EOPAS, the EthnoER online representation of interlinear text[1]

Ronald Schroeter (ITEE, University of Qld) and Nick Thieberger (Linguistics, University of Melbourne)

One of the goals of the ARC funded Eresearch project called *Sharing access and analytical tools for ethnographic digital media using high speed networks*, or simply EthnoER is to take outputs of normal linguistic analytical processes and present them online in a system we have called the EthnoER online presentation and annotation system, or EOPAS. EthnoER has twenty-three chief investigators from ten organizations, both Australian and international, at Universities and other agencies, including CSIRO, language centres and language archives. The project includes a set of testbed projects with varying requirements for online annotation and presentation of data. We aim to provide online mechanisms for annotating data in order that, for example, archival media files can be annotated by experts with local knowledge and that samples of that media, perhaps a three minute chunk, can be presented so that users can get a sense of the quality of the recording and other information that may influence their decision to download the whole file.

We were inspired by Michel Jacobson's work[2] in presenting interlinear text for the LACITO group in Paris, and wanted to make a similar tool available more generally. In this paper we will focus on the linguistic goals of the project using interlinear text and transcribed media as the inputs to an online delivery and querying system, and on the archival data curated by the Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).

Ultimately the model we are developing would allow diverse datatypes to interoperate, with grammar, dictionaries, texts and media all interlinked and navigable. In this first step we are focussing on a means for viewing, concordancing and searching simple transcriptions or interlinear text associated with media combined with the ability to access the media.

The main types of linguistic outputs we want to work with are time-aligned transcripts (from ELAN[3] and Transcriber[4]) and interlinear texts (IT) (from Toolbox[5]) which will be the focus of this paper.

An existing tool that provided a model for the project was *Audiamus*[6] developed for exploring media linked to text for an analysis of South Efate, an Oceanic language from Vanuatu, providing access to field recordings via a textual index, allowing navigation via search and concordance of the whole corpus, and citability of the data at the level of the

---

1 The work reported here was funded by the ARC Eresearch grant SR0566965. Thieberger is suppported by an ARC APD (DP0450342).

2 http://lacito.vjf.cnrs.fr/archivage/contents_fr.htm

3 http://www.mpi.nl/tools/elan.html

4 http://www.etca.fr/CTA/gip/Projets/Transcriber/

5 http://www.sil.org/toolbox

6 http://www.linguistics.unimelb.edu.au/thieberger/audiamus.htm

sentence, all linked to media. This approach further facilitates analysis in that a transcript is not treated as a complete and true representation of the media, but rather as an index of the media that can be constantly improved through interaction with the media. Errors in the transcript due to one's incremental learning of the language and changing understanding of constructions in the language can be corrected. Apparently aberrant constructions located in the transcripts of the language can immediately be queried and perhaps corrected via reference to the primary data. This methodology should always have been part of standard linguistic fieldwork, but to date has not been widely used, despite the technology being available for some years now.

One of the reasons for slow uptake has been the lack of tools and standards. While *Audiamus* is fine for the author's purposes, and has a small group of users, it is a standalone tool. The data (which consists of text with timecodes and a media file) can be exported in a number of forms, one of which is XML conformant to the LACITO schema. A sample output from *Audiamus* to LACITO's schema was developed in 2005 and plans were laid for ongoing collaboration. Michel Jacboson then attended the EthnoER conference in 2006 to present recent developments and to begin collaboration with Ronald Schroeter and Michael Henderson (of the School of Information Technology & Electrical Engineering at the University of Queensland).

For a number of reasons it is desirable to present linguistic data online as one of the outputs of a normal linguistic workflow. The painstaking effort required to prepare an interlinear text collection will benefit from the present process of data conversion and the rigour imposed by XML validation. Toolbox already provides a significant level of consistency in its use of a controlled lexicon, and partially constrained data structure, certainly much better than manual construction of IT. However, errors still occur and using different data visualisation tools allows us to locate these errors and correct them in the master file.

Online views of data may be made available to the general web-user, or may just be a means for the linguist to access a corpus of their data. Primary data in small languages is difficult to locate, and typically is only found in grammatical descriptions as example sentences with no access to the primary recordings. Until recently it has been virtually unheard of for fieldrecordings to be made available as part of a grammatical description, but this has changed, partly assisted by the development of digital linguistic archives which provide for citation of primary data. Web delivery renders primary data in a form that can be accessed derived from an archival form of the data housed in a secure repository. In the EthnoER project we aim to develop a method for presentation of textually-indexed media that can be adopted by others. Thus we are committed to developing or using tools that are freely available and in which the data is not locked up in proprietary software. We want to encourage creation of data in forms that can be reused and that can be archived, and we have to rely on practitioners to provide data in that form in order to obtain the benefits offered by the EOPAS system. EOPAS is part of a workflow that allows data to be imported if it is in good form and so is a carrot being offered to linguists. If we can build tools and processes that offer diverse ways of interacting with our data it should be incentive for using the workflow, with the spinoff that well-formed data will be produced for longterm curation.

There are two current approaches that take the output of Toolbox data or its equivalent and make an explicit data structure available for online viewing. The first of these, from LACITO, offers much of what we wanted to build, but was established as an in-house

method for archiving texts created by LACITO researchers using the tool called ITE[7] and is not generalisable as it currently stands. It will be clear that EOPAS owes much of its design to LACITO's approach. We also wanted to include the ability for users to upload files to the system remotely and to then have a concordance built over all files of that language in the collection, neither of which is available in the current LACITO framework.

The second way in which Toolbox data has been rendered into an online document is using the beta tools BoxReader and BoxWriter (Hellmuth, Myers and Nakhimovsky 2006). BoxReader creates XHTML from Toolbox data and BoxWriter converts back to Toolbox format, allowing editing of Toolbox data in XHTML for reuse in Toolbox. Why would you want to do that? In the case of homophones, for example, a search on explicitly structured data will only operate on those morphemes X glossed as Y and not those glossed as Z, whereas an unstructured search would only be able to find all morphemes X. However, the current version of BoxReader makes no provision for linkage to media. So, while we want to work with these tools, neither satisfies all the desiderata of the EthnoER project.

It soon became apparent from previous work (e.g., Bow, Hughes and Bird 2003) and as we worked with the test data sets that there is a great diversity in the way that linguistic data is encoded, due mainly to the lack of standards, itself partly flowing from the lack of centralised provision of advice, templates and example data sets. In order, therefore, to be able to present even a simple working model of an online presentation of this data we had to put constraints on the inputs to our system. In the first round of the process we elected to use a single line annotation in the transcription tools which could otherwise allow several lines or tiers (in Transcriber) or an arbitrary number of tiers (in ELAN). Similarly, users of Toolbox are able to use as many lines of interlinear text as they want (see an extreme example in Drude 2003), and can choose to create fieldmarkers (fieldnames) as they please. Again, we chose to limit the input to a simple model, with text, morpheme-level, gloss-level, free gloss, and media reference together with a set of metadata fields as we will see below. We provided a Toolbox template with preferred fieldnames and hierarchies for users who want to follow the model we have developed.[8]

Transcriptions

Transcription with time-alignment provides an index of field recordings at the level of the utterance unit, as defined by the linguist. There are a number of tools that present media so that it can be transcribed and which will then produce the transcriptions with timecodes as an output. We will address only two in this project, ELAN and Transcriber, mentioned earlier. They are widely used among the small but growing group of linguists engaging with these methods and are suitable for the purposes of our model as they both produce an XML output. In order to constrain the variability possible in data structures we have chosen to accept only a single line of annotation from these tools. While it would appear that we are losing the richness of the original source, we regard the output to EOPAS as being a representative derived form of the original data, just as an mp3 file is a derived form of a pcm audio file suitable for delivery and representation of the data. This is also the approach taken by Toolbox in its MDF (Multi Dictionary Formatter) output which specifies what

7 Interlinear Text Editor (http://michel.jacobson.free.fr/ITE/)

8 http://ethnoer.unimelb.edu.au/EOPAS.zip

fieldnames should be used in order that there be a coherent migration of parts of a lexical database to the structured output provided by MDF. Thus, in our system, an archival format of an interlinear text file can readily be converted to EOPAS and viewed, with the master file subsequently able to be corrected and re-converted to EOPAS if necessary. Any editing of the archival file can then be reflected in a subsequently converted EOPAS file, deleting previous versions.

One of the major motivations for developing EOPAS was the need to provide access to archival data. For recordings made in the past we want to be able to present the data in an accessible form for access by the speakers and their descendants, and in order that new research may be carried out. As an example, a body of recordings made in the Solomon Islands for Professor Stephen Wurm in the 1970s has been digitised by PARADISEC and the associated handwritten transcripts have been put online[9] for access by colleagues in Norway (see Næss this volume).

Interlinear Text

Interlinear text (IT) is a generally used and accepted method for annotating textual material. It is particularly useful in presentation of material arising from linguistic fieldwork as discussed by Bow, Hughes and Bird (2003) who provide a typology of IT types from various sources. Schmidt (2003) argues that IT is a method for visualising a data mode, that is, it is a graphic representation of the relationship that holds between text and its constituent parts. Toolbox uses spaces between words and their vertical alignment as the means for displaying the relationships. The XML export in recent versions of Toolbox (1.5 in particular) provides the potential to make this relationship explicit. However, for our purposes the structure given by the .typ file needs to be fairly tightly constrained to provide the hierarchy of the text line over morphemic and gloss lines which are sibling lines, with the free gloss a sibling of the text line.

EOPAS

EOPAS derives from the general model for interlinear text proposed by Bow, Hughes and Bird (2003), but introduces elements and attributes necessary for cataloguing and time-aligned synchronization between the text and video/audio files. EOPAS consists of a root element `<eopas>` which contains a `<header>` and an `<Interlinear-text>` element. The header consists of an `<olac>` element from the OLAC[10] namespace, which gives us access to a range of Dublin core elements and OLAC attributes, codes and vocabulary for archiving the linguistic data in an open and recognised manner.

The `<Interlinear-text>` element provides the hierarchical structure for representing different semantic layers/tiers of IT:

[9] http://paradisec.org.au/fieldnotes/SAW2.htm

[10] http://www.language-archives.org/

1. transcript	.... pusereki kusu go wit				
2a. phrase (orthographic)	pusereki kusu go wit				
2b. phrase (translation)	tell about the rat and the octopus				
3. words	pusereki		kusu	go	wit
4a. morpheme (morpheme)	puserek	-ki	kusu	go	wit
4b. morpheme (gloss)	talk	-TR	rat	and	octopus

1. This element could contain the whole text or transcript, 2. phrases or chunks (which can be in (a) an orthographic form or (b) a free translation, etc.), 3. words and 4. morphemes and glosses. Each tier can have an arbitrary number of `<text>` elements of different type, e.g. orthographic (or phonetic), translation, morphemic, etc (see Figure 1), and an arbitrary number of `<meta>` elements, which are space holders for key-value pairs such as speaker="Speaker Name". This structure gives maximum flexibility to possible extensions in the future. While the semantic structure of IT will always remain the same, other types of texts can easily be added to the format and will be recognised as just another track of text within the specified layer/tier.

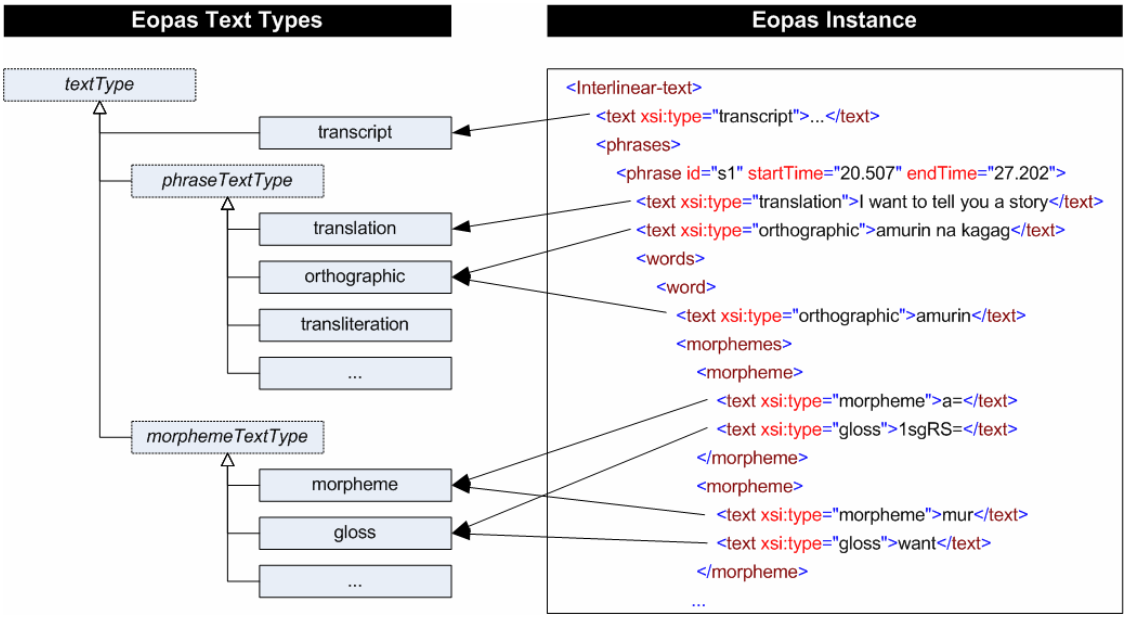

Figure 1: Eopas text types

Figure 1 also shows the attributes `id`, `startTime` and `endTime` that have been added to the `<phrase>` element. The start- and end-time-offsets are used to align the text with a sound or video file, the id is used to reference phrases. The minimum requirement to be a valid EOPAS file is to have one orthographic text element in each phrase. An example of a full EOPAS file can be viewed in Appendix A.

Input Formats

ELAN

The structure of the ELAN xml format is based around tiers (the `<TIER>` element). A tier is represented as a timeline item in the ELAN user-interface. It can have any user-defined meaning and carry structural information, e.g. the tier hierarchy can be defined. This gives ELAN the flexibility and customizability to be used in many different application domains,

however, it also creates user specific (rather than domain specific) XML outputs where the semantics of the structure is unknown to other systems. To be able to use ELAN within our domain, it is important to define a domain-specific ELAN template, in which the structure, semantics and constraints are well defined.

In ELAN, an entity within a tier is called an annotation (the `<ANNOTATION>` element). An annotation is defined by a text value, e.g. the orthographic text, the translation, etc, and a start and end time.

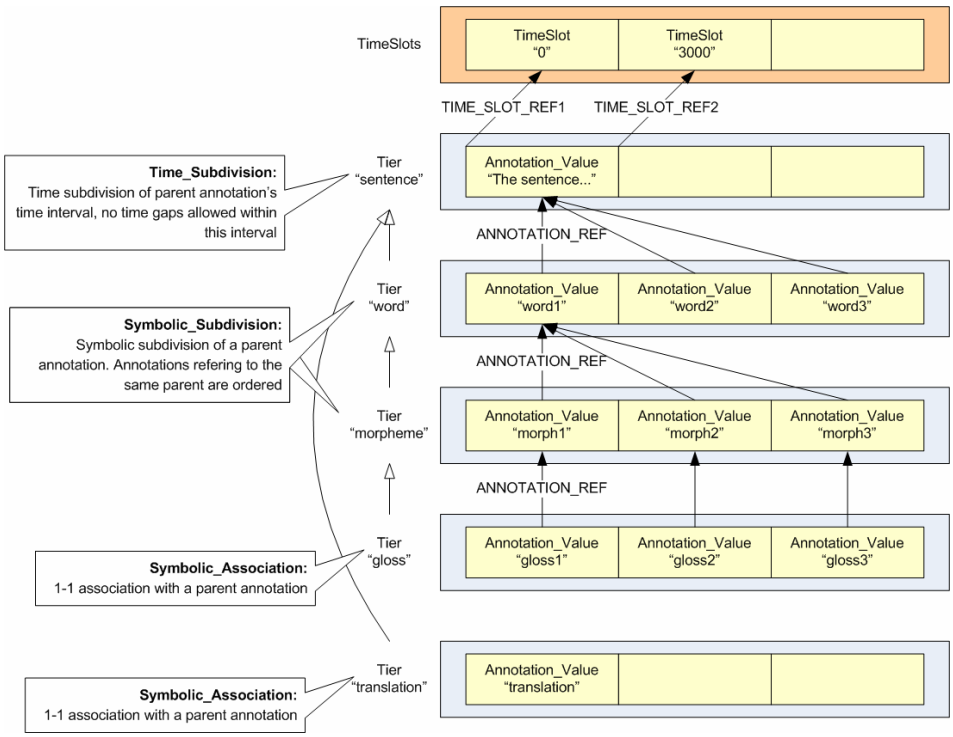

Figure 2: ELAN's tier structure

Figure 2 shows how the IT information represented in EOPAS could be represented in the ELAN format to ensure an easy mapping and transformation between the two. ELAN stores all timecode information separately inside the `<TIME_SLOT>` elements. To define a start and end time, the annotations might reference these timeslots. However, annotations might also reference annotations within a parent tier, in which case they "inherit" its time constraints. This makes the transformation a bit trickier, because we have to follow up a list of referenced elements within the inheritance hierarchy to retrieve the desired information.

At this stage we only support single line annotation of ELAN files, because we envisage a workflow in which users will use a more suitable format for creating IT, such as Toolbox, and so will only use ELAN for time-alignment.

However, the above example shows how a well-formed ELAN file with multiple tiers based on our template could be mapped to EOPAS and how the EOPAS format could be transformed back to ELAN without losing too much information.

Below is a concrete example of a single line ELAN transcript which has been simplified to an extent for presentation purposes. Some additional information provided by ELAN such as various linguistic types and controlled vocabularies are not shown here as EOPAS doesn't

offer to represent any of them. The example below also demonstrates the use of the OLAC header mentioned earlier, which would be semi-automatically populated by information extracted from the original transcript and information added through a web form during the upload process by the user.

ELAN

```
<annotation_document author="Nick Thieberger"
date="2006-08-05T13:21:47+10:00">

   <header media_file="200518.wav"
           time_units="milliseconds">
   …
   </header>

   <time_order>
      <time_slot time_slot_id="ts1" time_value="20507"/>
      <time_slot time_slot_id="ts2" time_value="27202"/>
      <time_slot time_slot_id="ts5" time_value="33212"/>
      ...
   </time_order>

   <tier default_locale="en"
         linguistic_type_ref="UtteranceType"
         participant="Tokelau Takau"
tier_id="Transcript">
      <annotation>
         <alignable_annotation annotation_id="a1"
           time_slot_ref1="ts1" time_slot_ref2="ts2">
            <annotation_value>Nick, amurin na ka gag
traus natrasuen nen kin-, natrausuen ni of.
            </annotation_value>
         </alignable_annotation>
      </annotation>
      <annotation>
         <alignable_annotation annotation_id="a2"
           time_slot_ref1="ts2" time_slot_ref2="ts3">
            <annotation_value>Tesa ni of
            </annotation_value>
         </alignable_annotation>
      </annotation>
      <annotation>
         <alignable_annotation annotation_id="a3"
           time_slot_ref1="ts3" time_slot_ref2="ts4">
            <annotation_value>akit tu tae na
            </annotation_value>
         </alignable_annotation>
      </annotation>
   </tier>
   ...
</annotation_document>
```

EOPAS

```
<eopas>
   <header>
      <olac:olac>
         <dc:creator>Nick Thieberger</dc:creator>
         <dcterms:requires xsi:type="dcterms:URI">
                  mediafiles/25717300.mp3</dcterms:requires>
         <dc:title>Of go wit, 'Heron and the octopus'</dc:title>
         <dc:contributor xsi:type="olac:role" olac:code="recorder">
                  Nick Thieberger</dc:contributor>
         <dc:contributor xsi:type="olac:role" olac:code="speaker">
                  Tokelau Takau</dc:contributor>
         <dc:coverage>VU</dc:coverage>
         <dc:subject xsi:type="olac:language" olac:code="erk"/>
         <dc:language xsi:type="olac:language" olac:code="erk"/>
         <dc:date>2006-08-05T13:21:47+10:00</dc:date>
         <dc:type xsi:type="olac:linguistic-type"
                  olac:code="Primary Text"/>
         <dc:type xsi:type="olac:discourse-type"
                  olac:code="Narrative"/>
      </olac:olac>
   </header>

   <Interlinear-text>
      <phrases>
         <phrase startTime="20.507" endTime="27.202" id="a1">
            <meta key="Speaker">tokelau takau</meta>
            <text xsi:type="orthographic"> Nick, amurin na ka gag
traus natrasuen nen kin-, natrausuen ni of..</text>
         </phrase>
         <phrase startTime="27.202" endTime="33.212" id="a3">
            <meta key="Speaker">tokelau takau</meta>
            <text xsi:type="orthographic">Tesa ni of</text>
         </phrase>
         <phrase startTime="33.212" endTime="36.582" id="a4">
            <meta key="Speaker">tokelau takau</meta>
            <text xsi:type="orthographic">akit tu tae na</text>
         </phrase>
         …
      </phrases>
   </Interlinear-text>
</eopas>
```

Transcriber

A Transcriber file consists of a list of topics, a list of speakers and finally the transcription text which makes references to the lists and is structured as follows:

- Episode (root of the transcript)
- Section (a segment defined by a topic, red box in screenshot Figure 3)
- Turn (a segment defined by a speaker, blue box in screenshot Figure 3)
- Sync (a time point defined by seconds and milliseconds, turquoise bullet points in screenshot Figure 3)

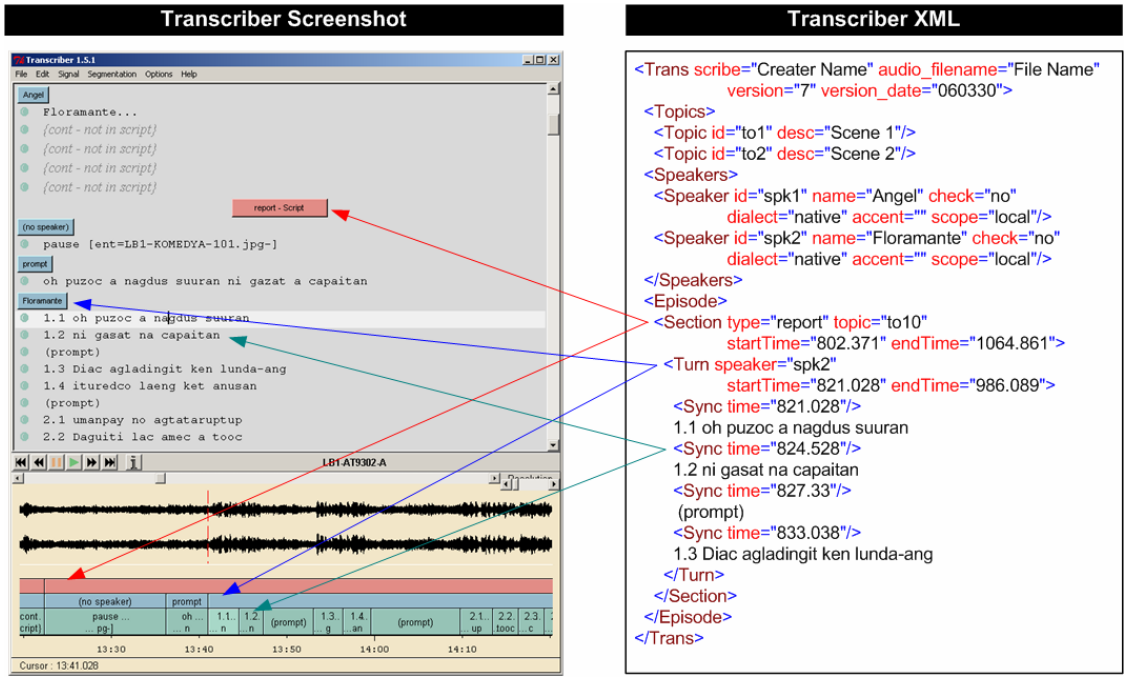

Figure 3: Screenshot and XML output of Transcriber

The text between two sync points is transformed into a phrase in EOPAS, with the previous `<Sync>` element delivering the start-time and the next `<Sync>` element the end-time. The XSL to do this is not trivial because the text is not always between two `<Sync>` elements, e.g. the last one labelled 1.3. in Figure 3, in which case the end time has to be derived from the endTime attribute of the superseding `<Turn>` element. Unfortunately Transcriber does not provide a way to label the text on the `<Sync>` level, which is why the creator of the example added the numbering (1.1, 1.2, 1.3, …) to the text. As every user has different habits in using Transcriber, this kind of additional information is very difficult to extract or filter out automatically. The semantics are only clear to a human user understanding the semantics of the numbering. In this case regular expressions can be used to disregard any numbers at the beginning of the text, but the point we are trying to make here is that users of tools such as Transcriber should be aware that any information they might add to the transcripts in an undefined manner might not be understood by automatic tools that further process that data.

As mentioned earlier, our approach is to keep the original transcription file and store it alongside the transformed EOPAS format, which might have lost some information during the transformation process. The search, browse and concordance view functionalities of the EOPAS website are built on top of the unified EOPAS format, so any queries based on that kind of information is not supported, but the information itself is still being archived and not lost.

Transcriber offers a range of options to provide detailed information or metadata about

- the episode, e.g. author's name, version, principal language, *program* and recording date, most of which we try to translate into the OPAC metadata header in EOPAS;
- the speaker, e.g. name, *global name*, *sex*, *dialect* and *accent*, which can be put into the `<meta>` elements of each phrase in EOPAS.

- the *section*, e.g. *topic*, *type (report, filler, nontrans)*, *start-* and *end-time*.

The metadata indicated in italic is usually disregarded during the transformation process into EOPAS. The same is true for the section information, which is essentially a grouping of phrases/sentences into different section types or topics. Below is a snippet of the IT section of the EOPAS file representing the Transcriber text in Figure 3.

```
<Interlinear-text>
  <phrases>
    <phrase id="s11" startTime="821.028" endTime="824.528">
      <meta key="speaker">Floramante</meta>
      <text xsi:type="orthographic">1.1 oh puzoc a nagdus suuran</text>
    </phrase>
    <phrase id="s12" startTime="824.528" endTime="827.33">
      <meta key="speaker">Floramante</meta>
      <text xsi:type="orthographic">1.2 ni gasat na capaitan</text>
    </phrase>
    <phrase id="s13" startTime="827.33" endTime="833.038">
      <meta key="speaker">Floramante</meta>
      <text xsi:type="orthographic">(prompt)</text>
    </phrase>
    <phrase id="s14" startTime="833.038" endTime="836.425">
      <meta key="speaker">Floramante</meta>
      <text xsi:type="orthographic">1.3 Diac agladingit ken lunda-ang</text>
    </phrase>
    ...
  </phrases>
</Interlinear-text>
```

Toolbox

Similarly to ELAN, Toolbox allows full customization of text labels and their hierarchies, which makes it impossible to write a general tool to import Toolbox data. However, a domain-specific template (Toolbox .typ file) can be provided that, if used consistently, allows us to read Toolbox markers into EOPAS through the standard Toolbox XML export function. In our experience, and following Bow, Hughes and Bird (2003:28) the minimal representation of IT that satisfies a linguist's needs includes four levels, *text, phrase, word* and *morpheme*. This is outlined in the list of markers and their hierarchy below (indicated by indentation), which is supported by EOPAS. We have added a fifth marker which is the reference to the time offsets in the media file from which the text is transcribed:

`\aud`	time reference in the form 'filename starttime endtime' in seconds and milliseconds
\tx	text line
\mr	morphemic
\mg	glosses of the morphemic line
\fg	free gloss

The hierarchy can be encoded in the Toolbox .typ file associated with the data (by including 'following field' information in the document properties). Toolbox will automatically group markers of the same hierarchical level. Below is an example of how the XML output of the Toolbox file should look (the schema is illustrated in Figure 4) and how it maps to EOPAS (the header is not shown here, but note that the title inside the `<itm>` element would go in there).

Toolbox

```
<database>
    <itmgroup>
        <itm>105</itm>
        <idgroup>
            <id>001</id>
            <aud>
                200518.aud 20.507 27.202
            </aud>
            <txgroup>
                    <tx>Amurin</tx>
                    <mr>a=</mr>
                    <mg>1sgRS=</mg>
                    <mr>mur</mr>
                    <mg>want</mg>
                    <mr>-i</mr>
                    <mg>-TS</mg>
                    <mr>-n</mr>
                    <mg>-3sgO</mg>
            </txgroup>
            …
            <fg>I want to tell you a story.</fg>
    </idgroup>
```

EOPAS

```
<Interlinear-text>
  <phrases>
    <phrase id="s1" startTime="20.507" endTime="27.202">
      <text xsi:type="translation">
        I want to tell you a story</text>
      <text xsi:type="orthographic">Amurin na kagag</text>
        <words>
          <word>
            <text xsi:type="orthographic">amurin</text>
            <morphemes>
              <morpheme>
                <text xsi:type="morpheme">a=</text>
                <text xsi:type="gloss">1sgRS=</text>
              </morpheme>
              <morpheme>
                <text xsi:type="morpheme">mur</text>
                <text xsi:type="gloss">want</text>
              </morpheme>
              <morpheme>
                <text xsi:type="morpheme">-i</text>
                <text xsi:type="gloss">-TS</text>
              </morpheme>
              <morpheme>
                <text xsi:type="morpheme">-n</text>
                <text xsi:type="gloss">-3sgO</text>
              </morpheme>
            </morphemes>
          </word>
          …
```

Metadata should be entered into the EOPAS online form, but a metadata header for a Toolbox file can include the following:

\itm	Text title
\sp	Speaker name
\sex	Sex of speaker
\age	Age of speaker when recorded
\lg	Language code (Ethnologue code or ISO/DIS 639-3)
\pi	Persistent identifier of media file
\da	date recorded (YYYY-MM-DD, conforming to ISO 8601)

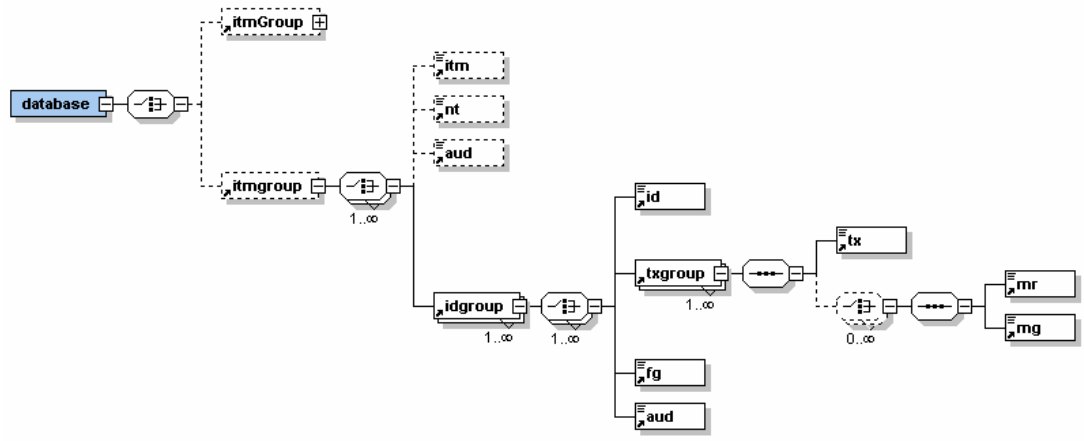

Figure 4: EOPAS specific Toolbox Schema (representation in XML Spy Schema View)

We have provided a template for Toolbox users who want to constrain their data to output to the EOPAS schema[11]. This template provides a sample lexical file and a text ready for interlinearising. Using this template to enter one's texts should allow the normal XML output of Toolbox to be produced in a form conformant to the EOPAS specific Toolbox Schema in Figure 4.

System Architecture

Upload Process

Figure illustrates the upload process of the different supported input files into the EOPAS system, while Figure shows the web form presented to the user. In the first step, the users have to upload their xml files from one of the tools. The server then checks the input files against the EOPAS specific schema, which describes the various restrictions explained in the previous section, for that particular input. If the files have been created using our template files (for Toolbox and ELAN) correctly, they will validate against the schema, otherwise the user is presented with an error message indicating the wrong use of an element or attribute, etc.

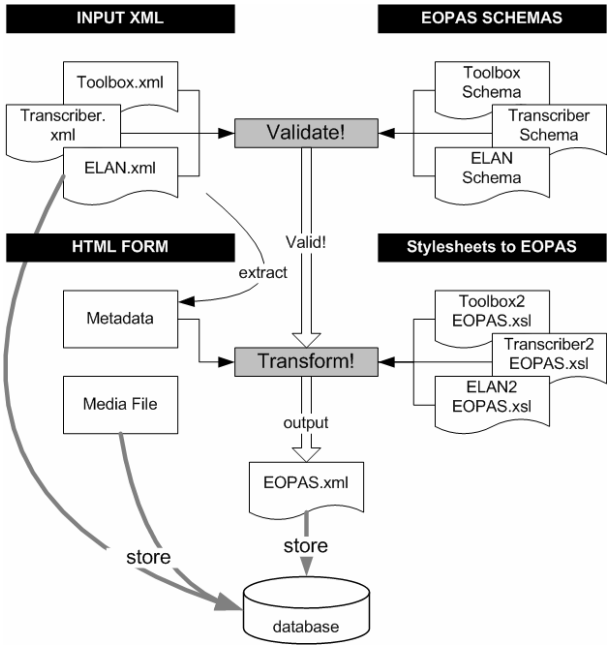

Figure 5: Upload process

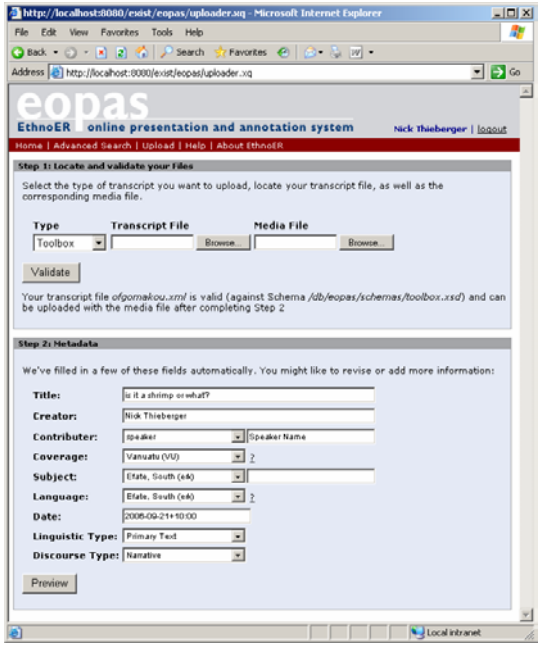

Figure 6: Screenshot of the upload process and the metadata form

Only when the file validates will the users get to step 2, which requires the user to enter additional metadata through a web form (see Figure). Some of the metadata fields are filled out automatically by extracting the information from the input files where applicable. The form also hooks into repositories of controlled vocabulary, e.g. to enter the proper country or language code. Finally, the user needs to specify the media file (audio or video) for the

11 http://ethnoer.unimelb.edu.au/EOPAS.zip

transcription. After confirming the provided information, the original input file, the media file and the EOPAS file (including the metadata and a link to the media file) are stored in the database as shown in Figure .

Download & Presentation

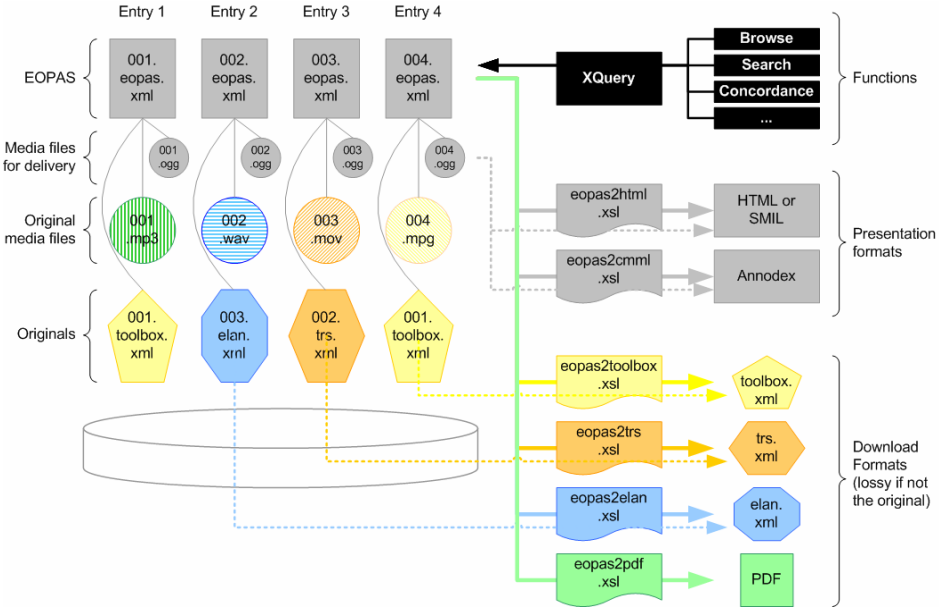

Figure 7: Download & Presentation

Figure illustrates the design for archiving and presenting the various transcription and IT files. As described in the previous section, the different original files (indicated by different shapes and colours) entered into the system are transformed into the EOPAS format but archived alongside so no information is lost. A similar strategy applies to the various media files, archival originals are transcoded to a streaming delivery format, currently Ogg vorbis (audio) and Ogg theora (video). CSIRO's Annodex[12] system is the delivery option being explored during this phase of EOPAS as it allows streaming delivery of media linked to text. Citation forms of the data can be accessed via standard HTML URLs. Annodex can be implemented in one of two ways at the moment. Annodex files are files that combine the media (in ogg format) and the transcript (in CMML format) into one file and are served by an Apache server. A URL in the form of http://mediafile.anx?[time] will jump to the specified timecode location and start

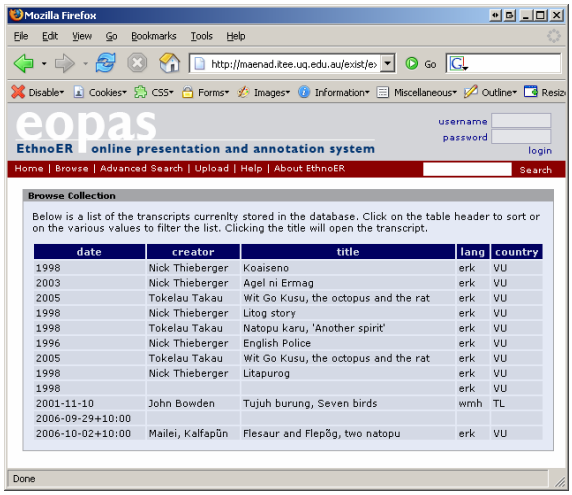

Figure 8: Screenshot of the browsing function over some test data

[12] http://www.annodex.net

streaming the content from that point onwards. An alternative is to host a CmmlWiki that serves media and CMML files, joining them on the fly.

Based on the EOPAS files, we can now perform a wide range of tasks over the whole collection of the uploaded data. For example we can browse the collection based on transcripts of a specific language, or from a specific author (Figure). Of course we can also search the collection based on these restrictions, plus do a text based search, e.g. find all phrases that use a specific word or even find all phrases where one word has a specific distance (number of words between two words) to another word.

Figure 9: Screenshot of viewing IT from the EOPAS repository

Another function is the concordance view (bottom left in Figure): every word or morpheme within a transcript can be clicked to directly retrieve and view its concordance within all transcripts of that particular language in the database.

Entire transcripts can be viewed in HTML (Figure) and each phrase (or chunk) can be clicked to play the actual media in the media player. Figure shows a transcript file viewed through the Annodex Player.

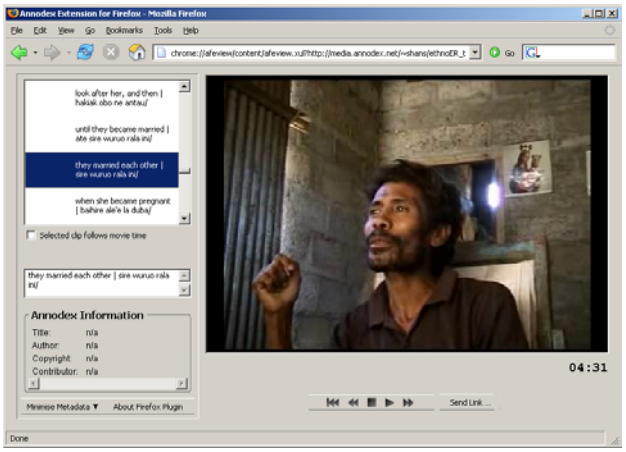

Figure 10: Screenshot of viewing annotations in the Annodex player

The transcripts are transformed from the EOPAS format to HTML or CMML on the fly using appropriate XSL style sheets. XSL style sheets are also used to convert EOPAS into file formats other than the original. So a Transcriber file uploaded to the EOPAS repository can be downloaded as a Toolbox file. Finally, using XSL:FO, the transcript can be converted into files more suitable for printing, such as PDF or RTF.

Sustainability of this data

The model we are building includes the option of depositing data with PARADISEC as part of the upload process. Not all data will be suitable for accession into PARADISEC, and we have already discussed the fact that the textual data in EOPAS may be a delivery version of more complex (archival) text. A future component of the project will be delivery of the EOPAS encoded file to a suitable repository, together with its metadata.

We need to consider how the files in this format can be located and accessed. We could establish the textual repository as one of the OLAC archives and then benefit from that existing infrastructure. Ideally, the EOPAS filetype would allow discovery via the existing OLAC repositories and be playable via our system. The HTML representation of the EOPAS files already includes the DC metadata in the HTML meta elements[13] as we envisage that web crawlers used by search engines will be able to harvest them.

Problems

Unicode characters have proven to cause problems. The eXist[14] database creates an index of keyword search strings for its full text search operators, e.g. "&=" and "|=". The keyword search strings are split into tokens using its default tokenizer function. This seems to work well for all European languages. However, in languages such as South Efate words like "natam̃ol" are being split into "natam" and "ol", because the combining character (tilde) ̃ after the m is not recognised by the tokenizer. Essentially this means that for the current implementation we need to avoid using the full text search operators of eXist (which perform up to 10 times faster) until the eXist's tokenizer has been modified accordingly.

On the presentation side of Unicode characters one needs to be careful what font to use, as not all fonts provide a means to display these special characters. Arial seems to work for us, while Verdana, a similar sans-serif font, and Times do not display them correctly.

Badly formed input data is a problem, but its failure to validate allows us to see inconsistencies in the primary data. However, if a line is just a single character out of alignment it will not be correctly exported by Toolbox, but will be validated by the import process, so it is necessary to check all uploaded files visually.

A final problem has been the wrapping of interlinear text. This is a notorious issue (see for example Schmidt 2003: 16 or Bow, Hughes and Bird 2003: 39) and one which causes real problems with delivery of IT, especially where pagination and an output to RTF or PDF is required. We have used an HTML rendering that allows the IT to wrap in table cells[15] but we still need to find a way of implementing wrapping using XSL:FO. The difficulty here is not only to find the appropriate way to do this, but also that different XSL:FO engines support the XSL:FO specifications to varying degrees.

---

[13] http://dublincore.org/documents/2000/08/15/dcq-html/#structure

[14] http://exist.sourceforge.net/

[15] This was provided to us by John Thomson of SIL.

Directions

We envisage EOPAS or something similar becoming a standard for online presentation which allows a linguist to place well-formed data into a repository and have the appropriate tool read and present the data in the file. At the moment we have the OLAC search mechanism which takes XML catalogues of data repositories and makes them available via a query interface. A similar approach could be taken to any file that declared itself to conform to the EOPAS schema and so became available for online presentation.

The current implementation provides for glossing in one language only, but one of our sample datasets, from Timor-Leste, includes morphemic glossing and free translations in Indonesian, English, Portuguese, and Tetun, and these should be available via EOPAS in a future implementation.

In the current implementation we have constrained input from Toolbox to contain specified field markers, but it should be possible to add an input function that queries the user for marker names and converts from them to the appropriate EOPAS equivalents and assigns the hierarchy to them rather than relying on the Toolbox .typ file to provide this.

Conclusion

Through the construction of the EOPAS system it has become clear that we must encourage standard data structures in whatever tool is used for transcription and annotation of linguistic data. While much of the data produced using current tools is reusable because it is in a non-proprietary format and has an explicit structure, it is not interoperable without considerable effort. The only method that seems to be capable of inducing linguists to use constrained data structures is the provision of templates as part of a clearly described workflow, resulting in data presentation that can be made available in a variety of views. Further, advocacy and training are critical to the uptake of suitable tools and methods in linguistics and more generally in the humanities.

References

Bow, Catherine, Baden Hughes and Steven Bird. 2003. Towards a General Model of Interlinear Text. Proceedings of the EMELD Language Digitization Project Conference 2003: Workshop on Digitizing and Annotating Texts and Field Recordings. [http://www.emeld.org/workshop/2003/papers03.htm]

Drude, Sebastian. 2003. Advanced Glossing — a language documentation format and its implementation with Shoebox. Proceedings of the LREC Workshop in May 2002. Las Palmas – International workshop on resources and tools in field linguistics.

Hellmuth, Chris, Tom Myers & Alexander Nakhimovsky. 2006. The Linguist's Toolbox and XML Technologies. Paper presented at the EMELD meeting, [http://emeld.org/workshop/2006/papers/hellmuth.html]

Hughes, Baden, Steven Bird and Catherine Bow. 2003. Encoding and Presenting Interlinear Text Using XML Technologies. In Alistair Knott and Dominique Estival (eds.) *Proceedings Australasian Language Technology Workshop,* Melbourne, Australia. 105-113.[http://eprints.unimelb.edu.au/archive/00000455/]

Jacobson, Michel, Boyd Michailovsky, and John B. Lowe. 2001. Linguistic documents synchronizing sound and text. *Speech Communication* 33, 1-2: 79-96

Schmidt, Thomas. 2003. Visualising Linguistic Annotation as Interlinear Text. Arbeiten zur Mehrsprachigkeit. Working papers in multilingualism. Series B. Hamburg: Univ. Hamburg

Appendix A: EOPAS file format

```
<?xml version="1.0" encoding="UTF-8"?>
<eopas xmlns="http://wiki.arts.unimelb.edu.au/ethnoer/EOPAS/1.0/" xmlns:olac="http://www.language-
archives.org/OLAC/1.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/">
  <header>
    <olac:olac>
      <dc:format xsi:type="dcterms:IMT">text/xml</dc:format>
      <dc:identifier xsi:type="dcterms:URI">18405154</dc:identifier>
      <dcterms:requires xsi:type="dcterms:URI">mediafiles/18405154.vob</dcterms:requires>
      <dc:creator>Nick Thieberger</dc:creator>
      <dc:title>Wit Go Kusu</dc:title>
      <dc:contributor xsi:type="olac:role" olac:code="annotator">Nick Thieberger</dc:contributor>
      <dc:contributor xsi:type="olac:role" olac:code="speaker">Tokelau Takau</dc:contributor>
      <dc:coverage>VU</dc:coverage>
      <dc:subject xsi:type="olac:language" olac:code="erk"/>
      <dc:language xsi:type="olac:language" olac:code="erk"/>
      <dc:date>2006-09-22+10:00</dc:date>
      <dc:type xsi:type="olac:linguistic-type" olac:code="Primary Text"/>
      <dc:type xsi:type="olac:discourse-type" olac:code="Narrative"/>
    </olac:olac>
  </header>
  <Interlinear-text>
    <phrases>
      <phrase id="s1" startTime="10.963" endTime=" 16.802">
        <text xsi:type="translation">I want to tell you, Nick, I'll tell you about the rat and the octopus.</text>
        <text xsi:type="orthographic">Amurin gag puserek, Nick, kafo gag pusereki kusu go wit. </text>
        <words>
          <word>
            <text xsi:type="orthographic">Amurin</text>
            <morphemes>
              <morpheme>
                <text xsi:type="morpheme">a=</text>
                <text xsi:type="gloss">1sgRS=</text>
              </morpheme>
              <morpheme>
                <text xsi:type="morpheme">mur</text>
                <text xsi:type="gloss">want</text>
              </morpheme>
              <morpheme>
                <text xsi:type="morpheme">-i</text>
                <text xsi:type="gloss">-TS</text>
              </morpheme>
              <morpheme>
                <text xsi:type="morpheme">-n</text>
                <text xsi:type="gloss">-3sgO</text>
              </morpheme>
            </morphemes>
          </word>
          <word>
            <text xsi:type="orthographic">gag</text>
            <morphemes>
              <morpheme>
                <text xsi:type="morpheme">gag</text>
                <text xsi:type="gloss">2sgBEN</text>
              </morpheme>
            </morphemes>
          </word>
          <word>
            <text xsi:type="orthographic">puserek,</text>
            <morphemes>
              <morpheme>
                <text xsi:type="morpheme">puserek</text>
                <text xsi:type="gloss">talk</text>
              </morpheme>
            </morphemes>
          </word>
```

```
<word>
  <text xsi:type="orthographic">Nick,</text>
  <morphemes>
    <morpheme>
      <text xsi:type="morpheme">Nick</text>
      <text xsi:type="gloss">Nick</text>
    </morpheme>
  </morphemes>
</word>
<word>
  <text xsi:type="orthographic">kafo</text>
  <morphemes>
    <morpheme>
      <text xsi:type="morpheme">ka=</text>
      <text xsi:type="gloss">1sgIRR=</text>
    </morpheme>
    <morpheme>
      <text xsi:type="morpheme">fo</text>
      <text xsi:type="gloss">PSP.IR</text>
    </morpheme>
  </morphemes>
</word>
<word>
  <text xsi:type="orthographic">gag</text>
  <morphemes>
    <morpheme>
      <text xsi:type="morpheme">ag</text>
      <text xsi:type="gloss">sgBEN</text>
    </morpheme>
  </morphemes>
</word>
<word>
  <text xsi:type="orthographic">pusereki</text>
  <morphemes>
    <morpheme>
      <text xsi:type="morpheme">puserek</text>
      <text xsi:type="gloss">talk</text>
    </morpheme>
    <morpheme>
      <text xsi:type="morpheme">-ki</text>
      <text xsi:type="gloss">-TR</text>
    </morpheme>
  </morphemes>
</word>
<word>
  <text xsi:type="orthographic">kusu</text>
  <morphemes>
    <morpheme>
      <text xsi:type="morpheme">kusu</text>
      <text xsi:type="gloss">rat</text>
    </morpheme>
  </morphemes>
</word>
<word>
  <text xsi:type="orthographic">go</text>
  <morphemes>
    <morpheme>
      <text xsi:type="morpheme">go</text>
      <text xsi:type="gloss">and</text>
    </morpheme>
  </morphemes>
</word>
<word>
  <text xsi:type="orthographic">wit.</text>
  <morphemes>
    <morpheme>
      <text xsi:type="morpheme">wit</text>
      <text xsi:type="gloss">octopus</text>
    </morpheme>
  </morphemes>
```

```
            </word>
          </words>
        </phrase>
        <phrase id="s2" startTime="16.802" endTime=" 20.465">
          <text xsi:type="translation">One day, it was low tide.</text>
          <text xsi:type="orthographic">Naliati iskei, elau imat. </text>
          <words>
            <word>
              <text xsi:type="orthographic">Naliati</text>
              <morphemes>
                <morpheme>
                  <text xsi:type="morpheme">naliati</text>
                  <text xsi:type="gloss">day</text>
                </morpheme>
              </morphemes>
            </word>
            <word>
              <text xsi:type="orthographic">iskei, elau imat.</text>
              <morphemes>
                <morpheme>
                  <text xsi:type="morpheme">i=</text>
                  <text xsi:type="gloss">3sg.R.SBJ=</text>
                </morpheme>
                <morpheme>
                  <text xsi:type="morpheme">skei</text>
                  <text xsi:type="gloss">one</text>
                </morpheme>
                <morpheme>
                  <text xsi:type="morpheme">elau</text>
                  <text xsi:type="gloss">sea</text>
                </morpheme>
                <morpheme>
                  <text xsi:type="morpheme">i=</text>
                  <text xsi:type="gloss">3sgRS=</text>
                </morpheme>
                <morpheme>
                  <text xsi:type="morpheme">mat</text>
                  <text xsi:type="gloss">low_tide</text>
                </morpheme>
              </morphemes>
            </word>
          </words>
        </phrase>
      </phrases>
    </Interlinear-text>

</eopas>
```