

Searchable speech/text corpora: a resource for spoken language

Boyd Michailovsky, LACITO/CNRS, Villejuif, France

Introduction

How can information technology adapt to less well-known, mainly unwritten languages, and contribute to meeting the needs and priorities of their speakers? As a linguist and not a technologist, I will attempt to address one aspect of this question, that of linking recorded speech to a searchable transcription. I will present techniques which have been developed for the linguistic study of spoken language and oral tradition, based on (1) a structured document architecture in which transcriptions are synchronized with the digitalized recordings of spontaneous speech and (2) implementations of standard tools of corpus linguistics giving simultaneous access to sound and text in response to queries. The system is implemented on a web site accessible to standard browsers.

Although the techniques described here can be applied to the study of material in any language, they are especially valuable in the study of unwritten, or little-written languages. Linguists' transcriptions of such languages are, at least in the initial stages, work in progress, subject to revision as the phonological and morphosyntactic hypotheses on which they are founded are revised; they benefit from constant confrontation with the recorded speech. For speakers of the languages, it is the spoken word and not the transcription, however well-designed, that carries authority in the transmission of oral traditions.

The present system is used by linguists, as an aid to linguistic research, on languages of Nepal among others, and in a museum media center, as a means of providing access to a collection of recorded and transcribed oral literature. It relies entirely on generic, open-source technology.

The LACITO Archive

The techniques presented have been developed at the LACITO (Oral Tradition: Languages and Civilizations), a research group of the French National Center for Scientific Research (C.N.R.S.) specializing in linguistic and anthropological field research. The LACITO Archive project was initiated with the goal of conserving linguistic documentation, both recordings and transcriptions made in the field, and making it available to researchers and others in a format in which (1) the recordings are synchronized with the annotation and (2) the data can be accessed using modern research tools.

The documents produced by the project are used at present in two ways. First, some 150 documents have been put on an intranet server, and 38 of them on a publicly accessible Internet server as well, where they are used by researchers -- mainly those who recorded and transcribed them in the first place -- for linguistic research and

lexicography. Secondly, about 100 of these documents, comprising corpora in a dozen languages of New Caledonia (a French territory in the Pacific) are accessible on an intranet at the multimedia library of the Tjibaou Cultural Center in Nouméa, where they are consulted by visitors including speakers of the languages.

The South Asian part of the archive consists currently of 10 mainly autobiographical narratives in Limbu (a Tibeto-Burman language with about 150.000 speakers in eastern Nepal and the neighboring Indian states of West Bengal and Sikkim) and three texts in Hayu (a Tibeto-Burman language with 200 speakers in Nepal).

The structure of the system

The documents, which include Unicode-coded IPA-based transcriptions and word- (or morpheme-) and sentence-level translations (into English, French, or both) are structured in XML. Associated with each sentence is an element tagged <AUDIO> indicating the location of the corresponding sound in a sound resource ([example](#).)

In accessing the archive, the user chooses a corpus (by language) and a document to browse. He then chooses one or more of the available "views" for the chosen document: transcription (all that is needed for speakers of the languages), transcription and free translation (English ([example](#)) or French), or transcription with both interlinear and free translations. In fact, the user passes parameters to an XSL stylesheet which is run on the server and generates HTML which is returned to the client browser. The chosen annotation appears on the user's screen, segmented into sentences.

If a sentence in the display is clicked on, the corresponding sound is played. Alternatively, the user may choose to listen to the whole recording as the annotation scrolls past on the screen, with each sentence highlighted in turn. These are the basic browsing interfaces. (The links in this document are to inert screen images -- to access the same data with the sound, please [connect](#) to the archive website.)

For research or language-learning purposes, the possibility of searching for all sentences containing a particular word or morpheme is provided. When this option is chosen, an alphabetical list of all words in the document is generated (once again by an XSL stylesheet) and displayed. This makes it possible for the user to choose an item without having to enter it on the keyboard (a problem when Unicoded phonetic characters are present). When a word is chosen (cok 'to do', in this [example](#)), each sentence in which it appears is displayed, with both free and interlinearized translations if available. Again, if one of the sentences is clicked on, the recorded sound is played.

The computer coding used for Limbu isolates morphemes -- that is, the meaningful parts of complex words -- and not only whole words as in some of our early codings. That is how mEcogu "they did it" and acogumba "which we did" can be identified as forms of the same verb cog- "to do" in the [example](#) above. Later we will take advantage of this coding to link the text to a computerized dictionary. ([Example](#) of coding segmented into S[entences] (roughly), W[ords] and M[orphemes]). In the Limbu documents, the sentence-level transcription is phonological, while the

morpheme-level transcription is morphophonemic, that is, each morpheme always has the same transcription, regardless of immediate context.)

Another option provided is concordancing. The concordance of the document being browsed (again defined by a stylesheet) is realized on the server -- not something one could do with the British National Corpus -- and returned to the client. Since it is generally more useful to have a concordance of a whole corpus than of a single text, I show [here](#) a (working, in principle) sample of the concordance of all the verb stems (actually past stems) appearing in all 10 Limbu texts in the corpus -- a concordance that I use in my research. The verb stems appear in blue, in the center column, in alphabetical order, surrounded by the context for each sentence in which they appear. We see that the verb ab- "shoot" appears twice, akt- 'serve (beer)' once, a:tt- "cut (something vertical)" five times, etc. At the far left is a reference to the document and the sentence, followed by the morpheme translation if available. Once again, if one clicks on a word in the center column, the recording of the sentence in which it occurs is heard.

Word searches and concordances are standard tools used in the study of linguistic or literary corpora. What is new here is the "talking concordance", which gives access to the sound recording at the same time as to the transcription and other textual material. It is easy to imagine the value of this for checking the accuracy of the transcription and the coherence of the annotation, for studying the meanings or uses of a particular word, or for finding example sentences for a dictionary. More detailed searches adapted to a particular corpus can be programmed in XSL.

Technical notes

The whole system has been implemented using open-source "plain vanilla" non-proprietary software. The data architecture is structured text, marked-up in standard XML according to a format defined in a [DTD](#). The user interface and all processing of the XML data is defined in XSL stylesheets. These are applied by an XSL processor, which is called by a servlet in response to client input. The servlet returns the generated HTML to the client. The only requirement on the client side is a standard browser, a Unicode font (PDF output with incorporated fonts is offered for Macintosh clients), and a standard multimedia player. All of the software developed at the project is open-source and freely available (together with the data) on the website. This software includes stylesheets, the servlet, an applet, CGI scripts which make it possible to extract parts of a sound resource, and a stand-alone program, SoundIndex, which aids in the time-alignment of an XML document with a sound resource ([screen](#)).

Conclusion

I have presented an architecture and tools which have served in the development of resources such as dictionaries and grammars of lesser-known languages. The system was designed primarily with the needs of linguists in mind, and it has proved useful in the development of transcriptions, grammars and dictionaries. But access to recorded speech as a searchable corpus can be of interest to a far wider community of users, not least to the speakers of unwritten languages. The system currently serves as the basis

of an oral literature archive which is accessible to speakers of New Caledonian languages and other visitors in a multimedia library at the Tjibaou Cultural Center in Nouméa, New Caledonia. This can be seen as an example of the adaptation of information technology to the needs of language communities. In New Caledonia, the functions served are the preservation and transmission of oral tradition, and the diffusion of newly developed orthographies for unwritten languages. These can certainly be transposed into the South Asian context.

NOTE:

The LACITO public web archive is accessible at the address <http://lacito.archivage.vjf.cnrs.fr>. Users with low bandwidth should first try the short Hayu text "CREDO". The markup of the Limbu texts is somewhat better developed. Complete information about the system is available on the website.

[Boyd Michailovsky](#), coordinator
[Michel Jacobson](#), computer specialist

The LACITO Archive
LACITO/CNRS
7 rue Guy Môquet
94800 Villejuif
FRANCE
<http://lacito.archivage.vjf.cnrs.fr>