chacune leurs particularités, qu'aucune version ne reflète intégralement. Mais il faut reconnaître que les documents de travail des linguistes ne méritent pas nécessairement d'être scrutés dans le même détail qu'un manuscrit de poésie : dans l'extrême majorité des cas, un historique détaillé des brouillons aboutissant à une transcription de texte, ou à un article scientifique, n'aurait aucun sens. Mieux vaut donc profiter de la constitution de l'« archive » pour achever de mettre au point une annotation bien faite, en impliquant le chercheur.

## Le format de la base de données

Dans le choix d'une structure informatique pour la base de données, il est très coûteux de ne pas adopter un standard. Cela demande beaucoup de travail à l'informaticien qui construit les programmes *ad hoc* et maintient la structure (dont il est prévisible qu'on souhaite la modifier au cours de son utilisation). Il faut qu'un informaticien soit là pour régler les problèmes à tous les stades ; en particulier, les interfaces d'interrogation sont lourdes à mettre en place. Si l'on s'adresse exclusivement à des utilisateurs qui connaissent le langage d'interrogation des bases de données (SQL), on leur laisse le soin de rechercher les informations par leurs propres moyens ; mais si l'on s'adresse à des personnes qui ne sont pas spécialistes d'informatique, il faut définir des interfaces raisonnablement conviviales et transparentes. Le problème des normes se posera de façon cruciale si on veut mettre une partie de la base de données sur Internet (perspective qu'il paraît dommage d'abandonner *a*

*priori*). Mieux vaut donc adopter un standard de bibliothèque et s'y conformer, quitte à placer de grandes quantités d'information dans les champs marginaux (« Commentaires », « Autres »).

### Les droits d'auteur

*Last but not least* parmi les problèmes d'analyse documentaire, les documents non publiés posent des problèmes de droits d'auteur : le chercheur peut souhaiter assurer la conservation de ses données tout en gardant ses prérogatives d'accès exclusif à son propre fonds. Les droits d'auteurs des informateurs sont également à prendre en compte, même si ceux-ci n'envisagent pas nécessairement les questions de « propriété intellectuelle » dans les mêmes termes que les enquêteurs qui viennent les voir : les questions les plus sensibles concernent le contenu des textes, l'enregistrement conservant fidèlement la trace de confidences qui ne seraient pas à leur place sur un site Internet. Dans la perspective d'un archivage pour le long terme, il peut être prudent de fixer une date pour la diffusion des enregistrements. Selon les cas, cela pourrait être : après le décès du chercheur, pour les documents dont il souhaite conserver les *droits exclusifs d'exploitation* de son vivant ; cinquante ans après le dépôt, voire plus, pour les documents contenant des informations de nature personnelle à ne pas publier... En théorie, il serait concevable que le LACITO accueille une copie de chaque document numérisé avec le matériel du programme Archivage, sans diffuser ces documents tant que toutes les personnes disposant du droit d'auteur ne l'ont pas autorisé. Mais en pratique, le programme Archivage du LACITO pourrait difficilement prendre en charge ce travail de garantie de la confidentialité des données : il faudrait une infrastructure adaptée, avec des procédures qui donnent de bonnes garanties de confidentialité. L'existence d'une phonothèque « semi-pérenne » accueillant les archives préparées en attendant qu'elles ne puissent être transférées à un lieu de consultation « pérenne » (la Bibliothèque Nationale ?) permettrait la mise en place d'un tel système, et apporterait par là un élément de réponse à ce problème épineux qui n'a pas de réponse claire à l'heure actuelle.

## *3. Le programme Archivage du LACITO*

Deux citations d'un article de M. Jacobson et B. Michailovsky résument la vocation de ce programme :

> « Le Programme Archivage du LACITO (Laboratoire de Langues et Civilisations à Tradition Orale du CNRS) a pour but la pérennisation, l'exploitation et la diffusion de documents

> linguistiques intégrant texte et son, en particulier les enregistrements faits et transcrits sur le terrain par les chercheurs du laboratoire. »
>
> « ...a major goal of the project is to make the recordings and their textual documentation available to researchers, including those who recorded them in the first place, using modern, computer-aided research methods. » (2001 p. 80)

Il n'est pas inutile de commencer par justifier le choix de ce format, qui au premier abord peut paraître à la fois complexe et spartiate d'apparence, comparé aux outils plus conviviaux que proposent les maisons d'édition de CD-ROM, par exemple les outils d'enseignement des langues. Les amateurs de langues sont peut-être familiers des outils informatiques que proposent les éditions Assimil. Assimil propose une collection de CD-ROM conviviaux d'apprentissage des langues. Cela permet d'étudier de façon plus dynamique et commode qu'une cassette. Les phrases sont affichées à l'écran une après l'autre. On peut s'en servir dans le sens "version": on voit le texte en langue étrangère, et l'on doit chercher à le traduire. On peut l'écouter le nombre de fois que l'on veut, toute la phrase ou un seul mot, ce qui est très commode pour s'imprégner du texte. On peut voir le texte en français et s'essayer au thème avant de demander la solution ; là encore, on peut écouter la phrase en langue étrangère autant de fois qu'on le souhaite. Pour le chinois, il est également possible de voir le tracé des caractères, à une vitesse plus ou moins élevée, ce qui est d'un grand secours pour retenir les caractères. Séduit par ce format, on peut concevoir l'idée d'un fonds documentaire sur les langues rares qui ait recours à ce type de présentation, à laquelle on ajouterait des lignes supplémentaires d'annotation, et des moteurs de recherche.

En réalité, les systèmes multimédia de ce type, qui sont assez nombreux, sont réalisés avec des programmes tels que QuickTime et Director, qui permettent de faire du multimédia sous Windows. Ces systèmes sont des systèmes « propriétaires » développés pour répondre à une demande précise, à un moment précis. Ils sont liés à MacroMedia, aux outils avec lesquels ils ont été produits : ils ne peuvent être partagés, et il y a un risque qu'ils cessent tout simplement d'être exploités, remplacés par de nouveaux outils. Ce n'est pas une solution sur le long terme. On comprend ici l'importance de recourir à des outils standard et évolutifs. 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 (Jacobson 1999). De fait, le programme Archivage, malgré sa petite taille, apporte une contribution très significative au travail de documentation, ainsi qu'au débat international sur les méthodes de structuration des données linguistiques (participation de B. Michailovsky et M. Jacobson aux congrès de Santa Barbara 2001 et La

Palmas 2002) ; le logiciel SoundIndex, créé par Michel Jacobson pour réaliser commodément la synchronisation entre son et annotation, a fait des adeptes à l'étranger, montrant que de petits programmes sont en effet légitimes, et efficaces. Une mise en forme conviviale est tout-à-fait envisageable, sur la base solide de cette architecture XML ; le meilleur exemple en est le CD-ROM sur les langues de Nouvelle-Calédonie réalisé en collaboration par le LACITO, pour le Centre culturel Jean-Marie Tjibaou de Nouméa.

La description du programme Archivage par les auteurs eux-mêmes est fournie sur le site internet du programme. Il n'est pas ici question de revenir sur ces descriptions pour en proposer un florilège de citations, mais de présenter les étapes du travail de constitution d'un document et de son archivage selon le format créé au LACITO.

La mise en forme des transcriptions selon le « format LACITO » consiste à coder le contenu de l'annotation (transcription, analyse, gloses interlinéaires, traductions) en langage XML. Les outils informatiques utilisés pour traiter cette information sont tous plus ou moins directement liés à XML (« outils génériques XML »), par exemple le logiciel SoundIndex qui permet la synchronisation du son avec l'annotation, l'applet JAVA qui permet aux browsers d'accéder au son, feuilles de style XSL qui définissent les vues sur les données. La large compatibilité d'XML permet, pour une diffusion par Internet, de traduire les données à transmettre en langage HTML (langage exclusif de dialogue par Internet). Cette opération est effectuée par le serveur du LACITO pour le "client" (la personne qui consulte le site) ; mais à partir de XML, on peut également créer des fichiers PostScript, RTF (format d'échange), PdF, ou .txt, en fonction des besoins.

Pour préparer une « archive LACITO », il faut connaître les rudiments d'XML :

- Le balisage

- La DTD, « Document Type Definition »

- Les feuilles de style.

XML est l'acronyme de « eXtensible Markup Language ». Pour comprendre la raison d'être de XML, il faut d'abord évoquer son « aîné », SGML (qui, dans le langage du Web, est son *ancêtre*, bien que seulement dix années les séparent).

SGML est « Standard Generalized Markup Language », *« an open, extensible technology that facilitates sound information management techniques »* (Dan Connolly 1997 p. 1) Les deux acronymes ont en commun « ML », « Markup Language » :

> Structure is markup. Markup is structure. To mark up a document is to describe its structure using metadata, also known as tags. Tags are meant to describe contents, not presentation. For example, the <P> tag denotes a paragraph. (...) Imagine all the things that go into a newspaper. Properly tagged, you could apply a layout engine, using a set of layout rules, to layout an entire newspaper automatically. (David Siegel 1997 p. 14)

Une notion centrale est celle du *balisage* (« tagging ») :

> "I'm writing this document in Microsoft Word. To tell you that *The New York Times* is a newspaper title, I've used the italic feature, as any good editor would. If you read it in HTML, someone has put it between <I> and </I> tags, so that the people who consult this paper over the Internet see the text in italics, too. Blasphemy! How dare we use italics, when we mean <newspapertitle>The New York Times</newspapertitle>, don't we? If we had a <newspapertitle> tag, then people with 24x80 terminals in Ximbabwe would see "The New York Times" rather than *The New York Times*, because on a 24x80 terminal, you can't display italics. The browser itself adds the quote marks--*they would not be part of the document.* In this case, the tag indicates meta-information, which the User Agent (also known as a browser) interprets however it can on the target display. Similarly, if this text were being spoken by a speaking browser for the blind (yes, there are some--and no, they're not very good), the <newspapertitle> tag would be a signal to the program to pause and then emphasize the name of the paper. Gripping, isn't it?"

David Siegel montre de façon frappante ce que permet le balisage de texte, en proposant une vision (bien sûr chimérique) d'une application généralisée et standardisée du principe du balisage :

> "In a perfectly tagged world, the big search engines do all the work for us, by searching the Web and storing not only the data, but also the metadata. In a perfectly tagged world, we would standardize our tags, so that everyone would use the same exact tag to denote a <MOVIEREVIEW CLASS=HORROR> or <RECIPE CLASS=VEGAN>. If one person decided to use a tag called <MOVIEREVIEW> and another used one called <FILMREVIEW>, the search engines would have a hard time keeping up with all the new tags (...). Hence, the need for standardized markup (...)." (ibid.)

Le balisage est solidaire d'une définition de la structure du document : quel type de données contient-il ? Comment sont-elles organisées ? C'est la question de la définition du type de document (DTD, en anglais : Document Type Definition), que David Siegel justifie de la façon suivante :

> Go one step further and say what kind of a document you are reading--a play, a newspaper, a Ph.D. thesis, a recipe book, a journal--and you soon need to generalize the language to include the document type, followed by the appropriate markup for that document type. Cookbooks contain recipes, report cards contain grades, plays contain dialogue, scene and action description, and so forth. (...) I've just performed a magic trick: Now we know what SGML is--Standard Generalized Markup Language. (ibid., p. 15)

La DTD définit : les noms des éléments autorisés, la fréquence d'apparition des éléments, leur ordre d'apparition, et leur architecture interne (c'est-à-dire le nom des éléments autorisés à apparaître à l'intérieur d'autres, jusqu'au niveau des données caractères). Elle précise les attributs des balises et leur valeur par défaut, et les noms de toutes les entités qui doivent être utilisées.

SGML est défini dans la norme ISO 8879:1986. En ce sens, c'est un point de référence stable. La langue du World Wide Web, HTML (HyperText Markup Language), se fonde sur SGML ; mais certaines difficultés techniques liées à HTML ont amené sa « dialectisation », ce qui nuit à sa vocation d'outil de communication partagé. XML permet de résoudre un certain nombre de ces problèmes.

XML est un « dialecte » simplifié de SGML. « SGML is too general and too complicated for the Web. Instead, we need a junior version, and that's called XML. » (David Siegel, op. cit., p. 18) Pour donner une idée des relations entre HTML et XML, Dan Connolly écrit :

> This is not to say that HTML will fade away: not everyone wants to develop a new document structure, stylesheet, or Java applet just to put up a Web page. HTML will always be there making the easy things easy. But with XML, if you want to go beyond the boundaries of HTML, it will be straightforward to do just that. (p. 3)

XML a vocation à être simple, et « lisible par l'homme ». Notre expérience de débutant confirme ce qu'affirme Norman Walsh (1997, p. 97) : « For the most part, reading and understanding the XML specification does not require extensive knowledge of SGML or any of the related technologies. »

## Les feuilles de style

On voit maintenant la façon dont est conçu XML, formalisme d'expression de la structure logique des données. Un autre principe important à comprendre est celui des feuilles de style.

C'est encore à David Siegel (1997 p. 18) que nous empruntons l'explication de ce principe : « If we're going to learn anything, it's that style sheets are the future and tag-based layout is the past. What can style sheets do? They can do a lot of typographic things, like set your margins, indents, drop caps, leading, and other niceties invented in the time of the Romans. »

Les feuilles de style permettent de rechercher un certain type d'information dans les documents XML, et de les afficher comme on le souhaite. L'avantage déterminant des feuilles de style comparées à une mise en page « traditionnelle » est leur faculté d'adaptation.

> « Style sheets have the capability to degrade gracefully. Today, you can specify which fonts to use for which sections of a site, and if those fonts aren't available, you can specify a second and third choice, and so on. Style sheets go further. If a style sheet is well written, it will contain instructions for what happens when you see a Web page under optimal conditions (big screen, fast modem, lots of colors, etc.), then how it should look under

> suboptimal conditions (...) and also how it should look under lousy conditions (Microsoft CE, black-and-white screen, etc.). (...) there are as many [levels] as the designer can specify. »

C'est au travers de XSL-T, langage de feuilles de style, que l'on définit des *vues* sur les données (c'est-à-dire diverses façons de présenter les données, soit en partie soit en totalité).

On comprend l'intérêt de ce système pour la conservation de très grandes quantités de données : « [XML] will let us build great libraries simply by building our own sites. It will let the average person put together very sophisticated and powerful applications, simply by tagging everything properly so it fits into the larger schema of the Web » (Siegel 1997). C'est là une vision *synchronique* des choses (assurer la bonne circulation des informations sur le Web, avec un idéal de partage "démocratique), plutôt que la vision *diachronique* d'un conservateur de bibliothèque, qui soulignerait aussitôt la nécessité, pour les données à conserver telles que les données linguistiques sur langues rares, d'un archivage indépendant des sites, où elles sont hébergées pour une durée non déterminée. Mais l'idée de D. Siegel est tout de même très importante pour la conservation. Elle rejoint la conclusion de M. Jacobson, B. Michailovsky et J.B. Lowe : « In our view, standardization is to be sought in adherence to the principle of logically structured text, which makes it relatively easy to transform one markup into another, rather than in particular markup conventions » (Jacobson, Michailovsky et Lowe 2001, p. 91). En d'autres termes, des équipes relativement petites peuvent constituer leurs propres bases de données linguistiques en évitant deux dangers opposés : se conformer, par souci de compatibilité, à un modèle fixe (une DTD fixe), lit de Procuste pour les données créées par cette équipe, ou, à l'opposé, créer des documents informatiques qui correspondent exactement aux souhaits des chercheurs de l'équipe mais qui ne peuvent être consultés qu'avec les outils propres à cette équipe.

Pour une équipe de linguistes convaincus du bien-fondé de ces techniques, la question qui se pose alors est la suivante :

### Comment créer un document XML ?

Pour produire le document XML, il existe quatre possibilités (voir Van Herwijnen 1995):

1) taper les balises à la main (sous NotePad, par exemple). XML a vocation à pouvoir être lu directement par des êtres humains ; on peut s’en servir comme cela. Il est utile de savoir le faire (voir les exercices proposés par Michel Jacobson sur son site personnel), pour comprendre certains des principes de XML, pour pouvoir le cas échéant faire des retouches en « mode texte ». Mais saisir toutes les balises à la main est particulièrement fastidieux.

2) si on a des documents sur papier, utiliser un système intelligent de balisage automatique (par exemple FastTAG d’Avalanche), scanner le résultat, et appliquer un programme de reconnaissance de caractères (par exemple Textbridge de Xerox).

Mais il faut que les documents aient une structure visuelle claire ; même dans ce cas, le taux de succès est situé entre 40% et 60%, ce qui signifie de nombreuses heures de travail pour corriger les documents. C'est plus rapide que de ressaisir tous les documents. Mais si c'est un linguiste qui fait lui-même la saisie de ses données, il est préférable pour lui de tout ressaisir avec un éditeur XML, occasion de « repasser » tout son corpus, plutôt que de corriger les erreurs aléatoires des logiciels de balisage et de reconnaissance, tâche qui n'est pas naturelle comme la saisie, où l'on est aux prises avec des erreurs techniques de divers ordres, qui ne correspondent pas à une logique humaine : erreurs qui sont des défis au bon sens. Par exemple, la proximité graphique qui induit en erreur un logiciel de reconnaissance de caractères ne gênerait pas un être humain, qui se servirait de l'environnement du mot pour aller à la solution juste. Ce n'est pas du tout le même processus que la fréquentation des données qui se fait à l'occasion d'une saisie systématique.

Même si le processus est plus long, il paraît donc moins pénible de tout reprendre.

3) Saisir les données sous un autre système et le convertir en écrivant un filtre de conversion. Un exemple de bon outil pour écrire ces filtres est OmniMark de Exoterica. En théorie, la substitution de codes peut se faire avec une fiabilité de 100%, n'étant pas soumises aux mêmes contingences que la reconnaissance de structures visuelles par le balisage intelligent et les outils de reconnaissance de caractères.

Mais l'environnement structuré dans le système de départ doit être très rigoureux. Plus le format de départ est contraignant, plus la transformation a de chances de se passer efficacement. Le document de départ doit se conformer à la structure de la DTD. Même dans ce cas, le taux de conversion réussie n'atteint pas les 100% (Van Herwijnen estime ce taux à 95%). En d'autres termes, les processus de conversion ne peuvent pas être automatisés de manière sûre. Si le document de départ pour la conversion a été composé en traitement de texte, avec une structure implicite du type « un espace signale une frontière de mot, un tiret signale une frontière de morphème », il est à peu près sûr qu'il y aura des erreurs, sauf si l'auteur a prêté une attention méticuleuse à la structure du document, caractère après caractère, ce que l'on peut difficilement demander à quelqu'un qui travaille sur des documents à l'annotation complexe : l'utilisateur aura oublié de mettre le bon caractère pour signaler l'absence d'une annotation ou au contraire une annotation double, et la conversion sera bloquée. Il faut alors un regard averti pour déceler d'où vient la panne.

A partir de Lexware ou Shoebox, la réussite beaucoup plus probable. Ce sont des traitements de texte avec règles explicites. Il existe déjà de petits programmes (créés par Michel Jacobson en langage Perl) pour opérer des conversions de ce type.

Il ne faut pas exclure cette possibilité, qui permet à chaque chercheur de travailler avec les outils dont il est familier. Mais les problèmes qui se poseront lors de la conversion ne

pourront pas être résolus par les linguistes « informaticiens amateurs » à qui s'adresse ce guide ; cette solution n'est donc adéquate que si le centre de recherche compte un informaticien.

4) La meilleure approche consiste à créer les documents directement dans un éditeur XML.

C'est ainsi que le corpus naxi réalisé en préparation d'une étude de terrain (juillet à octobre 2002) a été saisi sous XML Spy (voir partie 3, « Le corpus naxi »). Cet éditeur permet une vision en grille particulièrement commode, et permet également d'afficher les balises à l'écran.

Pour bien utiliser l'outil, l'utilisateur doit connaître certains de ses principes de fonctionnement. Il est conseillé de connaître un peu d'XML ; le fait d'effectuer la saisie directement par un éditeur XML est un moyen de pratiquer XML, et d'en comprendre la logique.

Pour préciser les notions concernant XML qui viennent d'être exposées, tout en voyant comment ce langage est employé pour un programme d'archivage linguistique en fonctionnement, voici des extraits de documents du programme Archivage du LACITO, avec des explications. La compréhension sera facilitée par la connaissance préalable que le lecteur a des données (linguistiques) qu'il s'agit de mettre en forme.

La DTD LACITO :

```
<!ELEMENT TEXT      (HEADER, BODY)      >
<!ELEMENT BODY      (S+)                >
<!ELEMENT S         (AUDIO?, TRANSCR, TRADUC+>
<!ELEMENT TRADUC    (#PCDATA)           >
<!ELEMENT TRANSCR   (W|PONCT)*          >
<!ELEMENT W         (AUDIO?, FORM, GLS?)>
<!ELEMENT FORM      (#PCDATA)           >
<!ELEMENT GLS       (#PCDATA)           >

<!ATTLIST TRADUC    lang CDATA     "Francais"     >
<!ATTLIST S         id   ID        #REQUIRED
                    who  CDATA     #IMPLIED>
```

Examinons la première ligne :

```
<!ELEMENT TEXT      (HEADER, BODY)      >
```

| Elément | Description |
|---|---|
| <! | délimiteur d'ouverture de déclaration de balise |
| ELEMENT | mot-clef indiquant le type de déclaration |
| TEXT (HEADER, BODY) | contenu de la déclaration |
| > | délimiteur de fermeture de déclaration de balise |

| Signe | Signification |
|---|---|
| \| | OR |
| & | AND |
| , | SEQUENCE |
| + | occurrence obligatoire et répétable : « required and repeatable » |
| * | occurrence optionnelle et répétable : « optional and repeatable » |
| ? | occurrence optionnelle |

Dans

```
<!ELEMENT S         (AUDIO?, TRANSCR, TRADUC+>
```

on constate que l'élément nommé S (« Sentence ») doit comporter au moins une traduction (c'est ce que signale le signe +). Le point d'interrogation après AUDIO signale que a composante audio est optionnelle. Dans la ligne qui suit,

```
<!ELEMENT TRANSCR   (W|PONCT)*          >
```

les éléments W et PONCT sont optionnels ; il peut y avoir l'un ou l'autre ; ces éléments peuvent être répétés.

Dans la ligne

```
<!ELEMENT TRANSCR   (W|PONCT)*          >
```

l'expression (W|PONCT) renvoie à la définition d'un groupe. Il existe trois types de groupes que l'on peut utiliser dans les déclarations de balisage :

- le groupe modèle (« model group ») qui est une liste d'éléments

- le groupe d'unités lexicales nominales (« name token group ») qui est une liste d'unités lexicales nominales

- le groupe de noms (« name group ») qui est une liste de noms.

Le contenu d'un groupe est constitué d'unités lexicales (« tokens »). Les objets inclus à l'intérieur des parenthèses du groupe sont reliés par les connecteurs (virgules, signes pourcentages % et barres verticales |).

Soient les deux lignes suivantes :

```
<!ELEMENT FORM      (#PCDATA)           >
<!ELEMENT GLS       (#PCDATA)           >
```

"...the special symbol #PCDATA is reserved to indicate charater data. The moniker PCDATA stands for 'parseable character data'." (Walsh 1997 p. 100)

Ne pas confondre avec CDATA. "CDATA attributes are strings; any text is allowed." (ibid.)

Suite de la DTD LACITO :

```
<!ATTLIST TRADUC    lang CDATA     "Francais"     >
<!ATTLIST S         id   ID        #REQUIRED
                    who  CDATA     #IMPLIED>
```

Les éléments qui commencent par `<!ATTLIST` et se terminent par un `>` sont des déclarations d'attributs. "Attribute declarations identify which elements may have attributes, what attributes they may have, what values the attributes may hold, and what default value each attribute has." (ibid.) L'élément TRADUC, déclaré plus haut comme composé de caractères (ligne <!ELEMENT TRADUC (#PCDATA)>), possède un attribut lang, qui est une chaîne, et dont la valeur par défaut est `Francais`.

```
<!ATTLIST S         id   ID        #REQUIRED
                    who  CDATA     #IMPLIED>
```

Cette séquence signifie que S (la phrase) possède pour attributs id et who. id est du type ID (identification), who est du type CDATA (chaîne). #REQUIRED indique que l'attribut doit avoir une valeur, spécifiée explicitement pour chaque occurrence de l'élément dans le document. #IMPLIED indique qu'il n'est pas indispensable que l'attribut ait une valeur, et qu'il n'y a pas de valeur par défaut.

Balise de début et balise de fin :

```
<GLS>dog</GLS>
```

Certains éléments sont à la fois début et fin : ils sont du type <NOMDEBALISE/>. *"While most elements in a document are wrappers around some content, empty elements are simply markers where something occurs. The trailing slash, <name/>, indicates to a program processing the XML document that the element is empty and no matching end-tag should be sought." (Norman Walsh 1997 p. 97)*

Attribut :

Il faut préciser le nom de l'attribut ainsi que la valeur assignée. La grammaire est : nomd'attribut = valeurattribut

ex.:

```
<GLS lang=English>dog</GLS>
```

Exemple donné dans Jacobson, Michailovsky et Lowe 2001 :

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE TEXT SYSTEM "Archives.dtd">
<TEXT id="hayu1" lang="hayu">
<HEAD>
     <TITLE>Two sisters</TITLE>
     <SOUNDFILE src="SOEURS_SOUND"/>
</HEAD>
<BODY>
...
<S id="hayu1s1">
     <TRANSCR>
          <W><FORM>nakpu</FORM><GLS>two</GLS></W>
          <W><FORM>nonotso</FORM><GLS>sisters</GLS></W>
          <W><FORM>si&#331;</FORM><GLS>wood</GLS></W>
          <W><FORM>pa</FORM><GLS>make</GLS></W>
          <W><FORM>la&#660;natshe-m</FORM>
               <GLS>go:REFL:3du-ASS</GLS></W>
          <W><FORM>are</FORM><GLS>REP</GLS></W>
          <PONCT>.</PONCT>
```

```
    </TRANSCR>
    <TRADUC lang="Francais">On raconte que deux soeurs
allèrent chercher du bois.</TRADUC>
    <TRADUC lang="Anglais">They say that two sisters went to
    fetch wood.</TRADUC>
</S>
...
</BODY>
</TEXT>
```

Observez la ligne

```
<SOUNDFILE src="SOEURS_SOUND"/>
```

Ce n'est pas une balise de fin, à laquelle correspondrait, plus haut dans le document, une balise de début

```
<SOUNDFILE src="SOEURS_SOUND">
```

Si c'était une balise de fin, la barre / se trouverait aussitôt après le délimiteur d'ouverture de balise <.

Il s'agit d'une balise unique, qui contient ouverture et fermeture. Une balise de ce type est ouverte par < et fermée par />.

Pour insérer des commentaires : le délimiteur d'ouverture de commentaire et le délimiteur de fermeture de commentaire sont formés de la façon suivante :

```
<!-- Ceci est un commentaire -->
```

Un élément qui rend complexe la création d'un document XML avec un traitement de texte est qu'il existe des glyphes ayant une signification en XML, et qui donc ne peuvent pas figurer telles quelles dans le contenu. Chaque fois que l'on souhaite utiliser ces caractères dans le texte en tant que donnée, il faut utiliser les références à entités. Par exemple, si on a besoin d'utiliser le signe inférieur à ("<") et être sûr que son rôle délimiteur ne sera pas appliqué, il faut utiliser la référence à entité suivante :

```
&lt;
```

| Glyphe | Rôle délimiteur | ISO 646:1983 | Nom abstrait |
|---|---|---|---|
| & | connecteur et | 38 | AND |
| &# | ouverture de référence à caractère | 38 35 | CRO |
| [ | ouverture de sous-ensemble de déclaration | 91 | DSO |
| ] | fermeture de sous-ensemble de déclaration | 93 | DSC |

| | | | |
|---|---|---|---|
| [ | ouverture de groupe de balises textuelles | 91 | DTGO |
| ] | fermeture de groupe de balises textuelles | 93 | DTGC |
| & | ouverture d'appel d'entité | 38 | ERO |
| </ | ouverture de balise de fin | 60 47 | ETAGO |
| ( | ouverture de groupe | 40 | GRPO |
| ) | fermeture de groupe | 41 | GRPC |
| " | délimiteur de début ou fin de contenu littéral | 34 | LIT |
| ' | délimiteur de début ou fin de contenu littéral (alternatif) | 39 | LITA |
| > | fermeture de déclaration de balisage | 62 | MDC |
| <! | ouverture de déclaration de balisage | 60 33 | MDO |
| - | signe moins ; exclusion | 45 | MINUS |
| ]] | fermeture de section marquée | 93 93 | MSC |
| / | balise de fin nulle | 47 | NET |
| ? | indicateur d'occurrence optionnelle | 63 | OPT |
| \| | connecteur ou | 124 | OR |
| % | ouverture d'appel d'entité paramètre | 37 | PERO |
| > | fermeture d'instruction de traitement | 62 | PIC |
| <? | ouverture d'instruction de traitement | 60 63 | PIO |
| + | obligatoire et répétable ; inclusion | 43 | PLUS |
| ; | fermeture d'appel | 59 | REFC |
| * | optionnel et répétable | 42 | REP |
| # | indicateur de nom réservé | 35 | RNI |
| , | connecteur de séquence | 44 | SEQ |
| < | ouverture de balise de début | 60 | STAGO |
| > | fermeture de balise | 62 | TAGC |
| = | indicateur de valeur | 61 | VI |

Mais un éditeur tel que XML Spy permet de saisir directement le texte : il sépare de lui-même les balises et leur contenu.

### Ce qu'il faut connaître du langage XSL

XSL est un langage de feuilles de style. Pour illustrer ses possibilités, prenons l'exemple d'une phase de la mise sur Internet du document tamang « moineau » du professeur Martine Mazaudon. Plusieurs feuilles de style sont utilisées (elles sont appliquées par SAXON.EXE, processeur de style écrit en langage C ; c'est un outil public).

Partant d'un document « texte » ne comportant aucune balise, et seulement une structure implicite, la feuille de style MORPH.XSL ajoute des balises <M></M> (morphèmes) : pour chaque mot, les chaînes séparées par un tiret sont placées à l'intérieur de balises <M> </M> successives.

La feuille de style TOUTF8.XSL change la norme de codage de caractères, la faisant passer à UTF-8.

TOHTML.XSL transforme le document en document HTML.
La feuille de style CONC.XSL est un outil de concordance qui permet de lister toutes les suites, de les trier par ordre alphabétique, d'observer leur contexte d'apparition. Tel contexte apparaît toujours à la même position, etc. Il permet donc de tester la cohérence interne du document. Il sert également pour la recherche.

CORE.XSL contient des fonctions appelées par d'autres feuilles de style. La conception de l'ensemble est modulaire. (CORE sert principalement pour l'APPLET en JAVA.)

Le script PERL 2XML.PL permet de passer en XML en partant d'un fichier en code ASCII, qui s'obtient, par exemple, en choisissant le format "Texte seul" dans le menu "Enregistrer sous" de Microsoft Word. M. Michailovsky avait créé un programme équivalent, dont les connaisseurs de Perl apprécieront la portée (ceux qui ne connaissent pas Perl pourront du moins admirer la **concision** de ce programme, qui tient en moins d'une page) :

```
# markup.pl -- Perl script to generate LACITO XML markup
$i=0;
$morphseps="\-\+\=";
while ($line = <STDIN>) {
     chomp $line;
     print "<S id=\"s" . ++$i . "\">\n";
     while ($line =~ s/^([^ \t]+)([ \t]+|$)//) {
          $word = $1;
          print "\t<W>\n";
          while ($word=~ s/^([^$morphseps]+)([$morphseps]|$)//)
{
                    print "\t\t<M><FORM>$1</FORM></M>\n";
          }
          print "\t</W>\n";
     }
     print "</S>\n";
}
```

En réalité, il faut que ces quelques lignes soient précédées d'instructions qui créent l'en-tête du document XML, indiquent la version de XML qui est utilisée, etc.:

```
print "<?xml version=\"1.0\" encoding=\"iso-8859-1\" ?>\n\n";
print "<TEXT>\n";
```

```
print "<HEADER>\n";
print "\t<SOUNDFILE href=\"KCL1.wav\"/>\n";
print "\t<TITLE>KCL1</TITLE>\n";
print "</HEADER>\n";
```

Le fichier résultant est un fichier XML découpé en phrases. Il ne se conforme pas à telle ou telle DTD.

On applique alors des feuilles de style pour le découpage en morphèmes. Même procédure : on passe par une ligne de commandes, mais cette fois on utilise un processeur XSL, qui applique une feuille de style XSL. En l'occurrence, la feuille de style s'appelle morph.xsl, le processeur de style utilisé est saxon.

```
saxon fichier.xml morph.xsl > nouveau.xml
```

*Bilan*

Tout format de codage a nécessairement ses limites, qui tiennent à sa finalité. Il n'existe pas de format universel adapté à toutes les utilisations, comme en témoigne l'évolution de l'énorme entreprise qu'est la TEI (Text Encoding Initiative), dont le format d'architecture des données, d'abord conçu comme unitaire, a ensuite dû évoluer vers une TEI « allégée » et surtout « modulaire » : voir *Guidelines for Electronic Text Encoding and Interchange* (TEI P3) (Sperberg-McQueen et Burnard 1994). L'orientation du format LACITO est l'archivage des textes longs. Il ne prévoit pas de moteur de recherche pour phonèmes isolés, ni une annotation par phonème. L'ajout de commentaires (« Cette séquence n'est pas très breathy, c'est peut-être à cause du contexte, et cela tient aussi au locuteur, qui a telle ou telle habitude »), la transcription prosodique avec visualisation de la ligne mélodique demanderaient une refonte du système. Le caractère obligatoire de la DTD une fois qu'elle est adoptée par un programme d'archivage peut donc limiter la liberté du linguiste dans l'annotation du document.

En outre, certaines contraintes sont liées au langage XML lui-même. XML décrit des structures nécessairement arborescentes. Cela pose problème pour un codage multi-lignes non arborescent : si l'on veut noter à la fois la syntaxe, la prosodie, le niveau fonctionnel, le niveau perceptif et le niveau pragmatique, les unités vont se chevaucher, ce qu'XML ne permet pas. Ce type de données remet en cause le principe d'XML. C'est d'ailleurs une critique qui a été formulée contre XML : voir l'article de Theodor Holm Nelson intitulé "Embedded Markup Considered Harmful", in Dan Connolly (1997). Des solutions sont imaginables pour contourner le problème, et utiliser XML pour l'encodage de données non arborescentes : il fait correspondre un triplet de valeurs à chaque unité. Mais, en principe, pour des données non arborescentes, mieux vaut utiliser une base de données.

### *4. Une base de données pour phonéticiens : la base de données phonétique de Kiel*

Au cours de la partie I, les limites des logiciels d'analyse du signal SNOORI et PRAAT ont été brièvement mentionnées : ces logiciels proposent des outils d'annotation, qui permettent d'établir divers types de transcriptions ; PRAAT permet de placer divers types de transcriptions sur des lignes séparées. Des moteurs de recherche peuvent ensuite opérer sur ces transcriptions. Il apparaît en fait très clairement que **ces logiciels ne sont pas conçus pour la création de corpus** : l'annotation créée par SNOORI est enregistrée directement dans le document sonore, ce qui peut l'endommager ; l'annotation par PRAAT est plus complète, mais est loin de régler tous les problèmes de codage de caractères.

Il existe un modèle de bases de données spécialement conçu pour un travail de recherche en phonétique (annotation segmentale pour l'étude de la réalisation des segments en contexte, et annotation prosodique), celui de l'Université de Kiel : **xassp**, « a platform for segmental and prosodic labelling and speech analysis, xassp : Advanced Speech Signal Processor under the X Window System » (Simpson, Kohler et Rettstadt (eds.) 1997, p. vii). La description qui suit se fonde principalement sur les démonstrations réalisées par le professeur Klaus Kohler à l'ILPGA, ainsi que sur les publications de l'Université de Kiel, notamment Kohler (1992). Des différences importantes séparent xassp des outils utilisés par le programme Archivage. Sa compatibilité est moindre. Il fonctionne sous LINUX, pas sous Windows ni Macintosh. Ses utilisateurs doivent donc connaître les rudiments de LINUX. Il n'est certes pas impossible d'étendre la compatibilité de xassp ; des efforts sont actuellement en cours à Kiel pour associer xassp, outil d'annotation, et EMU, outil créé à l'université de Sydney (Australie), et plus précisément destiné à la recherche à l'intérieur de corpus, et à l'analyse. Mais, comparé au format LACITO, xassp présente un certain nombre d'avantages pour la transcription des segments et de la prosodie, qui sont ses finalités premières.

#### a. Notation des segments

Le codage des caractères se fait selon la norme SAMPA (Wells, Barry et Fourcin 1989). Cette norme est assez largement utilisée pour les bases de données de langues européennes, et donc compréhensible par certaines personnes extérieures au projet de Kiel ; mais elle n'est pas transparente comme l'API utilisée par le programme Archivage. L'emploi de SAMPA demande un apprentissage signe par signe des conventions utilisées : @ pour ə, par exemple. En outre, tous les caractères de l'API ne sont pas dans SAMPA. Pour l'allemand, l'inventaire est globalement suffisant ; mais, dans le projet de Kiel, c'est une version « modifiée et augmentée » de SAMPA qui est utilisée (Simpson, Kohler et Rettstadt (eds.) 1997, p. 9), forme de « dialectalisation » qui complique les échanges entre formats. Par

exemple, parmi les quelques conventions ajoutées se trouve Q pour le coup de glotte [ʔ] et q pour la glottalisation [ ˞]. Si l'on souhaite étendre l'utilisation des outils créés à Kiel à d'autres langues, en particulier à des langues non européennes, la question se pose aussitôt des symboles qui ne sont pas présents dans le SAMPA « standard » : il ne paraît pas possible d'éviter une dialectalisation croissante.

Au plan des outils proposés, les avantages de xassp pour les phonéticiens sont évidents : l'annotation se fait au vu du signal et du spectrogramme, avec un grand confort d'écoute du fait que la fenêtre d'écoute peut être librement choisie. Partant d'une transcription canonique (constituée de la forme phonématique des mots isolés), le transcripteur indique pour chacun des phonèmes s'il est absent (notation : S-) ou non (notation : S), s'il a été remplacé par un autre (notation : S-S'), ou s'il y a eu insertion d'un segment (-S'). et quelles sont ses lieux de début et de fin ; un signe diacritique (%) signale les cas où il n'y a pas de critère convaincant pour fixer le point de début et de fin. Il peut être utile de donner une citation relativement longue à ce sujet :

> « ... the orthographic text files are automatically converted into segmental phonemic transcription using the grapheme-to-phoneme module and the pronunciation exceptions lexicon of the RULSYS/INFOVOX German TTS system (Klaus J. Kohler 1992). To cope with the expanded symbolic repertoire of spontaneous speech transliterations the transformation rules, originally devised for standard orthographic text, have had to be supplemented. The alphabet used is modified and augmented SAMPA (Wells, Barry et Fourcin 1989). For each orthographic text file input the module automatically generates a transcription file output. It is manually corrected and represents a lexical citation form pronunciation : canonical transcription.
>
> The canonical word pronunciations are nevertheless generated from continous text, not from derived word lists. (...) the Kiel processing system allows the automatic rule-governed derivation of canonical transcriptions from running orthographic text. » (Simpson, Kohler et Rettstadt (eds.) 1997, p. 9-10)
>
> « Phonetic labelling proceeds from a prototype label file and attributes its canonical symbols to the time scale of the related speech file. Thus a prototype label file is converted into a label file with real time values.

- One phonemic symbol after another of the prototype label file is offered by the program for manual positioning in the speech signal.
- Through visual inspection of speech wave and spectrogram displays (and other possible signal representations, e.g. Fo curves), supported by auditory control, the phonemic labels are manually aligned with the beginnings of the corresponding speech signal segments and are given their time (sample) values.

- The segmentation is thus strictly linear without overlap. » (*Ibid.*, p. 10)

Parmi les avantages de ce système, retenons l'existence de plusieurs niveaux de transcription, qui permettent de rechercher tel ou tel contexte *phonématique* pour étudier sa réalisation *phonétique*, ce qui ne serait pas possible avec une transcription « phonétique » sur la seule foi des impressions retirées de l'audition et de l'étude du spectrogramme. Si une séquence était purement et simplement étiquetée [ʃtlədi], dans un modèle mono-linéaire, il ne serait plus possible de se fonder sur un tel corpus pour étudier la réalisation des phonèmes en contexte, puisqu'on ne saurait plus quels sont les phonèmes en question (en l'occurrence : /ʒətələdi/)[10]. *Les choix faits par les transcripteurs comportent nécessairement une part d'arbitraire, mais cette petite marge d'incertitude n'est plus gênante dès lors que l'utilisateur a accès aux différents niveaux de la transcription*, et au signal lui-même.

Les recherches auxquelles est destiné le corpus réalisé à Kiel à l'aide de xassp concernent en premier lieu la réalisation des phonèmes en contexte. Par exemple : « vowel deletion and vowel devoicing in German and English spontaneous and read data from monolingual, multilingual and pluristylistic perspectives ; (...) the acoustic manifestation of the German vowel system in read speech and (...) the problem of speaker normalization » (*Ibid.*, p. viii). Mais ces recherches peuvent également concerner la prosodie.

### b. L'annotation prosodique.

Il n'existe pas actuellement de notation satisfaisante de l'intonation qui puisse être appliquée à diverses langues (Vaissière 2002). Le choix d'une annotation avec xassp paraît adaptée, dans ce contexte, au sens où l'annotation prosodique ne se limite pas à un inventaire fermé de symboles liés à tel ou tel cadre théorique, ce qui en fait une annotation adaptée à la transcription d'observations précises et fines. Jusqu'à présent, xassp a été utilisé principalement pour des corpus de langue allemande (et française), et la transcription prosodique adoptée est celle élaborée par Klaus Kohler (Klaus Kohler 1996; Simpson, Kohler et Rettstadt (eds.) 1997). Il serait imprudent de chercher à apprécier les possibilités d'extension de xassp à d'autres langues, et à d'autres modèles prosodiques que celui envisagé par ses concepteurs, avant d'avoir une expérience approfondie de l'étiquetage prosodique de bases de données avec xassp. Au plan technique, la comparaison avec le programme du LACITO fait ressortir un problème de compatibilité, dont il faut espérer que l'équipe de Kiel cherchera à le résoudre à l'avenir. Dans cette perspective, il faudrait créer

[10] Tel était le cas dans certaines bases de données sur l'anglais : le mot *international* était transcrit selon l'impression auditive, qui donnait [innaʃn̩] ; il n'y avait alors plus moyen de rechercher automatiquement dans le corpus toutes les réalisations de /t/, puisque la transcription ne portait plus trace des formes « sous-jacentes » (J. Vaissière, séminaire à l'ILPGA, 2001).

des possibilités d'exportation de l'annotation en format XML : cela n'est nullement exclu, car l'annotation sous xassp est strictement arborescente, « indexing [non-linear] phonetic parameters in the segmentally labelled environment » (*Ibid.*, p. 13).

Après la présentation de ces modèles de corpus et de leurs limites, la troisième partie sera consacrée à l'examen d'exemples.

# *Troisième partie :*

*Quelles données pour quelles recherches ?*

Ce dernier axe du travail illustre par des cas précis les questions soulevées dans les deux premiers axes. Il est très loin de représenter une typologie complète, son fil directeur étant une forme de « rapport d'activité », présentant certains travaux d'archivage réalisés au cours de l'année de rédaction du présent mémoire.

## *1. Exemple de données de chercheur : la donation René Gsell.*

En premier lieu, nous souhaiterions exposer un exemple qui montre la fragilité des « documents de chercheur ». Cet exemple concerne l a 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 et Paris 1974). La langue sur laquelle ils nous renseignent, et qui est aujourd'hui éteinte, était (de l'avis du professeur R. Gsell et de ses collègues) 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.

Voici un rapide inventaire du fonds audio (les informations ci-dessous sont regroupées dans un document gravé sur chacun des CD-R de l'archive dont la construction a commencé cette année au LACITO).

La voix est celle de Tevfik Esenç, enregistré par Georges Dumézil, avec l'assistance de Christine Leroy lors de certaines séances. Georges Dumézil disait le mot en turc et Tevfik Esenç en donnait la traduction oubykh.

Les fichiers sonores sont une version numérisée au LACITO des bobines magnétiques originales (type UHER, vitesse 7 1/2, enregistrement MONO). Ces bobines ont été regroupées au LACITO, où Mme **Dina Dabjen-Bailly** travaille à la poursuite des entreprises éditoriales du professeur **Catherine Paris** et à la conservation de ses collections.

Une partie des enregistrements se trouvait dans les collections du professeur Catherine Paris. La bande magnétique 3 se trouvait à la bibliothèque de l'Institut de Phonétique, parmi les enregistrements divers que Mme la bibliothécaire a confiés en 2001 à **Alexis Michaud** pour les écouter et les référencer.

Les transcriptions, ainsi que les originaux des bandes magnétiques 1 et 2, se trouvaient dans les collections particulières du professeur **René Gsell**, et ont été généreusement confiées aux soins de l'ILPGA par Mme **Agnès Gsell-Noy**.

Tous les documents ont été numérisés (par Alexis Michaud) à partir des originaux. Ce sont ces documents numériques qui sont désormais le point de référence pour la recherche.

Leur qualité est globalement bonne. L'effet d'écho perceptible sur certaines bandes (bande 3

en particulier) serait la conséquence d'un vieillissement des bandes (qui ont été numérisées en 2001-2002, une trentaine d'années après leur enregistrement), et non d'un enregistrement dans une pièce à l'acoustique mauvaise. (Communication personnelle de **Bernard Gautheron**, ingénieur responsable du Laboratoire de l'ILPGA, 19 rue des Bernardins, 75005 Paris, qui a connaissance des conditions dans lesquelles l'enregistrement a eu lieu.) Des pertes dans les fréquences aiguës sont également très probables : ce sont les fréquences aiguës qui sont les premières affectées par le vieillissement du support.

Après numérisation, le son n'a subi aucun traitement. Les fichiers ont été vérifiés (pour éviter toute fausse manipulation lors du transfert) ; la correspondance entre les originaux et les fichiers numérisés a été vérifiée. Lorsqu'un redécoupage a eu lieu, l'ordre des documents sur l'original est respecté. Exemple : les histoires 2, 3, 4, 5, 6 et 7 de la bande 3. Les "parasites" au début et à la fin des documents n'ont pas été supprimés non plus : "oui", "ça va" en français, soupir de l'informateur lorsque la séance s'achève, et cela jusqu'au déclic de l'enregistreur qu'on interrompt.

Mais ces documents, qui représentent toute une bande magnétique numérisée, sont très peu maniables. Pour permettre un accès simple et précis aux données, et notamment en vue de la création de spectrogrammes portant sur tel ou tel phénomène (réalisation d'un groupe de consonnes, comparaison des réalisations d'un même phonème...), il a paru utile de proposer, aux côtés du document original, une version retravaillée. L'objectif était de construire une base de données de haute qualité, qui dispense du long travail de dépouillement. Les documents retravaillés sont d'un accès beaucoup plus rapide, du fait de leur petite taille. Au bilan, l'ensemble des fichiers retravaillés occupe 47 Mo, contre 81,1 Mo pour le document numérisé original « oubykh1.wav ».

Les opérations pratiquées sur le signal d'origine ont été intégralement réalisées avec le logiciel SoundForge (version : SoundForge 5.0 XP). Les opérations ont été les suivantes :

1. pour le fichier OUBYKH1.wav (listes de mots) :

a) découpage par séries de trois mots, ou par paires de mot, en respectant le découpage du document écrit. Sur l'enregistrement, cet ordre est suivi de façon rigoureuse. Les mots se succèdent par séries de trois, puis par séries de deux. Les chiffres correspondants sont dits en français (« un, deux, trois », « trois, quatre, cinq », « six, sept, huit », etc.).

Je n'ai pas corrigé la petite erreur des enquêteurs qui ont redoublé le n°3, faisant succéder une série « trois, quatre, cinq » à la série « un, deux, trois ». **Le 3 qui clôt la première série et le 3 qui ouvre la deuxième série sont des mots différents.** Mais le fait de redresser toute la numérotation aurait entraîné un risque d'erreur, et de confusion lors des références à tel ou tel mot. Les numéros employés ici sont donc point pour point ceux qui figurent dans les transcriptions originales.

b) amplification du signal de chacune des phrases, par la commande « Normalize », dans le menu « Process ». Amplification jusqu'à 0 dB. Cette opération augmente le volume du son à la sortie de l'ordinateur ; en format 16 bit, **cette manipulation n'apporte pas de modification au son, mais fausse les comparaisons de mesure d'intensité qui pourraient être faites entre fichiers**, puisque le degré d'amplification varie d'un fichier à l'autre (il dépend de la distance entre le 0 et le pic d'intensité le plus élevé présent dans le fichier). Pour des mesures d'intensité, paramètre important mais délicat à mesurer pour de nombreuses raisons, il faut repartir du document original « Oubykh1.wav ».

c) suppression des nombres français correspondant aux mots, prononcés par l'enquêteur.

d) suppression des interventions de l'enquêteur. Celui-ci joue un rôle de souffleur : il indique discrètement à l'informateur l'initiale du mot qui va suivre, ʃ ou ʒ ou dʒ ou χ, etc. Il glisse parfois quelques mots en turc ou en français, du type « Encore » ou « Bon, là ça va ». (Pour entendre clairement et distinctement ces commentaires chuchotés, ouvrir le document « Oubykh1.wav » dans SoundForge, sélectionner une portion du signal apparemment vide entre deux mots, et choisir « Normalize » dans le menu « Process ». Le programme amplifie jusqu'à 0 dB cette portion du signal, sinon inaudible, et on entend la voix de l'enquêteur (souffleur attentionné) qui guide son informateur. Ces portions du signal ont été supprimées.

e) suppression des mots présentant un problème de **superposition des commentaires de l'enquêteur avec la parole de l'informateur**, qui dénature le son. Ces opérations ont été effectuées avec la plus grande minutie ! Dans aucun cas il n'a été nécessaire d'effacer un mot qui était en exemplaire unique dans l'enregistrement. Dans la plupart des cas, l'enquêteur, conscient du problème d'enregistrement, avait demandé une répétition supplémentaire, et la transcription originale porte la mention « Ne garder que la 2$^{e}$ fois ».

f) pour segmenter le document sonore, **une annotation propre à SoundForge a été ajoutée** : un découpage en « regions list », une à gauche de chaque mot, avec le numéro du mot en question. Par exemple, le mot qui se trouve entre la « region list » 7 et la « region list » 8 est le mot numéro 7. (La position exacte des balises, à quelques centièmes de secondes près, n'est pas significative, le propos étant simplement d'encadrer le mot par des balises.) Ces indications sont particulièrement importantes pour certaines séquences de mots : la séquence est généralement répétée, et l'informateur peut se reprendre sur un mot ou un autre. Certains mots ne sont pas utilisables : l'enquêteur souffle régulièrement l'initiale du mot à l'informateur, ou lui donne une consigne, messages qui se superposent avec la parole de l'informateur. Ainsi, le fichier 6-8.wav comporte les mots : 6, 7, 6, 8, 6, 8. Il est essentiel d'avoir cette information. Elle est donnée avec le texte de la transcription, accompagnée de l'indication de sa position temporelle dans le fichier ;

mais il est commode de pouvoir la voir à l'écran. Important : **un fichier .wav annoté dans SoundForge perd son annotation lorsqu'il est lu par un autre logiciel (WinSno par exemple)** : non seulement le logiciel d'analyse n'affiche pas l'annotation, mais il l'efface du fichier. Lorsqu'on ré-ouvre le fichier avec SoundForge, l'annotation a disparu. Il est donc nécessaire d'**établir une copie des fichiers pour les ouvrir avec un logiciel autre que SoundForge**. Utiliser alors les indications temporelles données avec le texte de la transcription.

La mise en forme du corpus oubykh a également été l'occasion de redécouvrir une communication du professeur René Gsell au sujet des données cinéradiographiques.

Lors du IVe colloque de caucasologie, les 27, 28 et 29 juin 1988, le professeur René Gsell a prononcé une communication intitulée « Points controversés de la phonétique de l'oubykh ». Le sujet est esquissé dans le volume résumés de ce Colloque international du CNRS de caucasologie et de mythologie comparée, à la mémoire de Georges Dumézil pour le 90e anniversaire de sa naissance, et IV colloque de caucasologie (brochure diffusée par l'UA 390 du CNRS « Caucase et monde indo-européen » et le LP3-121 du CNRS « Section Euscaro-caucasique et langues paléosibériennes, LACITO »). Page 18 : « GSELL, R. Points controversés de la phonétique de l'oubykh. Il s'agit principalement de l'étude des différentes sifflantes, labialisées et non-labialisées, qui ont donné lieu à diverses interprétations. L'examen est fait d'après le seul film radiographique existant actuellement. / Controversial Issues in the Phonetics of Ubykh. The various sibilants of Ubykh—both rounded and unrounded—have as yet been granted diverging interpretations. This paper introduces the results of a further examination, performed using the sole radiographic film known to exist. »

La même brochure annonce une communication de G. Charachidze : « Le dictionnaire oubykh de G. Dumézil et G. Charachidzé » (pas de résumé), et un bref aperçu du « Dictionnaire abzakh de C. Paris et N. Batouka, en tant que 'prototype' d'un dictionnaire dialectologique du tcherkesse ». Au cours du même colloque, A.S. Özsoy présentait « Quelques contraintes syntaxiques de relativisation en oubykh ».

Voici le texte de cette communication (qui figurait dans la donation René Gsell), qui n'a à notre connaissance fait l'objet d'aucune publication.

> Comme on le sait, la phonétique de l'oubykh est particulièrement compliquée et son système consonantique extrêmement riche. En 1959 Dumézil (Etudes Oubykhs) estimait que la distinction des 78 espèces consonantiques est maintenant certaine, mais que la description et l'interprétation de plusieurs présentent encore des difficultés. En 1963, dans son dictionnaire de la langue oubykh, H. Vogt fournit un tableau de 80 consonnes. Ni lui, ni G. Dumézil ne donnent le statut de phonème aux unités phoniques [g], [k], [k']

(vélaires) qui élevaient le nombre total à 83. Koumakhov (1967) ne retient que 80 phonèmes consonantiques. Typologiquement, l'oubykh serait avec le margui (une langue du groupe chadique parlée au Cameroun) l'une des deux langues les plus riches en consonnes. Les questions non résolues concernant le consonantisme portent sur trois points :

1) la nature et le rôle fonctionnel des différentes « sifflantes et chuintantes », et celle de leurs partenaires labialisés.

2) le problème des vélaires : Dumézil distingue des occlusives palatales palatalisées (g', k', k'$^{?}$) ou labialisées (g$^{\circ}$, k$^{\circ}$, k$^{?\circ}$) avec des fricatives palatales simples (g NOTATION A PRECISER : AJOUTER UN TRAIT « BREF » AU-DESSUS DU G, $\chi$) et des arrières vélaires ( = uvulaires) à la fois labialisées et pharyngalisées. Koumakhov et Vogt s'entendent pour considérer g', k' et k'$^{?}$ comme des palatales palatalisées (voir exemples et références) et intègrent y comme « spirante dentale » classée avec n et p dans la série dentale, alors que Dumézil, plus intelligemment, place r, y, lambda hors système. Les vélaires simples qui apparaissent ne sont pas des phonèmes, mais des variantes tantôt de palatales palatalisées, tantôt d'uvulaires. En réalité le système n'est pas homogène et son interprétation n'est pas aisée. C. Leroy et C. Paris, « Etude articulatoire de quelques sons de l'oubykh d'après film aux rayons X », BSL LXIX, 1974, fascicule 1 distinguent dans la série occlusive des vélaires palatalisées et des vélaires labialisées, et dans celle des constrictives des vélaires simples. Peut-on équilibrer ce système ?

3) le troisième problème à résoudre c'est celui de la nature exacte du trait pharyngal : les « consonnes pharyngales »de Vogt, Koumakhov, Leroy et Paris ont-elles une existence autonome ? ou bien la pharyngalisation n'est-elle pas plutôt un trait de résonance secondaire réservé aux deux ordres extrêmes du système consonantique, l'ordre labial et l'ordre uvulaire ? Telle était en effet la présentation de Dumézil en 1959 (et c'était probablement la bonne, mais il l'a abandonnée en 1975, sans doute sous l'influence de Vogt et de Koumakhov). De toute façon à l'audition l'occlusive $\overline{q}$ (pharyngalisée) et $\overline{q}^{?}$ (pharyngalisée glottalisée) semblent voisines, sinon identiques à diverses variétés de q ɑ f des dialectes (oidis ? aides ? ardis ?) modernes

Catford d'autre part, in « Articulatory possibilities... », met en doute l'existence de véritables occlusives « pharyngales ».

Les questions 2 et 3 ne peuvent être résolues en réalité que si l'on a tout d'abord identifié avec soin les points d'articulation des régions antérieures : des alvéolaires aux palatales. C'est donc cette dernière question que nous traiterons ici, et ele est de taille : « Sifflantes et chuintantes (sibilantes) en oubykh : en tout 12 unités. » Auparavant,

rapide présentation des tableaux de Dumézil, Vogt, Koumakhov et Leroy/Paris.

Il est évident que la solution « phonologique » doit être conforme aux réalités « phonétiques » : à la fois acoustiques, articulatoires et perceptives. Certes, les traits phonologiques ne sont pas des traits phonétiques : un trait distinctif n'est rien d'autre que la marque d'une différenciation significative, c'est-à-dire le paramètre d'une fonction d'opposition. Un trait distinctif cerets a une réalité abstraite, ou si l'on préfère logique, d'élément différenciateur, d'une fonction d'opposition ; cependant il a pour être perçu comme support (ou [illisible] comme corrélat) une réalité phonétique (acoustique, articulatoire, perceptive) qui elle est « matérielle et observable », et mis à part les cas de superposition de phonèmes et de variation combinatoire d'allophones, celle-ci est quasiment toujours identique. C'est donc d'après les méthodes de la phonétique instrumentale qu'il y a lieu de définir les « lieux d'articulation » des différentes séries de « sifflantes », « chuintantes » et « semi-chuintantes » simples ou labialisées.

Point de départ :

1) les enregistrements de la prononciation de Tevfik Esenç (généralement de paires minimales) faits par G. Dumézil à l'Institut de Phonétique de Paris, généralement avec la collaboration de Christine Leroy, ou de B. Gautheron, enregistrements dont on a tiré des spectrogrammes
2) deux films radio-cinéma : le premier de 1968 (voir C. Leroy et C. Paris, art. cit.) fait par le professeur Chévigié à l'Hôpital Saint-Antoine à Paris avec l'aide de G. Dumézil et de C. Leroy : G. Dumézil disant les mots turcs que Tervfik Esenç répétait ensuite en oubykh. Le deuxième en mai 1973 fait à l'Institut de Phonétique (devenu alors Institut d'Etudes Linguistiques et Phonétiques de la Sorbonne Nouvelle) par Bernard Gautheron, ingénieur attaché à cet Institut, avec la collaboration de G. Dumézil et de Ch. Leroy. Ces deux films sont les documents de base de l'article cité de Ch. Leroy et C Paris.

Après la parution de cet article qui fit grand bruit (cf. les réactions du professeur G.B. Hewitt et de I. Lucatsen, ces deux documents uniques (illisible exceptionnels) ont eu des destinées diverses, Mme Leroy ayant en effet changé d'orientation avait quitté la phonétique expérimentale et le Caucase pour la grammaire et la linguistique française, sa spécialité actuelle. Le premier film, grâce à Mme Paris et à M. Siegfried a été il y a un mois retrouvé dans les archives du film scientifique du C.N.R.S. et monté ; le 2$^{e}$ existe actuellement en vidéo de contrôle pour magnétoscope, une version pour projection est en cours de montage. D'autre part dans 1 an environ les 2 films seront réunis et reconditionnés pour être diffusés par les soins du Film scientifique du C.N.R.S.

Dans le cours de cet exposé, on utilisera systématiquement les transcriptions de Dumézil

1959, ce sont celles qui figurent dans ses enregistrements. Dans les documents définitifs on a intérêt à utiliser la transcription de l'IPA autant que possible, quitte à y ajouter quelques modifications, ou bien si l'on veut rester fidèle à une certaine tradition linguistique donner les équivalences.

La question des lieux d'articulation est avant tout une question de définition. L'IPA : dental, alvéo-dental, alvéolaire, post-alvéolaire (généralement rétroflexe), palato-alvéolaire (prépalatal), alvéo-palatal, palatal : lieux d'articulation antérieurs.

L'articulateur inférieur (langue) permet des distinctions supplémentaires : apical (pointe), apical rétroflexe (pointe recourbée), laminal (dos de la pointe), pré-dorsal (partie antérieure du dos), dorsal (dos).

Exemples d'enregistrements (transcription Dumézil 1968) :

1) la série des sourdes non labialisées

ṡɑ~šɑ~š'ɑ

šɑ~ ṡɑ~š'ɑ

sɑ~ṡɑ~ š'ɑ

2) Sourdes labialisées :

1. s̊ɑ mer
2. ṡɑ trois
3. s̊ɑ blanc
4. s̊ɑ blanc
5. s̊ɑ mer
6. ṡɑ trois
7. s̊ɑ mer
8. s̊ɑ blanc

3) mer blanche et cent ans : XVII bis et XIII ter. 2e bande. [Correspond à oubykh2, NdEditeur.]

s̊ɑs̊ɑ mer blanche

s̊ɑs̊ɑ cent ans

Les longues discussions sur l'interprétation de ces sons, leur phonématicisation, les avatars et les hésitations dans les transcriptions de G. Dumézil, excellent connaisseur des systèmes phonétiques des langues du Caucase Nord-Ouest et qui avait l'oreille très fine

(un remarquable Ohrenphonetiker), ont été très bien résumés et discutés par Ch. Leroy et C. Paris dans l'article cité. L'examen du film de 1968 permet à ces deux auteurs de proposer des définitions plus exactes des lieux et des modes d'articulation de ces « sibilantes » et « sibilantes labialisées ».

Outre le corpus oubykh, les enregistrements de la donation René Gsell contiennent des enregistrements de langue sifflée turque réalisés par le Prof. Busnel, dont la transcription intégrale (mot en turc, et traduction) se trouvait dans le boîtier de la bande magnétique ; en outre les mots sont dits une fois en turc avant d'être sifflés, de sorte que le corpus est tout-à-fait exploitable.

Le reste du fonds n'est pas utilisable tel quel, faute de transcriptions ; il a tout de même été décidé de numériser tous les documents qui pouvaient présenter un intérêt documentaire. Les documents numérisés occupent trois CD-R ; en voici un inventaire (celui qui est gravé sur chacun des trois CD-R) :

*Numérisation à partir des bandes magnétiques.*

**Amharic :**

Une fiche dans le boîtier donne des renseignements sur le document selon le format de l'Institut de Phonétique de Grenoble (fiche cartonnée) :

Langue : Amharic. Région d'origine : Addis-Abeba.

Circonstances d'exécution : chambre sourde

Locuteur : FANTAYE Achagari, sexe : M, année de naissance : 1911, origine géographique : Addis-Abeba.

Lieu d'exécution : Institut de Phonétique (Grenoble). Date : 29 août 1967

Collecteur : M. DURAND

Genre : Texte spontané

Magnétophone : NAGRA III. Microphone : AKG. Vitesse : 19 cm/s. Bande : BASF

Original.

Durée : environ 10 minutes. 1 bobine de 13 cm.

**birman :**

enregistrement court. Qualité médiocre. Le son de la bande originale sature très fortement.

**kenga :**

enregistrement de terrain (avec chants de coqs, etc.), de qualité correcte. Vitesse 3 1/2. Enregistrement original non saturé. Contenu très complet : paires minimales et listes de mots. Les mots proviennent d'une grammaire (qu'il faudrait retrouver). Tous les mots sont dit en français puis répétés/traduits en

kenga par l'informateur. Enregistrement long (plus d'une heure). Comprend de très nombreuses paires minimales par la longueur et le ton, et des phrases.

**Ketchuan, kabre**

Les deux enregistrements se succèdent sur la bande magnétique. Vitesse 3 3/4, une piste : "KETCHUAN de BOLIVIE, KABRE (AFRIQUE)" (mentions portées sur le boîtier). L'enregistrement sature très fortement. Le texte ketchuan (quechua?) est présenté en espagnol. L'enregistrement réalisé en Afrique, par un pasteur dont le nom est donné dans l'enregistrement, date de 1963. Il consiste en mots isolés non accompagnés de leur traduction.

Titre : dans le dossier ketchuankabré, document n°3.wav.

**khmer :**

15 minutes en tout. Bonne qualité d'enregistrement ; stéréo, vitesse 7 1/2. Contenu : présentation des sons de la langue au travers d'exemples, puis longue liste de mots empruntés au français (bactérie, écrou, électron, épiderme, béton...). Il s'agit d'un document réalisé par des chercheurs et non d'une copie de méthode d'apprentissage, comme l'indique le petit bruit de papier tout au début de l'enregistrement, et la longueur de la liste de mots empruntés au français.

**khün :**

original : petite cassette. Son de qualité moyenne. Le boîtier de la cassette porte pour seule mention "khün", mais il y a une rapide présentation en français au début du document. Il s'agit d'un sermon bouddhique. La face A et la face B de la cassette sont dans deux documents séparés, khünA et khünB.

**kikongo :**

Indications sur le boîtier de la bande magnétique : Chanson populaire kikongo

Faits prosodiques en kikongo : ton, durée

Conte "La jeune fille qui épouse un diable" en kikongo

Conversation entre deux étudiants **bamiléké**

Le document sonore comporte des indications à propos des enregistrements. Ils ont été réalisés dans les années 1960 à l'Institut de phonétique de Grenoble.

Vitesse 3 3/4.

La chanson populaire kikongo est d'abord commentée (ou lue) en kikongo, puis chantée. Date : 8 juillet 1965. Enregistré par M. Makoula.

Enregistrements bamiléké : datent de 1964.

Sur la même bande (piste de droite, vitesse 7 1/2) se trouvent des syllabes d'une langue monosyllabique à tons (une même syllabe lue avec plusieurs tons). Prononciation hyper-articulée, son très bon (pas saturé). Langue : ce n'est ni du mandarin ni du vietnamien.

Tous les documents sont en MONO. Titres : kikongobamiléké(=1).wav, et pairesminiminconnu.wav pour les paires minimales.

moore :

le début de l'enregistrement indique "Texte tiré de la page 50 du syllabaire moore [möre?] de 1960". Une mention sur la bobine précise : "J.B. Bunkungu, 30/7/1968". Enregistrement de bonne qualité, vitesse 7 1/2. Durée : 45 secondes.

moore2: auteur : "Ab. Emmanuel KALMOGHO, 6 rue du Regard, Paris 6e". La bande est intitulée "Pour une transcription phonétique du Môoré. Le son est de qualité très médiocre : écho.

munzambo

(mention portée sur la boîte, de la main du Professeur Gsell : "Munzambo, parler de Boyi") : enregistrement de 6 mn seulement, mais de très bonne qualité acoustique. Vitesse 7 1/2. Présentation du locuteur, en français : Jean BONI, né en 1948 à Brazzaville. Présentation des tons, et de phrases, avec chaque fois un bref commentaire en français. Le nom de la langue est dit au début de l'enregistrement.

turc sifflé

Mots dits en turc puis répétée en « langue sifflée ». Enregistrement de très bonne qualité, avec transcription complète. Auteur : Prof. BUSNEL.

L'original du document magnétique et des transcriptions, ainsi qu'un exemplaire du document numérisé et un jeu de photocopies des transcriptions, ont été remis à Mme Annie Rialland (voir ci-dessous, « Exemplaires existants »).

uldeme

35 minutes de mots, phrases et textes.

Réalisé à la fin des années 1960 par le Pr. Pierre PROVOOST. De brèves informations sur l'enregistrement sont enregistrées en français sur le document.

Plusieurs informateurs (jeune, âgé...) disent les mots à tour de rôle. Qualité globalement correcte ; bruit de sifflement sur une grande partie de la bande. Deuxième partie de la bande : liste de mots d'une autre langue (illisible : DONGLADEAH?) en vitesse 4,75 cm par seconde (vitesse 1 7/8), que l'on ne peut écouter au LACITO pour l'instant. N'a donc pas été numérisé.

Dernièrement, les enregistrements de vietnamien qui ont donné lieu aux *Remarques sur la structure de l'espace tonal en vietnamien du sud (parler de Saigon)* (Gsell 1980) ont également été numérisés ; ils fournissent un complément très intéressant à cette publication.

Les questions qui se posent pour la suite du travail concernent :

- l'achèvement du travail de numérisation : une décision doit être prise concernant les enregistrements thaï, qui sans être des documents irremplaçables pourraient illustrer les travaux du professeur Gsell sur le thaï lors d'une éventuelle publication groupée
- les modalités de conservation et de diffusion des documents numérisés (qui pourraient par exemple représenter les premiers éléments d'une phonothèque de l'ILPGA).

### 2. Autres tâches de numérisation menées en 2001-2002

Les premiers travaux de numérisation réalisés au LACITO au cours de cette année ont concerné nos propres données : sur le diu miên, langue hmông-miên parlée au Vietnam, les enregistrements ayant été réalisés au cours de cinq brefs séjours sur le terrain (ces enregistrements numérisés occupent sept CD-R au total) ; sur le teko (langue tupi de Guyane française), données collectées en 1998, au cours d'une première mission de terrain. Il s'agit d'enquêtes de vocabulaire et surtout de récits traditionnels, dont une transcription partielle avait été réalisée sur place. L'essentiel de ces données réparties sur plus d'une dizaine de cassettes tient sur trois CD-R. Ce fonds incomplètement renseigné a été transmis à la personne qui a repris depuis les recherches sur la langue teko, et est à même de les exploiter.

Des tâches de numérisation d'enregistrements appartenant à des chercheurs travaillant dans le domaine des langues d'Asie du Sud-Est sont également en cours. Elles concernent de petits volumes : trois CD-R ont été réalisés (en quatre exemplaires), huit autres sont en préparations. Une poursuite de ce travail est envisagée, selon les urgences documentaires constatées par les chercheurs eux-mêmes. La procédure, dans l'état actuel des choses, est très informelle. Lorsqu'un chercheur formule le souhait de transférer des données de cassettes analogiques sur un support numérique, la numérisation est effectuée par nos soins au LACITO, éventuellement en présence du chercheur. Les chercheurs sont encouragés à fournir en même temps les documents écrits renseignant sur les documents audio. Ceux-ci sont saisis par traitement de texte ou scannés, et gravés sur le même CD que les documents sonores. Une fois les données reproduites sur CD-R, le chercheur dispose de plusieurs exemplaires, qui leur assurent de bonnes chances de pérennité dans l'attente d'une solution plus durable.

## a. Le corpus vietnamien

Au cours de deux années au Vietnam, un corpus illustrant l'ensemble des oppositions phonologiques du vietnamien a été réalisé.

## Procédure de réalisation du corpus

La base du corpus consiste en un ensemble de listes de phonèmes, de syllabes et de mots, qui est prononcé par les douze locuteurs. Les textes lus ont été sélectionnés en fonction des domaines d'intérêt des informateurs, le choix se faisant en définitive avec eux, pour être assuré qu'ils portent de l'intérêt au texte et ne considèrent pas sa lecture comme un pensum. L'idée sous-jacente est qu'il faut être respectueux des *genres* de parole existants (de même que René Gsell observait, à propos du thai, que la voix chuchotée est employée par les femmes, qui en acquièrent la maîtrise au cours d'un apprentissage, tandis que les hommes ne la pratiquent pas) pour bien cibler le corpus. Ainsi, Mme Thu Thê, dont la profession consistait à traduire des films étrangers et à en lire la voix off, nous a expliqué qu'elle pratiquait deux types de lecture bien différenciés : la lecture de documentaire, et l'interprétation de dialogues de films. Le corpus enregistré correspond à ces deux styles ; s'agissant de l'interprétation de dialogues de films, ce sont ses propres traductions qu'elle interprète. Sans préjuger de l'utilisation qui sera faite du corpus, on a ainsi de solides gages du caractère authentique des productions.

Le fait de parler la langue des informateurs (qui en outre étaient tous des amis ou connaissances depuis plusieurs mois au moins) s'est avéré important pour pouvoir accompagner le travail de l'informateur pendant les séances d'enregistrement, et le reprendre en cas d'erreur de lecture ou d'interprétation. Pouvant contrôler avec une grande certitude si l'informateur se trompait ou non dans l'identification du mot écrit, l'enquêteur est à même d'établir son corpus de façon particulièrement fiable.

## Intérêt scientifique

L'écriture actuellement employée au Vietnam est directement issue de la transcription, de nature phonétique, élaborée au XVIIe siècle par des Européens (portugais, italiens et français)[11]. Elle offre au regard une analyse phonématique de la langue, que l'on est quotidiennement amené, au cours de l'apprentissage de la langue, à comparer avec les réalisations phonétiques du vietnamien contemporain. Plus précisément, l'alphabet représente un type idéal de langue. Aucun dialecte ne possède toutes les oppositions. Pour aborder la langue réelle, il est nécessaire de partir d'un dialecte homogène, et d'abord de l'idiolecte d'une seule personne. Principe que défend avec énergie XUE Fengsheng 1999 p. 2, mais qui n'est pas systématiquement appliqué dans la collecte des dialectes vietnamiens (travail entrepris par le département de linguistique de l'Université Nationale de Hanoi). Les linguistes utilisent généralement l'orthographe vietnamienne pour noter la langue, norme commune qui permet de se comprendre aisément entre chercheurs, et se dérobent ainsi à la

[11] Pour une présentation détaillée de « L'origine des particularités de l'alphabet vietnamien » voir A.G. Haudricourt (1949).

nécessité d'élaborer une notation phonétique adéquate. C'est ainsi que Doan (2000) ne propose à aucun moment de tableau en écriture A.P.I. des rimes et initiales existant en vietnamien, moins encore de tableau complet de toutes les syllabes tonalisées existant en vietnamien de Hanoi. Au cours de 2 ans à Hanoi (1999-2001), curieux de la phonétique de la langue réelle, nous avons souhaiter dépasser le stade de la collection de curiosités éparses, et recueillir les données qui permettraient une analyse des sons en système. Nous avons demandé à des amis et connaissances d'enregistrer des éléments de corpus. Pour le volet phonématique/ syllabique, un large corpus a été réalisé ; les tensions du système ont fait l'objet d'une attention particulière : confusions de phonèmes et syllabes distingués par l'orthographe ; transphonologisations ; traits secondaires des phonèmes ; différenciations en cours entre rimes qui appartiennent à une même série étymologique.

Parmi les constats que le corpus permet d'établir avec certitude, se trouve une vérification de phénomènes établis depuis longtemps, expliqués dans les manuels, tels que la confusion entre s, ʃ et ʒ (orthographiés x, s, gi), mais aussi, de façon plus originale, des tensions constatées à Hanoi, telles que la confusion de n et l en l (tandis qu'ils se confondent en n dans la région de Nam Dinh). Le phénomène, surtout répandu dans le parler « vulgaire », est un sujet de plaisanterie pour les puristes, mais le phénomène se répand.

L'opposition entre les rimes uə et wœ (notées –ua et –uơ) s'est perdue. Les informateurs ayant un diplôme universitaire en langues et lettres étaient encore capables de les réaliser de façon claire (mais précisaient qu'ils ne la réalisent pas dans la langue de tous les jours) : dans le premier cas c'est u qui est noyau de syllabe, dans le second cas c'est œ. Parmi les « moins de 30 ans », seule Hanh, particulièrement conservatrice et attachée à la lettre, sait lire sans hésitation le mot *quở* (dans le mot *quở trách*) en choisissant le deuxième élément vocalique, œ, comme noyau de syllabe. Mme Li (née à Nam Dinh, éducation primaire) ne sait pas réaliser l'opposition, et ne distingue plus les « paires minimales ». Acoustiquement, les mots en finale –ua évoluent librement dans l'ensemble de l'espace précédemment partagé entre les deux finales : l'un ou l'autre des deux éléments vocaliques peut être noyau de syllabe. La syllabe của, qui signifie « de », est très fréquente, et permet donc une vérification commode des réalisations en discours spontané (enregistrements à la radio, textes suivis lus par des informateurs, conversations). Elle est parfois prononcée kwœ, le noyau de syllabe étant alors œ, et non u comme cela devrait nécessairement être le cas s'il y avait une opposition phonologique.

En conclusion, il ne paraît pas inutile de disposer de ces enregistrements qui à la fois établissent le caractère vestigial de l'opposition et donnent la prononciation des rares locuteurs qui savent encore la réaliser.

De même pour –ươu [ɯəw], lequel n'est plus distinct de –iêu [iow].

-ưu ([ɤu]) est en train de disparaître, identifié à –iu. A l'heure actuelle, on entend trois variantes : ɤu, yu, iu. En lecture soignée, les informateurs qui ont une éducation universitaire insistent pour lire [ɤu], et sont gênés par –yu et –iu, tout en reconnaissant qu'ils l'entendent souvent prononcé ainsi. En parole spontanée, les réalisations oscillent entre –yu (rime différente de l'ancien -ɤu, mais qui reste distincte de –iu) et –iu.

L'étude de ces changements n'est pas indifférente pour l'étude des contacts entre langues, puisque Michel Ferlus (communication personnelle) remarque que ces oppositions qui semblent être en train de disparaître n'apparaissent que dans des emprunts anciens au chinois.

Parallèlement à ces phénomènes, on constate des confusions (peut-être anciennes) entre rimes très bien représentées (qui représentent donc une perte d'oppositions pourtant très importantes au plan fonctionnel par rapport au système étymologique reconstruit par les spécialistes, Maspéro, Haudricourt, Ferlus) : -ay, -ây (=ăj, ɛj). Ainsi, dạy et dậy (« enseigner » et « se réveiller ») sont homophones. L'école entretient une conscience de la variante « correcte » chez les locuteurs, qui d'ordinaire ne font pas la différence. Plus généralement, dans la langue relâchée (langue de tous les jours), le son [ɛ] est très représenté, en remplacement des finales –ai et même –ay : des mots comme phai « devoir », ngay « jour » deviennent ŋɛ, fɛ, le [j] final étant à peine perceptible. Mais le phénomène n'est sensible qu'en parole continue et en contexte adverse : quand ils s'appliquent, les locuteurs rétablissent les distinctions.

Une des motivations principales de cette entreprise était de pouvoir par la suite visualiser sur spectrogramme ces sons, en un mot passer de l'apprentissage à l'analyse. C'est ainsi qu'a pu être réalisée une *Etude acoustique des finales occlusives et nasales du vietnamien au ton bas glottalisé* (voir dossier VIETNAMIEN), présentée aux 7e Rencontres des doctorants de l'université Paris 7 (mai 2002) et aux Journées d'Etudes Linguistiques sur l'Asie Orientale du CRLAO (juin 2002).

Au-delà du domaine phonématique, ce corpus sera en outre employé dans notre recherche de thèse sur la prosodie du naxi (langue tibéto-birmane de Chine), du vietnamien et du mandarin.

## b. Le corpus naxi, support d'un travail de thèse[12]

Dans la réalisation du travail de thèse que nous envisageons au sujet de l'intonation de langues à tons, la constitution d'un corpus de langue naxi (langue tibéto-birmane du Yunnan, Chine) occupe une place importante. Dans l'esprit du programme Archivage du LACITO, ce corpus est préparé directement avec le langage XML (logiciel : XML Spy). Plusieurs documents sont en cours de réalisation, sur la base de la documentation existante sur la langue naxi. L'un d'eux rassemble des séries de syllabes aux quatres tons et les phénomènes tonaux signalés (rôle grammatical des tons, ajustements entre tons, sandhi) ; le dernier contient un corpus précisément ciblé pour l'étude des phénomènes de focalisation ; un autre, utilisé pour l'apprentissage de la langue, reprend l'ensemble des données du premier livre chinois consacré à la langue naxi ([He85]), sous une forme extrêmement souple, qui se prête à de nombreuses utilisations. Projet documentaire et projet de recherche apparaissent ainsi étroitement liés.

### *Nécessité de corpus comparables dans le domaine de la prosodie*

Une citation de J. Vaissière montre l'importance de corpus calibrés :

> The lack of uniformity in the way F0 features and durational features are described in the literature has rendered the search for invariants more difficult. It is however of the highest importance to set up a reliable method for differentiating between prosodic phenomena specific to a particular language (or shared by several languages), from other, more general phenomena inherent to human speech.
>
> The use of the same linguistic material (translated into different languages, to the extent that it is possible to translate), and read under the same laboratory conditions

[12] Notre travail sur la langue naxi, qui en est actuellement à ses premières étapes, bénéficie de données déjà existantes, généreusement communiquées par le chercheur qui les a réalisés. Le propos de ce chapitre étant d'illustrer l'importance d'un corpus ciblé pour la bonne réalisation de la thèse envisagée, nous passons aussitôt aux projets d'enregistrements qui seront réalisés à partir de l'été 2002 ; mais nous souhaitions témoigner du fait qu'une collaboration documentaire telle que celle dont nous avons ainsi bénéficié est extrêmement profitable au débutant, qui peut se familiariser avec la réalité acoustique de la langue tout en faisant « oeuvre utile » (en effectuant la saisie informatique des transcriptions, en précisant la position des mots sur le signal audio...). Sans doute l'existence d'une phonothèque au sein des laboratoires de recherche encouragerait-elle les convergences de ce type.

> by native speakers, should also ease the comparison. There are very few cross-linguistic data available, and most descriptions concern only one of the prosodic parameters at a time (there are almost no cross-linguistic, pluriparametric descriptions). (Vaissière 1983)

S'il est permis d'ajouter une seconde citation :

> ...this paper may look in some respects premature. It is intended as an invitation for scholars of different languages to collaborate on the realization of a common corpus, and to work jointly on a uniform way of description.

Dans ce but, et dans la perspective d'un séjour d'étude de la prosodie de la langue naxi (langue tibéto-birmane du Yunnan, Chine) sur le terrain de juillet à octobre 2002, il paraissait utile de poser les bases d'un corpus ciblé.

Dans l'article qui vient d'être cité, le choix des énoncés du corpus n'est pas expliqué ; les traits essentiels qui ont motivé le choix ne sont pas indiqués. La phrase la plus utilisée est :

*Il a contribué à la majorité des progrès technologiques des vingt dernières années.*

phrase traduite dans trois langues. Les langues étudiées sont essentiellement l'anglais, le français et l'espagnol ; l'énoncé est aisément transposable entre ces trois langues. Si l'on souhaite tenter une traduction en chinois ou en vietnamien, pour le propos qui est le nôtre d'une comparaison étendue, l'énoncé devient moins similaire en surface à ce qu'il était dans les trois langues relativement proches que sont l'espagnol, le français et l'anglais ; en chinois, avec un minimum d'aménagement stylistique, on peut proposer 他是二十年来大多数科技发展成果的见证者与参与者, par exemple : « il a été témoin et acteur des progrès technologiques des vingt dernières années ». Quant à la traduction de cet énoncé en naxi, langue parlée à divers degrés par 200.000 locuteurs en concurrence avec le chinois, elle amènerait à utiliser une majorité de mots récemment empruntés au chinois, et faiblement intégrés phonétiquement (la situation pouvant bien sûr être variable d'un informateur à l'autre, selon la relation qu'il entretient avec sa seconde langue, le chinois), de sorte que l'énoncé risquerait de ne pas renseigner sur les patrons intonatifs usuels de la langue. Il est donc nécessaire de franchir un pas supplémentaire dans l'abstraction, en construisant un corpus fondé, non sur la traduction littérale, mais sur la conservation d'un ensemble de traits formels de l'énoncé d'origine, abandonnant la fidélité de détail pour conserver l'essentiel de la structure. En l'occurrence, l'énoncé

*Il a contribué à la majorité des progrès technologiques des vingt dernières années.*

permet par sa **longueur** d'étudier de nombreux phénomènes tels que **les pauses, les phénomènes semi-globaux, le découpage en groupes (« hat-patterns »), et les phénomènes globaux (« declination line »)**. Pour le reste, le corpus employé par J. Vaissière dans l'article

cité ne formule pas de propositions détaillées quant à la forme que pourrait prendre un corpus voué à la comparaison prosodique entre langues.

## Corpus en cours d'élaboration sur la base des travaux de Beyssade, Delais-Roussarie, Doetjes, Marandin et Rialland

Un corpus permettant d'étudier la prosodie du français et la structure de l'information a été réalisé dans le cadre du travail mené en commun par Claire Beyssade, Elizabeth Delais-Roussarie, Jenny Doetjes, Jean-Marie Marandin et Annie Rialland (Delais-Roussarie et al. 2002; Marandin et al. 2002). Nous avons été témoin, en qualité d'informateur, d'une partie du travail de création de corpus. Les exemples, dont certains sont volontairement calqués sur des « figures imposées » des travaux sur le focus, apparaissent bien improbables aux yeux du linguiste soucieux du naturel des données. Par exemple, confronté à la séquence

Qui a tué César ?
Brutus a tué César.

l'informateur remarquera peut-être que la réponse ne ressemble pas à ce que l'on trouve dans l'usage réel (quel que soit le registre) : on répondrait plutôt *Brutus* (dans le contexte d'un jeu de culture générale), ou *C'est Brutus*, ou, selon les contextes : *C'est Brutus ! Tu savais pas ?* ou une réponse explicative du type *Hé bien c'est Brutus, son protégé*. Si l'on insiste pour que l'informateur lise question et réponse telles quelles (hors contexte), comme c'était le cas lors des séances d'enregistrement, un informateur qui connaît l'anglais (ou l'allemand) reconnaîtra le modèle « germanique » des *pitch accents* que calque cette séquence question-réponse ; il aura naturellement tendance à recourir aux moyens intonatifs que proposent l'anglais et l'allemand pour mettre en valeur le mot *Brutus*, même si cela est contraire au mode de fonctionnement habituel du français. Il ne faut pas sous-estimer les facteurs de cet ordre, tous les chercheurs en linguistique étant immergés dans un vocabulaire anglais, soucieux de s'exprimer correctement dans cette langue dans leurs articles et dans leurs colloques, au point qu'ils remarquent eux-mêmes cumuler les emprunts à l'anglais lorsqu'ils s'expriment en français.

Néanmoins, au cours de la brève « enquête », nous avons eu l'occasion de revenir sur ces solides résistances personnelles à l'encontre des « faux corpus » réalisés en laboratoire, et de nous convaincre que les procédures appliquées dans la constitution de ce corpus créé de toutes pièces permettaient en fait de mettre au jour des phénomènes intonatifs bien réels (réalité qui peut ensuite être vérifiée par le dépouillement de corpus de parole moins contrainte).

Nous avons donc souhaité faire appel aux conseils du professeur Annie Rialland dans le choix de certaines phrases à intégrer à notre propre corpus de recherche. Ces énoncés ont pour la plupart subi des modifications visant à leur donner un contexte minimal permettant de stabiliser les interprétations possibles par l'informateur. Il va de soi que cette entreprise n'engage que la responsabilité de l'auteur.

Pour le français, les hypothèses que le corpus a permis aux auteurs de formuler et d'étayer sont, en particulier : (i) La partition fond-focus et la partition entre contenu informatif et non informatif sont distinctes (« orthogonales »). (ii) Le marquage prosodique de l'ajout d'information (« update informatif ») caractérise l'assertion, non les autres types illocutoires. Soulignons d'emblée la définition qui, dans cette perspective, est donnée de la notion de focus : *le focus, c'est ce qui est interrogé dans une question, et ce qui est indiqué dans une réponse (ce qu'on asserte dans une assertion)* (ces formulations appartiennent à A. Rialland) ; en d'autres termes, le focus est la cible de la modalité illocutoire, ou « force illocutoire » pour reprendre la formule de BEYSSADE C., MARANDIN J.M. & RIALLAND Annie (2002). En français, le focus serait marqué par un « illocutionary boundary tone », qui peut éventuellement se trouver répété en fin de groupe :

C'est Jean-Pierre(TON DE FOCUS) qui est venu hier(TON DE FOCUS) ?

C'est Jean-Pierre(TON DE FOCUS) qui est venu hier(TON DE FOCUS).

(Dans le premier énoncé, le « ton » est Haut, si on souhaite le noter comme un ton ponctuel, en suivant les modèles de tonologie africaine récemment repris par le courant de la « phonologie intonative » ; dans le second énoncé, il est Bas.)

La présence d'un focus suppose donc qu'il y ait une alternative ouverte :

Apporte-moi *trois* livres de grammaire. *(et pas quatre, ou plus, ou moins...)*

La focalisation se distingue de l'emphase, insistance quantitative :

Tu sais combien il a lu de livres en une seule journée ? Il en a lu *dix-sept* ! / Il a lu *dix-sept* romans policiers ! (c'est vraiment impressionnant)

Pour mettre en lumière les phénomènes de cet ordre, il est nécessaire de contrôler le contexte, par un jeu de questions et réponses :

Tu as pris le train de 18 heures ?

Non, c'est celui de 19 heures que j'ai pris.

Les énoncés du corpus original sont particulièrement artificiels, dans le sens où la réponse reprend l'intégralité des informations contenues dans la question. Ce type d'échange, pour le moins peu naturel en français, est peut-être moins naturel encore dans certaines langues qui évitent la répétition de référents déjà mentionnés : par exemple le tamang, dont les habitudes

d'enchaînement du discours ont été décrites par Martine Mazaudon[13] ; la situation semble similaire en naxi (Hé et Jiāng 1985, p. 100-101)[14]. Le caractère artificiel de ce « dialogue » peut parfois ne pas être gênant ; néanmoins, il paraît utile de laisser ouvert à l'informateur un choix plus large de réponses ; pour le français, on peut par exemple suggérer :

> Tu as pris le train de 18 heures ?
>
> Non, c'est celui de 19 heures que j'ai pris. / Non, celui de 19 heures. / Non, le train de 19 heures.

A ce stade du travail, nous nous bornerons à indiquer certains des exemples retenus, qui donneront une idée du type de contrastes dont un tel corpus permet l'exploration. La recherche d'équivalents acceptables de ces phrases en naxi sera l'une des tâches de l'été 2002.

a) Le domaine focal :

Qu'est-ce qui se passe ?
C'est le petit qui est tombé dans l'escalier.

Qu'est-ce qui se passe ?
Y'a sa valise qui est tombée dans l'escalier.

Qu'est-ce que c'est que ce bruit ?
C'est sa valise qui est tombée dans l'escalier.

Qu'est-ce qu'il a montré au juge pendant sa garde ?
Il a montré son agenda au juge. [15]

Qu'est-ce que tu lui as demandé de faire ?
Je lui ai demandé de ranger sa chambre.

---

[13] Intervention au sujet de la langue tamang dans le Séminaire « Langues tibéto-birmanes », Université Paris 3, 2001-2002. [Ces réflexions ont été publiées postérieurement au présent mémoire (Mazaudon 2003a; Mazaudon 2003b).]

[14] Les exemples fournis par He Jiren sont (traduits littéralement) : « Où est-il parti ? – Voir du théâtre » (« élision » du sujet), « Tu as mangé ? – Mangé (V + ASPECT ACCOMPLI) » (« élision » du sujet et de l'objet du verbe), « Qui veut y aller ? – Moi » (« élision » du prédicat). Mais He Jiren juge que les deux solutions, avec et sans « élision », sont acceptables. En naxi, ces phrases se disent :

$t^{h}$ɯ˧ ze˩ xɯ˧ ? – ($t^{h}$ɯ˧) ɕi˥ ly˩ xə˩ .

ŋv˥ xɑ˧ ə˥ ndzɯ˧ sɯ˧ ? – (ŋə˩ xɑ˧) ndzɯ˧ se˩ .

ə˧ne˩ bɯ˧ le˧ ? – ŋə˩ (bɯ˧).

[15] Réalisation en français :

**Il a montré son agendaL% au jugeL%.**

Le premier est un ton de frontière qui marque le domaine focal, le second est une répétition du premier, il lui fait écho.

C’est ma soeur qui va rigoler.

C’est au petit que ça va faire plaisir !

b) Contraste (réponse négative) :

(4) C’est Jean-Pierre qui est sorti avec Marie ?
(5) Tu pars dans la voiture de Jean-Bernard dimanche prochain ?
Réalisation en français :
C’est Jean-PierreL% qui est sorti avec MarieL% ?
Tu pars dans la voiture de Jean-Bernard% dimanche prochain% ?

Ces questions appellent une réponse portant sur l’élément qui se trouve dans le domaine focal. La réponse attendue à (5) est du type « Non, dans celle de Marina », non « Non, samedi prochain ».

Tu as pris le train de 18 heures ?
Non, c’est celui de 19 heures que j’ai pris. / Non, celui de 19 heures. / Non, le train de 19 heures.

Vous allez bien à Copenhague ?
Non, je vais à Amsterdam.

Vous allez bien à Budapest ?
Non, je vais à Bucarest.

C’est un violon qu’il s’est acheté ?
Non, c’est un violoncelle qu’il a acheté. / Non, un violoncelle.

C’est pour Jospin que Mathilde a voté ?
Non, c’est pour Chirac que Mathilde a voté. / Non, c’est pour Chirac qu’elle a voté.

C’est le sac de Jean-Pierre qui est tombé ?
Non, c’est la valise de Jean-Pierre qui est tombée.

c) Réponses à plusieurs termes :
Qui est venu à la soirée ?
François, Bernard et Marina.

Qui est venu à la soirée ?
Jean-François, Marina et Valentin.

Qu’est-ce que fumaient les chanteurs de rock dans les années 60 ?
Les chanteurs de rock fumaient de la marijuana.
Les chanteurs de rock anglais fumaient de la marijuana.

Qui a refusé de participer ?
Marie a refusé de donner de l’argent, et Jean-Pierre d’écrire un discours.

Qu’est-ce qu’il a donné aux enfants ?
Il a donné des ballons aux garçons et des poupées aux filles.

Quels sont les projets de Jean-Pierre pour cette semaine ?
Il veut aller au cinéma lundi soir et dîner au restaurant mercredi.

Qu'est-ce que Jean-Pierre a l'intention de faire ?
Il a l'intention de vendre sa maison, mais pas de quitter la ville.

(s'agissant d'un mari qui rechigne à s'occuper des enfants :)
Qu'est-ce qu'il refuse de faire ?
Il refuse de chercher les enfants le soir, mais pas de les amener le matin.

Que sont devenus les étudiants dont Bernard s'est occupé ?
Les étudiants que Bernard a entraînés ont intégré l'équipe de France, ceux qu'il a dissuadés de poursuivre une carrière sportive ont repris un cursus normal.

Que sont devenus les enfants de Marie ?
Ses garçons sont à la fac et sa fille est encore au lycée.

Vous avez mangé quoi ?
François a pris du saumon, Bernard du gigot, et Marina des huîtres.

Vous avez mangé quoi ?
Marie-Pierre a pris des spaghettis, Jean-François du poulet rôti, et Marina des croquettes de poisson.

Vous avez mangé quoi ?
Jean-François a mangé des spaghettis, et Marie-Pierre a pris de la paëlla.

Qui est à la fac ?
Les enfants de Bernadette sont à la fac de Nanterre.

Qui est à la fac ?
Les enfants de Bernadette sont à la fac de Nanterre, et les garçons de Jean-Bernard sont à Jussieu.

Qu'est-ce que les enfants ont fait cet après-midi ?
Valentin a travaillé sa clarinette et Marianne a relu ses cours.

Je vais aller en Belgique en décembre, et en Hollande en mars.
Moi, c'est en Angleterre que j'irai en janvier, et en Allemagne que j'irai en février.

Je vais aller en Belgique en décembre, et en Hollande en mars.
Moi, c'est en janvier que j'irai en Angleterre, et en février que j'irai en Allemagne.

c) « Question à choix multiples »

Comment tu viens, en métro ou en autobus ?
En autobus.

e) Réplique en décalage avec le premier énoncé

Dans ces exemples, la réponse réoriente la question. Ce type de réponse s'accompagne en français d'une intonation marquée ; on peut s'attendre à retrouver un contour particulier dans une autre langue :
Tu as vu Jean-Pierre ?
J'ai vu son blouson dans l'entrée.

Qui a pris mes affaires ?
Ton cartable est dans l'entrée.

Tu as vu mon frère ?
J'ai vu son manteau dans l'entrée.

Tu as vu Bernadette ?
J'ai vu sa voiture sur le parking.

Qu'est-ce qui te déplaît ?
C'est pour Tournier qu'elle va voter.

Qui a repeint la table ?
Marie l'a poncée.

Qu'est-ce qu'il aime manger, Jean-Pierre ?
Il ne mange jamais de viande.

Est-ce que Jean-Pierre a acheté l'appartement ?
Bernadette n'a pas voulu.

Qui a pris mon stylo ?
Ta trousse est sur la table.

Je suis allée au Printemps avec Marie-Claire aujourd'hui. J'ai acheté plein de trucs.
Et Marie-Claire, qu'est-ce qu'elle a acheté ?

Il est également prévu d'étudier les mots « associés au focus » selon l'expression de Jackendoff (1972) : certaines expressions telles que *seul (only)* sont « associés au focus ». « Mary only invited BILL for dinner »/ « Mary only invited Bill for DINNER ».

# *Conclusion et perspectives*

## *1. Des apprentissages à approfondir*

La documentation est un travail qui peut occuper à plein temps. Or le chercheur (aspirant ou confirmé) est rapidement rejoint par la nécessité d'avancer ses recherches. Au cours de la présente année, nous n'avons pas été au bout des questions que nous souhaitions soulever dans le travail documentaire, du fait de l'investissement de temps demandé par notre formation en phonétique et la préparation d'un travail de thèse (concernant l'intonation pragmatique d'une langue à ton, la langue naxi). En particulier, nous n'avons pas encore eu l'occasion d'apprendre le maniement des outils employés aux laboratoires de Grenoble et d'Aix-en-Provence, ni de consulter la base de données mentionnée par Mario Rossi dans son ouvrage sur l'intonation (Rossi 1999), non plus que d'utiliser xassp, outil employé au laboratoire de Kiel. Ces apprentissages seront sans aucun doute importants à la fois pour nos recherches sur l'intonation et pour le choix de formats documentaires adaptés.

## *2. Perspectives institutionnelles*

Pour ce qui est du travail documentaire, qui met en oeuvre, à petite échelle, les principes qui ressortent du présent travail, il est permis d'espérer qu'il permette d'aboutir, à moyen terme, à la constitution d'une « phonothèque phonétique » au sein de l'**ILPGA**. Le soutien apporté par le Conseil Scientifique de Paris 3 au « projet innovant » intitulé « Travail de première main sur des langues rares : description, archivage, transmission » nous encourage à espérer que des structures puissent être mises en place pour répondre à certains des besoins documentaires constatés à l'échelle du laboratoire de phonétique de Paris 3.

Au sein du **programme Archivage du LACITO**, nous espérons

- contribuer à une plus grande visibilité du programme, notamment en participant à la création d'une interface conviviale mettant en valeur les données de très grande qualité librement mises à disposition sur Internet

- faciliter la mise en place de coopérations avec le département des Collections sonores de la Bibliothèque nationale, pour que le programme Archivage soit adossé à une institution pérenne[16].

Nous avons également le projet d'entrer en contact avec la phonothèque du Musée de l'Homme, et avec la Société des Missions Etrangères de Paris, qui possède des fonds documentaires sur de nombreuses langues, mais dont nous ne savons pas si elle possède des collections d'enregistrements sonores.

Le présent mémoire souhaitait poser quelques jalons dans ce travail, qui ne peut prendre tout son sens que sur le long terme.

[16] Ces coopérations pourraient par exemple conduire à ce que le programme Archivage du LACITO devienne « pôle associé » à la Bibliothèque nationale.

*ANNEXE : projet formulé au Vietnam*

Projet de

# Constitution d'une phonothèque des langues du Vietnam

*A) Grandes lignes du projet*

I. Résumé du projet :

- Constituer un fonds d'enregistrements aussi exhaustif que possible : langues et traditions orales des diverses ethnies. (1) rassembler les documents existants ; (2) **fournir un équipement de grande qualité aux linguistes qui vont sur le terrain (et aux étudiants de linguistique qui effectuent des stages de terrain)**.
- Assurer la conservation de ces enregistrements, par transfert sur CD-ROM, comme le fait déjà l'Institut de Recherche sur le Folklore pour les enregistrements vidéo de rites des ethnies minoritaires. Réaliser une collection de CD-ROM présentant les données accompagnées de leur analyse.
- Assurer la diffusion de ces documents dans la communauté scientifique (voir en page 7 la liste des programmes et organismes similaires existant à l'heure actuelle, avec lesquels des échanges très fructueux pourront avoir lieu). Ce corpus pourrait aisément être diffusé sur l'Internet (en veillant aux droits d'auteur des chercheurs qui ont réalisé le travail).

II. Echelle du projet :

A long terme : les langues minoritaire du Nord du Vietnam, et les dialectes des diverses régions.

Dans l'immédiat, le travail peut se faire dans deux directions :

1) En concertation avec les chercheurs (vietnamiens et étrangers) qui travaillent d'ores et déjà dans ce domaine, rassembler, à Hanoi, les enregistrements déjà réalisés. Les conserver (la dégradation des cassettes magnétiques est très rapide au Vietnam, d'où la nécessité du passage sur disque compact). Etendre et systématiser le travail d'enregistrement dans le champ d'étude de ces chercheurs.
2) Constituer un grand corpus des langues les plus menacées, en mettant à contribution les étudiants du Département de linguistique de l'Université des Sciences Sociales et Humaines (qui pourront investir beaucoup de temps et d'énergie dans le travail de documentation, lequel fait partie de leur cursus).

III. Partenaires principaux :

Département de linguistique de l'Université des Sciences Sociales et Humaines (Khoa Ngôn ngữ học, Trường Đại học Xã hội và Nhân văn)

Centre de Recherches Linguistiques sur l'Asie Orientale (54 bd Raspail, 75006 Paris)

et (lorsque le projet aura déjà été mis en route à l'Université des Sciences Sociales) : Musée d'Ethnologie du Vietnam ; centres de recherche en linguistique et en ethnologie vietnamiens et étrangers.

Pour l'expertise technique : travail avec le laboratoire de phonétique de l'Institut de Linguistique et Phonétique Générale et Appliquée, Université Paris III-Sorbonne Nouvelle (laboratoire où Alexis Michaud travaillera à partir de l'automne 2001).

IV. Personnes à qui la direction du projet pourrait être confiée :
M. Trần Trí Dõi, Vice-Doyen du Département de Linguistique de l'Université des Sciences Sociales

et Mme Barbara NIEDERER, spécialiste des hmông-miên, CRLAO-CNRS.

V. Collaborateurs (liste ouverte) :
Chercheurs-linguistes vietnamiens et étrangers ;

M. Michel FERLUS, Centre de Recherche Linguistique sur l'Asie Orientale, EHESS-CNRS

avec la contribution d'ethnologues qui travaillent sur le terrain, pour enrichir les collections d'enregistrements : M. Christian CULAS, Institut de Recherche sur le Sud-Est Asiatique (UMR 6571 du CNRS) ; M. Jean MICHAUD, Université de Hull (G.-B.).

VI. Durée envisagée :
Le projet dans toute son ampleur demandera un travail de très longue haleine. La première phase concerne les 5 années à venir : automne 2000-automne 2005.

## B) Contenu scientifique du projet

### I. Enjeu scientifique du travail de conservation

Un corpus fiable et abondant, établi avec soin, est précieux pour les spécialistes de divers domaines : pour les linguistes, mais aussi pour les ethnologues. En particulier, lorsque le projet aura atteint toute son ampleur, il permettra l'édition des textes de littérature orale selon les critères scientifiques. Les recueils de littérature orale des ethnies minoritaires sont jusqu'à maintenant publiés en traduction vietnamienne seulement. Les auteurs de ces collections soulignent qu'ils souhaitent réaliser une édition bilingue, mais que les conditions de travail ne leur permettent pas de présenter le texte en langue originale. L'extension du travail d'enregistrement, offrant une meilleure connaissance des langues du Vietnam, permettra un travail d'édition scientifique de cette composante très riche du patrimoine culturel vietnamien.

S'agissant de la langue vietnamienne (*kinh*), le travail d'enregistrement est également la condition d'une édition selon les critères scientifiques des oeuvres de la littérature orale. A l'Institut d'Etude de la Culture Populaire, M. le Professeur Ngô Đức Thịnh (Directeur du centre) projette de reprendre le travail de collecte de contes vietnamiens que ses prédécesseurs avaient mené dans une optique différente (les contes étaient réécrits dans l'idée d'en « élever » le registre). Le travail d'enregistrement, étendu à ce domaine, permettra de faire avancer à la fois la connaissance du patrimoine culturel vietnamien (culture populaire) et du patrimoine linguistique (langue populaire "réelle", différente de la langue écrite).

Ce travail est d'une grande importance pour les langues du monde entier, et du Vietnam en particulier. Dans le passé, des chercheurs ont effectué avec rigueur un travail de collecte et d'établissement de corpus. Par exemple, il existe des recueils de texte de la langue aztèque ancienne, l' "aztèque classique" (langue qui était parlée au Mexique), recueils constitués au XVIIe siècle. Aujourd'hui, les chercheurs qui travaillent sur cette langue disposent d'un corpus fiable, qui permet des études très approfondies. Mais de tels corpus sont rares, et ne semblent pas être très à la mode, la "grammaire générative", par exemple, n'utilisant pas de grands corpus, seulement un petit nombre de phrases. Or le phénomène de disparition de langues s'accélère[17].

Aujourd'hui, le travail de conservation n'est plus aussi difficile qu'auparavant. Les techniques modernes d'enregistrement facilitent pourtant beaucoup le travail de conservation. La conservation des enregistrements ajoute une dimension supplémentaire à un corpus de langue "rare". L'enregistrement est porteur de renseignements précieux : intonation, rythme, jeux vocaux... S'agissant de cultures à tradition orale, cette composante n'est pas seulement pittoresque : elle renseigne sur l'art de la parole dans ces cultures, qui mérite des études très fines dont un tel corpus fournit la base.

### II. Importance d'un cadre général de recherche

Le présent projet vise à l'exhaustivité. Il permet de regrouper les documents (dans un centre situé au Vietnam), et donc d'évaluer précisément le travail qui reste à faire.

**Faute d'un tel projet global, les recherches (si intéressantes soient-elles) n'aboutissent pas à la constitution d'un corpus qui reflète toute la richesse du patrimoine linguistique et culturel du pays.** Les documents originaux ne sont pas diffusés, et les données recueillies ne sont communiquées aux autres chercheurs qu'au travers des publications qui exploitent ces données. Il ne s'agit pas là d'un problème spécifiquement vietnamien : c'est la méthode d'enquête qui est en cause. Les universités qui disposent de gros moyens pour la recherche en

[17] voir l'ouvrage collectif *Endangered Languages*, éd. Berg, 1991.

linguistique n'évitent pas ce problème. Le souci d'établir un corpus complet (intéressant phonétique et phonologie, morphosyntaxe, lexique...) n'est pas toujours la première préoccupation du linguiste, qui s'intéresse souvent à un problème très délimité. A fortiori, la conservation de la tradition orale dépasse son champ de recherche, le bon ethnolinguiste étant d'abord linguiste, et un peu ethnologue pour mieux faire son travail de linguiste. Lorsqu'une langue s'éteint, s'il n'en reste que les quelques minutes (parfois quelques secondes) d'enregistrements réalisés par tel ou tel spécialiste, n'est-ce pas une perte pour les autres branches de la linguistique, et un appauvrissement considérable pour le patrimoine culturel du pays ? Or quelques heures d'enregistrements de mythes et légendes, avec transcription, contiennent plus d'information sur la langue que la meilleure thèse de phonologie ou de syntaxe : d'un côté le document, absolument irremplaçable, de l'autre son analyse, nécessairement incomplète. Dans les faits, les bons corpus sur les "petites langues" ("langues rares") sont plus rares que les bons travaux théoriques. Le présent projet **valorise le travail de documentation, long et ingrat, mais particulièrement urgent aujourd'hui. Installée dans une université, la phonothèque, avec ses équipements de pointe (matériel d'enregistrement et station de travail informatique), constituera un encouragement considérable pour les étudiants** (qui actuellement doivent travailler avec des équipements vétustes, ce qui est très décourageant). En s'appuyant sur ces jeunes collaborateurs, le projet de phonothèque pourra rapidement prendre de l'ampleur.

Pour leur part, ethnologues et géographes n'ont pas toujours le temps d'apprendre les langues concernées. A plus forte raison, les chercheurs qui s'intéressent à une région où coexistent plusieurs minorités ne peuvent approfondir toutes ces langues. S'agissant des enregistrements, ces chercheurs ne connaissent pas nécessairement les quelques notions en prise de son qui permettent de réaliser de bons enregistrements (types de micro...) et de les conserver (utiliser des boîtes étanches, avec des sels spéciaux pour éviter l'humidité).

Grâce au présent projet, au croisement de plusieurs disciplines, on peut envisager un travail auquel chacun des chercheurs (au Vietnam et à l'étranger) contribue, et qui profite à l'ensemble des chercheurs concernés, au Vietnam et à l'étranger.

Le bénéfice pour les chercheurs vietnamiens et étrangers qui travaillent sur le terrain sera également considérable. L'existence de ce cadre de recherches encouragera les linguistes qui, depuis longtemps, consacrent beaucoup de temps et d'énergie au travail de collecte. Le travail de terrain présente beaucoup de difficultés ; l'assurance que la documentation recueillie sera conservée, diffusée, qu'elle recevra l'attention qu'elle mérite dans la communauté des linguistes, qu'elle sera reconnue comme une part importante du patrimoine culturel de la République Socialiste du Vietnam, est importante pour entretenir la motivation des linguistes. En outre, dans le cadre du projet, les droits d'auteur seront clairement pris en compte lors de la diffusion et de l'exploitation. La coopération internationale (et franco-vietnamienne au premier chef) dans ce domaine, toujours active (avec notamment l'ouvrage *Giao l-u v¨n ho¸ vµ ng«n ng ÷ ViÖt-Ph¸p* publié il y a tout juste un an), pourra, avec ce projet, trouver un élan considérable.

## III. Projet détaillé

Principes scientifiques :

1) réaliser **une heure minimum d'enregistrement sur DAT (Digital Audio Tape) pour chaque langue, qui comprenne les sons isolés (logatomes) ; les syllabes ; des paradigmes de verbes et des listes de mots; des phrases ; et un peu d'oral "spontané".**

La durée d'une heure paraît dérisoire. En réalité, un tel inventaire, dans la perspective de la comparaison et de la reconstruction de langues, aurait une valeur considérable. Il permettrait l'élaboration d'une transcription rigoureuse, ouvrant la voie à une conservation des traditions orales répondant aux normes les plus exigeantes.

Le patrimoine linguistique du Vietnam est extrêmement riche, mais il serait faux de penser que le Vietnam soit "en retard" dans le travail de conservation. Au contraire, il n'est pas exclu, lorsque le projet sera en place, que le Vietnam fasse figure de pionnier.

Pour prendre un exemple proche, celui de la Chine, la constitution de grands corpus avec les moyens d'enregistrement les plus modernes fait partie des préoccupations principales des chercheurs, mais n'est pas encore très avancée. Ainsi, elle figure dans le projet pour les cinq années à venir du Centre de recherches sur la culture **DongBa** (culture et religion d'une ethnie minoritaire du sud de la Chine, les NaXi), installé dans la ville de Lijiang (province du Yunnan). (Nous tenons à remercier le directeur du centre, ZHAO Hongshi, qui a bien voulu, en août 2000, nous exposer le détail de son projet, et les difficultés rencontrées, pour que le présent projet bénéficie de son expérience.)

Les enquêtes se feront sur la base d'un questionnaire, rédigé en vietnamien, en français et en anglais (éventuellement en d'autres langues), pour qu'il puisse être utilisé par des enquêteurs de divers pays. Il sera assorti de consignes détaillées pour l'enquêteur. (Mais la procédure pourra être modifiée en fonction des connaissances qui existent déjà sur la langue en question.)

2) Sur la base de ce premier travail :

**Etablir un corpus aussi complet que possible : mythes, contes, légendes, chants... avec une grande rigueur philologique**. Enregistrer avec du bon matériel : un micro de qualité professionnelle, et un enregistreur à cassettes D.A.T. (à défaut, enregistreur à mini-disques, nettement moins coûteux, ou enregistreur à cassettes analogiques de qualité professionnelle).

Procédure pratique :

**Les frais de mission des chercheurs ne seront pas pris en charge**. Le projet consiste **à doter d'un matériel de pointe** les personnes qui, dans le cadre de leurs études ou de leurs recherches, **à leurs propres frais ou missionnés par une institution** (Université, Centre de Recherche...), se rendent sur le terrain. Le matériel d'enregistrement sera conservé au Département de Linguistique de l'Université des Sciences Sociales et Humaines. Le matériel d'enregistrement sera prêté (gratuitement) aux personnes qui partent sur le terrain.

Le corpus sera, petit à petit (le travail peut prendre plusieurs mois), mis en forme sur CD-ROM, à la station de travail installée à Hanoï. La personne qui aura réalisé l'enquête sera invitée à annoter lui-même le corpus. Il pourra travailler avec l'aide d'un membre du comité scientifique pour apprendre le maniement du matériel, et lever les difficultés rencontrées. Idéalement, il s'agit d'ajouter, en regard du *signal* (le son proprement dit) :

- une transcription phonème par phonème ;
- une transcription mot par mot, avec traduction ;
- une traduction littérale phrase par phrase ;
- éventuellement, une traduction plus libre, avec des commentaires d'ordre ethnologique.

Pour cela, il faudra un équipement informatique : **ordinateur (P.C.), logiciels, lecteur de cassettes D.A.T., équipement pour graver des CD**. Cet équipement sera installé au Département de linguistique d'une université disposée à fournir un local adapté (Département de Linguistique de l'Université des Sciences Sociales et Humaines si possible), ou au Musée d'Ethnologie du Vietnam.

A terme, l'idéal serait la constitution, dans une université de Hanoï (avec les crédits du gouvernement de la République Socialiste du Vietnam), d'un laboratoire de phonétique comme celui de l'Institut de Phonétique Générale et Appliquée à Paris (Université Paris-III), avec une chambre sourde qui permettrait d'améliorer considérablement la qualité des enregistrements, et des appareils spécialisés ; mais cela demanderait un budget **considérable**. Il faut souhaiter que le développement du pays permette bientôt la création d'un tel centre.

## C) Questions pratiques et calendrier.

## I. Calendrier (indicatif)

Automne 2000 :
Mise en route de l'inventaire des enregistrements existants.
Choix des priorités pour la réalisation d'enregistrements nouveaux.
Constitution d'un comité scientifique (autour de M. TrÇn TrÝ Dâi, Vice-Doyen du Département de Langues de l'Université des Sciences Sociales et Humaines) pour diriger le projet sur le long terme.
Novembre-décembre 2000 :
Commande du matériel d'enregistrement et du matériel de laboratoire.
Janvier-février-mars 2001 :
Début du prêt gratuit du matériel de terrain.
Session de présentation du mini-laboratoire, et formation à l'utilisation du matériel et des logiciels : numérisation des données ; analyse du signal ; annotation du corpus. (Formation coordonnée par Alexis Michaud.)
(à partir d'avril 2001 si possible) : ouverture du mini-laboratoire aux chercheurs, qui viennent dépouiller le corpus qu'ils ont enregistré.
Mai 2001 :
Premier bilan, et choix d'orientations. Discussions avec les autres projets d'archivage sonore du monde entier, pour des échanges et un partenariat :
Allemagne (Gesellschaft für bedrohte Sprachen, http://www.uni-koeln.de/gbs)
Hong Kong-République Populaire de Chine : Language Information Sciences Research Center, City University of Hong-Kong (M. William S.Y. Wang)
LACITO, Centre A.-G. Haudricourt, bâtiment 23 de l'hôpital P. Brousse, 7 rue Guy Moquet, 94...Villejuif, France (M. Mazaudon, B. Michailovsky, et M. Jacobson)
CELIA, Paris (équipe de M. Launey)
Australie : projet de M. Stephen Wurm
Etat-Unis : Endangered Language Fund, http://www.ling.yale.edu/~elf
Grande-Bretagne : Foundation for Endangered Languages (http://www.unizh.ch/spw/aspw/dang)
Phonothèque du Musée de l'Homme
Unesco : CIPSH/ICHEL/LINGUA PAX/Unesco Education

## II. Matériel nécessaire, et budget prévisionnel

### Pour les enquêtes de terrain :

2 ou 3 équipements d'enregistrement DAT complet : microphones ; enregistreur D.A.T. portable ; cassettes..... 50.000 F
3 appareils à mini-disques, et micros..... 7.000 F
2 ou 3 enregistreurs de cassettes audio de qualité professionnelle, avec microphone, et cassettes..... 40.000 F
SOUS-TOTAL..... **97.000 F**

### Pour le poste de travail fixe, servant à l'archivage :

Locaux *(à la charge de l'Université)*
Armoire blindée pour conserver les appareils et les cassettes *(à la charge de l'Université)*
Déshumidificateur (pour éviter les moisissures) *(à la charge de l'Université)*
Ordinateur (P.C.)..... 25.000 F
Logiciels (Unice, Snorri, Winpitch, Signalize)..... 25.000 F
Lecteur de cassettes D.A.T. ; équipement pour graver des CD..... 25.000 F

"Fonds de caisse" pour imprévus (dont : assurances pour le matériel)..... 25.000 F

TOTAL.....**200.000 F**

# Références citées

BONNEMASON Bénédicte, GINOUVÈS Véronique et PÉRENNOU Véronique, 2001, *Guide d'analyse documentaire du son inédit, pour la mise en place de banques de données*, Parthenay, France, Modal - AFAS.

BOUQUIAUX Luc et THOMAS Jacqueline, 1971, *Enquête et description des langues à tradition orale. Volume I : l'enquête de terrain et l'analyse grammaticale*, 2nd edition 1976., Paris, Société d'études linguistiques et anthropologiques de France.

BURNARD Lou et SPERBERG-MCQUEEN Michael, 1995, « The design of the TEI Encoding Scheme », *Computers and the Humanities*, 1995, vol. 29, n° 1, p. 17-39.

CALAS Marie-France et FONTAINE Jean-Marc, 1996, *La Conservation des documents sonores*, CNRS Editions., s.l.

CORDEREIX Pascal, 2001, « Ferdinand Brunot, le phonographe et les "patois" », *Le Monde alpin et rhodanien*, 2001, vol. 1er-3e trimestre, p. 39-54.

CULIOLI Antoine, 1976, *« Recherches en linguistique: théorie des opérations énonciatives »: transcription du séminaire de DEA de M. Antoine Culioli*, s.l., Centre de documentation sciences humaines.

DELAIS-ROUSSARIE Elisabeth, RIALLAND Annie, DOETJES Jenny et MARANDIN Jean-Marie, 2002, « The prosody of post-focus sequences in French », Aix-en-Provence.

DÔ THÊ DUNG, TRÂN THIEN HUONG et BOULAKIA Georges, 1998, « Intonation in Vietnamese » dans Daniel Hirst et Albert Di Cristo (eds.), *Intonation systems: a survey of twenty languages*, Cambridge, Cambridge University Press, p. 395-416.

DYBKJÆR Laila et BERNSEN Niels Ole, 2000, « The MATE markup framework », s.l., Association for Computational Linguistics, vol.10.

GRIMES Barbara, 1992, *Ethnologue: Languages of the World*, Dallas, Texas, Summer Institute of Linguistics.

GSELL René, 1980, « Remarques sur la structure de l'espace tonal en vietnamien du sud (parler de Saïgon) », *Cahier d'études vietnamiennes, Département de Langues et Civilisations de l'Asie Orientale de l'Université Paris 7*, 1980, vol. 4.

HÉ Jírén 和即仁 et JIĀNG Zhúyí 姜竹仪, 1985, *Nàxīyǔ jiǎnzhì 纳西语简志 (A brief*

*description of the Naxi language)*, Beijing 北京, The Ethnic Publishing House 民族出版社.

JACOBSON Michel, MICHAILOVSKY Boyd et LOWE John B., 2001, « Linguistic documents synchronizing sound and text », *Speech Communication*, 2001, 33 [special issue: « Speech Annotation and Corpus Tools »], p. 79-96.

KOHLER Klaus, 1996, « Modellgesteuerte Prosodiegenerierung. Die Implementation des Kieler Intonationsmodells (KIM) in der TTS-Synthese fuer das Deutsche », *Fortschritte der Akustik*, 1996, vol. 22, p. 90-91.

KOHLER Klaus J., 1992, « Prosodisches Transkriptionssystem für die Etikettierung von Sprachsignaldaten », *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK)*, 1992, vol. 26, p. 238-252.

LEROY Christine et PARIS Catherine, 1974, « Etude articulatoire de quelques sons de l'oubykh d'après film aux rayons X », *Bulletin de la Société de Linguistique de Paris*, 1974, LXIX, nº 1, p. 255-286.

MADDIESON Ian, 1984, *Patterns of Sounds*, Cambridge, Cambridge University Press.

MARANDIN Jean-Marie, BEYSSADE Claire, DELAIS-ROUSSARIE Elisabeth et RIALLAND Annie, 2002, « Discourse marking in French: C accents and discourse moves », s.l.

MAZAUDON Martine, 2003a, « Tamang » dans Graham Thurgood et Randy LaPolla (eds.), *The Sino-Tibetan languages*, London, Routledge, p. 291-314.

MAZAUDON Martine, 2003b, « From discourse to grammar in Tamang: topic, focus, intensifiers and subordination » dans David Bradley, Randy LaPolla, Boyd Michailovsky et Graham Thurgood (eds.), *Language Variation: Papers on variation and change in the Sinosphere and in the Indosphere in honour of James A. Matisoff*, Canberra, A.N.U. (coll. « Pacific Linguistics »), p. 145-157.

ROSSI Mario, 1999, *L'intonation, le système du français: description et modélisation*, Gap/Paris, Ophrys.

SIEGEL David, 1997, « The Web is ruined... and I ruined it! », *World Wide Web Journal*, 1997, vol. 2, nº 4, p. 13-21.

SIMPSON Adrian, KOHLER Klaus J. et RETTSTADT Tobias (eds.), 1997, *The Kiel Corpus of Read/Spontaneous Speech : acoustic data base, processing tools, and analysis results*, Kiel., Institut für Phonetik und digitale Sprachverarbeitung der Universität Kiel, (coll. « Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK) »).

SPERBERG-MCQUEEN C. Michael et BURNARD Lou, 1994, *Guidelines for electronic text encoding and interchange*, s.l., Text Encoding Initiative Chicago and Oxford, vol.1.

VAISSIÈRE Jacqueline, 2002, « Cross-linguistic prosodic transcription: French vs. English » dans N.B. Volskaya, N.D. Svetozarova et P.A. Skrelin (eds.), *Problems and methods of experimental phonetics. In honour of the 70th anniversary of Pr. L.V. Bondarko*, St Petersburg, St Petersburg State University Press, p. 147-164.

VAISSIÈRE Jacqueline, 1983, « Language-independent prosodic features » dans Anne Cutler et Robert Ladd (eds.), *Prosody: Models and Measurements*, Berlin, Springer Verlag, p. 53-66.

WELLS John C., BARRY William et FOURCIN Adrian, 1989, *Transcription, labelling and reference. Multilingual methods and standards*, Chichester, Ellis Horwood.