Documenting and researching endangered languages: the Pangloss Collection

Boyd Michailovsky, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François, Evangelia Adamou

LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 8 (2014), pp. 119-135
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/4621

Documenting and Researching Endangered Languages: The Pangloss Collection

Boyd Michailovsky[1], Martine Mazaudon[1], Alexis Michaud[1,2], Séverine Guillaume[1], Alexandre François[1,3], Evangelia Adamou[1]
CNRS-LACITO[1], MICA Institute (HUST-CNRS/UMI2954-Grenoble INP)[2], Australian National University[3]

The Pangloss Collection is a language archive developed since 1994 at the Langues et Civilisations à Tradition Orale (LACITO) research group of the French Centre National de la Recherche Scientifique (CNRS). It contributes to the documentation and study of the world's languages by providing free access to documents of connected, spontaneous speech, mostly in endangered or under-resourced languages, recorded in their cultural context and transcribed in consultation with native speakers. The Collection is an Open Archive containing media files (recordings), text annotations, and metadata; it currently contains over 1,400 recordings in 70 languages, including more than 400 transcribed and annotated documents. The annotations consist of transcription, free translation in English, French and/or other languages, and, in many cases, word or morpheme glosses; they are time-aligned with the recordings, usually at the utterance level. A web interface makes these annotations accessible online in an interlinear display format, in synchrony with the sound, using any standard browser. The structure of the XML documents makes them accessible to searching and indexing, always preserving the links to the recordings. Long-term preservation is guaranteed through a partnership with a digital archive. A guiding principle of the Pangloss Collection is that a close association between documentation and research is highly profitable to both. This article presents the collections currently available; it also aims to convey a sense of the range of possibilities they offer to the scientific and speaker communities and to the general public.

1. INTRODUCTION.[1] The Pangloss Collection (http://lacito.vjf.cnrs.fr/pangloss/index_en.htm) is a language archive developed at the Langues et Civilisations à Tradition Orale (LACITO) research group of the French Centre National de la Recherche Scientifique (CNRS). The goal of this archive is to preserve and disseminate recorded and transcribed oral literature and other linguistic materials in (mainly) endangered or poorly documented languages, giving simultaneous access to sound recordings and text annotation.

[1] The authors are grateful to the various structures whose support makes this project possible (CNRS-InSHS, LACITO, Humanités Numériques (Huma-Num; formerly Adonis), CINES and CC-IN2P3), and to Bernard Bel and two anonymous reviewers for useful comments on a draft version. Many thanks to Jean-Michel Roynard for his help with editorial matters, and to the *Language Documentation and Conservation* editorial team and the anonymous reviewers for felicitous suggestions. Special thanks are also due to Anne Behaghel-Dindorf for help in the design and maintenance of the web interface, and to the various contributors to the Pangloss Collection for their participation, feedback and encouragement. This work is related to the research projects EuroSlav (ANR-09-FASHS-025 & DFG-BR 1228/4-1) and HimalCo (ANR-12-corp-0006), and to the research strand 'Phonetics and Phonology' of the Paris-based LabEx *Empirical Foundations of Linguistics* (funded by the ANR/CGI).

E-ISSN 1934-5275

The necessity to document the world's languages is now well known to linguists and the general public. Fewer people are aware of the dismal current state of multimedia linguistic documentation. Looking back at a century of speech recording, the legacy is not as extensive – and nowhere near as tidy – as the layman would think. If it is true, as Whalen (2004) puts it, that "the study of endangered languages will revolutionize linguistics" (p. 321), and that "the vanguard of the revolution will be those who study endangered languages" (p. 340), then it is all the more unfortunate that "enormous amounts of data – often the only information we have on disappearing languages – remain inaccessible both to the language community itself, and to ongoing linguistic research" (Thieberger & Nordlinger 2006; see also Woodbury 2003, 2011).

In recent years, a number of archives have been created to address this major need of the linguistic community – including the Academia Sinica Collections in Taipei, the Archive of the Indigenous Languages of Latin America (AILLA) at the University of Texas in Austin, the Endangered Languages ARchive (ELAR) at the School of Oriental and African Studies (SOAS) in London, the Language Archive at the Max Planck Institute for Psycholinguistics in Nijmegen, and the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), jointly run through the Australian National University in Canberra, the University of Melbourne, and the University of Sydney. The Open Language Archives Community (OLAC) and the Language Archive at the Max Planck Institute for Psycholinguistics list many other repositories. Each of these archives makes a contribution to the world-scale effort of documenting the diversity of spoken languages. The LACITO in Paris began to take part in this collective endeavor as early as 1994, and an Open Archive of audio recordings and texts in endangered languages went online in 2001 (Jacobson, Michailovsky & Lowe 2001). This archive, now known as the *Pangloss Collection* (etymologically 'all languages'), continues to contribute to knowledge of endangered languages and cultures, by sharing annotated spoken texts of lesser-studied languages. Compared to other existing archives, the main original component of Pangloss is its easy and open access to media resources, as well as its emphasis on text transcription and annotation as a tool for further research. This article presents the collections currently available; it also aims to convey a sense of the range of possibilities they offer to scientific and speaker communities as well as to the general public.

2. THE PRESENT STATE OF THE COLLECTION

2.0 GENERAL PRINCIPLES. The Pangloss Collection started as an answer to the increasing need to archive and access the recordings of the LACITO researchers and their colleagues. More recently, externally funded national and international research programs involving LACITO researchers have also become associated with the Collection. Pangloss welcomes audio recordings on endangered languages regardless of their geographic or institutional origin. It has no *a priori* goals or policy regarding material to be archived, beyond a general expectation of scholarly and (minimal) technical quality. Our initial policy was to accept only audio files that were fully transcribed and annotated, but limiting online collections to fully annotated documents slowed down the process of archiving the original recordings safely and making them available. So while providing annotation remains the ultimate goal, we now also accept deposits of recordings accompanied only by metadata (cataloguing information on the speakers, the researchers, the content, date and place, etc.).

The data collectors/depositors retain rights to the materials, and may receive advice and some direct help in preparing their data, but Pangloss itself has no funding for fieldwork or the transcription and annotation of data. This is provided by research institutions or granting agencies, either through programs specifically directed at documentation or, increasingly, by insisting that all research programs provide for conservation of field data and access to it. Pangloss can provide expertise in digitizing obsolescent media like quarter-inch reel-to-reel magnetic tape and compact cassettes and for capturing minidiscs (which are already digital) as well as access to the necessary equipment. As of 2014, the Pangloss Collection has become a sizeable collection of more than 1,400 recordings in 70 languages, totalling 193 hours; more than 400 of these recordings (about 60 hours) have a full transcription and annotation. Figure 1 shows the English version of the main web page giving access to the Collection. The survey presented below focuses on the geographical areas that constitute the main strengths of the Pangloss Collection, providing some detail on the history of these data sets.

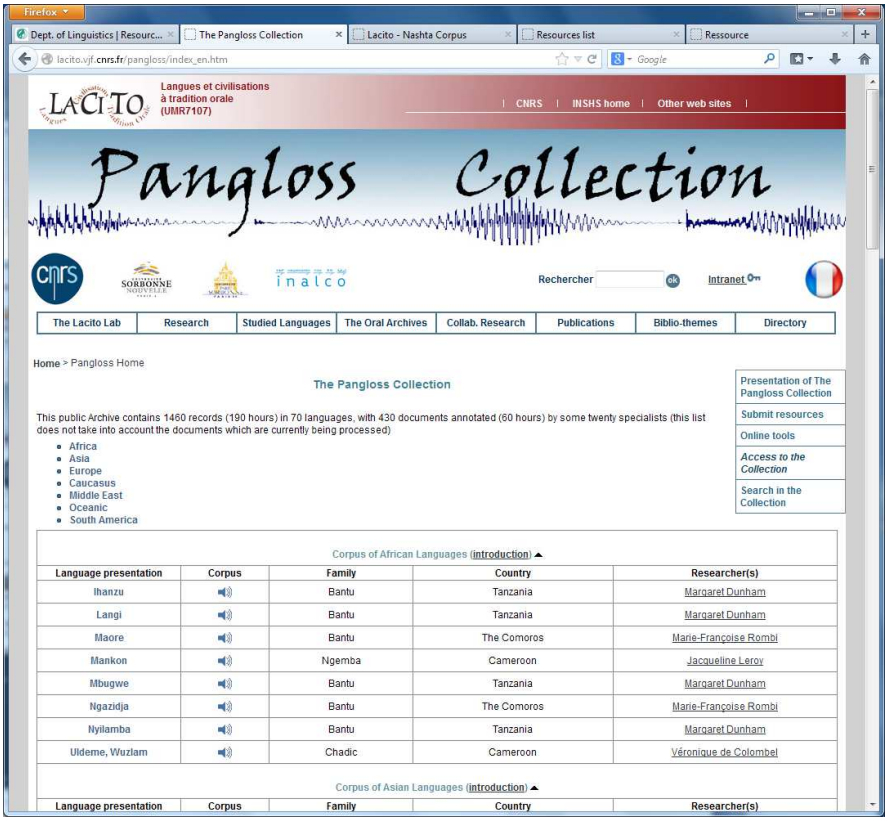

FIGURE 1. Web interface of the Pangloss Collection

2.1 ASIAN LANGUAGES. Languages of Nepal are especially well-represented in the Pangloss Collection, offering access to recordings in Hayu, Limbu, Tamang, Bahing and Nepali by Boyd Michailovsky and Martine Mazaudon, as well as Koyi Rai and Thulung Rai data collected by Aimée Lahaussois. The original 'Languages of Nepal' section of the archive has recently expanded into an 'Asian languages' section (http://lacito.vjf.cnrs.fr/pangloss/index_en.htm#asia). The recordings featured there were collected in India, China, Vietnam and Thailand, and include Chang Naga, Japhug (Rgyalrong), and Prinmi (Pumi) (Guillaume Jacques); Naxi, Na, Laze (Alexis Michaud); and Vietnamese (Thi-Lan Nguyen).

2.2. OCEANIC LANGUAGES. Oceanic languages have been another major component of the collection since its beginnings (http://lacito.vjf.cnrs.fr/pangloss/index_en.htm#oceanic). The first of these annotated and translated documents belonged to the rich set of indigenous languages of New Caledonia, also known as *Kanak* languages, including Ajië, Bwatoo, Cèmuhî, Drehu, Iaai, Kwênyïï, Nêlêmwa, Nemi, Numèè, Paicî, Pije, West Uvean, Xârâcùù, Xârâgurè, and Yuanga/Zuanga. These recordings were collected, transcribed, and annotated by Françoise Ozanne-Rivierre, Jean-Claude Rivierre, Jacqueline de La Fontinelle, Claire Moyse-Faurie, and Isabelle Bril. Moyse-Faurie has archived documents in languages of Wallis and Futuna (East Futunan and East Uvean).

The Oceanic section also holds Alexandre François' recordings of four languages of the Solomon Islands (Lovono, Tanema, Teanu, Tikopia) and nineteen languages of Vanuatu (Araki, Bislama, Dorig, Hiw, Koro, Lakon, Lehali, Lemerig, Lo-Toga, Löyöp, Mota, Mwerlap, Mwesen, Mwotlap, Nume, Olrat, Vera'a, Volow, and Vurës). Of these 23 languages, seven (Araki, Lemerig, Lovono, Mwesen, Olrat, Tanema, Volow) are moribund – with between one and ten speakers for each – and may not survive the current generation. The other languages are more vital today, but may be soon threatened by the rapid expansion of Bislama, the national creole of Vanuatu (François 2012). François' archived recordings total about 93 hours of sound – including 23 hours of musical recordings and 70 of spontaneous speech (oral literature, ethnographical or procedural explanations, life histories, interviews, conversations).

2.3. LANGUAGES OF THE CAUCASUS. The best-represented language of the Caucasus in the Pangloss Collection is Ubykh, studied notably by the Caucasologist Georges Dumézil (1898-1986). This language achieved a measure of celebrity as the poster-child of endangered languages before the demise of the last speaker in 1992. Its rich consonant inventory is legendary among linguists (see in particular Vogt 1963; Dumézil 1965; Charachidzé 1989; and Colarusso 1994). In 1968 Dumézil made ten tapes, including wordlists and seven unpublished stories, with the last speaker, Mr. Tevfic Esenç (or Tevfic Saniç) in Paris. These were digitalized at LACITO in 2003 by Alexis Michaud, who undertook to locate, catalogue, and digitize Parisian Ubykh materials as part of a pre-doctoral project on language documentation and archiving. The tapes included recordings of seven short narratives. Transcriptions and word-by-word French glosses in Dumézil's own handwriting were found for four of these and initially made available in the Pangloss Collection as scanned images.

In March 2008, Brian Fell, a student of linguistics working on Ubykh, found these materials on the web and proposed to prepare transcriptions and English translations of all

seven stories for us, an unexpected offer we were glad to accept; his annotations of the four stories for which we had Dumézil's transcriptions are now available online. During a visit to Paris in 2010 at our invitation, Fell helped us complete the annotation of Dumézil's recordings of Ubykh vocabulary, elicited (usually in Turkish) to illustrate Ubykh phonology and verify the perception of some of the phonemic oppositions. These were computerized and time-aligned with the recordings by Tanguy Solliec in 2010. This shows how the free access to data online can lead to its enrichment by new collaborators.

The Pangloss Collection also hosts documents in Abzakh, Bjedug and Shapsug collected by Catherine Paris and digitized and formatted at LACITO after her death by her former consultant, Dina Dabjen-Bailly, as well as new recordings of words and sentences in Bjedug (Adyge) by the same consultant. The latter are accompanied by an electroglottographic signal, which is especially useful for the study of the glottalized consonants, which are abundant in North Caucasian languages.

The most recent addition to the Caucasus Languages section is a tale in Arhavi Laz collected in 2007 by René Lacroix, who is currently completing a documentation program on Laz (sponsored by the Hans Rausing Endangered Languages Project) with more than 250 hours of recordings.

2.4. EUROPEAN LANGUAGES. The European languages documented in the Pangloss Collection are Romani (Indo-Aryan) spoken in Greece (collected by Evangelia Adamou) and a set of endangered Slavic language varieties found in non-Slavic-speaking European countries: Burgenland Croatian (Austria) collected by Lenka Scholze and Maria Utschitel; Colloquial Upper Sorbian (Germany) collected by Lenka Scholze; three dialects of Nanašu (Italy) collected by Walter Breu; two dialects of Balkan Slavic collected in Greece by Evangelia Adamou in 2003-2006 (Nashta) and by Georges Drettas in 1976 (Bulgarian-Macedonian). Each variety is represented by about one hour of recordings or approximately 5,000 words, which were elaborated for the Pangloss Collection with funding from the Agence Nationale de la Recherche (ANR) in France and the Deutsche Forschungsgemeinschaft (DFG) in Germany.

2.5 OTHER LANGUAGES IN THE ARCHIVE. The Pangloss collection also hosts language data from other parts of the world:

- ◊ languages of Africa: Ihanzu, Langi, Mbugwe, Nyilamba (Tanzania; coll. Margaret Dunham); Maore, Ngazidja (Comoro Islands; coll. Marie-Françoise Rombi); Mankon (Cameroon; coll. Jacqueline Leroy); and Uldeme/Wuzlam (Chad; coll. Véronique de Colombel);
- ◊ languages of the Middle East: Yemeni Arabic (coll. Samia Naïm);
- ◊ languages of South America: Wayana (French Guyane, coll. Hervé Rivière); Yucuna (Colombia, coll. Laurent Fontaine).

We refer readers to the Pangloss Collection for details about these data sets.

3. DATA STRUCTURE AND INTERFACE

3.0 GENERAL PRINCIPLES. The LACITO Archive was conceived in the mid-1990s, as the web revolution was taking place. An American collaborator, J. B. Lowe, alerted Martine Mazaudon and Boyd Michailovsky of LACITO to the existence of the Web and to the possibility of integrating sound and text, and developed our first prototypes. With his help, and that of another computer engineer, Michel Jacobson, we adopted the new technologies and standards as they appeared: Unicode, HTML, XML, XSLT, Dublin Core Metadata, the Open Archives framework (Jacobson et al. 2001). The current LACITO engineer, Séverine Guillaume, has continued this process (e.g. with HTML5) in developing the Pangloss collection (Michailovsky et al. 2011).

The XML format that we developed for continuous speech, and which became de facto the LACITO (and now Pangloss) format, is essentially a representation of the logical structure underlying traditional linguistic interlinear text publication, of which Boas was a celebrated early exponent (e.g. Boas 1902). It provides for a basic document type, TEXT, which may be divided into S(entences or utterances), W(ords), and M(orphemes), down to the desired granularity, in a hierarchical structure. Each level may contain transcriptions (FORM elements: phonetic, phonological, and/or orthographic), TRANSL(ations) into different languages, NOTE(s), and elements of the next lower level. Each level also may contain a unique ID(entifier) and an AUDIO element indicating its beginning and end points (in seconds) in a media file. In practice, for continuous speech, we have provided time-alignment data only for S elements.

The Pangloss system was designed from the start for a web interface, to be accessed using standard browsers. The interface on the Pangloss website allows users to choose which elements of the markup they wish to see, and then composes an HTML document which is displayed in the user's browser window in traditional interlinear format. Simultaneously, the user can choose to hear the sound corresponding to a single element, or to the rest of the text. Since our adoption of HTML5, this requires no software installation on the part of the user.

The data structure has immediate appeal to linguists because of its homology with traditional interlinear text representation. But the structure is logical, not typographical: the function of each element is explicitly defined, not just inferred from its position on a page. Thus it is accessible for searching, indexing, concordance-making, etc., using computer tools. (We currently use our own XSLT scripts for these purposes off-line; a few examples are posted on the site. We are developing online versions.) To the extent that elements returned in response to queries (or their containing elements) are time-aligned with a recording, any such response (e.g. a line in a concordance) can include access to the corresponding sound.

Our XML markup (defined by a DTD/Schema) also provides for a WORDLIST document type in parallel with the TEXT type, and for an ARCHIVE document type made up of TEXT elements that one wishes to process together, as in making a concordance.

From the start, we made a point of not proposing a 'LACITO markup,' because we expected that users with different research interests would demand different, incompatible modifications and additions. But in fact, after more than a decade, such demands have been very few, and we have usually been able to accommodate them by adding XML 'attributes'

to existing elements: for example, adding an optional attribute to indicate the part of speech of a W or M (e.g. <M class="verbstem">). In spite of its theoretical limitations (Bird and Liberman 2001), the basic, single-hierarchy, text/utterance/word/morpheme structure has allowed us — and many other researchers and research groups — to archive large amounts of data and to develop straightforward web interfaces. We believe that such markup and interfaces can also serve as display vehicles for the corresponding elements of a globally more complex or multi-hierarchical markup.

3.1 XML DATA STRUCTURE: EXAMPLES. As mentioned above, the annotation structure reflects traditional practice in interlinear glossing, which will be familiar to linguists. In linguistic publications, this structure is implicit in the typographical format, as in (1) below.

(1)
(…) dy˥-ku˩lu˩ | mæ˧ʏ˥ ɖɯ˧-ʝi˧ lɑ˧-ʂɯ˧ ||

dy˥	ku˩lu˥	mæ˧ʏ˥	ɖɯ˧	ʝi˧	lɑ	ʂɯ˧
地	里面	女婿	一	量词	实施	寻找
earth	inside	son_in_law	one	CL	ACCOMP	look_for

[王母娘娘] 在地（天下）找了一个女婿。
'[The Heavenly Mother] looked for a son-in-law down on earth.'

The basic annotation is the transcription of the original language (here: Laze, a Sino-Tibetan language). The transcription of an entire text is divided into sentences, and sentences into words. Translations into any number of languages (here: Mandarin and English) may be provided, aligned with the transcription at different levels: glosses for words or morphemes, and free translations at the sentence and text levels. Many annotators are guided by the detailed glossing conventions proposed by Bickel et al. (2008) and Lehmann (2004). Multiple transcriptions (e.g. phonetic, phonological, morphophonemic, orthographic) can also be shown, either on a single level or on different levels. In example (1), the word-by-word transcription makes it clear that /lɑ/ (glossed as ACCOMPLISHED marker) does not have a lexical tone of its own, unlike the other words in the excerpt. In the sentence context, it surfaces with a Mid tone (indicated by the IPA mark for Mid tone : ˧). Also, the postposition /ku˥lu˥/ 'inside', which has a lexical High tone on both syllables, surfaces with a changed tone (Low tone) in the sentence transcription: /ku˩lu˩/.

The (abridged) XML markup corresponding to example (1) is shown in (2). The language of the text is an attribute of the root element TEXT and is inherited by default in S and W elements. (French translations have been omitted from both (1) and (2) to save space.)

(2)

```
<S id="FemmeCelesteS3">
<AUDIO start="4.2" end="6.9"/>
<FORM>dy˥-kuɹluɹ, mæ˦y˥ ɖɯ˦-ji˦ lɑ˦-ʂɯ˦.</FORM>
    <TRANSL xml:lang="cn">在地（天下）找了一个女婿。</TRANSL>
    <TRANSL xml:lang="en">[The Heavenly Mother] looked for a son-in-law down on
earth.</TRANSL>
    <W>
        <FORM>dy˥</FORM>
        <TRANSL xml:lang="cn">地</TRANSL>
        <TRANSL xml:lang="en">earth</TRANSL>
    </W>
    <W>
        <FORM>ku˥lu˥</FORM>
        <TRANSL xml:lang="cn">里面</TRANSL>
        <TRANSL xml:lang="en">inside</TRANSL>
    </W>
    …
</S>
```

The time alignment markup in the AUDIO element indicates that the containing segment S3 is pronounced between 4.2 seconds and 6.9 seconds into the recording. It is used to play the sound when the user requests it through the web interface. Sound recordings can be segmented as finely or as coarsely as desired. The main decision to be made in aligning sound and text annotation is the *granularity*, that is, the length of the smallest text elements to be time-anchored, and hence the length of the smallest segments of the sound resource which can be accessed. Since the documents under discussion here consist of connected text, it was decided to anchor units longer than words or phonemes.

A second example, from Limbu (Tibeto-Burman, Nepal) shows slightly different markup possibilities underlying the interlinear display of a conversation, with glossing at the morpheme level.

(3)
H: attiha^{ʔ}rɛ cəĩ kɛmɛmmɛttinni nurik?

atti-ha^{ʔ}-rɛ	cəĩ	kɛmɛn-mɛtt-in-i	nurik?
which.one-PL-ERG	TOP	3NSG>2.NEG-do.S2-12PL.SO.NEG-Q	well

'and which ones didn't treat you well?'

The source markup shows the S-level attribute "who" identifying the speaker; we do not have a structural speaker-turn or paragraph level. Included in the markup, but not

displayed, are the "kindOf" attribute to indicate the kind of transcription, the "xml:lang" attribute used to indicate the source of a loan word, and the "class" and "s[ub]class" attributes used to indicate word-classes (for verb stems and affix-strings only in this case).

(4)

```
<S xml:lang="x-sil-LIF" id="DANCEs11" who="H">
  <AUDIO start="60.8600" end="67.5999"/>
  <FORM kindOf="phono">attihaʔrɛ cəĩ kɛmɛmmɛttinni nurik?</FORM>
  <TRANSL xml:lang="en">and which ones didn't treat you well?</TRANSL>
  <W>
    <M class="misc">
      <FORM kindOf="phono">atti</FORM>
      <TRANSL xml:lang="en">which.one</TRANSL>
    </M>
    <M class="misc">
      <FORM kindOf="phono">haʔ</FORM>
      <TRANSL xml:lang="en">PL</TRANSL>
    </M>
    <M class="postposition">
      <FORM kindOf="phono">rɛ</FORM>
      <TRANSL xml:lang="en">ERG</TRANSL>
    </M>
  </W>
  <W>
    <M xml:lang="ne">
      <FORM kindOf="phono">cəĩ</FORM>
      <TRANSL xml:lang="en">TOP</TRANSL>
    </M>
  </W>
  <W>
    <M class="vprefix">
      <FORM kindOf="phono">kɛmɛn</FORM>
      <TRANSL xml:lang="en">3NSG&amp;gt;2.NEG</TRANSL>
    </M>
    <M class="v" sclass="pastem">
      <FORM kindOf="phono">mɛtt</FORM>
      <TRANSL xml:lang="en">do.S2</TRANSL>
    </M>
    <M class="vsuffix">
      <FORM kindOf="phono">in</FORM>
      <TRANSL xml:lang="en">12PL.NAGT.NEG</TRANSL>
    </M>
    <M class="misc">
      <FORM kindOf="phono">i</FORM>
      <TRANSL xml:lang="en">Q</TRANSL>
    </M>
  </W>
  <W>
    <M class="misc">
      <FORM kindOf="phono">nurik</FORM>
      <TRANSL xml:lang="en">well</TRANSL>
    </M>
  </W>
</S>
```

One feature of our XML markup is that it is relatively human-readable. Readers are invited to think of improvements.

3.2. METADATA. Metadata provide cataloguing information for each dataset in the archive: the language, the subject matter, the time and place of creation, the participants involved, the technical characteristics, links to other datasets (e.g. between a recording and its annotation), etc. Metadata make it possible for a dataset to be located reliably and precisely among all datasets available on the web, for consultation by users or referencing by search engines. Our metadata is also used locally in managing the archive and web interface.

The Pangloss Collection has adopted the metadata standard defined by the Open Language Archives Community (OLAC) (http://www.language-archives.org/) for linguistic documents of all kinds. The OLAC standard is a domain-specific adaptation of the most widely used general metadata standard for digital documents, the Dublin Core (DC) (http://dublincore.org/). As an example, where the DC category, 'participant,' could be used to identify the speaker or the annotator or any other participant in the creation of a linguistic resource, the OLAC standard defines a fixed number of 'participant' roles ('speaker,' 'researcher,' 'depositor,' 'annotator,' etc.) agreed upon by members of the OLAC community, while remaining within the DC norm.

The Pangloss metadata are stored in an 'Open Archive' (as defined by the Open Archives Initiative (OAI)) called Collection de Corpus Oraux Numériques (CoCoON); i.e. 'Collection of Digital Audio Corpora,' housed by the CNRS Humanités Numériques (HUMA-NUM) program (http://cocoon.tge-adonis.fr/). As a 'data provider' CoCoON presents OLAC metadata on the web in a form compatible with the OAI Protocol for Metadata Harvesting. The metadata from archives around the OAI world are 'harvested' by 'service providers.' For example, OLAC harvests metadata from all OLAC-registered linguistic archives and makes available online a consolidated catalogue of linguistic documents.

4. SOFTWARE TOOLS AND IMPLEMENTATION

4.1. AUTHORING TOOLS. There are several solutions for preparing a text transcription/annotation in XML format.

(i) Users who are familiar with writing scripts can type the text and interlinear glosses as plain text (for example, with sentences separated by carriage-return, words separated by spaces, and morphemes by hyphens, roughly as seen in (3) above) and then run a script that adds the XML markup. Perl scripts are available from the Pangloss Collection's 'Tools' page.

(ii) Users of Toolbox, ELAN or other authoring software can convert annotations to the Pangloss document format by means of scripts.

(iii) The authoring tool Interlinear Text Editor (ITE), developed by Michel Jacobson at LACITO, can be used for composing a complete annotation (other than sound alignment) in the Pangloss format, or for tokenizing and glossing an existing transcription. It still has users, but it is no longer maintained. Many of our depositors use Toolbox and ELAN.

(iv) SoundIndex was developed by Michel Jacobson as a tool for time-aligning XML-formatted text annotation. Pangloss AUDIO elements are inserted into the

XML. Although the process of time-alignment might seem tedious, field linguists who use SoundIndex generally profit from it to improve their transcriptions.

4.2. BROWSING. Pangloss documents, including annotations with synchronized sound, are consulted on the website using a standard browser. Direct access to most resources is through the menu item 'Archive access,' which lists the languages in which documents are available (Figure 1). Once a language is selected (by clicking on the loudpeaker icon in the 'Archive' column), an index of the texts available in the language is displayed. Icons on each line give access either to the sound only (loudspeaker), to metadata (an encircled 'i'), or to the annotation with the sound (scroll+loudspeaker). The partial metadata display is illustrated in Figure 2. The XML and WAV icons in the display are linked to data files available for downloading.

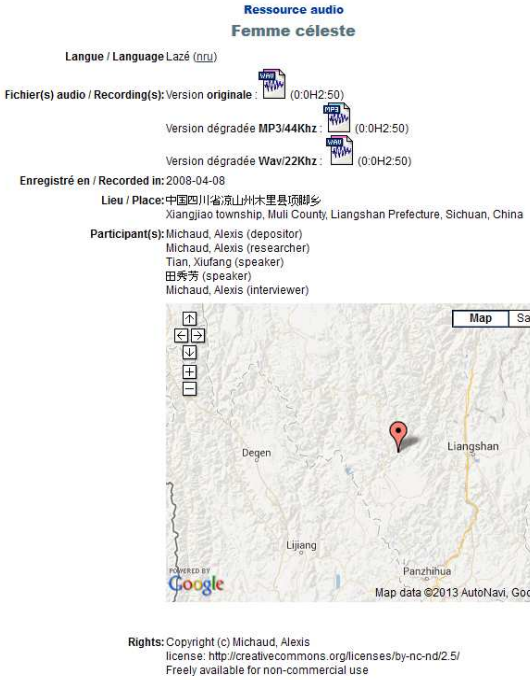

FIGURE 2. A view on the metadata of the recording corresponding to annotation (1) above

Clicking on the scroll+loudspeaker icon brings up an HTML page generated from the XML annotation document, showing text annotations and a HTML5 audio player (Figure 3). Through the use of stop and play buttons placed before each sentence, the user can choose to listen to one sentence at a time or to the remainder of the text. (This choice is made at the top of the page: 'continuous play.') The sentence currently being played is highlighted.

Text annotations can include many kinds of information: several transcriptions (phonetic, phonological, orthographic, etc.), translations in any number of languages, glosses

for words and morphemes, and notes. The interface offers the option of hiding unneeded parts of the annotations. If there is more than one translation, the user can choose which one(s) he or she wishes to see. When glosses are provided in several languages, each one appears on a separate line. Figure 3 shows the HTML display generated from a document in Balkan Slavic Nashta.

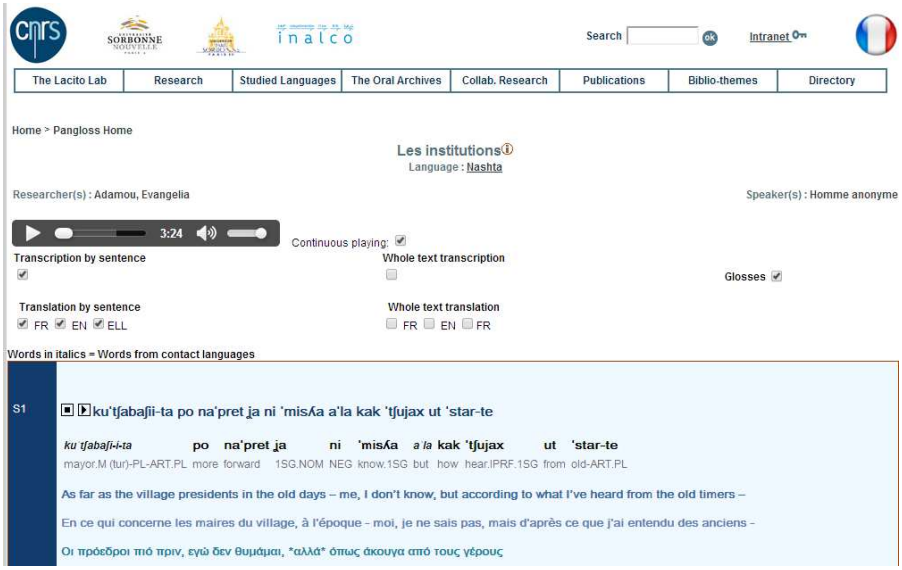

FIGURE 3. HTML display of a text in Balkan Slavic Nashta

5. WEB HOSTING, DATA PRESERVATION AND COMMUNITY ACCESS

5.1. LONG-TERM PRESERVATION. Finding solutions for perennial archiving (long-term data preservation) and web hosting is a central concern for the creators of digital open archives. Our first milestone in this direction was passed when the French National Library agreed to accept and maintain a one-time deposit from the LACITO in 2006. Since that time, the CNRS has shown interest in digital data in the humanities and social sciences by creating HUMA-NUM (http://www.huma-num.fr/, formerly ADONIS), and our data is currently housed on servers managed by this structure. For perennial archiving, HUMA-NUM has an agreement with the French National Higher Education Computing Center (CINES), which is certified by the French National Archives for the conservation of public archives. CINES conducts basic technical verifications on the documents submitted for long-term archiving; the data formats that we use (XML for text data and WAV or AIFF for audio) are among the accepted formats (see Schmidt and Bennöhr 2008 for examples of the difficulties encountered in recovering legacy data from outdated formats). After this technical verification, our materials are in a 10-year pipeline, which should lead to 'permanent' archiving. We have no experience of any but the initial stages of this process. For current access and diffusion, HUMA-NUM provides accessible data storage, including copies of material archived at CINES, and that is where CoCoON data is housed. The Pangloss web interface is currently housed with the LACITO website on a CNRS campus website at Villejuif.

5.2 ACCESS FOR THE SPEECH COMMUNITIES. The Pangloss Collection serves as a point of entry for long-term preservation, an online interface for consultation, and a meeting-point where engineers and linguists can discuss tools, formats, and documentation and research projects. Local initiatives have contributed to enabling speaker communities to gain effective access to the archived resources. In 1997-2000, the Agency for the Development of Kanak Culture in New Caledonia financed the preparation of an initial corpus of recordings in a dozen New Caledonian languages with time-aligned annotation, under the direction of Jean-Claude Rivierre. These texts, which constitute the core of the Pangloss New Caledonia collection, were made available for public consultation at the Tjibaou Cultural Center in Noumea, with a specially designed graphic interface.

During the following decade, Alexandre François' recordings of languages of Vanuatu were made available to the general public through the online interface of Pangloss. But this could not satisfy the demand of the speaker communities, who live on remote islands lacking internet access and often electricity. So it was decided to establish a multimedia library (the first in rural Vanuatu) and distribution point on Motalava in the Banks Islands. The library and its computer are managed by Edgar Woleg Howard, a community leader active in cultural preservation; the solar-powered computer was purchased with the help of the Alliance Française in Port Vila.

CDs were judged unsatisfactory as a distribution media, as the whole collection would have taken up 86 audio CDs, and players are rare in the islands. Mobile phones, however, had begun to be used in 2009 in the Banks Islands. Internet access remains costly, but these phones are widely used as mp3 players. So we undertook to produce mp3 versions of the 1000+ sound files. Key elements of the Pangloss OLAC metadata were transformed by script and integrated (as id3 tags) into the mp3 files. These audio files, with their embedded metadata, display perfectly in a local interface such as iTunes, or on mp3 players that were donated to the library. Even computer-illiterate local islanders were able to master the iTunes search engine – whether they searched by location, speaker, village name, or topic – and retrieve the recordings they desired.[2] They then could use their own mobile phones to further disseminate the recordings. These materials were welcomed by the community leaders, as a contribution to preserving not only the languages, but also the oral literature, songs and musical arts of a whole region.

This experience illustrates how the standard digital formats used by the Pangloss Collection and other archives can be transformed and adapted to new user communities. As technologies become more widespread locally, we hope that new channels can be opened, allowing community members to access existing resources more conveniently, and also to contribute new resources. Good-quality smartphones could be used as a device not only for playing previously recorded material, but also for creating new language recordings, potentially turning every speaker into a data contributor. When communities become connected to the internet in the future, user-friendly smartphone interfaces for our audio archives could potentially broaden the circle of their contributors by allowing community members to send feedback, upload their own recordings into the Pangloss Collection, and enrich existing resources through the addition of alternative versions, comments, or translations.

[2] The iTunes interface which was created on that occasion can be seen in a video, available at http://www.youtube.com/watch?v=hZGm0CLzxU8&hd=1.

This appears as a promising perspective for 'community-based research' (Rice 2011; Mosel 2006) in the digital age.

6. EXPLOITATION OF LANGUAGE CORPORA FOR RESEARCH

6.1. TALKING CONCORDANCE. The fragment shown below is from a concordance of over 20 texts in Hayu, made offline using an XSLT script by Michel Jacobson. The keywords (in the next to last column) are linked to sound resources located on a server, which may either be local (on the researcher's computer) or located on the HUMA-NUM grid. The left and right contexts of the keyword (3rd to last and last columns) are defined so as to cover the entire time-aligned utterance. When opened, these links cause the sound of the utterance to be downloaded and played. This kind of 'talking concordance' has proved useful for verifying text transcriptions and for finding and verifying example sentences in dictionary-making.

reference	word class	gloss	left context	keywd	right context
HAYUAAs42	postposition	EXTENT	ima-mʊ yeksa ko c^hə yeksa-	boŋ	bon-caŋ po-yi-ha bŏ-bon-ha t^hek dak-ta no-m ɪt-tse
HAYUGs42	verbstem	scold	teri ma bhansa dʊt-mi pi	bot	-ŋo-m paha dzot-tse-m re
HAYUDs10	verbstem	carry	k^həi tə ga	bu	-ŋ dɪyʊ-əi pa bu-ko-m ɪt-tse a nono-ha
HAYUUs11	verbstem	carry	ga-ha	bu	-no-m səmdhini paha bu-ko-m ɪt-tse

FIGURE 4. Lines of a 'talking concordance'

The corpus containing the TEXT elements to process is constituted by defining an ARCHIVE (as defined in the Pangloss XML). User-friendly online tools for indexing such corpora (by morpheme or by gloss) are among our current priorities for software development.

6.2 DICTIONARY/TEXT LINKS. A web dictionary of Limbu (Nepal), adapted from a print dictionary, is included in the collection as a prototype (http://lacito.vjf.cnrs.fr/pangloss/dico/). Example sentences in the dictionary contain references, most of which are linked to utterances in texts in the Pangloss Collection. Clicking on these links brings up the original source texts (the dictionary examples are often abbreviated) and the corresponding sound. A prototype system, in which morphemes in the online texts link dynamically to entries (and occasionally to homonyms) in the online dictionary is described in Jacobson and Michailovsky 2002 and was online for a few years, but no longer functions. At the same time, entries for affixes were added to the dictionary, so that every morpheme in the texts would link to a dictionary entry.

6.3. QUANTITATIVE DATA ANALYSIS. Despite the relatively small size of corpora of lesser-known languages, it is possible to exploit them for quantitative analysis of a wealth of topics; collaborations between linguists and engineers hold great potential (some reflec-

tions on this topic are proposed in Besacier 2012 and Michaud et al. 2012). Conversions have been successfully implemented from Pangloss documents in XML to the phonetic software Praat – for acoustic analysis – and the LaTeX typesetting system – to obtain a high-quality, ready-to-print version of the documents. Our XML markup also allows for a fully automated conversion to a relational database in the R software environment (http://www.r-project.org/), and for a number of other specific developments. The Limbu corpus was the first of our data sets to be integrated into the database of the corpus query processor CQPweb (Hardie 2012). It was possible to do an automatic conversion to the format used in CQPweb, thanks to the fully explicit nature of the XML markup. On the basis of this successful experiment, we plan to integrate other languages from the Pangloss Collection to the CQPweb database in future, to benefit from its powerful corpus-query tools.

Another example of how research goals shape the Pangloss Collection and benefit from it, comes from the team of researchers who contributed data on the minority Slavic languages of Europe. Since language contact was a central focus of the research program, loanwords were marked up with a specific attribute in the XML annotations, allowing them to be singled out visually in the online display (in italics – see figure 3 above) and retrieved through an XML query. The annotation of word tokens in a multilingual corpus with respect to the language of origin allows for an overall evaluation of the corpus in terms of overt language mixing. For example, Balkan Slavic texts recorded in the 1970s in Greece contain less than 1% of borrowings from Greek, whereas data collected from 2002 to 2011 show up to 6% of Greek tokens (Adamou & Breu 2013).

Finally, a study based on the Romani data illustrates the application of digital speech analysis software to documents of connected, spontaneous speech (Arvaniti & Adamou 2011). The Romani data were exploited for the study of information structure, including the prosodic marking of focus and topic, after conversion of the data from Pangloss format to PRAAT. Analysis of the data revealed frequent concurrent use of several focus marking strategies. This finding crucially relied on the use of unscripted conversations, which provide more lively and varied speech than is typically obtained in the highly controlled data sets (such as scripted question-and-answer pairs) often used for the study of information structure.

7. CONCLUSION. Building corpora is known to be extremely time-consuming yet highly rewarding. For widely-known languages of international communication, researchers can easily find assistants, but this is more complex – although not impossible – for lesser-known and endangered languages. Increased institutional recognition is indispensable for academics to invest as much time as they should (and in many cases would like to) in the preparation of data for long-term preservation, online circulation, and further computer processing. Signs of change include the recognition of the scholarly merit of language documentation by the Linguistic Society of America in 2010 (LSA 2010). The existence of calls for projects specifically intended for corpus preparation is also instrumental in raising the status of documentation work, allowing linguists to deal with data preparation tasks without being made to feel that they are off on a tangent from their legitimate occupations. We hope that the Pangloss Collection will help linguists meet the challenges of the digital age, and bring a useful contribution to the worldwide effort of linguists and language

documentation communities, in making endangered languages more readily available to everyone.

References

Adamou, Evangelia & Walter Breu. 2013. Présentation du programme Euroslav 2010 : Base de données électronique de variétés slaves menacées dans des pays européens non slavophones. XV International Congress of Slavists. Minsk, August 20-27, 2013.

Arvaniti, Amalia & Evangelia Adamou. 2011. Focus expression in Romani. In Mary Byram Washburn, Katherine McKinney-Bock et al. (eds), *Proceedings of the 28th West Coast Conference on Formal Linguistics* (WCCFL 28). Somerville, MA: Cascadilla Proceedings Project. 240-248. (www.lingref.com, document #2456)

Besacier, Laurent. 2012. A multi-disciplinary approach for processing under-resourced languages. In Xiong Deyi, Eric Castelli et al. (eds.), *Proceedings of IALP 2012 (2012 International Conference on Asian Language Processing)*. Hanoi: MICA Institute, Hanoi University of Science and Technology.

Bickel, Balthasar, Bernard Comrie & Martin Haspelmath. 2008. Leipzig Glossing Rules. (http://www.eva.mpg.de/lingua/resources/glossing-rules.php)

Bird, Steven & Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication* 33(1-2). 23-60. (http://arxiv.org/pdf/cs/0010033v1.pdf)

Boas, Franz, 1902. Tshimshan Texts. Smithsonian Institution, Bureau of American Ethnology, *Bulletin* 27. (http://gallica.bnf.fr/ark:/12148/bpt6k27476d)

Charachidzé, Georges. 1989. Ubykh. In John Greppin (ed.), *The Indigenous languages of the Caucasus*, vol. 2: George Hewitt (ed.) *The North West Caucasian Languages*. Delmar: Caravan. 357-459.

Colarusso, John. 1994. How many consonants does Ubykh have? In George Hewitt (ed.), *Caucasian Perspectives*. Munich: Lincom Europa.

Dumézil, Georges. 1965. *Documents anatoliens sur les langues et les traditions du Caucase III, Nouvelles Études Oubykhs*. Paris: Institut d'Ethnologie.

François, Alexandre. 2012. The dynamics of linguistic diversity. Egalitarian multilingualism and power imbalance among northern Vanuatu languages. *International Journal of the Sociology of Language* 214, 85–110.

Hardie, Andrew. 2012. CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3). 380–409.

Jacobson, Michel, Boyd Michailovsky & John B. Lowe. 2001. Linguistic documents synchronizing sound and text. *Speech Communication* 33 [special issue: "Speech Annotation and Corpus Tools"]. 79–96.

Jacobson, Michel & Boyd Michailovsky. 2002. Linking linguistic resources: time aligned corpus and dictionary. *International Workshop on Resources and Tools in Field Linguistics*, Las Palmas, Canary Islands, Spain, 26-27 May 2002 (http://www.mpi.nl/lrec/2002/papers/lrec-pap-27-JACMICv2.pdf).

Lehmann, Christian. Interlinear morphemic glossing. 2004. In Geert Booij, Christian Lehmann et al. (eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung. 2. Halbband*. (Handbücher Der Sprach- Und Kommunikationswissenschaft 17.2). Berlin: de Gruyter.

Leroy, Christine & Catherine Paris. 1974. Étude articulatoire de quelques sons de l'oubykh d'après film aux rayons X. *Bulletin de la Société de Linguistique de Paris* LXIX(1). 255–286.

LSA (Linguistic Society of America). 2010. Resolution Recognizing the Scholarly Merit of Language Documentation [*http://www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation*, retrieved 19 Dec 2013].

Michailovsky, Boyd, Alexis Michaud & Séverine Guillaume. 2011. A simple architecture for the fine-grained documentation of endangered languages: the LACITO multimedia archive. Keynote speech at Oriental-COCOSDA 2011, October 26th-28th, 2011, Hsinchu, Taiwan.

Michaud, Alexis, Andrew Hardie, Séverine Guillaume & Martine Toda. 2012. Combining documentation and research: Ongoing work on an endangered language. In Xiong Deyi, Eric Castelli et al. (eds.), *Proceedings of IALP 2012 (2012 International Conference on Asian Language Processing)*, 169–172. Hanoi: MICA Institute, Hanoi University of Science and Technology.

Mosel, Ulrike. 2006. Field work and community language work. In J. Gippert, N.P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 67–83. Berlin/New York: de Gruyter.

Rice, Keren. 2011. Documentary linguistics and community relations. *Language Documentation & Conservation* 5. 187–207.

Schmidt, Thomas & Jasmine Bennöhr. 2008. Rescuing legacy data. *Language Documentation & Conservation* 2(1). 109–129.

Thieberger, Nick & Rachel Nordlinger. 2006. Doing Great Things with Small Languages (Australian Research Council grant DP0984419). (http://linguistics.unimelb.edu.au/research/projects/greatthings.html)

Vogt, Hans. 1963. *Dictionnaire de la Langue Oubykh*. Oslo: Universitets Forlaget.

Whalen, Doug. 2004. How the study of endangered languages will revolutionize linguistics. In Piet van Sterkernburg (ed.), *Linguistics today: Facing a greater challenge*. Amsterdam/Philadelphia: John Benjamins. 321–344.

Woodbury, Tony. 2003. Defining documentary linguistics. In Peter Austin (ed.), *Language documentation and description*, vol. 1. London: School of African and Oriental Studies. 35–51.

Woodbury, Tony. 2011. Language documentation. In Peter Austin & Julia Sallabank (eds.), *The handbook of endangered languages*, vol. 1, 35–51. Cambridge: Cambridge University Press.

Boyd Michailovsky
boyd.michailovsky@vjf.cnrs.fr