Les outils modernes pour la transcription de corpus de parole

Michel Jacobson

Les outils modernes pour la transcription de corpus de parole

Michel Jacobson CNRS/LACITO
7 rue Guy Môquet, Bât. D, 94800 Villejuif
jacobson@idf.ext.jussieu.fr

Mots clefs :

annotation de la parole, linguistique de terrain, archivage, diffusion web

Résumé.

Nous présentons ici une revue des différents outils et formalismes informatiques récents qui peuvent aider le linguiste à faire de la transcription, et plus généralement à faire de l'annotation sur des corpus de parole. La standardisation des ces outils et de ces formalismes facilite le codage, l'échange et la diffusion de l'information.

Nous présentons à titre d'illustration une méthode d'annotation de corpus de parole mise au point dans le cadre d'un programme d'archivage d'enregistrements de terrain. Cette méthode utilise le plus possible les standards émergeants (Unicode et XML). Nous décrivons dans cet article à la fois la structure des données (enregistrements et annotations) et les outils de manipulation de ces dernières (parseurs, éditeurs, browsers, etc.).

Abstract.

Computer tools and formats for linguistic transcription and for the annotation of linguistic corpora are reviewed. Standardization of these tools and formats will facilitate the coding, exchange, and dissemination of information.

A method of annotation for corpora of spoken language, developed as part of a program to archive linguistic field recordings, is presented as an example. The method relies as far as possible on emerging standards for structured text (XML, Unicode). Data formats for both sound and annotation and processing tools (editors, parsers, browsers) are discussed.

1. Annotations de corpus

1.1. Transcription vs. notation.

Le linguiste qui analyse un corpus enregistré utilise principalement des méthodes d'observation. Son activité peut être aidée par des instruments qui lui donnent des mesures ou des estimations sur les indices de son choix (acoustiques, articulatoires, perceptifs, etc.). Une autre catégorie d'instruments largement utilisée est la grille d'observation. Contrairement aux mesures d'indices acoustiques qui fournissent des valeurs continues (fréquence, intensité, amplitude,…), les grilles proposent des catégories discrètes.

Une grille peut être vue comme un métalangage qui permet d'exprimer les observations dans le cadre d'un formalisme particulier. En effet, grilles et théories linguistiques sont intimement liées puisque les catégories définies dans les premières sont issues, pour être plus pertinentes, des connaissances en linguistique générale ou des connaissances plus particulières sur des langues ou familles de langues.

Nous rangeons dans les grilles d'observation les systèmes de notation orthographique ou phonologique de la parole. La notation phonologique permet de décrire les sons du langage de manière fonctionnelle, c'est-à-dire en ne retenant que les traits (articulatoires, acoustiques, etc.) dont on soupçonne qu'ils participent à la fonction distinctive. La notation orthographique permet, elle, de noter de manière normative des unités lexicales. Dans le premier cas, les critères d'analyse retenus dans la grille sont les traits, dans le second cas, il s'agit d'entrées dans un lexique. On peut parler aussi de systèmes de notation de la gestuelle, de la mimique, de la situation d'interlocution etc. Tous ces systèmes de notation renvoient à des grilles d'observations différentes.

Parallèlement à cette appellation de *notation* phonologique et orthographique, on parle, pour recouvrir souvent la même signification, de *transcription* phonologique ou orthographique. La parole est vue dans ce cas comme un système de codage de l'information ; transcrire consiste alors à passer d'un système de notation à caractère oral vers un système de notation à caractère écrit. Nous envisageons donc la transcription comme une activité qui permet le passage d'un système de notation à un autre.

1.2. Annotations

Si la transcription peut être considérée comme une activité de description qui cherche à reproduire le plus fidèlement possible la forme, la traduction peut être définie, elle, comme l'activité qui vise à trouver des équivalences (de sens et parfois aussi, accessoirement, de forme) dans une autre langue.

A tous ces termes de *transcription,* de *traduction* et d'autres encore, nous préférons le terme *d'annotation*, plus neutre et plus général. Sous ce terme nous regroupons toutes les activités qui consistent à caractériser les observations. L'annotation peut donc comprendre des transcriptions de la parole qu'elles soient phonétiques, phonologiques ou orthographiques, des translitérations, des traductions, ou d'autres transcriptions utilisant des systèmes de notation de la mimique, de la gestuelle, etc.

2. Les différents systèmes de notation de la parole.

Les systèmes de notation de la parole proposent des grilles d'analyse associant des valeurs d'indices à des catégories d'analyse. Ces associations forment autant d'hypothèses sur la langue.

Une des grilles les plus connues est l'alphabet de l'Association de Phonétique Internationale (créée en 1885 par les linguistes Paul PASSY et Daniel JONES). Ce système de notation offre l'avantage d'être standardisé et d'être largement utilisé. L'alphabet phonétique international (ou API)

représente en fait l'état d'avancée de la recherche descriptive en matière phonologique, puisqu'il est censé permettre la représentation de toutes les oppositions fonctionnelles déjà rencontrées dans les langues décrites. A côté de l'API, il existe aussi des systèmes de notation phonétique sous forme non alphabétique, comme par exemple ceux notant les états et changements observés sur l'onde (début de voisement, explosion, début de friction, etc.).

Les systèmes de notation orthographique, eux, sont presque aussi divers que les cultures qui les emploient. On comptera à ce titre tous les systèmes alphabétiques (latin, cyrillique, hébreu, etc.), syllabiques (katakana, hiragana, etc.), logographiques (chinois).

Enfin, les systèmes de notation de la gestuelle sont soit peu répandus, soit souffrent d'un manque de standardisation. Nous pouvons citer pour la langue des signes: la notation en alphabet signé, la notation de Bébian, de Stokoe[1].

3. Les standards informatiques utiles pour la notation de la parole.

3.1. Le codage des caractères

La plupart des systèmes de notation orthographique est déjà normalisée dans des codes caractères plus ou moins compatibles entre eux (ISO-Latin-n, JIS, etc.). Il est cependant difficile de les faire coexister dans un même document, ce qui rend compliquée la saisie de textes multilingues. Ce problème est en partie résolu par l'utilisation d'un codage des caractères défini par le consortium Unicode[2]. Actuellement dans sa version 3.2, ce code comporte 95.221 caractère[3] issus de nombreux systèmes de codage (alphabet latin, alphabet arabe, syllabaires japonais, caractères idéographiques chinois, etc.). Ce codage, connu sous le nom de « Unicode », entretient une compatibilité avec la norme ISO-10646, et a été adopté à ce titre par de nombreux fabricants de logiciels et organisations informatiques.

Tous les caractères des tableaux de l'A.P.I. existent dans Unicode, répartis dans plusieurs blocs (latin, grec, extensions A.P.I., diacritiques, modificateurs, etc.). Chaque caractère est identifié par sa position dans le code et par une définition. Par exemple : le glyphe entouré par un cercle dans la figure 1 correspond au 643ème caractère d'Unicode et est défini comme " LATIN SMALL LETTER ESH " ou avec une définition articulatoire non normative comme " voiceless postalveolar fricative ". La représentation typographique des caractères n'est pas du ressort d'Unicode et est laissée aux polices de caractères et autres outils d'édition de texte.

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex
Plosive	p b			t d		ʈ ɖ
Nasal	m	ɱ		n		ɳ
Trill	ʙ			r		
Tap or Flap				ɾ		ɽ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ

Figure 1 : extrait du tableau de l'Alphabet Phonétique International (version révisée de 1993)

[1] Pour une revue des différents systèmes de codage pour les langues des signes : Martin-Dupont, X. 1994. *Les modalités d'évaluations objectives en communication non verbale*. Notes et Documents LIMSI, 8 mars 1995, 115 p.
[2] Site web du consortium Unicode: http://www.unicode.org
[3] Calcul extrait de http://www.i18nguy.com/unicode/char-count.html

L'usage d'un tel standard met un terme aux "bricolages" informatiques, en général incompatibles les uns avec les autres et qui la plupart du temps empêchaient ou limitaient les échanges.

Cependant, les notations de certains événements phonétiques, gestuels, de la mimique etc. qui dans leur formalisation ne font pas appel à la notion de caractère, ne sont pas prévus dans Unicode. Leur codification doit donc passer par un autre mécanisme que celui du codage des caractères.

3.2. Structuration des informations

Jusqu'à présent le système de structuration de l'information le plus courant était sans doute la base de données. Depuis quelques années les techniques à base de langages de balisage de texte se sont considérablement répandus, jusqu'à devenir (avec XML[4]) des standards d'échange de documents utilisés dans de nombreux domaines, allant de l'expression des chaînes d'atomes à la notation musicale, en passant par les expressions mathématiques, etc.

Tous les types d'annotation (transcriptions, traductions, marquages temporels ou spatiaux, indications scénographiques …) peuvent s'exprimer en XML par l'utilisation de balises définies par l'utilisateur. Schématiquement, le balisage représente la structure des données, et le contenu des éléments balisés les données elles-mêmes.

Par exemple : analyser une phrase en mots consiste à entourer la phrase de balises l'identifiant comme telle, puis d'entourer chaque mot de balises l'identifiant comme mot (cf. figure 2).

```
<phrase>
      <mot>Ceci</mot>
      <mot>est</mot>
      <mot>une</mot>
      <mot>phrase</mot>
</phrase>
```

Figure 2 : exemple de balisage XML

Pour contraindre l'annotation, il est possible de définir une syntaxe formelle (DTD[5] ou XML Schema[6]) qui indique quelles sont les balises autorisées et dans quels contextes elles peuvent apparaître. Par exemple, pour exprimer que les phrases d'un texte contiennent des mots et empêcher que les mots puissent se trouver en-dehors d'une phrase, on pourrait définir la mini syntaxe suivante:

```
<!ELEMENT   texte       (phrase+)>
<!ELEMENT   phrase      (mot+)>
<!ELEMENT   mot         (#PCDATA)>
```

Figure 3 : exemple de DTD

3.2.1. Une application particulière : Le programme « Archivage »

Le programme « Archivage » du Laboratoire de Langues et Civilisations à Tradition Orale (LACITO[7]) du CNRS a pour but de sauvegarder et diffuser les enregistrements et les analyses effectués sur le terrain par les linguistes et anthropologues de ce laboratoire depuis plusieurs dizaines d'années. La syntaxe utilisée pour ce programme permet un découpage en quatre niveaux hiérarchiques : le texte, la phrase, le mot et le morphème. A chacun de ces niveaux, il peut y avoir des

[4] eXtensible Markup Language (XML) version 1.0 est une recommandation du Consortium W3C datant du 10 février 1998 (http://www.w3.org/TR/REC-xml)
[5] Document Type Definition
[6] XML Schemas version 1.0 est une recommandation du Consortium W3C datant du 2 mai 2001 (http://www.w3.org/TR/xmlschema-0 -1 et –2)
[7] site web du programme « Archivage » : http://lacito.vjf.cnrs.fr/archivage

transcriptions et des traductions. Au niveau du mot et du morphème, les traductions (dans une langue cible) correspondent à ce que l'on appelle la *glose*[8].

Ce découpage en niveaux permet de faire des transcriptions où l'on peut noter la forme prototypique d'un morphème différemment de sa forme en contexte. Il permet aussi de transcrire une forme de mot qui ne tient pas compte des phénomènes de sandhi, lesquels ne seraient transcrits qu'au niveau de la phrase.

Pour les traductions, le problème est le même. Il n'est pas possible, en mettant bout à bout les traductions des mots ou des morphèmes (gloses), d'obtenir une phrase syntaxiquement correcte, ce qui justifie l'existence de ces différents niveaux.

Chaque niveau peut aussi faire l'objet d'autres types d'annotations (scénographique, typologique, etc.) comme par exemple au niveau de la phrase l'indication du locuteur. Enfin, chaque niveau de texte, phrase, mot et morphème peut faire l'objet d'un ancrage temporel dans un fichier son ou vidéo.

4. Les outils informatiques capables de traiter ces standards

4.1. Les *parsers*

L'outil incontournable pour manipuler des documents XML est le *parser* (vérificateur syntaxique). Il permet deux niveaux de vérification des documents: 1) la bonne formation (*well-formedness*), qui garantit que le document est bien un document XML et 2) la validation, qui garantit la conformité du document à une DTD ou à un XML-Schema particulier. Il existe de nombreux *parseurs* incorporés dans d'autres outils de gestion tels que les éditeurs XML, ou alors utilisables de manière indépendante.

4.2. Les éditeurs

Pour la saisie de documents XML, n'importe quel éditeur de texte peut suffire, mais il existe maintenant de nombreux éditeurs XML qui rendent la saisie et la maintenance de ce type de document plus conviviales:

- en permettant la vision du document sous une forme arborescente,
- en permettant l'affichage de tous les caractères Unicode,
- en proposant seulement les balises autorisées, en fonction de la syntaxe, au moment de leur saisie.
- en proposant des facilités pour effectuer des changements systématiques

[8] Une glose est une note destinée à éclairer le sens d'un mot ou morphème

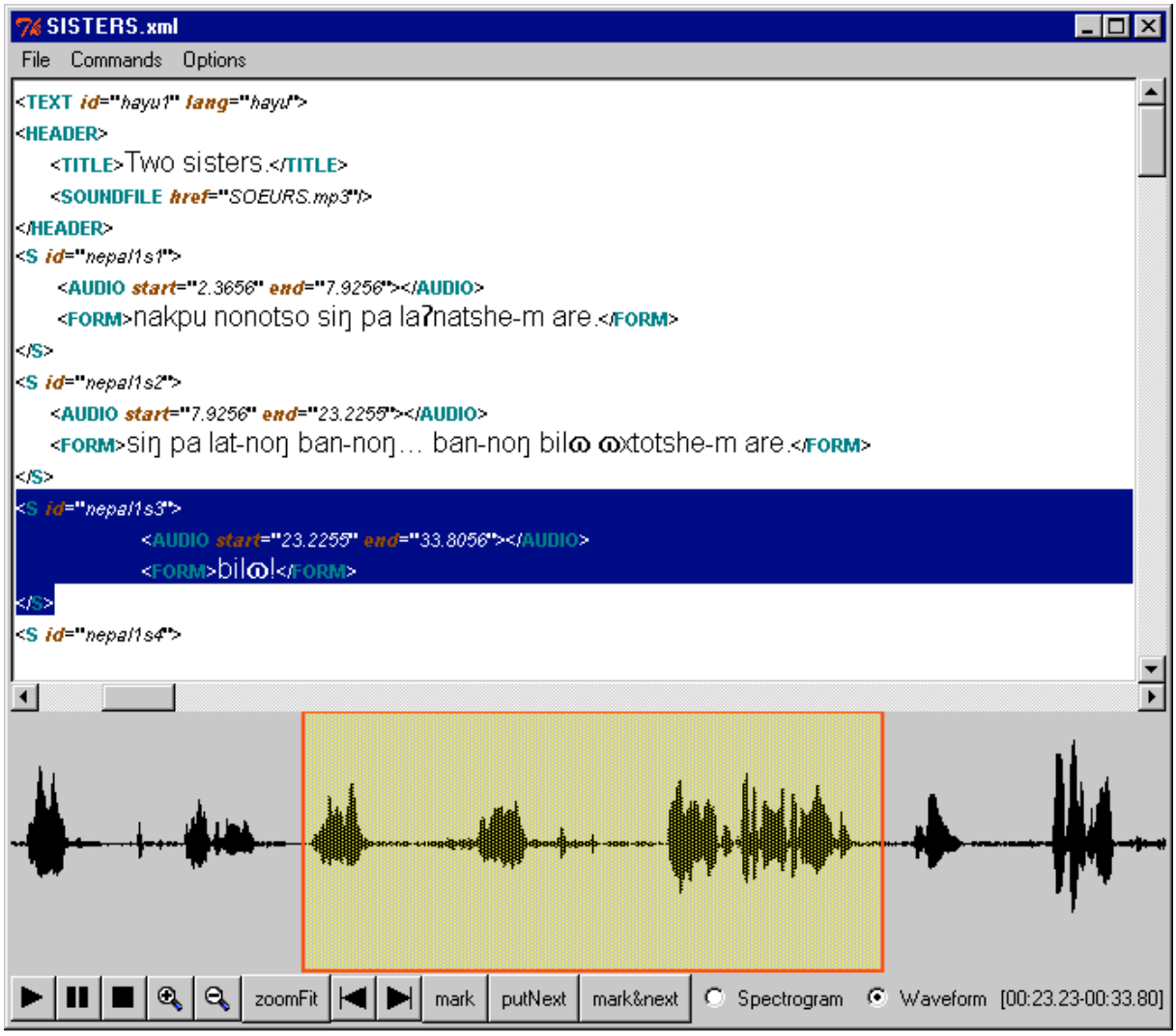

Figure 4 : image d'écran du logiciel SoundIndex

Comme XML ne se préoccupe que des données textuelles, nous avons mis au point pour répondre aux besoins spécifiques du programme « Archivage » du LACITO, un outil de création : SoundIndex (cf. figure 4), qui associe un éditeur de texte avec un éditeur de son, afin de faciliter l'ancrage temporel des différentes unités d'analyse. Cet outil utilise un *parser* pour garantir que le document d'annotation est, et reste, un document XML.

4.3. Les processeurs de styles

XML ne code que la structure logique des données. Si l'on veut définir une interprétation typographique ou multimédia des données, il faut le faire avec un langage de formatage. XSL (XML Stylesheet Language) est le langage de feuilles de styles dédié à XML proposé par le consortium W3 (World Wide Web). Ces feuilles de styles sont de simples documents XML qui utilisent une syntaxe et des noms de balises particuliers (cf. figure 6 : *exemple de feuille de styles pour l'extraction de la liste des morphèmes)*.

Nous avons créé pour les archives du LACITO des feuilles de styles qui définissent pour chaque niveau comment afficher les transcriptions et les traductions, comment restituer les indications scénographiques, etc. Par exemple :

- Les mots sont toujours séparés par des espaces blancs ;
- Les morphèmes d'un même mot sont séparés entre eux par des tirets ;
- Les mots ou parties de mots empruntés à d'autres langues que celle du texte sont présentés en italique ; …

Pour appliquer une feuille de styles à un document XML nous avons besoin d'un outil que l'on appellera un processeur de styles ou un processeur XSL. Il en existe maintenant de nombreuses implémentations.

4.4. Les langages de requêtes

Un document XML peut être vu comme une base de données qui contient beaucoup plus de d'informations que ce que l'on souhaiterait voir en même temps. Nous avons donc besoin d'un mécanisme, permettant de faire des sélections dans cette base, afin de n'afficher que des morceaux ou des vues particulières de celle-ci. Par exemple, pour afficher uniquement la transcription au niveau de la phrase, ou pour afficher seulement la transcription et la traduction en français, etc.

Nous avons aussi besoin d'un mécanisme permettant de réorganiser la structure des données pour présenter le texte ou les morceaux de texte de différentes manières. Par exemple, pour présenter les transcriptions des mots avec leur glose juste en dessous (présentation « interlinéaire » ou « mot sous mot »).

337.

yammu thɛpt-u yuks-u-aŋ lam-ɛtna *cha* *sāt* *ṭhāū*
again plant.S2-3O put.down.S2-3O-and road-along six seven place
thɛpt-u
plant.S2-3O

338.

khɔmbhɛaŋ him-mu phɛr-ɛ-aŋ yammu yɛba-n-lɛ ɛn
then house-at come.S2-PA-and again shaman-DEF-ERG this
mundhum-ɛn khunɛ sur-u
ritual-DEF 3SG finish.S2-3O

339.

tyatinai *ho*
that.much.EMPH be:3SG.PR

Figure 5 : Présentation interlinéaire d'un texte limbu : « An untimely death »[9]

XSL-T[10] est la partie de XSL qui permet d'effectuer des *Transformations* correspondant aux opérations d'extraction, de tris, de comptage, etc., que l'on avait l'habitude de faire avec des langages comme SQL (Structured Query Language) dans des bases de données. Un document XML peut être vu comme une structure strictement arborescente. Dans un tel cadre, effectuer des transformations avec XSL-T revient à définir une manière de parcourir cet arbre pour réécrire une nouvelle structure arborescente.

[9] Le texte entier est consultable à l'adresse : (http://lacito.archivage.vjf.cnrs.fr/cgi-bin/archives.pl?ID=oai:lacito:LIF:SOGHA)

[10] XSL Transformations (XSLT) version 1.0. est une recommandation du Consortium W3C datant du 16 novembre 1999 (http://www.w3.org/TR/xslt).

Pour les archives du LACITO nous avons défini des feuilles de styles qui construisent pour un texte :

- La liste triée des formes phonétiques des morphèmes ou de leurs gloses ;
- Une concordance (liste de toutes les occurrences des morphèmes présentés avec leur contexte d'apparition) ;
- La liste de toutes les phrases qui comportent un mot ou un gabarit de mot particulier ;
- …

```
<?xml version="1.0"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
       version="1.0">
      <xsl:template match="/">
            <xsl:for-each select=".//morpheme/forme">
                  <xsl:copy-of select='.'/>
            </xsl:for-each>
      </xsl:template>
</xsl:stylesheet>
```

Figure 6 : exemple de feuille de styles pour l'extraction de la liste des morphèmes

4.5. Les browsers

Pour en finir avec la revue des principaux outils, nous avons besoin, pour la consultation des données, d'un outil qui nous permette d'effectuer un rendu typographique correct de nos textes ou morceaux de texte, ainsi qu'un rendu adéquat de nos enregistrements (audio ou vidéo) tout en conservant le lien de synchronisation qui existe entre les deux.

Nous avons choisi le langage HTML[11] (HyperText Markup Language) comme formalisme pour exprimer nos exigences en matière de présentation. Ceci met à notre disposition de nombreuses balises à signification typographique (paragraphe, ligne, interlignage, sens d'écriture, etc.), ainsi qu'une interprétation correcte des caractères Unicode.

Les capacités d'HTML en matière de son ou de vidéo étant très réduites et non normalisées, cette tâche est en règle générale confiée à des plugins[12] ou des applets[13].

Avec HTML, nous pouvons aussi bénéficier du mécanisme hypertexte, et surtout de la possibilité d'une interaction avec l'utilisateur. C'est ainsi que nous avons défini pour le programme « Archivage » du LACITO, un script (en Javascript) qui permet a) lorsque l'utilisateur clique sur une phrase d'entendre celle-ci, b) de mettre en relief (par exemple en changeant sa couleur) la ou les phrases actives, au fur et à mesure de l'écoute d'un texte.

[11] HyperText Markup Language (HTML) version 4.01 est une recommandation du Consortium W3C datant du 24 décembre 1999 (http://www.w3.org/TR/html401)
[12] Un plugin est un logiciel qui permet d'ajouter des fonctionnalités au logiciel pour lequel il est destiné
[13] Une applet est un programme écrit en Java et qui peut s'exécuter au sein même d'une page HTML.

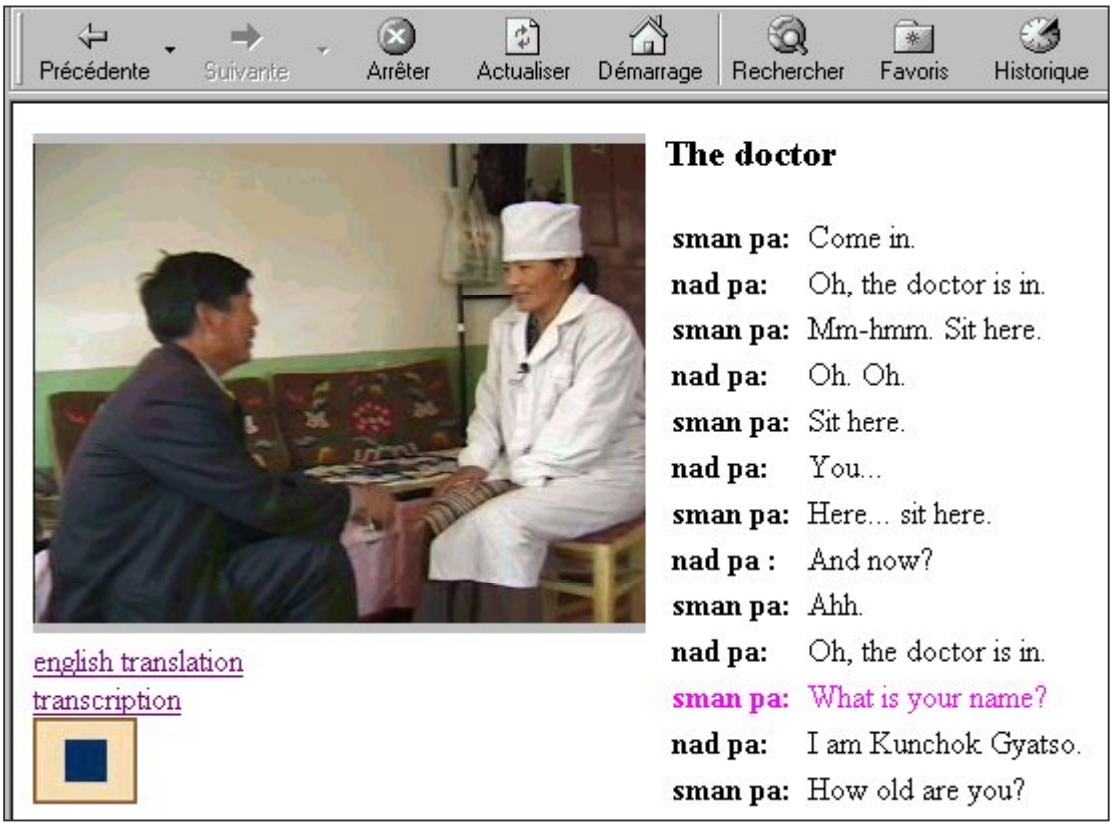

Figure 7 : exemple d'écran de navigateur (enregistrement vidéo d'un dialogue en tibétain entre un docteur et son patient)[14]

L'outil d'accès pour la consultation et l'interrogation des données devient alors le navigateur web. C'est lui qui prend en charge l'affichage des textes en interprétant le code HTML. Encore une fois la production de ces documents HTML se fera par l'intermédiaire de feuilles de styles. Il suffit pour cela d'ajouter dans les feuilles de styles déjà définies des balises propres au langage HTML.

5. Perspectives

Nous avons vu comment l'utilisation des formalismes et des technologies du web qui les implémentent nous a permis de définir une structure particulière de données (l'archive Texte/son) ainsi qu'une application qui permet d'interroger ces données via Internet. Nos meilleurs alliés dans cette tâche auront certainement été les mouvements du 'logiciel libre' et ceux du 'logiciel à sources ouvertes'. Grâce à eux, les développements informatiques que l'on a effectué ont pu se concentrer sur des tâches où les connaissances du linguiste sont une plus-value. Pour les autres tâches plus génériques ou moins spécifiquement linguistiques, nous avons pu nous reposer sur les autres membres de la communauté.

Notre travail consistera maintenant à réfléchir sur les modalités qui permettront de rendre accessibles ces données. En effet, mettre à disposition des données et des outils pour consulter ces données sur le web ne suffit pas. Il faut aussi fournir des informations sur ces données pour que les utilisateurs sachent qu'elles existent, et puissent extraire de celles-ci uniquement les parties qui sont pertinentes pour leurs besoins. Ceci devra être fait en concertation avec les autres fournisseurs de données afin que l'utilisateur puisse rechercher de l'information sur les langues de manière standardisée sans se préoccuper de la répartition géographique de ces données entre tous les fournisseurs. Il s'agit là d'un vaste domaine de travail qui concerne cette fois autant les fournisseurs

[14] Le texte est extrait du projet de l'Université de Virginie « Tibetan and Himalayan Digital Library » (THDL) consultable à l'adresse : (http://iris.lib.virginia.edu/tibet)

de ressources (linguistes, ethnologues,…) que leurs utilisateurs (enseignants, chercheurs,…) et que leurs distributeurs (archivistes, bibliothécaires…).

Références Bibliographiques

BRAY, T., PAOLI, J., SPERBERG-McQUEEN, C.M. (Eds.), 1998. Extensible Markup Language (XML) 1.0. Recommandation du consortium W3C du 10 février 1998 (http://www.w3.org/TR/REC-xml)

CLARK, J. (Ed.), 1999. XSL Transformations (XSLT) version 1.0. Recommandation du consortium W3C du 16 novembre 1999 (http://www.w3.org/TR/xslt).

JACOBSON, M., MICHAILOVSKY, B., LOWE, J.B. Linguistic documents synchronizing sound and text in *Speech Communication* 33 a special issue on Speech Annotation and Corpus Tools, 2001.

MARTIN-DUPONT, X. *Les modalités d'évaluations objectives en communication non verbale*. Notes et Documents LIMSI, 8 mars 1995, 115 p.

Programme « Archivage » du LACITO (http://lacito.vjf.cnrs.fr/ archivage)

RAGGET, D., Le HORS, A., JACOBS, I. HyperText Markup Language Specification version 4.01 Recommandation du consortium W3C du 24 décembre 1999 (http://www.w3.org/TR/html401).

Tibetan and Himalayan Digital Library (http://iris.lib.virginia.edu/tibet/)
UNICODE Consortium, 2000. The Unicode Standard, Version 3.0. Addison-Wesley, Reading, MA (http://www.unicode.org)

XML Schema 1.0. Recommandation du consortium W3C du 2 mai 2001
(http://www.w3.org/TR/xmlschema-0 –1 et -2)