Digitized resources for languages of Nepal

Boyd Michailovsky
LACITO, CNRS, Villejuif, France

1. Introduction

The object of the present paper is to describe some currently available resources reflecting the application of information technology (IT) to languages of Nepal, with particular emphasis on linguistic documentation and research. Three categories of resources will be considered:

(1) tools for the coding and rendering of Nepal languages and scripts
(2) spoken and written corpora, in particular annotated speech corpora
(3) dictionaries and wordlists.

The first category comprises general-purpose software tools. The second and third categories, which will be my main focus, cover properly linguistic resources. Since one of my aims is to show the possibilities of information technology, resources that take advantage of this technology beyond text processing will be given most attention. In particular, I have given a rather detailed description of the Lacito Archive, for which I am responsible with my colleague Michel Jacobson. Given the evident conflict of interest, I do not pretend to evaluate the resources covered.

Resources are considered here from the point of view of linguistic research and language study. Readers who are interested in language engineering in the South Asian context can refer to the proceedings of the SCALLA ("Sharing Capability in Localisation and Human Language Technologies") conference, held in Kathmandu in January 2004. Another source of information is the journal *Vishva Bharati* of the Indian Ministry of Communications and Information Technology's Technology Development for Indian Languages (TDIL) project.

I apologize in advance for any omissions. Very few of the resources described have metadata, the cyberspace equivalent of library catalogue information, available on the web (see §3.1 below).

2. Tools for Nepal languages and scripts

The IT industry was slow to establish standards for coding character sets beyond ASCII (96 printable characters) and some European language extensions, so that users of phonetic characters and of Devanagari and many other scripts adopted a variety of unstandardized codings and associated fonts. Nonstandard Devanagari fonts like Preeti, Kantipur, Himal, etc., are still very widely used in Nepal and in India, but new development is generally based on the Unicode standard (Unicode Consortium 2003), which has been adopted by the World Wide Web Consortium (W3C). Some developments relevant to the Devanagari and the Limbu (Sirijonga) scripts, to transliteration, and to the International Phonetic Alphabet in the context of Unicode will be mentioned below.

2.1 Nepali Unicode

The Madan Puraskar Library (MPP) in Nepal has developed and made freely available a software package facilitating the use of standardized Unicode coding for Devanagari according to Nepali typographic usage. This includes:

- Installation instructions
- TrueType fonts covering the relevant portions of Unicode
- two Windows keyboard layouts, based roughly on:
  - the Nepali Remington layout familiar to Nepali typists (but inevitably requiring adaptation on the part of typists)
  - romanization
- a utility for converting to Unicode from existing, non-standard fonts

These developments are part of an ambitious program of software localisation in Nepali which is outside the scope of the present article. See Chalmers and Gurung (2004) for a progress report on Nepali Unicode and further activities designed to promote its adoption.

2.2 Using Devanagari for lesser-known languages

Other Nepalese languages which use Devanagari, like Newari, and, more recently, Tamang and Wambule, can take advantage of Devanagari Unicode. However, certain combinations which do not occur in the languages on which the standard is based may not be handled correctly by rendering software.

2.3 Limbu script

The Limbu, or Sirijanga script has been included in Unicode version 4.0 (Unicode consortium 2003:260-262, based on Michailovsky and Everson 2002), although no Unicode font appears to have implemented it as yet. Current fonts are based on earlier Unicode versions. Limbu Unicode defines codepoints for all current Limbu characters and for some that are obsolete.

2.4 Romanisation; Phonetic fonts

The Unicode standard provides for the coding of the International Phonetic Alphabet and other characters commonly used by linguists, including a wide variety of spacing and non-spacing diacritics. Roman transliteration of Devanagari orthography can be coded using Unicode diacritics and combinations of diacritics, but certain combinations, like macron and tilde on the same letter (used for Nepali nasalized long vowels), may not be handled correctly by rendering software.

3. Corpora

The design of annotated speech corpora has been the object of considerable interest in recent years, having spread from the language engineering world, where digitized corpora are used

to test automatic speech processing, to language research and documentation. Two large current programmes for the study of endangered languages, the Volkswagen Foundation Dobes project and the Hans Rausing Endangered Languages Project, require grantees to prepare and make available digitized speech corpora; both have research projects in Nepal. At the same time, fifty years after the use of sound recording became widespread in field research, a number of research institutions have taken steps to conserve and make available existing speech recordings and transcriptions made in the course of field research. This activity is often referred to as "archiving", with the understanding that digitized archives should be more accessible and less dusty and expensive to maintain than traditional archives. The Lacito Archive, which the I coordinate with Michel Jacobson at the French CNRS, is one such project, the only one to offer material in Nepal languages. The architecture of this site is described in some detail below.

Written language corpora are very useful for many kinds of linguistic research and have become indispensable for lexicography. Large corpora, important for the study of relatively infrequent phenomena, are relatively easily attained since transcription is not required. Unfortunately, there do not appear to be any purpose-built corpora of Nepal languages available at present. A few computerized sets of journalistic and literary material in Nepali will be mentioned below.

3.1 The Lacito archive

The purpose of the Lacito Archive is to (1) conserve and to make available speech recordings in little-known languages, with synchronized transcriptions, translations, and other annotation and (2) to develop an architecture for such documents and tools for their exploitation, using standard information technology. Four of the 18 languages currently covered by the Archive are Nepalese. Two corpora, in Limbu (10 texts) and in Hayu (26 texts), two texts in Tamang, and one text (less scientifically annotated than the others) in a western dialect of Nepali are currently available on the Internet.

The Lacito Archive is a fairly representative example of current thinking on the design of speech archives and the annotation of recorded speech. It is structured in a client-server architecture and accessed using a standard browser. The underlying data is coded in Unicode and marked-up in XML, the W3C's metalanguage for structured text. In response to client requests, the data is processed on the server and the response is furnished to the client, along with associated digitized sound data.

The user interface on the Lacito Archive website proposes a number of "views", which do not exhaust the possibilities of the archived data. In this interface, the user chooses the language, the document to browse, and a "view" on the data. If he chooses the "text" view, he can choose among different transcriptions, and, for the more thoroughly annotated documents, among translations at different levels (e.g. utterance-level "free" translations, morpheme-level glosses, etc.) or in different languages. When the document is displayed, morpheme-level transcriptions and translations are displayed in aligned interlinear format. The user can choose to hear the recording corresponding to a single sentence, or to hear the whole remainder of the text while scrolling through the annotation. Fig. 1 shows an interlinear view of a Limbu document on the Lacito Archive site. The user can request a "search list" of all morphemes or glosses occurring in the text. Selecting an item from this list (which spares the user having to enter Unicode characters) brings up all the utterances in which it appears. If he chooses the

“concordance” view, a concordance of the entire text is displayed. Any time that an utterance is displayed, the recorded sound is immediately accessible.

Figure 2 shows a fragment of a concordance of the document shown in fig. 1, accessed by selecting the “concordance” view. Five concordance entries, including the three occurrences of the verb stem *pa:r* (including one in utterance 31, shown in fig. 1) are shown. The preceding and following context are shown to the left and right of the concorded items (underlined), with a reference identifying the text and utterance number. Clicking on the concorded item causes the sound recording of the utterance in which it occurs to be played. “Talking concordances” of this kind, particularly of whole corpora, have proved a useful tool for verifying transcriptions.

All of the data displays described — straight transcription, interlinear glossed text, morpheme lists, lists of utterances containing a particular morpheme, concordance — are simply different “views” on a single set of structured data — the “annotation” — associated with a recording. Figure 3 shows an annotation document, in the form of structured text marked up in XML, whose contents have been reduced to contain only the first sentence (no. 31) seen in fig. 1. The annotation is structured hierarchically into logical units, text, utterance (“S”), word (“W”) and morpheme (“M”), and elements at each level marked up as transcriptions (“FORM”), translations, etc. Views are produced by selection and rearrangement of these logical data elements, a classic IT paradigm.

Although the immediate print-medium ancestors of such documents (without the sound recordings) are collections of texts in interlinear format, going back to early examples like Boas 1911, it is misleading to characterize the computerized documents as “interlinear”. The interlinear display may be the view that shows the largest proportion of the data at one time, but the data structure is logical, not typographical. If the underlying data were in fact marked up typographically, in lines and aligned tabulations or table cells, as in a word processor or in the HTML which is actually furnished by the Lacito server in answer to requests, it could not be searched intelligently or transformed automatically into concordances, wordlists, etc. for linguistic research.

Metadata, or cataloguing information, has been mentioned above, in the context of locating linguistic resources on the web. The Lacito Archive is one of 29 “data providers” which are members of the Open Language Archives Community (OLAC, part of the more general Open Archives Initiative) and provide metadata in the OLAC format. This metadata is “harvested” by two OLAC “service providers”, the Linguist List and the Linguistic Data Consortium, which provide search interfaces and serve as portals to the participating archives. Thus linguists do not need to know where a resource is archived in order to find it.

The term “open” in the context of “open archives” refers to metadata, not to the data itself: in fact, data providers are only obliged to provide metadata. However, in the case of the Lacito Archive, the metadata for each catalogued resource includes a web address (URL) where the data resource itself is accessible. The Lacito Archive website provides an interface with a limited number of “views” on this data, as described above. But since the data itself is accessible, any user can write a script providing a view on it to satisfy a particular research need. A script accessing the Lacito annotation — like those on the Lacito server — would normally be written in XSLT, the W3C’s transformation language for XML. The data can also be downloaded, but downloaded copies risk becoming obsolete if the data is updated.

3.2 Written language corpora

The online archives of back issues of Nepali periodicals constitute corpora which could be used for linguistic research, although they are not designed for this purpose. Collections of periodicals in non-standard codings for easily available fonts can be found on the NepalNews, KantipurOnline, and Gorkhapatra sites, among others.

An interesting corpus of early twentieth century Nepali, unfortunately practically impervious to automatic treatment because it is in the form of scanned images, is the complete run of the periodical *Gorkha Sansar* (Dehra Dun 1926-1929) on the DSAL site.

A corpus of Nepali literary works is available on a Nepali site at the Indian Institute of Technology (IIT), Kanpur, a regional outpost of the TDIL project mentioned above. Data on the site is stored in MySQL, coded in the Indian ISCII standard coding (see *Vishwa Bharat* 10:22, July 2003).

4 Lexical resources

A number of dictionaries and wordlists of Nepalese languages are available in digital form. Some are coded as databases, some as structured text, and some in display formats like PDF or HTML, or as word-processing (e.g. Word) files. Most use non-standard character codings.

4.1 The Digital South Asia Library (DSAL)

The Digital Dictionaries of South Asia (DDSA) project, which is part of the Digital South Asia Library at the University of Chicago, has dictionaries of a large number of South Asian languages available or in preparation. Most are computerized versions of large, authoritative print dictionaries. Nepali is well-represented by digitized versions of Turner's classic *Nepali Dictionary* (1931) and of Schmidt (1993), the only recent large Nepali-English dictionary specifically addressed to non-speakers. Turner's comparative Indo-Aryan dictionary (1962-1966) is also available.

Users of the digital dictionaries can choose between Unicode and non-Unicode data display. The non-Unicode display omits diacritics in roman transliteration, so that serious users will want to use Unicode. Headwords in Devanagari, coded according to the Indian ISCII standard, are present in the underlying digitized data for the Turner dictionary (of which James Nye, director of the DDSA project, kindly provided a sample), but Devanagari is not displayed by the current interface. Headwords in Devanagari script are displayed by the Unicode interface to the Schmidt dictionary. I am not sure whether their underlying encoding is ISCII or Unicode; the two codings are in principle interconvertible. (On ISCII, see "standards" on the TDIL site.)

The search interface proposes to match strings of characters (with the options "starts with", "ends with" and "exact word match") either (1) in the entry headword or (2) anywhere in entries, without regard to the type of information searched for. The underlying data markup is essentially typographical (its notation resembles HTML) rather than logical, which would seem to rule out more precise search options.

4.2 The Newari Lexicon

As a non-Sanskritist I cannot fully appreciate the *Newari Lexicon*, but I cannot leave it out of a review of computerized lexical resources. It is marked as "under construction". Here is how it is described by its authors, the Nepal Bhasa Dictionary Committee:

> "The *Newari Lexicon* is compiled from a group of Nepalese manuscripts of a single text, the *Amarakosa*. The manuscripts date from the end of the fourteenth to the early nineteenth centuries. Each manuscript contains the text of the *Amarakosa*, a popular Sanskrit thesaurus written in verse, with notes in Newari containing brief Newari glosses of each group of Sanskrit synonyms. The *Newari Lexicon* is a compilation of words from the Newari glosses, with reference to the original Sanskrit and to English glosses."

The lexicon is in the form of a database in which every Newari word cited in the manuscripts is indexed as to its manuscript source and to a standard printed edition of the *Amarakosa*. The sources can be browsed online. All the manuscripts are indexed so that they can be compared and referred to the standard edition. The underlying data has been entered into a MySQL database in two roman transliterations, one idiosyncratic using only ASCII characters, the other using diacritics in the Sanskritist tradition, in a non-standard coding for display with a supplied font.

4.3 The Tower of Babel

Sergei Starostin's Tower of Babel site holds a vast and growing array of lexical resources including bilingual and comparative dictionaries covering a significant proportion of the world's languages in a database format. The Nepal material is the following:

> Limbu dictionary (van Driem, 1987)
> Dumi dictionary (van Driem 1993)
> Yamphu dictionary (Rutgers 2001)
> Kulung dictionary (Tolsma, to appear)
> Kiranti comparative dictionary (an original set of reconstructions by Starostin)

The bilingual dictionaries are computerized versions of glossaries included in published descriptions of Kiranti languages written by members of the Himalayan Languages Project at the University of Leiden. The data fields are the following (with some variation): headword, part of speech, English definition, Nepali gloss, comments (which may include example sentences), and cross-references to Starostin's comparative Kiranti dictionary.

The comparative Kiranti dictionary draws on the bilingual dictionaries mentioned above and on published sources for Khaling (Toba and Toba 1975) Sunwar (Bieri and Schulze 1971) and Thulung (Allen 1975). It is not a dictionary of synonyms but of etymologies, relating words which continue hypothesized protoforms. The data fields are: protoform, English gloss, forms by language [drawn from the 7 languages], comments. No table of phonological correspondences or justification of the reconstructions is available.

The lexical data for both the bilingual and comparative dictionaries are in Starostin's StarLing database format, in non standard character codings. The search interface is quite complete:

string matches can be specified for any field or independently for any combination of fields. This is particularly useful for comparative databases. Data is displayed either in Unicode or in a special font. The complete databases and the database software itself (for DOS or Windows platforms) can be downloaded.

4.4 Limbu dictionary of the Mewa Khola dialect

The Limbu dictionary of the Mewa Khola dialect is the online version of Michailovsky (2002), coded in Unicode and marked up in XML. The markup, loosely inspired by the Text Encoding Initiative (Sperberg-McQueen and Burnard 2001), distinguishes a large number of data fields, including headword, part of speech, morphological and derived forms, grammatical information, definitions, English glosses for inverse indexing, Nepali glosses, botanical binomials, example sentences with references to a text corpus, etymological and derivational relations (coded as hyperlinks to other items in the dictionary), etc. Roughly half of the example sentences are drawn from the Lacito Archive Limbu corpus (see above). References to the corpus are displayed as links giving acces to the original recorded sound and glossed transcriptions. The dictionary is also accessible from the corpus: clicking on a morpheme in the Limbu corpus brings up the corresponding entry in the computerized dictionary.

The XML formalism seems well-adapted to a rich structure in which one entry does not necessarily resemble the next. The current Internet interface provides for sequential browsing, or for selection of articles via Limbu and English indexes.

4.5 Wambule and Jero dictionaries

Dictionaries of the closely related Kiranti languages Wambule and Jero have been put online as MySQL databases. These are comptuerized versions of the dictionaries accompanying linguistic analyses by Jean-Robert Opgenort (2004, forthcoming) of the Himalayan Languages Project. Language forms are coded (1) in a Unicode-coded phonological transcription and (2) in a coding corresponding to the Kantipur Devanagari font. The interface provides for searching by string match in the English definition field and optionally by part of speech. Fields displayed are the definition, the Wambule or Jero form in Unicode IPA, dialect identification, and part of speech or verb class. Example sentences and a few other fields present in the printed dictionaries are not displayed.

4.6 Thangmi

Mark Turin's (2004) 1700-word *Thangmi-Nepali-English Dictionary* can be consulted online or downloaded in pdf format. The online interface provides for string-matching in any or all of the three language fields, Thangmi (Devanagari Unicode), Nepali (Devanagari Unicode), English.

4.7 Other lexical resources

An 800-word monolingual Nepali dictionary and a synonym dictionary (thesaurus) are available for download on the MPP site. These are coded in Nepali Unicode.

A computer version of Oja et al. 2004, a 1400 entry Nepali-English and English-Nepali glossary keyed to Oja and Oja 1992, a language primer is available on the Tibetan and Himalayan Digital Library site. Fields in the Nepali-English alphabetical browsing view are: headword (Nepali Unicode), part of speech, English gloss, subject area, primer lesson reference. The glossary can be viewed in English, Nepali, or primer reference order. String search on the English gloss field is available.

The Nepal Research website offers a number of dictionaries in pdf format. These are bilingual dictionaries (1400 entries) of Sherpa in English, French, and German, compiled by Lhakpa Doma Salaka-Binasa Sherpa and Chhiri Tendi Salaka Sherpa in cooperation with Karl-Heinz Kraemer, and Nepali-English and Nepali-German dictionaries (4300 entries) compiled by Karl-Heinz Kraemer.

References

Allen, N. 1975. *Sketch of Thulung Grammar*. Ithaca, N.Y.

Bieri, Dora and Marlene Schulze. 1971. A vocabulary of the Sunwar language. Kirtipur. SIL [mimeo].

Boas, Franz. 1911. *Handbook of American Indian Languages*. Government Printing Office. Washington.

Driem, George van. 1987. *A Grammar of Limbu*. Mouton de Gruyter. Berlin. [Glossary: http://www.starling.rinet.ru]

—. 1993. *A Grammar of Dumi*. Mouton de Gruyter. Berlin. [Glossary: http://www.starling.rinet.ru]

Michailovsky, Boyd. 2002. *Limbu-English dictionary* with English-Limbu index and notes on the Maiwa-Mewa dialect. Kathmandu. Mandala Book Point. [http://lacito.vjf.cnrs.fr/archivage/dico].

Michailovsky, Boyd and Michael Everson. 2002. Revised proposal to encode the Limbu script in the UCS. ISO Working Group Document/Expert Contribution. [http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2410.pdf].

Oja, Banu and Shambhu Oja. 1992. *Nepali: A Beginner's Primer, Conversation and Grammar*. Cornell University South Asia Program. Ithaca.

Oja, Banu, Shambhu Oja, Mark Turin, Elisabeth Uphoff. 2004. Nepali-English, English-Nepali Glossary. 2nd edition. Cornell University South Asia Program. Ithaca. [http://iris.lib.virginia.edu/tibet/reference/dictionary/nepali/index.php]

Jean Robert Opgenort. 2004. *A Grammar of Wambule*. Leiden. Brill. [http://www.opgenort.nl/dictionary.php]

—. forthcoming. *A Grammatical sketch of Jero*. Leiden. Brill.

Rutgers, R. 1998. Yamphu. CNWS. Leiden. [Glossary: http://www.starling.rinet.ru]

Sperberg-McQueen, C. M. and Lou Burnard. 2001. *Guidelines for Electronic Text Encoding and Interchange.* 4th edition. TEI Consortium. [http://etext.virginia.edu/teip4/]

Toba, Sueyoshi and Ingrid Toba. 1975. *A Khaling-English English-Khaling Glossary*. SIL. Tribhuvan University. Kirtipur, Nepal.

Tolsma, Gerard. n.d. [?1999] Kulung dictionary. [http://www.starling.rinet.ru]

Turin, Mark, with Bir Bahadur Thami. 2004. *Nepali-Thami-English Dictionary*. Kathmandu. Martin Chautari. [http://iris.lib.virginia.edu/tibet/reference/dictionary/thangmi/index.php]
Turner, Ralph Lilley, Sir. 1931. *A comparative and etymological dictionary of the Nepali language*. London: K. Paul, Trench, Trubner. [http://dsal.uchicago.edu/dictionaries]
—. 1962-1966. *A comparative dictionary of Indo-Aryan languages*. London: Oxford University Press. Includes three supplements, published 1969-1985. [http://dsal.uchicago.edu/dictionaries]
Schmidt, Ruth Laila. 1993. *A Practical dictionary of modern Nepali*. Delhi. Ratna Sagar. [http://www.dsal.uchicago.edu/dictionaries/schmidt]
Unicode Consortium. 2003. *The Unicode Standard, version 4.0*. Boston. Addison Wesley. [http://www.unicode.org]
Vishwa Bharat. Technology Development for Indian Languages (TDIL). Ministry of Communications and Information Techonology. Government of India. ISSN 0972-6454.

Sites and abreviations/acronyms of institutions and projects:
DSAL (Digital South Asia Library): http://dsal.uchicago.edu
Gorkhapatra: http://www.gorkhapatra.org.np
IIT (Indian Institute of Technology) Kanpur Nepali site: http://www.nepali.iitk.ac.in
ISCII (Indian Script Code for Information Interchange): http://tdil.mit.gov.in/standards.htm
Kantipur Online: http://www.kantipuronline.com
Lacito Archivage: http://lacito.vjf.cnrs.fr/archivage
LDC (Linguistic Data Consortium): http://www.ldc.upenn.edu
Linguist List: http://linguistlist.org
MPP (Madan Puraskar Library): http://www.mpp.org.np
Nepal Bhasa Dictionary Committee: http://www2.pair.com/webart/mysqllex
Nepalnews: http://www.nepalnews.com
Nepal Research: http://nepalresearch.org/index.htm
OAI (Open Archive Initiative): http://www.openarchives.org
OLAC (Open Language Archives Consortium): http://www.language-archives.org
SCALLA (Sharing Capability in Localisation and Human Language Technologies): http://www.elda.org/rubrique70.html
TDIL Technology Development for Indian Languages: http://tdil.mit.gov.in. Nepali site: http://www.nepali.iitk.ac.in
THDL (Tibetan and Himalayan Digital Library): http://iris.lib.virginia.edu/tibet/
Tower of Babel: http://starling.rinet.ru
Unicode Consortium: http://www.unicode.org
W3C (World Wide Web Consortium): http://www.w3c.org

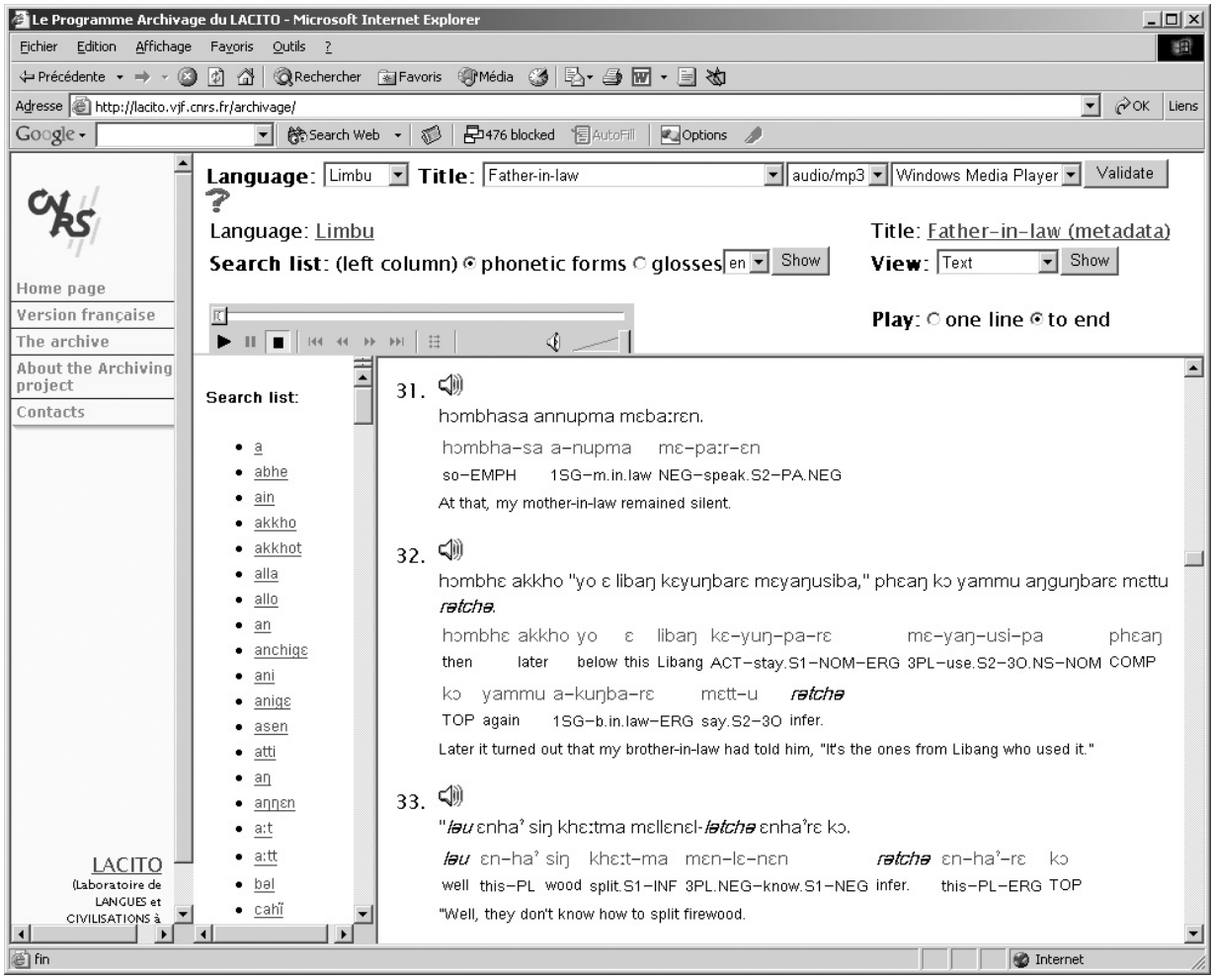

Fig. 1: Interlinear view of a text in the Lacito Archive

NUPPAs98	ləu alla kɔ peg-innɛ lɔˀr-ɛ yammu khunɛ	paŋs	-i yus-igɛ
NUPPAs31	hɔmbha-sa a-nupma mɛ	-paːr	-ɛn
NUPPAs82	na rɔt kɛ-tha-pa lɔˀr-ɛ-aŋ kɔ anigɛ sor cuːk-pa	paːr	-i-aŋ yuŋ-igɛ
NUPPAs93	alla kɔ hara ni phiːŋ-ma-siŋ-aŋ pek-ma him-mu pek-ma phɛaŋ anigɛ kɔ yammu anigɛ s'waʔ	paːr	-igɛ mɛn-khɛm-ɛ khunchi
NUPPAs21	na th'y-ɛ-aŋ kɔ mɛn	-paːt	-ɛ s'waʔ yammu kɛŋs-usigɛ-rɔ cupcuppə yo sɔŋmaŋ yo la-si pe-sigɛ

Fig. 2: Concordance view showing the stem *paːr* and adjoining lines.

```
<?xml version="1.0" encoding="UTF-16"?>
<!-- copyright 1977-2000 B.Michailovsky -->
<!DOCTYPE TEXT SYSTEM "Archive.dtd">
<?xml-stylesheet href="http://lacito.archivage.vjf.cnrs.fr/archives/styles/default.xsl" type="text/xsl"?>
<TEXT id="LIF:NUPPA" xml:lang="x-sil-LIF">
        <HEADER>
                <TITLE xml:lang="en">Father-in-law</TITLE>
                <TITLE xml:lang="fr">Beau-père</TITLE>
                <SOUNDFILE href="Nepal/Limbu/NUPPA.wav"/>
        </HEADER>
        <S xml:lang="x-sil-LIF" id="NUPPAs31">
                <AUDIO start="148.5400" end="150.6601"/>
                <FORM kindOf="phono">hɔmbhasa annupma mɛba:rɛn.</FORM>
                <TRANSL xml:lang="en">At that, my mother-in-law remained silent.</TRANSL>
                <TRANSL xml:lang="fr">Du coup, ma belle-mère n'a plus rien dit.</TRANSL>
                        <W>
                                <M class="misc">
                                        <FORM kindOf="phono">hɔmbha</FORM>
                                        <TRANSL xml:lang="en">so</TRANSL>
                                </M>
                                <M class="misc">
                                        <FORM kindOf="phono">sa</FORM>
                                        <TRANSL xml:lang="en">EMPH</TRANSL>
                                </M>
                        </W>
                        <W>
                                <M class="misc">
                                        <FORM kindOf="phono">a</FORM>
                                        <TRANSL xml:lang="en">1SG</TRANSL>
                                </M>
                                <M class="misc">
                                        <FORM kindOf="phono">nupma</FORM>
                                        <TRANSL xml:lang="en">m.in.law</TRANSL>
                                </M>
                        </W>
                        <W>
                                <M class="vprefix">
                                        <FORM kindOf="phono">mɛ</FORM>
                                        <TRANSL xml:lang="en">NEG</TRANSL>
                                </M>
                                <M class="v" sclass="pastem">
                                        <FORM kindOf="phono">pa:r</FORM>
                                        <TRANSL xml:lang="en">speak.S2</TRANSL>
                                </M>
                                <M class="vsuffix">
                                        <FORM kindOf="phono">ɛn</FORM>
                                        <TRANSL xml:lang="en">PA.NEG</TRANSL>
                                </M>
                        </W>
        </S>
        </TEXT>
```

Figure 3 : XML source for “Father-in-Law” utterance 31 (see fig. 1).