

The LACITO Archive project markup

NOTE to markup group participants: I would very much appreciate receiving questions or comments before the workshop. -- Boyd Michailovsky (boydm@vjf.cnrs.fr).

The LACITO Archive project markup is designed for marking up linguistic documents in which a digitized sound recording is time-aligned with annotation such as transcription(s) and translation(s).

The data structure is described in an XML Document Type Definition (see DTD, below). The datatype described by the DTD is an archive or corpus (<ARCHIVE>) composed of one or more documents (<TEXT>). The sole function of the <ARCHIVE> element is to assemble a corpus of one or more texts; it will not be discussed further here.

The <TEXT> element requires a header (<HEADER>) containing metadata. Our current metadata is a bare minimum, pending the working out of standards. The only required elements are a <TITLE> and the <SOUNDFILE> element, whose "href" attribute (normally an URL) identifies the digitized sound resource.

A <TEXT>, according to the DTD, may be segmented at a maximum of three hierarchically ordered levels. The units of segmentation at the highest level are labeled <S>. In our usage, these generally correspond to utterances, which may be segmented into words (<W>), which in turn may be segmented into morphemes (<M>). In any case, if only a single level of segmentation is required, <S> elements are used; if two are required, <W> elements are added, etc. These basic units of segmentation will be referred to as "structural elements" below.

Each item of data (transcription, translation, etc.) appears as the content of a labeled element which itself is contained in one of the structural elements. For example, the free translation of an <S> will be labeled <TRANSL> and included directly in the appropriate <S> element; a word-gloss will also be labeled <TRANSL> but will be included in a <W> element. Each data element is thus attached to the tree structure at the level of the lowest-level containing structural element.

Structural elements at any level may contain labeled elements containing data of the following types: transcription (<FORM>), translation (<TRANSL>), time-anchoring (<AUDIO>), and (except at the lowest level) structural elements of the next lower level.

Time-anchoring data consists of start- and end-time offsets into a digitized sound resource. These values (in seconds) are expressed as the "start" and "end" attributes of the technically empty <AUDIO> element, e.g. <AUDIO start="10.5280" end="12.6912"/>. An <AUDIO> element immediately included in an <S> element (the usual case in our usage) indicates the start- and end-times of the sound corresponding to that element in the sound resource. Time-anchoring may be provided at more than one level, e.g., both to utterances and to words.

Transcriptions may be provided at different levels. Our usual practice is to provide a word-level phonological or orthographic transcription, and for certain corpora a morpheme-level morphophonological transcription to facilitate lexical searches. The choices here depend on the nature of the language transcribed and on the aims and energy of the transcriber.

In our usual practice, free translations (<TRANSL>) are provided at the utterance level, immediately included in <S> elements, and glosses (also labeled <TRANSL>) in <W> or <M>

elements. There may be more than one translation at any level, into different languages. It is generally assumed that the concatenation of utterance-level translations will serve as a translation of the whole document, although a more literary translation could be provided at the <TEXT> level. A concatenation of morpheme- or word-glosses obviously could not serve as an utterance-level translation.

<TEXT> and <S> elements have unique identifier attributes ("id") -- essentially a short name for the <TEXT> and a string containing a serial number for the <S>. The <TEXT> and <TRANSL> elements have a language attribute ("lang"). (This attribute may later be generalized to other structural elements, for multilingual texts or unassimilated loans.) <S> elements may have a "who" attribute to identify the speaker in documents with more than one speaker.

The remaining items in the DTD are rather idiosyncratic: the <FOREIGN> element has to do with the annotation of unassimilated loanwords in my Nepal texts; the <PUNC> element is one way of allowing for punctuation in the transcription (another method is illustrated in the Limbu example below); the "type" attribute of the <M> element provides for just those form-classes that can be identified automatically by the script I use to generate the XML markup of my Limbu texts. (This brings up a point which should not be imitated -- I have not yet made the leap to using XML/Unicode as my basic data format.)

I will discuss "authoring" issues at the workshop. SoundIndex, the time-alignment tool written by Michel Jacobson, is available on the project website.

The DTD

```
<!-- DTD des documents d'archives (04/10/2000) -->
<!ELEMENT ARCHIVE (TEXT)+ >
<!--
***
*** The levels
***
-->
<!ELEMENT TEXT (HEADER, (FORM|TRANSL|AUDIO|S)*) >
<!ATTLIST TEXT lang CDATA #REQUIRED
id ID #REQUIRED >
<!ELEMENT S (FORM|TRANSL|AUDIO|W|PUNC)* >
<!ATTLIST S id ID #REQUIRED
who CDATA #IMPLIED >
<!ELEMENT W (FORM|TRANSL|AUDIO|M)* >
<!ELEMENT PUNC EMPTY >
<!ATTLIST PUNC type
(period|excl|quot|quest|emdash|comma|hellip|colon|unclear) #REQUIRED
place (right|left|free) #REQUIRED >
<!ELEMENT M (FORM|TRANSL|AUDIO)* >
<!ATTLIST M type
(prstem|pastem|stem|vprefix|vsuffix|preverb|redup) #IMPLIED>
<!--
***
*** The metadata
***
-->
<!ELEMENT HEADER (TITLE+, SOUNDFILE, RECORDING?, SPEAKER?) >
<!ELEMENT TITLE (#PCDATA) >
<!ATTLIST TITLE lang CDATA "English" >
<!ELEMENT SOUNDFILE EMPTY >
<!ATTLIST SOUNDFILE href CDATA #REQUIRED >
<!ELEMENT RECORDING EMPTY >
<!ATTLIST RECORDING date CDATA #REQUIRED
place CDATA #REQUIRED >
<!ELEMENT SPEAKER (#PCDATA) >
<!--
***
*** The data
***
-->
<!ELEMENT TRANSL (#PCDATA) >
<!ATTLIST TRANSL lang CDATA "English"
type (meta) #IMPLIED >
<!ELEMENT FORM (#PCDATA|FOREIGN)* >
<!ELEMENT FOREIGN (#PCDATA) >
<!ATTLIST FOREIGN lang CDATA #REQUIRED >
<!ELEMENT AUDIO EMPTY >
<!ATTLIST AUDIO start CDATA #REQUIRED
end CDATA #REQUIRED >
```

Examples

The following examples are taken, slightly edited, from documents which may be consulted (with synchronized sound) on the archive project website (<http://lacito.archivage.vjf.cnrs.fr>).

The first example, in the New Caledonia language Nemi, shows segmentation at <S> and <W> levels. Note, for example, that the lexeme *hingoo* 'story' is not explicitly marked up as such, which complicates searching. (Prepared by F. Ozanne-Rivierre.)

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE TEXT SYSTEM "../All/Archive.dtd">
<TEXT id="BAC" lang="nemi">
  <HEADER>
    <TITLE lang="French">Bac et Dangem, les deux lochons de rivière</TITLE>
    <TITLE lang="English">Bac and Dangem</TITLE>
    <SOUNDFILE href="../Nouvelle_Caledonie/Nemi/BAC.mp3"/>
  </HEADER>

  <S id="nemi13s1">
    <AUDIO start="0.0800" end="4.3201"></AUDIO>
    <W><FORM>Hingoo-ng</FORM><TRANSL lang="French">histoire-à moi</TRANSL></W>
    <W><FORM>bac</FORM><TRANSL lang="French">lochon bac</TRANSL></W>
    <W><FORM>ma</FORM><TRANSL lang="French">et</TRANSL></W>
    <W><FORM>dangem</FORM><TRANSL lang="French">lochon dangem</TRANSL></W>
    <PUNC place="right" type="period"/>
    <TRANSL lang="French">Je vais raconter l'histoire de deux lochons, bac et dangem.</TRANSL>
  </S>
  ...
</TEXT>
```

The following is a fragment of a text in Limbu, showing segmentation at utterance (here breath-group), word, and morpheme levels. The morpheme-level transcription facilitates word-search and concordancing. The S-level transcription includes punctuation. It is otherwise redundant, and it makes maintenance (correction) of the text more difficult. It could be also used to show sandhi between words.

```
<S id="s17">
  <AUDIO start="83.7199" end="86.8400"/>
  <FORM>annuppa phereaj kə, ɛtna omettesigeaj kə</FORM>
  <TRANSL lang="English">When my father-in-law came and looked at us,</TRANSL>
  <W>
    <FORM>annuppa</FORM><TRANSL>lsg.-f.in.law</TRANSL>
    <M><FORM>a</FORM><TRANSL>lsg.</TRANSL></M>
    <M><FORM>nuppa</FORM><TRANSL>f.in.law</TRANSL></M>
  </W>
  <W>
    <FORM>phereaj</FORM><TRANSL>come-3.pa.-conj.</TRANSL>
    <M type="pastem"><FORM>pher</FORM><TRANSL>come</TRANSL></M>
    <M type="vsuffix"><FORM>ɛ</FORM><TRANSL>pa.</TRANSL></M>
    <M><FORM>aj</FORM><TRANSL>conj.</TRANSL></M>
  </W>
  ...
</S>
```

In the following fragment from a Bantu language, an <M> level transcription shows lexical structure and an <S> level transcription fast speech and sandhi phenomena, which may cross word boundaries. Phonetic characters are represented by entities specifying decimal Unicode codepoints; the entity ɔ represents "ɔ". (Prepared by Margaret Dunham)

```

<S id="langi3s10">
  <AUDIO start="53.2814" end="55.1373"/>
  <FORM>n ir&#596; &#331; g&#596; rasatu</FORM>
  <TRANSL lang="French">avec de l'excrément de boa</TRANSL>
  <TRANSL lang="English">with boa excrement</TRANSL>
  <W>
    <M>
      <FORM>na</FORM>
      <TRANSL lang="French">conn</TRANSL>
      <TRANSL lang="English">conn</TRANSL>
    </M>
  </W>
  <W>
    <M>
      <FORM>i</FORM>
      <TRANSL lang="French">PI5</TRANSL>
      <TRANSL lang="English">IP5</TRANSL>
    </M>
    <M>
      <FORM>r&#596; Ng&#596; </FORM>
      <TRANSL lang="French">boue</TRANSL>
      <TRANSL lang="English">mud</TRANSL>
    </M>
  </W>
  <W>
    <M>
      <FORM>ri</FORM>
      <TRANSL lang="French">PD5</TRANSL>
      <TRANSL lang="English">DP5</TRANSL>
    </M>
    <M>
      <FORM>a</FORM>
      <TRANSL lang="French">dét</TRANSL>
      <TRANSL lang="English">det</TRANSL>
    </M>
    <M>
      <FORM>&#8709;</FORM>
      <TRANSL lang="French">PI9</TRANSL>
      <TRANSL lang="English">IP9</TRANSL>
    </M>
    <M>
      <FORM>satu</FORM>
      <TRANSL lang="French">boa</TRANSL>
      <TRANSL lang="English">boa</TRANSL>
    </M>
  </W>
</S>

```

In this annotation of a song in Wayana (French Guyana), only morpheme-level transcription is used, although a word-level of segmentation is recognized: (Prepared by Hervé Rivière.)

```
<S id="arrows9">
  <AUDIO start="45.6034" end="49.5400"/>
  <TRANSL lang="French">Allez, les braves ! Allez-y vaillamment !</TRANSL>
  <W>
    <M>
      <FORM>kajali</FORM>
      <TRANSL lang="French">vaillant</TRANSL>
    </M>
    <M>
      <FORM>tomo</FORM>
      <TRANSL lang="French">PL</TRANSL>
    </M>
  </W>
  ...
  <W>
    <M>
      <FORM>itë</FORM>
      <TRANSL lang="French">aller</TRANSL>
    </M>
    <M>
      <FORM>kë</FORM>
      <TRANSL lang="French">IMPER</TRANSL>
    </M>
  </W>
</S>
```

Finally, a made-up example shows speaker identification and overlap. In the hierarchical XML structure there is no way that two <S> elements -- or two S-level <AUDIO> elements -- can overlap. However, the values of the time offsets of the <AUDIO> elements can overlap, because they refer to a non-XML entity, the sound recording, and are not checked by the XML parser. This is shown in the example. Indicating the precise synchronisation of the overlapping or partly overlapping elements would require a more complex markup.

```
<S who="A">
  <AUDIO start="0.00" end="2.00"/>
  <FORM>I haven't finished talking.</FORM>
</S>
<S who="B">
  <AUDIO start="1.00" end="3.00"/>
  <FORM>And I have already started.</FORM>
</S>
```