Corpus oraux en linguistique de terrain

Michel Jacobson

Corpus oraux en linguistique de terrain

Michel Jacobson

LACITO - CNRS
7, rue Guy Môquet
Bâtiment D
94801Villejuif cedex
jacobson@idf.ext.jussieu.fr

RÉSUMÉ. *Les corpus de parole construits en linguistique de terrain sont en général caractérisés par leur faible volume. Ils concernent la plupart du temps des langues peu décrites, sur lesquelles les données sont rares. Le travail de terrain comporte lui aussi des particularités qui font que beaucoup d'enregistrements se rapprochent plus de la parole « naturelle » que de corpus de laboratoire. L'ensemble de ces caractéristiques influence les méthodes utilisées pour la constitution et la gestion de ces corpus.*
Pour illustrer ces méthodes, nous présentons ici un programme d'archivage de documents de linguistique de terrain visant à sauvegarder, pérenniser, normaliser et diffuser ces documents. Nous examinerons en particulier quels ont été et quels sont actuellement les formalismes et les outils qui existent pour aider à la gestion de tels corpus. Nous évoquerons enfin les principaux problèmes d'organisation et de droit concernant la diffusion de telles ressources.

ABSTRACT. *Speech corpora constructed in field linguistics are usually characterised by their small volume. They mostly concern little described languages, for which data are scarce. Field work also has its particularities, which means that recordings are closer to "spontaneous" than to laboratory speech. These characteristics influence the methods used for constituting and managing these corpora.*

To illustrate these specific methods, we will present a program for archiving field linguistics documents, aiming to preserve, perpetuate, normalise and circulate these documents. In particular, we will examine past and present formalisms and tools for managing corpora of this type. To finish, we will evoke some problems in organisation and law concerning the circulation of these resources.

MOTS-CLÉS : *linguistique de terrain, texte interlinéaire, langues minoritaires, corpus oraux, normalisation.*

KEYWORDS: *field linguistic, interlinear text, minority languages, oral corpora, normalisation*

1. Les corpus oraux en linguistique de terrain

Les corpus de linguistique de terrain peuvent être caractérisés par un petit ensemble de critères portant sur les conditions de récolte des données, ainsi que sur la nature de ces dernières. Les outils et méthodes employés pour leur constitution, leur enrichissement et leur étude sont empruntés à diverses disciplines comme les études textuelles ou la gestion documentaire. Les linguistes font notamment grand usage de concordances et de statistiques. En revanche, certains développements effectués pour la reconnaissance automatique de la parole, des outils de type analyseurs syntaxiques ou lématiseurs, issus d'autres disciplines qui pourtant semblent proches, sont généralement délaissés. Les raisons de ce désintérêt peuvent être de nature économique, l'investissement en temps pouvant être disproportionné par rapport au gain apporté quand on travaille sur des petits volumes de données. Ces raisons peuvent également être de nature idéologique. En effet, les linguistes considèrent la transcription, et de manière plus générale toutes les annotations, comme reflétant une analyse et donc tout un échafaudage d'hypothèses, et non pas comme des données qui peuvent se substituer à l'enregistrement. Ils sont donc bien souvent des partisans acharnés d'un étiquetage manuel.

1.1. La nature des données

Les langues sur lesquelles les linguistes de terrain travaillent sont souvent des langues « rares », « en danger » ou « minoritaires ». En fait, il s'agit de langues souvent sans tradition d'écriture, parlées par peu de locuteurs, dans des endroits où ce sont d'autres langues qui sont utilisées par les administrations, les écoles, les médias, etc. Le problème de la transcription est alors ici central, contrairement aux corpus sur les langues qui possèdent écriture et orthographe.

Les corpus sont pour la plupart composés d'enregistrements de parole : récits, dialogues, chants, cérémonies, questionnaires, élicitations, etc. Ces enregistrements sont par ailleurs transcrits, traduits et analysés au minimum jusqu'à un niveau morpho-phonologique.

La quantité de données que l'on peut posséder sur une langue particulière est aussi une caractéristique de la linguistique de terrain. En général, les corpus sur une langue ne sont constitués que de quelques heures d'enregistrements. Souvent, seule une partie des enregistrements a été transcrite, ou bien l'analyse de certaines parties n'a pas été poussée au même niveau que celle des autres. Par exemple, il est fréquent qu'un corpus soit composé de textes inégalement décrits, certains ne comportant qu'une transcription « au kilomètre », d'autres étant traduits en différentes langues et glosés morphème par morphème. Quelquefois, on possède un lexique ou un dictionnaire de la langue contenant entre trois et quatre mille mots. Enfin, avec un peu de chance, il est possible que d'autres linguistes possèdent des

ressources sur cette même langue, ou qu'il existe des descriptions anciennes de celle-ci dans de vieux ouvrages.

Par comparaison, le TLFi[1] *Trésor de la langue française informatisé* comporte quelque 100 000 mots et 270 000 définitions ; le corpus du quotidien *Le Monde*[2], qui est enrichi tous les mois de nouveaux articles, comportait déjà en 1997 environ 500 000 articles, soit plus de treize millions de mots.

1.2. Les conditions de la récolte

Les terrains d'enquêtes sont parfois difficiles d'accès (plusieurs journées de marche) et sous-équipés (en électricité par exemple). Les linguistes doivent alors se munir en batteries ou piles afin d'être le plus autonome possible une fois sur place. Le choix du matériel à emporter se fait alors en fonction du poids, de l'encombrement et de la consommation d'énergie et parfois au détriment de la qualité. Actuellement, de nombreux linguistes choisissent des enregistreurs de Minidisc qui sont souvent de moins bonne qualité que les matériels qu'ils utilisaient par le passé (Nagra, Uher, DAT, etc.), mais qui sont beaucoup moins encombrants, moins fragiles et moins lourds.

Les enregistrements étant faits « sur le terrain », en milieu naturel (par opposition aux travaux faits en laboratoire), ils sont donc bruités par des éléments extérieurs de l'environnement. Cet aspect des conditions d'enregistrement rapproche ainsi les données de ce que l'on pourrait appeler la parole spontanée par opposition au côté artificiel des données que l'on obtient dans des conditions de laboratoire.

1.2. Exemple : les corpus du LACITO

Le LACITO[3] (Laboratoire de langues et civilisations à traditions orales) est un laboratoire du CNRS dont les chercheurs (linguistes, anthropologues et ethno-musicologues) travaillent depuis plus d'une trentaine d'années à la description de langues pour la plupart sans écritures. De leurs enquêtes de terrain, ils ont ramené des enregistrements audio, plus rarement vidéo, ainsi que des transcriptions, des traductions… faites sur place avec l'aide de locuteurs. Ces enregistrements et analyses constituent les matériaux de base qui vont servir aux chercheurs pour poursuivre leurs recherches une fois revenus de leur mission.

Les analyses de ces enregistrements sont composées principalement de transcriptions phonético-phonologiques. Celles-ci sont notées, la plupart du temps,

---

[1] TLFi sur le site du laboratoire ATILF http://www.atilf.fr
[2] Corpus *Le monde* distribué par l'ELDA http://www.elda.fr
[3] Site web du laboratoire CNRS/LACITO : http://lacito.vjf.cnrs.fr

en utilisant les conventions typographiques de l'alphabet de l'Association de phonétique internationale[4] (API). Nous rencontrons aussi des systèmes de transcription basés sur d'autres alphabets, comme le cyrillique pour les langues des pays de l'Est. Enfin, on trouve aussi des conventions de codage propres à certains linguistes. En outre, ces analyses comportent, pour une bonne part, des traductions et des gloses. L'ensemble de ces annotations est en général présenté sous forme « interlinéaire ». Dans ce style de présentation, classique en linguistique, quelques conventions de notation sont utilisées conjointement à l'API, tels les espaces pour séparer les mots, les tirets et points pour marquer les jointures de morphèmes. Chaque mot dans une structure interlinéaire est noté dans un paradigme pouvant comporter plusieurs lignes ; chaque ligne d'un paradigme représente une vision, une analyse du mot en question. En premier figure souvent la transcription, puis suivent la glose, la partie du discours, d'autres transcriptions, translitérations, etc. On peut ainsi coder à la fois des transcriptions phonétiques, d'autres plus phonologiques tenant compte ou non des phénomènes de sandhi, des gloses dans plusieurs langues cibles, etc. Si dans une ligne particulière, un caractère est utilisé pour marquer le découpage du mot en morphèmes, ce même caractère sera repris dans les lignes suivantes pour rappeler ce découpage. Par exemple, dans la figure 1, le tiret utilisé dans la ligne 3 indique le découpage de la transcription en morphèmes. Ce tiret est repris systématiquement dans la ligne 4 pour séparer les gloses de chaque morphème.

[4] Cette association a été créée en 1885 par les linguistes Paul Passy et Daniel Jones. Son organe de diffusion « le maître phonétique » a permis de répandre et de standardiser un alphabet destiné à offrir aux linguistes un instrument de notation « phonétique » des langues orales. Site web de l'association : http://www.arts.gla.ac.uk/ipa/ipa.html

Hale halele, vhuka mutru zama izo woho, mabakoko yatru ya shimurima

hale	**halele**	**vhuka**	**mutru**
autrefois	autrefois	il y avait	quelqu'un
Ø-hale	Ø-hale-le	vhu-Ø-k-a	mu-tru
pn9-autrefois	pn9-autrefois-éloi	pv16-pas-être-sfx	pn1-personne

zama	**izo**	**woho**	**mabakoko**
temps anciens	ceux-là	dans le lointain	grands-pères
Ø-zama	i-zo	wo-ho	ma-bakoko
pn10-temps anciens	pd10-dréf	pd17-dréf, loc.temp	pn6-grands-pères

yatru	**ya**	**shimurima**
nos	d'	Afrique
i-a-tru	i-a	shi-murima
pd6-dét-pos1p1	pd6-dét	pn7-Afrique

Il y a très longtemps, il y avait une personne dans ces temps anciens, ceux de nos grands-pères d'Afrique.

Figure 1. *Exemple de texte interlinéaire (langue maore de Mayotte)*

À coté des textes glosés ou interlinéaires, les corpus du LACITO se composent de listes de mots (élicitations), de dictionnaires, de chants, etc. L'ensemble de ces données représente en tout cas des centaines d'heures d'enregistrement.

2. La gestion des corpus oraux en linguistique de terrain

Pour la gestion des corpus existants, nous pouvons distinguer trois types de préoccupations : celles liées à la conservation des supports et à la pérennisation des données, celles liées à la normalisation des données, enfin celles liées à leur diffusion. Ces trois problématiques étant interdépendantes, nous les présenterons ci-après en les illustrant d'exemples tirés du programme « Archivage[5] ».

2.1. Le programme « Archivage » du LACITO

Une partie des informations recueillies par le LACITO depuis sa création a été publiée, notamment des descriptions de langues, mais la majorité des matériaux d'enquête n'a jamais fait l'objet d'une publication ni même d'une gestion

[5] Site web qui décrit le programme « Archivage » du LACITO et qui propose un accès aux archives publiques de ce programme : http://lacito.vjf.cnrs.fr/archivage

centralisée ou organisée collectivement. Ces matériaux se dégradent au fil du temps, alors que leur valeur patrimoniale tend à augmenter. Un autre problème à résoudre est le manque ou la difficulté d'accès, pour les linguistes, à leur propres données, et encore plus à celles des autres. Pour toutes ces raisons, un vaste projet d'archivage a été entrepris avec comme objectifs principaux : sauvegarde, normalisation et diffusion.

2.2. La conservation et la pérennisation des données

Selon les époques, les enregistrements ont été stockés sur divers supports : bobines de bandes magnétiques, Compact Cassettes, cassettes DAT et plus récemment Minidisc. Tous ces supports et principalement ceux qui sont magnétiques et analogiques se détériorent avec le temps. Ce vieillissement des supports risque d'altérer les données stockées, ou bien de compromettre leur intégrité. Pour éviter de trop manipuler les originaux, des copies sont faites mais, hormis pour les nouveaux supports digitaux, la copie résultante est de moins bonne qualité que l'original. Le maniement de ces supports nécessite parfois du matériel que les chercheurs ne possèdent plus ou difficile d'accès tels que certains lecteurs de bandes. Enfin, l'écoute d'un moment particulier de l'enregistrement nécessite, mis à part pour les nouveaux supports optiques ou magnéto-optiques, un déroulement séquentiel du support, voire pour certains, une indexation *ad hoc* (par exemple un compteur de tours).

2.2.1. La numérisation

Une bonne réponse à cette inquiétude est apportée par la numérisation. Il s'agit d'une conversion d'un mode de représentation analogique (qui prévalait jusqu'à récemment) vers un mode de représentation digital (apporté principalement par l'informatique). Cette conversion analogique/digitale est de nos jours une opération courante demandant un équipement très répandu et facile d'emploi.

Au moment de la numérisation, les choix de la fréquence d'échantillonnage, de la taille de l'échantillon, du nombre de canaux et du codage déterminent la qualité du document numérique résultant. Le théorème de Nyquist, par exemple, nous dit que la fréquence la plus élevée bien représentée ne peut être supérieure à la moitié de la valeur de la fréquence d'échantillonnage. Pour le programme archivage, nous avons choisi de nous inspirer de la norme CD-audio (44 100 Hz, 16 bits). Cela peut paraître un luxe inutile pour certain documents, la qualité audio de l'original étant médiocre. Mais étant donné les capacités actuelles de stockage, il paraît raisonnable de choisir ces caractéristiques dans tous les cas. La copie se faisant à l'identique et à l'infini, il n'y a donc plus de notion d'original et de copie. La pérennisation des données n'est alors plus assurée que par une bonne gestion des supports de stockage, qui eux restent soumis à un processus de vieillissement.

2.2.2. Le catalogage

La mauvaise conservation des supports n'est pas la seule cause de disparition des données. Il faut aussi que ces données soient identifiées, décrites et cataloguées. La gestion des données d'une enquête de terrain (enregistrements et annotations), dans notre laboratoire, est du ressort des linguistes qui ont fait cette enquête. Les bandes, cassettes, etc. sont conservées dans leur bureau, ou bien stockées dans une pièce à température et hygrométrie contrôlées conçue à cet effet. L'étiquetage des supports ainsi que leur inventaire est aussi de la responsabilité du linguiste. La plupart du temps, ces derniers connaissent suffisamment leurs données pour ne pas ressentir le besoin d'une description complète et explicite des contenus. Ils se contentent souvent de mentionner la date, la langue, et quelques autres informations mnémotechniques qui leur permettront de se souvenir du reste.

Les annotations de ces enregistrements sont en partie pratiquées sur le terrain avec l'aide des informateurs, en partie construites au retour du terrain et enrichies au fur et à mesure des nouvelles connaissances et hypothèses. Elles étaient jusqu'à récemment, consignées sous forme manuscrite dans des cahiers. Aujourd'hui, elles sont plus fréquemment saisies sur ordinateurs portables. Certains établissent des catalogues à l'aide de simples fichiers texte, de bases de données, de tableurs, de fiches bristol. D'autres sont eux-mêmes leur propre catalogue. La plus grosse partie de ces annotations ne fait jamais l'objet d'une publication. Cette absence de description systématique des ressources que l'on engrange pose à terme le problème de la transmission des connaissances. Lorsqu'un linguiste part à la retraite, change de sujet, meurt, la documentation qu'il a accumulée pendant des années peut devenir inexploitable. D'une part, parce qu'il se peut que personne ne sache qu'il a laissé des documents, d'autre part parce que sachant qu'il existe des documents il se peut que personne ne sache de quoi il s'agisse. Un autre facteur aggravant est que les langues évoluent, qu'un état de langue ne se reproduit pas, enfin que parfois les langues meurent faute de locuteurs.

Pour illustrer ces problèmes de transmission, nous allons décrire brièvement une situation que nous avons rencontrée dans notre laboratoire. Catherine Paris, linguiste spécialiste des langues du Caucase, est décédée il y a quelques années, laissant derrière elle, outre d'abondantes publications, des travaux inachevés ainsi que des caisses de bandes magnétiques et de cahiers de terrain plus ou moins bien étiquetées. Parmi ces matériaux, des enregistrements étaient sans transcriptions et des transcriptions sans enregistrements. Une personne (Mme Dabjen-Bailly) a été affectée, au LACITO, sur ce fond documentaire afin de terminer les publications en cours, de mettre de l'ordre dans les matériaux et de numériser les enregistrements. Parallèlement à ce travail, l'Institut de linguistique et phonétique générales et appliquées[6] (ILPGA) s'est vu confier la donation de René Gsell, lequel avait travaillé dans les années 1960 avec Catherine Paris, Georges Dumézil, Christine Leroy, Georges Charachidzé et Bernard Gautheron sur une langue du Caucase,

[6] Site web de l'ILPGA : http://www.cavi.univ-paris3.fr/ilpga/ilpga/

aujourd'hui disparue : l'oubykh. Un certain nombre de ces enregistrements était déjà référencé dans d'autres bibliothèques, mais un de ceux-ci restait inédit. C'est grâce au travail de fouilles quasi archéologiques d'Alexis Michaud[7] (Michaud, 2002), que nous avons pu rassembler au LACITO enregistrements et transcriptions. Ces documents étaient jusqu'alors physiquement éloignés et *de facto* « orphelins ». Aujourd'hui tout le monde peut consulter librement un des documents de ce fonds sur le Web du LACITO.

2.3. La normalisation des données

Conserver des objets est une bonne chose, mais il faut aussi que ces objets puissent être regardés et compris par tout le monde de la même manière. Il faut donc pour cela qu'ils soient intrinsèquement explicites ou alors qu'ils soient décrits avec un métalangage partagé par tous. Une norme peut être considérée comme une sorte de métalangage dont la description est l'objet d'une concertation à un niveau international qui donne lieu à une description faisant référence. Dire que l'on se sert d'une norme particulière permet de ne pas avoir besoin de redéfinir les concepts que l'on manipule. Une norme est une abstraction, indépendante des implémentations qui peuvent en être faites. C'est cette indépendance qui rend possible le partage et les échanges de données dans des environnements différents. Une autre conséquence de l'utilisation de normes est qu'elles permettent la pérennisation car la norme étant publique, il est toujours possible (au plan technique et juridique) d'interpréter correctement (au sens de la norme) des données normalisées dès lors que l'on sait de quelle norme il s'agit.

2.3.1. Les données audio

La normalisation des données audio pose relativement peu de problèmes et les choix en ce domaine sont assez bien acceptés et partagés. Hormis les caractéristiques propres à la numérisation, reste à choisir un codage des données et un format de fichier. Le codage de type PCM[8] (ou étendu), c'est-à-dire sans compression, semble le plus adapté à l'archivage à long terme, même si ce n'est pas le codage que l'on veut utiliser aujourd'hui pour une diffusion. Les formats de fichiers sont eux assez nombreux. L'important est d'éviter les formats propriétaires et peu répandus. Nous avons opté, en ce qui concerne les archives du LACITO, pour le format RIFF[9] (fichiers wav) en raison de sa popularité, de sa simplicité et de la qualité de la documentation à son sujet.

---

[7] http://www.loria.fr/projets/JEP/JEP2002/papiers/26.pdf
[8] Pulse Code Modulation
[9] Rich Interchange File Format

2.3.2. Le codage des caractères

La normalisation des annotations est beaucoup plus délicate dans la mesure où les choix ne font pas toujours l'unanimité. La plus grosse partie des annotations est de nature textuelle (traductions, transcriptions, gloses, etc.). Celle-ci requiert l'utilisation de caractères et de scripts d'écriture qui dépendent de traditions elles-mêmes liées à des cultures différentes. Pour les traductions, le codage peut utiliser les orthographes nationales des langues concernées. Selon les langues, ce codage sera : alphabétique (latin, arabe, cyrillique, devanagari, etc.) avec des sens d'écriture différents, syllabique (katakana) ou idéographique (chinois). Pour les transcriptions phonétiques, l'API est souvent utilisé.

Nous avons en outre choisi d'utiliser Unicode (Unicode Consortium 2000), standard synchronisée sur la norme ISO/IEC 10646, pour le codage des caractères. Ce choix est aujourd'hui le seul possible, il est bien accepté par les utilisateurs comme par les constructeurs et les fabriquants de logiciels, mais surtout il est le seul code à visée universaliste à être normalisé. Dans sa version 4.0 de 2003, il comporte déjà quelque 96 382 caractères. Le premier plan (BMP) dit « multilingue » comporte la plupart des écritures utilisées actuellement dans les langues modernes ; d'autres plans comportent des écritures historiques, la notation musicale, etc. Enfin, il reste encore beaucoup de places non affectées dans ce code, afin d'y accueillir à l'avenir, encore plus d'écritures.

Unicode se veut un code universel définissant les caractères de toutes les écritures du monde. Seules les écritures sont définies dans ce code et non pas les langues. Ainsi, les caractères du latin ne sont codés qu'une seule fois, alors qu'ils sont utilisés par de nombreuses langues (français, anglais, italien, espagnol, etc.). Unicode fixe pour chaque position dans le code une définition ainsi qu'un certain nombre de propriétés telles que les équivalences entre caractères précomposés et les suites de caractères simples. La représentation physique d'un caractère, que l'on appelle œil ou glyphe n'est pas une préoccupation d'Unicode. Il ne s'agit pas d'un problème de codage mais d'un problème qui doit être pris en charge par les outils de rendu.

Un système de rendu peut, par exemple, représenter un même caractère par différents glyphes en utilisant des polices de caractères aux styles différents. Dans certaines langues, comme en arabe, les caractères peuvent avoir plusieurs glyphes suivant la place qu'ils occupent dans le texte.

Position	Isolée	Finale	Initiale	Médiane
Glyphe	ڄ	ڄ	ڄ	ڄ

Figure 2. *Glyphes du caractère ARABIC LETTER DYA'(position U+0684)*

A l'inverse, un même glyphe peut parfois être partagé par différents caractères tel le glyphe " ' " qui est utilisé à la fois pour la quote simple droite ainsi que pour l'apostrophe. Enfin quelques caractères tel que la parenthèse ouvrante sera représenté tantôt par le glyphe " (" comme en français, tantôt par le glyphe ") " comme en arabe qui s'écrit de droite à gauche. Dans tous ces cas, le choix du bon glyphe est du ressort du système de rendu (polices, éditeurs).

Unicode prévoit aussi un mécanisme d'empilement des diacritiques qui doit être implémenté dans les systèmes de rendu. Les systèmes classiques qui ne prévoient par exemple qu'une seule position au-dessus du caractère principal écrivent tout les diacritiques les uns par-dessus les autres, les rendant à la fois indistincts les uns des autres et non reconnaissables.

Figure 3. *Rendu correct d'un empilement de diacritiques au-dessus et en-dessous d'un caractère [extrait de la figure2.7 de l'ouvrage « The Unicode Standard, Version 3.0 » (Unicode Consortium 2000)]*

A ce jour et à notre connaissance, seuls deux outils permettent de coder et d'exploiter ces comportements particuliers des caractères : le format de police OpenType[10] et Uniscribe, ainsi que Graphite[11] et son système de « smart font ».

2.3.2. Le codage de l'annotation linguistique

Un caractère ne nous renseigne pas sur son contexte d'utilisation. Il peut faire partie d'une transcription, d'une traduction, d'un mot, d'un morphème, du nom du locuteur, de son âge. Selon la situation, le même caractère "a" pourra être interprété différemment comme la lettre latine minuscule "A" du français, de l'italien, de l'espagnol, etc., comme la voyelle orale d'avant la plus ouverte d'après l'API, ou

[10] Format de fichier multi-plateforme pour le codage de polices de caractères développé par les Société Adobe et Microsoft.
[11] Graphite est un projet du *Summer Institut of Linguistics* pour le développement d'outils de rendu pour des systèmes d'écriture complexes.

encore comme une forme conjuguée du verbe "avoir" en français. Nous avons donc besoin d'un langage de haut niveau pour expliciter ce contexte. Depuis longtemps, des formalismes de structuration des données existent. On peut distinguer deux grandes familles que sont les bases de données et les langages de balisage de texte que l'on peut aussi opposer, d'après leurs origines, en ceux issus des traditions de calcul et ceux issus des traditions textuelles. Sans faire un historique exhaustif[12] de ces formalismes, nous mentionnerons tout de même ceux qui ont été ou qui sont encore utilisés dans notre laboratoire : Lexware, Shoebox, divers logiciels de traitement de texte et XML.

Lexware

Lexware est une boite à outils logiciels créée par Robert Hsu (Hsu, 1989), qui permet de traiter des données structurées dans un format *ad hoc*. Les traitements prévus sont de nature linguistique puisqu'à l'origine cet outil était destiné à gérer des lexiques. Lexware dispose de fonctions d'extraction, de tri, de contrôle, etc. Il est possible d'augmenter le nombre de fonctions en les écrivant soi-même avec le langage Spitbol qui, entre autres avantages, comporte des fonctions d'expressions régulières. Le format de fichier utilisé est du « plain text » dont certaines conventions typographiques ont été re-spécifiées pour implémenter de manière formelle les notions d'enregistrement et de champ.

Un enregistrement commence par un caractère "." en début de ligne et finit au début du prochain enregistrement ou à la fin du fichier. Un enregistrement peut comporter des sous-enregistrements, ces derniers commençant par deux caractères "." en début de ligne et finissant au début du prochain sous-enregistrement de même niveau ou à la fin de l'enregistrement de niveau supérieur. De manière générale, le niveau d'imbrication d'un enregistrement est donné par le nombre de caractères "." en début de ligne.

Un champ est un ensemble comportant un label et une valeur. Il commence toujours en début de ligne. Le label est toujours le premier mot du champ (caractère "." excepté). Il est séparé de la valeur du champ par des caractères d'espacement (espaces blancs ou tabulations). Les noms des labels sont libres et sont définis par l'utilisateur. Les valeurs des champs peuvent être elles-mêmes structurées en fonction d'un certain nombre de conventions typographiques.

Shoebox

Shoebox[13] (Buseman et Buseman 1998) est un outil qui permet de saisir des lexiques ou bien des textes interlinéaires, tout en développant parallèlement un dictionnaire. Le formalisme défini par la Summer Institut of Linguistics[14] (SIL) et

[12] Pour un inventaire plus complet voire par exemple :
(http://www.ldc.upenn.edu/exploration)
[13] http://www.sil.org/computing/catalog/shoebox.html
[14] Site web du *Summer Institut of Linguistics* : http://www.sil.org

adopté par cet outil est connu sous le nom de « SIL standard format ». Ce formalisme très proche de celui de Lexware, ajoute quelques conventions supplémentaires comme l'utilisation des espaces blancs pour la notation interlinéaire. Comme avantages ce logiciel nous offre une interface utilisateur assez conviviale, ainsi que des traitements spécifiquement linguistiques. Ce système a été assez populaire chez les linguistes et l'on trouve fréquemment des corpus encodés de cette manière.

```
\ref 1
\t   nakpu  nonotso  siŋ    pa    la?natshe-m          are
\m   nakpu  nonotso  siŋ    pa    la?+natshe-m         are
\g   two    sisters  wood   make  go+REFL:3du-ASS      REP
\f   They say that two sisters went to fetch wood
```

Figure 4. *Exemple d'enregistrement d'un texte interlinéaire (langue hayu du Népal)*

Shoebox et Lexware souffrent tous deux des mêmes limitations, principalement dues à l'absence de prise en charge d'un codage de caractères autre que celui des codes propriétaires de MS-Windows et de MacOS. De manière plus généralement, l'outil et le formalisme sous-jacent entretiennent, dans les deux cas, des liens trop étroits.

Les logiciels de traitement de texte

On trouve de nos jours de nombreux traitements de texte, même si beaucoup d'utilisateurs sont persuadés qu'il n'existe que MS-Word. Ces outils permettent non seulement la saisie de contenus textuels, mais aussi leur structuration typographique en vue d'une mise en page. Les formalismes sous-jacents sont très divers et souvent propriétaires. Il existe aussi quelques alternatives publiques comme RTF ou TeX. De nombreux linguistes sont utilisateurs de ces outils, s'en servent pour écrire des articles, mais surtout pour noter leurs corpus, malgré l'absence presque totale d'outil de recherche. La facilité d'utilisation de ces outils a largement prévalu sur leurs fonctionnalités.

Les langages de balisage de texte : XML

XML[15] (Bray et Co. 1998) est un langage de balisage de texte. Il s'agit d'une recommandation du World Wide Web Consortium (W3C), ce n'est donc pas une norme au sens strict mais un standard. En revanche, son ancêtre SGML (Standard

[15] Extensible Markup Language

Generalized Markup Language) a été normalisé en 1986 sous le nom de ISO-8879. XML est en fait un avatar moderne et simplifié de SGML.

XML est accompagné de tout un ensemble de technologies elles aussi standardisées au sein du W3C. XSL pour les feuilles de styles, DOM comme interface de programmation, Xlink et Xpointer pour les liens, Xquery comme langage de requête, mais aussi des applications à des domaines d'activité comme SMIL pour l'animation multimédia, SVG pour le dessin vectoriel, etc. Toutes ces recommandations sont consultables sur le site du consortium W3C[16].

Un certain nombre d'outils a aussi vu le jour tels que les parsers, les éditeurs, les processeurs de styles, etc. Beaucoup de domaines d'activité ont défini des structures de données en XML, à l'aide de DTD ou de XML-Schema, pour représenter par exemple des expressions mathématiques, des chaînes de molécules, des notations musicales, des métadonnées, etc.

2.3.3. La normalisation des annotations dans le cadre des archives du LACITO

Le choix de XML comme formalisme de représentation pour l'ensemble des annotations des documents d'archives du LACITO, a été guidé par tout un ensemble d'aspects : le codage des caractères est Unicode ; ce langage s'intègre très facilement dans une architecture Web ; il existe un large panel d'outils l'implémentant ; il existe de nombreux utilisateurs ; l'accord sur son adoption est quasi unanime toutes disciplines confondues.

Le laboratoire ayant par le passé utilisé tous les formats précédemment cités, nous avons été conduit à développer un certain nombre d'outils de conversion ou de rétroconversion pour tous les normaliser en XML. En ce qui concerne la normalisation des données manuscrites, celle-ci doit avant tout passer par une étape de numérisation. Il est possible en théorie de scanner ces documents puis d'appliquer sur le résultat des outils de reconnaissance d'écriture. En pratique, l'aide ainsi fournie n'est pas plus avantageuse que de ressaisir entièrement le document, surtout quand il s'agit de faibles volumes de données comme ceux que l'on traite en général en linguistique de terrain.

Une fois choisi le formalisme XML, reste à décrire la structure logique de nos annotations, puis à alimenter cette structure de contenus. Avec XML, cette structure peut s'exprimer sous forme de DTD ou de XML-Schema. L'un comme l'autre expriment un certain nombre de contraintes sur les composants structurels (balises, attributs), entre autres les noms et les types des composants autorisés. D'autres contraintes portent sur les contenus des composants (nombre et ordre d'apparition, valeurs contrôlées par des listes, statut optionnel ou obligatoire, etc.). Cette syntaxe formelle doit refléter l'analyse que l'on souhaite faire des données. La normalisation en la matière est assez difficile dans la mesure où l'on se rapproche des théories utilisées dans l'analyse. Or, il n'y a pratiquement pas d'accord dans la

[16] http://www.w3.org/TR/

communauté des linguistes sur les objets manipulés. Des tentatives ont été faites, notamment la Text Encoding Initiative[17] (TEI) ou le Corpus Encoding Standard[18] (CES), et de multiples ontologies ont été proposées. Actuellement, au sein de l'ISO, un groupe de travail TC37/SC4[19] travaille sur la question. Ne trouvant rien d'intellectuellement satisfaisant, nous avons choisi de créer une DTD spécifique extrêmement simple et ressemblant à la TEI pour faciliter l'interopérabilité entre les deux formats.

Dans la DTD du programme Archivage, il existe cinq niveaux hiérarchiques qui sont définis par les balises de noms – ARCHIVES, TEXT, S, W et M – et qui correspondent respectivement à : corpus, texte, phrase, mot et morphème. Chaque niveau peut contenir un ou plusieurs items du niveau immédiatement inférieur. Ainsi une phrase peut être composée de mots, et ces mots de morphèmes, mais il n'est pas possible d'avoir un morphème d'une phrase qui ne soit pas encapsulé dans un mot. Chacun de ces niveaux peut comporter des transcriptions (FORM), des traductions (TRANSL) et un ancrage temporel (AUDIO). Les traductions doivent préciser la langue cible. Aux niveaux mots et morphèmes, la traduction correspond à ce qu'on a l'habitude d'appeler glose. Les transcriptions doivent préciser leur type (phonétique, phonologique, orthographique, translitéré…), et éventuellement le nom du transcripteur et la date de la transcription. Les phrases peuvent contenir des indications scénographiques comme le nom du locuteur, dans le cas de dialogues. Les mots et morphèmes comportent si besoin des indications typologiques comme la partie du discours, la classe, etc., ces dernières indications restant libres car chaque linguiste tient à utiliser son propre système. Enfin il est possible d'insérer des notes un peu partout.

L'ancrage temporel se fait par l'insertion au niveau que l'on souhaite d'une balise AUDIO comportant deux attributs 'start' et 'end', qui indiquent les valeurs temporelles de début et de fin (en milliseconde par rapport au début de l'enregistrement). Cet ancrage permet d'exprimer des événements temporels de type :

1. enchaînements : les phrases d'un récit se suivent l'une après l'autre ;
2. enchâssements : les mots d'une phrase sont compris entre le début et la fin de celle-ci ;
3. chevauchements : plusieurs locuteurs dans un dialogue peuvent parler en même temps.

L'annotation de l'ancrage temporel que nous proposons ici est inspirée de celle des recommandations de la *Text Encoding Initiative*. Elle repose sur la structure du langage XML en utilisant la hiérarchique des éléments pour représenter l'inclusion

[17] http://www.tei-c.org
[18] http://www.cs.vassar.edu/CES/
[19] http://www.tc37sc4.org/

temporelle et l'ordre des éléments pour représenter leur succession dans le temps. Tous les éléments des différents niveaux hiérarchiques, qu'il s'agisse de textes, de phrases, de mots ou de morphèmes peuvent être ancrés dans le temps, mais ne le sont pas obligatoirement ni systématiquement. Ceux qui ne le sont pas, sont alors situés entre les bornes de début et de fin de l'élément de niveau immédiatement supérieur, mécanisme récursivement applicable jusqu'au niveau le plus haut. Par exemple, un mot est toujours situé entre le début et la fin de la phrase à laquelle il appartient. Un élément non ancré se situe par ailleurs entre la fin de l'élément de même niveau immédiatement précédent et le début de l'élément de même niveau immédiatement suivant dans l'ordre de leur écriture dans le document. Contrairement à la hiérarchie des niveaux, les éléments d'un même niveau peuvent briser leur linéarité en le spécifiant dans des ancres temporelles. Par exemple, dans un récit, les phrases se suivent, la fin de l'une correspondant en général au début de la suivante, sauf cas de pauses silencieuses ou d'interruptions repérables par les plages de temps non contigus des deux ancres en question. Dans un dialogue, en revanche, il est fréquent de rencontrer des moments où plusieurs locuteurs s'expriment en même temps. Les plages de temps définies par les ancres des phrases des différents locuteurs présentent alors des périodes de recouvrement. Enfin un événement ponctuel est identifiable par des valeurs identiques des attributs début et de fin de son ancre.

Il existe bien sur, d'autres représentations de la structure temporelle. Par exemple SMIL qui est un langage XML dédié au codage d'animations multimédias définit deux grands types de structures temporelles : les séquences et les parallèles. Il prévoit aussi, comme dans la *Text Encoding Initiative*, la possibilité de fixer des bornes de début, de fin ou de durée. Sur la trame temporelle composée de séquences et de parallèles, peuvent se fixer, sur les zones géographiques de l'écran que l'on définit, des ressources de différents types (du son, de la vidéo, des images, du texte, etc.). Un autre modèle pour la représentation des annotations linguistiques de signal de parole a été proposé par Bird et Liberman 2001. Ce modèle abstrait, appelé « Annotation Graph », se veut indépendant des formats de fichiers, des codages et des interfaces graphiques. Il distingue les annotations ancrées dans le temps et celles qui sont flottantes. Dans l'implémentation XML proposée, la structure temporelle est décrite dans un premier bloc au début du document, puis les annotations linguistiques correspondants aux différents nœuds de cette structure temporelle sont ensuite décrits. Les relations entre les deux sont notées par un système d'identifiants et de pointeurs sur ces identifiants. La différence entre notre modèle et l'implémentation XML de celui-ci nous semble être principalement notationnelle, si l'on excepte les éventuels outils qui permettent de travailler avec cette structure (AGTK[20]). De manière générale, XML tout comme d'autres structures informatiques d'ordre génériques comme les bases de données n'implémentent pas

[20] Annotation Graph Tool Kit (AGTK) est un ensemble de composants logiciels de manipulation de données sous forme de graphes d'annotation, dédiées à la création d'outils d'annotation linguistique alignés sur du signal de parole.

directement de structure temporelle. Avec XML il est possible pour définir une telle structure de jouer sur son aspect hiérarchique et ordonné, ainsi que sur la sémantique des balises et des attributs que l'on définit. Enfin pour briser l'aspect strictement arborescent de XML il existe des systèmes de relations ID/IDREF et ceux de Xlink qui sont utilisés par exemple dans les recommandations CES.

Nous avons choisi de limiter les annotations au codage du niveau morpho-phonologique car l'analyse sur le terrain s'arrête généralement à celui-ci. Cette analyse peut être considérée comme « brute » ou « de bas niveau ». D'autres analyses de ces documents « bruts » peuvent être faites dans des documents externes au moyen de liens Xlink pointant sur des parties de documents. Cette solution nous permet de maintenir une DTD simple et centrée sur les matériaux de terrain, tout en laissant le champ libre pour d'autres analyses, voire éventuellement pour des analyses contradictoires.

2.3.4. Les métadonnées

Les archives du LACITO comportent des ressources linguistiques que sont les enregistrements et leurs annotations. Toutes ces ressources sont référencées dans un catalogue qui les décrit le plus finement possible à l'aide de descripteurs. L'ensemble de ces descripteurs est appelé métadonnées. Ces métadonnées sont elles aussi codées en XML. N'étant ni bibliothécaire ni documentaliste, nous étions prêts à accepter une norme en la matière à partir du moment où elle permettait de coder tout ce qu'on voulait exprimer et qu'elle était acceptée par une large communauté. Nous avons choisi les recommandations de l'Open Language Archives Community[21] (OLAC) pour le codage de nos métadonnées. Il existe d'autres propositions de codage telles que IMDI[22] ou MARC[23] qui auraient également pu répondre à nos exigences, mais qui sont plus compliquées à mettre en œuvre. OLAC est une organisation regroupant tous les détenteurs d'archives linguistiques qui le souhaitent (AILLA[24], LDC[25], etc.) pour décrire et diffuser leurs ressources de manière homogène. Le codage proposé est en fait une adaptation du Dublin Core Metadata Initiative[26] (DCMI) et le protocole d'échange est celui définit par l'Open Archive Initiative[27] (OAI). Nous allons brièvement présenter ces deux groupes.

La base de Dublin-Core se compose d'une quinzaine d'étiquettes (*title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage* et *rights*) dédiées à la description des documents

[21] Site web d'OLAC http://www.language-archives.org/
[22] EAGLES/ISLE Metadata Initiative : http://www.mpi.nl/IMDI/
[23] Description du format MARC sur le site de *Library of Congress* : http://lcweb.loc.gov/marc/
[24] Archive of Indigenous Languages of Latin America http://www.ailla.org
[25] Linguistic Data Consortium http://www.ldc.upenn.edu
[26] Site web du DMCI : http://dublincore.org/
[27] Site web de l'OAI : http://www.openarchives.org/

électroniques (en fait tout ce qui peut être identifié par un URI). Ce jeu de descripteurs a été normalisé en 2003 auprès de l'ISO sous le nom « Information and documentation - The Dublin Core metadata element set » (ISO-15836). Chaque descripteur est défini au sein de cette norme dans un sens assez général et est utilisable dans n'importe quel domaine d'application.

Pour distinguer des sous catégories dans les champs, assez larges, occupés par les quinze descripteurs de base, DCMI a défini de nouvelles étiquettes. Ces dernières sont actuellement au nombre de quarante trois. Par exemple pour l'élément *relation* treize nouveaux éléments ont été définis distinguant différents type de relation possibles (*isVersionOf, hasVersion, isReplacedBy, replaces, isRequiredBy, requires, isPartOf, hasPart, isReferencedBy, references, isFormatOf, hasFormat* et *conformsTo*). Une autre amélioration apportée est la possibilité d'utiliser des schémas de codage. Par exemple le codage préconisé pour indiquer des dates quelles que soient leur type (*created, valid, available, issued, modified, dateAccepted, dateCopyrighted* ou *dateSubmitted*), est celui du consortium World Wide Web , lui-même basé sur la norme ISO-8601.

Par dessus ces conventions, OLAC est venu ajouter un mécanisme permettant de préciser le sens d'un descripteur, de contrôler la valeur de son contenu et d'ajouter d'autres informations supplémentaires. En terme d'XML, ce mécanisme repose sur les XML-Schemas. Plus concrètement OLAC a défini cinq extensions en précisant pour chacune d'elles sa signification dans le cadre du domaine des ressources linguistiques :

- Les types de discours : dialogue, drame, …, narration, chant,…
- Les domaines linguistiques : lexicographie, phonologie, morphologie…
- Les types linguistiques : lexique, texte primaire, description de langue
- Les rôles : narrateur, signeur, traducteur, transcripteur…
- L'identification langues : utilisation du code d'Ethnologue[28] dans un encodage conforme à la norme ISO-639

La diffusion des métadonnées se fait en suivant les règles de l'OAI. Cette organisation définit entre autre un protocole relativement simple d'échange de métadonnées qui comprend un petit nombre de requêtes qu'il est possible d'adresser à un détenteur d'archives. Par exemple on peut demander à un détenteur d'archives son identification, la liste de ses identifiants de ressources, la liste des encodages qu'il utilise pour ses metadonnées, etc. Ce protocole fixe aussi la syntaxe des réponses que peut émettre un fournisseur d'archives. Un certain nombre de règles de politesses doivent être implémentées par le fournisseur comme l'envoi de codes

[28] Ethnologue est un catalogue identifiant de plus de 6000 langues et donnant accès à des informations à leur propos (localisation, nombre de locuteurs, apparentements génétiques, bibliographie, etc.) http://www.ethnologue.org/

particuliers pour signaler les erreurs de syntaxe des requêtes. A l'exception des codes d'erreurs, toutes les réponses émises sont encodées en XML.

2.3. La diffusion des données

Une fois le travail de pérennisation et de normalisation des données effectué, il devient possible de partager ces ressources avec une communauté plus large. Nous avons choisi le Web comme support de publication. En « dématérialisant » le support, le Web permet une publication pratiquement sans coût financier. Les données étant stockées sur des serveurs et non chez l'utilisateur, elles peuvent être régulièrement corrigées. Il est aussi possible de fournir des outils de consultation dynamiques et, comme pour les données, de maintenir ces outils plus facilement. Le Web s'adresse en outre au plus grand nombre de manière « démocratique ». C'est actuellement la plate-forme multimédia la plus pertinente si l'on ne veut pas réécrire soi-même tous les outils.

L'accès aux données de l'archive se fait à travers la consultation d'un catalogue. Ce catalogue contient toutes les métadonnées concernant les ressources disponibles. L'architecture adoptée pour la diffusion et l'interrogation de ce catalogue est celle définie par l'OAI que nous avons présenté plus haut. Le protocole d'échange définit par l'OAI est notamment utilisé par des moteurs de recherche et d'indexation. Par exemple, certaines bibliothèques consultent régulièrement notre catalogue pour alimenter leur base de données en ligne avec les nouvelles entrées.

Le catalogue recense l'ensemble des ressources (fichier d'enregistrement audio, fichier XML d'annotations) que l'on s'engage à maintenir, les décrit et en précise les conditions d'accès.

3. L'organisation du système d'archivage au LACITO

L'archivage des données s'articule autour de quatre grands pôles que sont l'auteur, l'éditeur, le conservateur et l'utilisateur final. Cette séparation des rôles reflète d'une part l'organisation institutionnelle, d'autre part la répartition des tâches entre les acteurs. Elle est aussi le reflet de l'organisation physique et logique des traitements informatiques.

Pour le moment, la seule institution en jeu est le LACITO. Elle doit donc assurer tous les rôles. La plupart des contributions à cette archive viennent de membres du laboratoire. C'est aussi le LACITO qui conserve sous forme de CD-ROM l'ensemble des archives numériques et qui offre un accès public à une partie de ces données sous la forme d'un serveur Web. Enfin, les auteurs font bien sûr partie des utilisateurs finaux. Le fait que ce soit la même institution qui chapeaute l'ensemble des rôles est problématique, car celle-ci n'est pas toujours la mieux placée pour les assurer (celui du conservateur en est un exemple). En effet, un laboratoire CNRS

étant créé pour quatre ans, renouvelable, on ne peut pas parler de conservation à long terme, ni de pérennisation. Un laboratoire de linguistique n'est peut-être pas non plus le mieux placé pour offrir des interfaces d'accès à des ressources linguistiques autres qu'une interface destinée à l'usage des linguistes.

L'organisation physique mise en place au LACITO pour la diffusion est simple : un serveur Web héberge l'ensemble des données publiques (documents d'annotations, documents sonores et métadonnées). Les annotations comme les métadonnées sont des fichiers XML. Les enregistrements sont diffusés dans un format étendu (wav) ainsi que dans un format compressé (mp3). Le format mp3 a été choisi pour sa capacité de compression et sa qualité, il constitue un compromis raisonnable entre qualité audio et taille des fichiers à télécharger. Des problèmes de droit sur ce format nous font aujourd'hui reconsidérer notre choix et envisager son remplacement par le format libre Ogg/Vorbis.

La consultation des documents se fait en deux étapes : identification de la ressource, puis consultation en ligne ou téléchargement de cette ressource. L'identification se fait en consultant le catalogue par le protocole OAI. Puis, une fois identifiée(s) la ou les ressources que l'on souhaite consulter, l'accès aux données se fait soit en téléchargeant les fichiers correspondants, soit en utilisant une interface de consultation multimédia qui transforme à la volée sur le serveur les annotations XML en documents HTML. L'interaction entre l'utilisateur et le document est assurée par du code Javascript. Celui-ci modifie dynamiquement l'affichage à l'écran de certaines parties du document (par exemple la couleur ou la graisse des caractères) pour indiquer à l'utilisateur quelle partie de l'annotation correspond au moment écouté. C'est aussi lui qui transmet les événements d'activation et d'inactivation de l'utilisateur vers un player (applet Java ou plugin sachant lire les fichiers audio et vidéo) incorporé dans le document HTML et inversement[29]. Ce mécanisme est décrit de manière plus approfondi dans Jacobson, Lowe et Michailovsky, 2001.

3.1. Le rôle de l'auteur

La plus grande partie des opérations d'archivage est effectuée directement par le linguiste qui a fait l'enquête, sauf cas exceptionnel où il disposerait de personnel technique pour l'aider.

Le laboratoire s'est doté d'un équipement audio : lecteurs de bandes, de cassettes, de DAT et de Minidisc tous ces appareils étant reliés à la carte de conversion d'un ordinateur. Cet équipement est en libre-service et ne nécessite qu'un minimum de formation. Les linguistes arrivent avec leurs anciens supports audio et repartent avec des CD-ROM.

[29] L'ensemble du code Java et Javascript est disponible sur le site http://fieldling.sourceforge.net/

Pour les annotations, nous distinguons deux situations : d'un côté, l'annotation n'existe pas encore ou est sous forme manuscrite, de l'autre, l'annotation a déjà été saisie sous une forme informatisée. Dans le dernier cas, suivant le type de format utilisé (Shoebox, traitements de texte, tableurs, etc.), un informaticien examine au cas par cas les données pour voir s'il peut aider à les transformer en XML, soit en utilisant et paramétrant des outils de conversion, soit par l'application de séries d'expressions régulières, enfin soit en donnant simplement quelques consignes de balisage au linguiste.

Si l'annotation n'existe pas encore où qu'il faille la ressaisir entièrement, nous avons créé un logiciel Interlinear Text Editor (ITE) aidant à la saisie en présentant les données sous forme interlinéaire (Jacobson 2004). Ce logiciel enrichit un lexique, au fur et à mesure des gloses saisies. Ce lexique est consulté à chaque fois que l'utilisateur veut saisir une nouvelle glose, afin de lui établir une liste de propositions (gloses déjà rencontrées) classées par ordre de probabilité. Ce mécanisme aide le linguiste à rester cohérent dans le choix de ses gloses du début à la fin de l'opération.

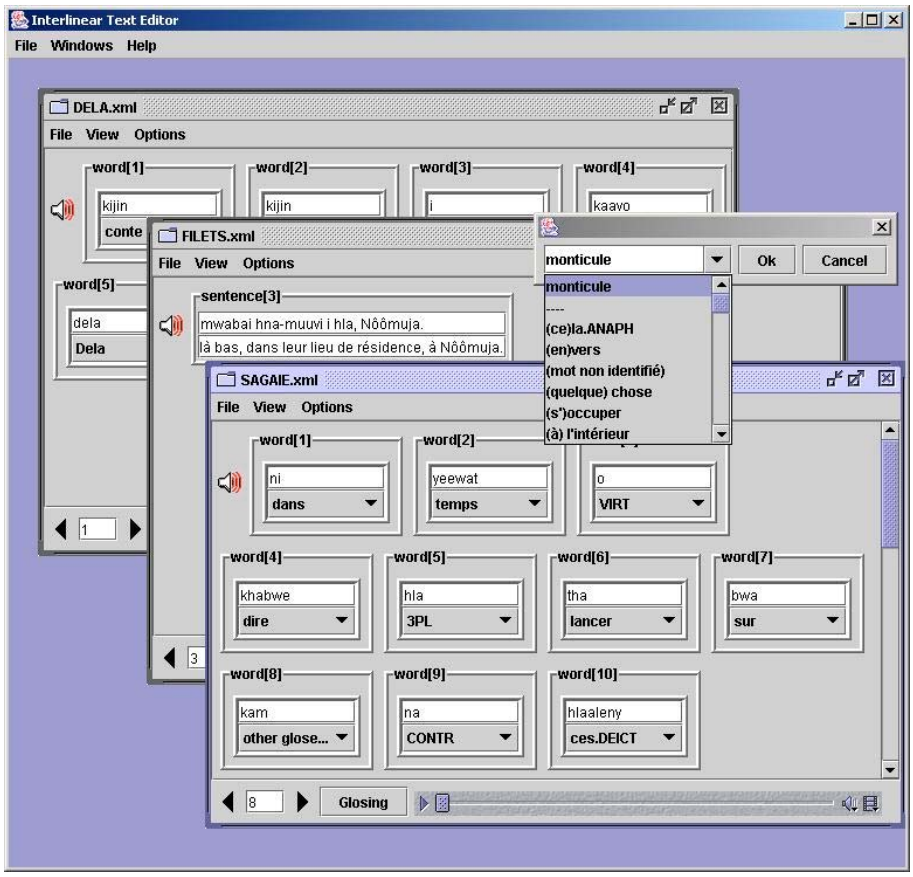

Figure 5. *Image écran du logiciel ITE (Interlinear Text Editor)*

Le document d'annotation peut/doit aussi être synchronisé avec l'enregistrement sonore correspondant. Pour cette tâche, nous avons créé un outil SoundIndex[30] associant un éditeur de son et un éditeur de texte XML. Il permet au linguiste d'ancrer dans le temps les unités qu'il souhaite (texte, phrase, mot, morphème). À la création de cet outil, il n'existait pas ou peu d'outils permettant de faciliter ce travail, ce qui n'est maintenant plus le cas, puisque l'on trouve par exemple des outils comme : Transcriber[31], Elan[32], Praat[33], etc.

[30] http://michel.jacobson.free.fr
[31] http://www.etca.fr/CTA/gip/Projets/Transcriber/
[32] http://www.mpi.nl/tools/elan.html
[33] http://www.fon.hum.uva.nl/praat/

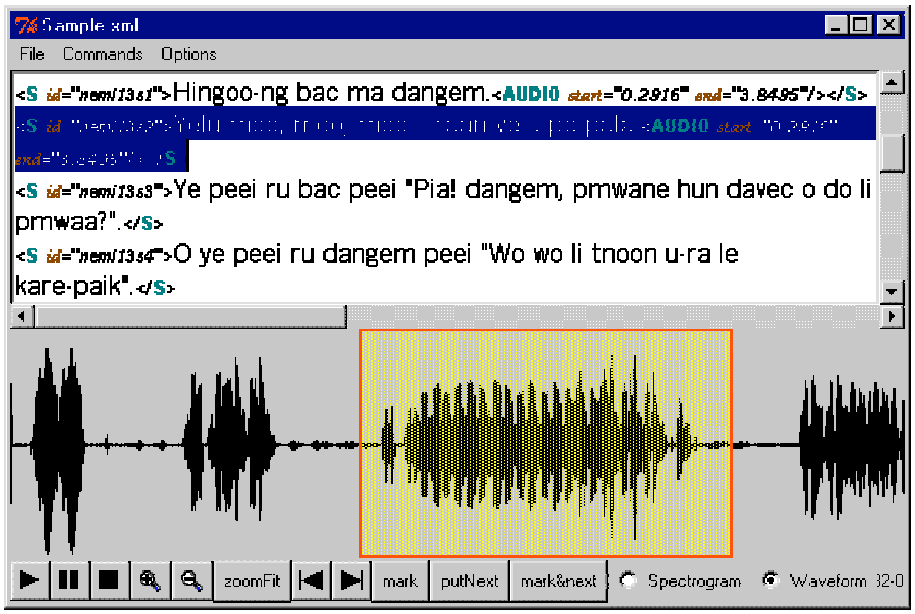

Figure 6. *Image écran du logiciel SoundIndex*

3.2. Le rôle de l'éditeur

Le point de départ du programme archivage était, à l'origine, une préoccupation interne au laboratoire, à savoir s'occuper des matériaux d'enquête du LACITO qui s'abîment dans les armoires et auxquels il est de plus en plus difficile d'avoir accès. Le succès de ce programme a dépassé cet objectif, dans la mesure où d'autres fonds documentaires veulent utiliser les mêmes outils, analyses et architecture de diffusion que ceux mis en place pour ce programme. La ligne éditoriale initiale qui consistait à encourager tous les membres du laboratoire à numériser, normaliser et diffuser ses données d'enquête de cette manière est peut-être maintenant dépassée. En effet, une auto-censure s'exerce déjà, qui écarte des enregistrements de mauvaise qualité, des analyses non finies, afin de ne pas nuire à l'image de l'archive. La prochaine étape sera certainement la constitution de comités comme pour une revue scientifique pour y définir des critères d'appréciations « objectifs ».

Le rôle de l'éditeur consiste à fournir une interface d'accès aux ressources qu'il publie. Nous distinguons dans la publication informatique de ressources : la simple mise à disposition des ressources brutes et la mise à disposition d'une interface de consultation de ces ressources.

La mise à disposition des ressources brutes consiste à déposer sur un serveur de fichiers accessible par Internet l'ensemble des enregistrements et des annotations. L'éditeur se doit aussi d'appliquer des conditions de restriction d'accès si certains documents ne peuvent être diffusés librement. Enfin, l'éditeur se doit de garantir des références stables aux ressources afin que les organisations qui se servent de ces ressources (bibliothèques, moteurs de recherche, fournisseurs d'interface) puissent pointer dessus.

Pour la gestion éditoriale des archives du LACITO, nous avons construit ce que l'Open Archives Initiative appelle un « data provider ». Pour cela, il faut s'engager à respecter un certain nombre de principes pour l'ajout et la modification de ressources dans notre catalogue. Des règles de politesse sont définies pour les suppressions afin de ne pas pénaliser les utilisateurs. Il faut enfin implémenter un protocole de communication autorisant un certain nombre de requêtes. Grâce à ces requêtes, il est possible par exemple, de communiquer l'ensemble des identifiants des ressources disponibles ou bien la liste des nouvelles ressources depuis une date donnée.

Ce protocole permettant de consulter l'ensemble des métadonnées pour chaque ressource, il ne reste plus, à l'utilisateur qui veut télécharger une ressource particulière, qu'à suivre l'URL donnée dans la valeur de l'étiquette <location>, et à consulter l'étiquette <right> pour connaître les conditions de droits qui lui sont attachées.

L'interface d'accès au catalogue du LACITO se fait par l'intermédiaire de requêtes formulées au « data provider », qui y répond en XML. Pour faciliter la navigation dans le catalogue nous avons construit une interface Web qui autorise la recherche sur les seuls critères des noms de langue et des titres des documents, ou bien sous forme multi-critères en remplissant les champs d'un formulaire. D'autre portails sur le Web offrent des interfaces d'accès en consolidant un certain nombre de catalogues, comme la LinguistList[34] qui permet une recherche sur l'ensemble des catalogues référencés dans OLAC.

Une fois la ressource identifiée, il est possible de la télécharger ou de la consulter en utilisant une interface créée à cet effet par un fournisseur d'interface. Le site Web du LACITO propose une consultation multimédia de ses données d'archives. Cette interface est entièrement implémentée en XML via des feuilles de styles XSLT (Clark Ed. 1999) appliquées aux documents d'annotation. Il est prévu entre autres de présenter les documents sous forme interlinéaire, de pouvoir lister l'ensemble des mots ou morphèmes distincts d'un document, de faire une concordance, de rechercher toutes les occurrences d'un mot ou morphème particulier. Toute consultation conserve les liens entre les ressources liées. Il est donc toujours possible d'écouter le segment correspondant à une annotation.

[34] Site web de la Linguist List : http://linguistlist.org/

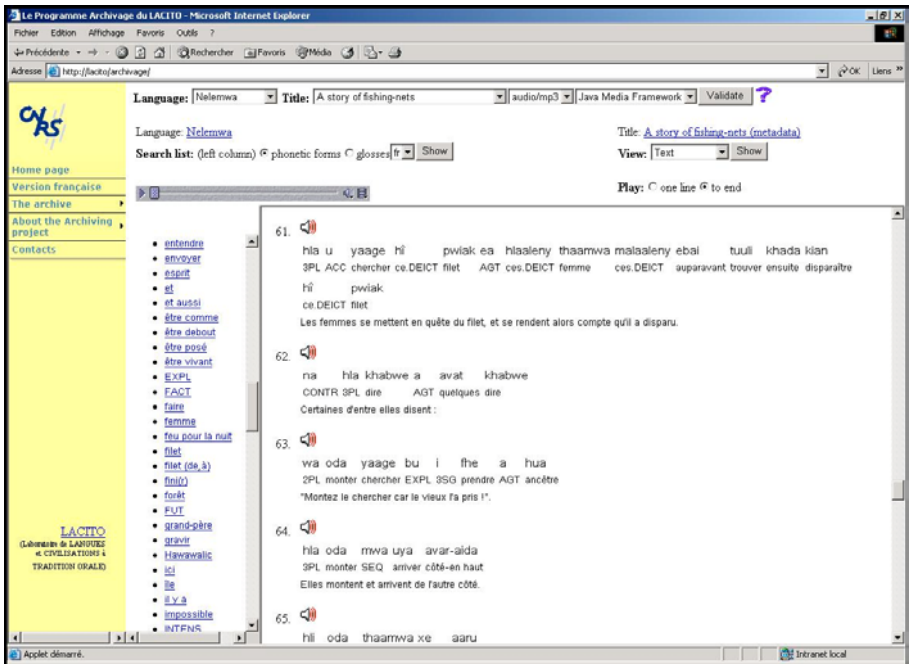

Figure 7. *Image écran du site Web du programme Archivage (consultation d'un conte en nelemwa, langue de Nouvelle-Calédonie)*

3.3. Le rôle du conservateur

Il a pour rôle d'assurer la bonne conservation des données. C'est à lui que revient la veille technologique en ce qui concerne les supports de stockage. C'est aussi lui qui est le garant de l'intégrité des données. Il doit conserver toutes les données y compris celles qui sont confidentielles. Les données doivent être dans le format le plus riche (format étendu pour les enregistrements). En outre, il doit garder un historique des éventuelles versions d'un même document. Au vu du contenu des archives, qui sont des archives de matériaux de terrain récoltés sur des deniers publics par des fonctionnaires du CNRS ou de l'université, avec une analyse scientifique, il serait logique que ce rôle soit confié à une institution publique de type Bibliothèque nationale de France (BnF) ou Archives nationales.

3.4. Les utilisateurs finaux

Les premiers utilisateurs sont les mêmes chercheurs qui ont récolté les données sur le terrain. Un des buts du programme était bien de rendre plus facilement accessible leurs données aux chercheurs. En deuxième lieu, les utilisateurs sont les autres chercheurs en linguistique. Ce qui est d'ailleurs un autre aspect du programme : la mutualisation de l'effort. Si le travail que l'on a fait sur des ressources peut servir à d'autres en agrandissant leur corpus d'étude, c'est autant d'économie et de temps gagné.

Enfin, les utilisateurs peuvent aussi être les informateurs ou les locuteurs des langues étudiées, car ces ressources sont souvent considérées comme faisant partie du patrimoine culturel. Il est également possible d'envisager des utilisateurs pédagogues, touristes, etc. L'architecture de diffusion distinguant les ressources de l'interface de consultation, il serait aisé de définir par exemple une interface pour une application de nature pédagogique ou ludo-éducative.

Les choix que nous avons fait ont porté principalement sur les formalismes de représentation des données plutôt que sur les outils de présentation. Les échanges avec les utilisateurs se font donc à travers le Web au niveau des formats de données. Les annotations textuelles sont rendues par du code HTML qu'il s'agisse de structures de type interlinéaire, de listes, de concordances, etc. Le mécanisme hypertextuel d'HTML permet par ailleurs de ne pas surcharger la présentation d'un document avec toutes les annotations possibles existantes, mais de pouvoir en cacher une partie qui ne sera consultée qu'à la demande de l'utilisateur en cliquant sur des ancres. L'utilisateur n'a donc besoin pour accéder et consulter le site que d'un navigateur Web. Les enregistrements sonores sont eux des fichiers de format wav ou mp3 ou encore des flux de données de types MIME[35] correspondants. Pour lire ces données audio, il faut un outil supplémentaire (un player). Cet outil peut être complètement séparé du navigateur ou bien être utilisé sous forme encapsulée directement dans les pages HTML (plugin, applet). L'aspect multimédia d'une consultation est rendu par les possibilités dynamiques de l'HTML lorsque l'utilisateur choisit une version encapsulée d'un player. Par exemple, la transcription d'une phrase change de couleur quand elle est prononcée. Nous avons préféré cette technique plutôt que l'utilisation d'un formalisme dédié comme SMIL, d'une part pour ne pas avoir à changer d'outil et continuer à utiliser un navigateur, d'autre part parce qu'il existe encore peu d'implémentations de ce formalisme et que les possibilités d'interaction avec l'utilisateur sont plus restreintes qu'avec HTML. SMIL est principalement fait pour la définition de présentations linéaires mélangeant texte, son, vidéo, image dans un scénario.

[35] Multipurpose Internet Mail Extensions

4. Les perspectives

Un rapport de l'Unesco (Wurm, 2001) sur les langues montre que le nombre de celles-ci va en s'amenuisant de plus en plus rapidement. On peut voir dans ce constat un danger face à la diversité culturelle, comme pour la disparition de certaines espèces animales ou végétales. Pour non pas enrayer ce phénomène, mais plutôt s'ériger en conservateurs de la diversité, certains programmes se sont montés récemment. Ils encouragent et financent la récolte, la conservation, la description et la diffusion d'enregistrements sur des « langues en danger ». On peut citer parmi eux le programme DOBES[36] de la fondation Volkswagen ou le « Hans Rausing Endangered Languages Project[37] » (HRELP) de la « School of Oriental and African Studies » (SOAS) de l'Université de Londre. La volonté de conservation de la diversité du patrimoine culturel, associée aux problèmes de conservation des anciens enregistrements analogiques et à l'aspect mature des technologies du Web, ont conduit à la constitution d'une communauté autour des corpus de parole. Assez informelle pour le moment, cette communauté se pose maintenant des questions de normalisation, d'organisation et de droit.

De nombreuses avancées restent à faire notamment pour régler les problèmes de statut juridique des enregistrements de parole et de telles archives de documents. Actuellement nous ne savons pas quelle est l'institution la plus adéquate pour conserver, héberger et diffuser des données de nature à la fois culturelle et scientifique. Ce n'est pour le moment une mission explicite d'aucune institution comme cela l'est avec l'INA pour les données radios et télévisées nationales.

Actuellement, il n'existe pas de dépôt légal pour ce type de données, et les contacts que nous avons pris avec d'autres institutions que le CNRS n'ont pas encore abouti. À noter qu'à l'initiative de la DGLFLF[38], un groupe de travail réunissant des juristes, des linguistes et des conservateurs, étudie actuellement ces questions dans le cadre d'un programme en faveur de la constitution, l'exploitation, la diffusion et la conservation des corpus oraux.

Une autre avancée pourrait venir de la reconnaissance de la constitution de corpus en tant que publication au même statut que celle d'une revue scientifique. Pour cela, nous devrions certainement nous organiser nous aussi comme le font les revues en établissant des comités scientifiques et de rédaction chargés d'évaluer sur des critères objectifs la qualité des contributions. Une telle reconnaissance serait de nature à encourager les chercheurs à rendre publiques leurs données et donc, par voie de conséquence, à faire bénéficier la communauté entière des chercheurs du travail de chacun.

---

[36] http://www.mpi.nl/DOBES/
[37] http://www.hrelp.org/home.htm
[38] Délégation générale à la langue française et aux langues de France (http://www.dglflf.culture.gouv.fr)

Bien sûr, les choix fixés pour la constitution des archives de nos documents d'enquête doivent être régulièrement réexaminés afin de suivre les évolutions technologiques puisqu'en matières de conservation et de normalisation les processus évoluent très rapidement.

5. Bibliographie

André, J., « Caractères, codage et normalisation - de Chappe à Unicode », *Document numérique*, vol. 6, n° 3-4, 2002, pages 13 à 49.

André, J. et Goosens, M., « Codage des caractères et multi-linguisme: de l'ASCII à Unicode et ISO/IEC-10646 », *Les cahiers de GUTenberg*, n°20, mai 1995.

Andries, P., « Entretien avec Ken Whistler, directeur technique du consortium Unicode », *Document numérique*, vol. 6, n° 3-4, 2002.

Andries, P., « Introduction à Unicode et à l'ISO 10646 », *Document numérique*, vol 6, n°3-4, 2002, pages 51-88.

Benoit, J.-L., Bernet, Ch, Bonhomme, P., Romary, L. et Viscogliosi N, « Du document électronique à son usage : le rôle central de la normalisation », *Solaris*, n°6, janvier 2000.

Bird, S. et Liberman, M., « A formal framework for linguistic annotation », *Speech Communication*, vol. 33, n° 1-2, 2001, p. 23-60.

Bird, S. et Simon, G., « Seven dimensions of portability for language documentation and description », *Language*, vol. 79, n° 3, 2003.

Borenstein, N., « MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies », *Request for Comments*, 1521, septembre 1993.

Bray, T., Paoli, J. et Sperberg-McQueen, C. M. (Eds), « Extensible Markup Language (XML) Version 1.0 », recommandation du Word Wide Web Consortium, 10 février 1998.

Buseman, A. et Buseman, K., *The Linguists SHOEBOX*, Summer Institut of Linguistics, 1998.

Calas, M-F et Fontaine, J-M, *La conservation des documents sonores*, Paris, CNRS Editions, 1996.

Clark, J. (Ed.), « XSL Transformations (XSLT) version 1.0 », recommandation du Word Wide Web Consortium, 16 novembre 1999.

DeRose, S., Maler, E. et Orchard, D. (Eds), « XML Linkind Language (XLink) Version 1.0 », recommandation du Word Wide Web Consortium, 27 juin 2001.

Hellwig, B., EUDICO Linguistic Annotator (ELAN) Version 1.4 – Manual, www.npi.nl/tools/, 23 mai 2003.

Hsu, R., *Lexware Manual*, Linguist Dept., University of Hawaii, Honolulu, Deuxième édition, 1989.

Jacobson, M., Corpus oraux glosés : outils logiciels d'aide à l'analyse, *7e Journées d'analyses statistiques des données textuelles*, Louvain-la-Neuve, 10-12 mars 2004.

Jacobson, M. Lowe, J. B. et Michailovsky, B., « Linguistic documents synchronizing sound and text », *Speech Communication*, vol. 33, n° 1-2, 2001, p. 79-96.

Sperberg-McQueen, C. M., et Burnard, L., « TEI Guidelines for Electronic Text Encoding and Interchange (P3) », Chicago and Oxford: ACH/ACL/ALLC Text Encoding Initiative, 1994.

Michaud A., « Conservation des langues et partage des ressources : le rôle des chercheurs dans la mise en place de banques de données », *Actes des 24e Journées d'étude sur la parole*, Nancy, 24-27 juin 2002, p.153-156.

Unicode Consortium. « The Unicode Standard, Version 3.0 », Addison-Wesley, Reading, MA, 2000.

Wurm, S. A., *Atlas of the World's Languages in Danger of Disappearing*, UNESCO Publishing, 2001, 90 p.