IGT-XML: an XML format for interlinearized glossed texts

Alexis Palmer, Katrin Erk
Department of Linguistics
University of Texas at Austin
{alexispalmer,katrin.erk}@mail.utexas.edu

Abstract

We propose a new XML format for representing interlinearized glossed text (IGT), particularly in the context of the documentation and description of endangered languages. The proposed representation, which we call IGT-XML, builds on previous models but provides a more loosely coupled and flexible representation of different annotation layers. Designed to accommodate both selective manual reannotation of individual layers and semi-automatic extension of annotation, IGT-XML is a first step toward partial automation of the production of IGT.

1 Introduction

Much previous work on linguistic annotation has necessarily focused on resource-rich languages, as it is these languages for which we have large corpora in need of linguistic annotation. In contrast, development of annotation schemata and methodologies to be used with language data from endangered languages has been left largely to individual documentary and/or descriptive linguists working with particular languages.

This paper addresses linguistic annotation in the context of the documentation and description of endangered languages. One interesting feature of language documentation projects is that, while the languages studied differ widely, there is a quasi-standard for presenting the material, in the form of **interlinearized glossed text (IGT)**. IGT typically comprises at least four levels: (1) the original text, (2) a separation of the original text into individual morphemes, (3) a detailed morpheme-by-morpheme gloss, and (4) a free translation of each sentence. Another characteristic of language documentation projects is the tentative nature of many analyses, given that linguistic analysis is often occurring in tandem with the annotation process, sometimes for the first time in the recorded history of the language. Furthermore, language documentation projects require long-term accessibility of the collected language data as well as easy accessibility to community members as well as to linguists.

In this paper we propose a new XML format for representing IGT, which we call **IGT-XML**. We build on the model of Hughes et al (2003) (the **BHB model** from now on), who first proposed using the IGT structure directly as a basis for an XML format. While their format shows closely integrated annotation layers using XML embedding, our model has a more loosely coupled and flexible representation of different annotation layers, to accommodate (a) selective manual reannotation of individual layers, and (b) the (semi-)automatic extension of annotation, without the format posing an a priori restriction on the annotation levels that can be added. The IGT-XML representation is thus a first step toward partial automation of the production of IGT, which in turn is part of a larger project using techniques from machine learning and natural language processing to significantly reduce the time and money required to produce annotated texts.

Besides the BHB model, we build on the Open Languages Archiving Community (OLAC)[1] metadata standard. OLAC is developing best practice guidelines for archiving language resources digitally, including a list of metadata entries to record

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

Proceedings of the Linguistic Annotation Workshop, pages 176–183,
Prague, June 2007.

with language data.

Plan of the paper. After discussing interlinearized glosses in Section 2, we show the BHB model and corresponding XML format in Section 3. Section 4 presents the IGT-XML format that we propose. Section 5 demonstrates the applicability of IGT-XML to data from different languages and different documentation projects, and Section 6 concludes.

2 Interlinearized glossed text

IGT is a way of encoding linguistic data commonly used to present linguistic examples. The example below is a segment of IGT taken from Kuhn and Mateo-Toledo (2004). The language is Q'anjob'al, a Mayan language of Guatemala.

(1) `Maxab' ek'elteq ix unin yet sq'inib'alil tu.`

(2) `max-ab' ek'-el-teq ix unin y-et`
COM-EV pass-DIR-DIR CL child E3S-when
`s-q'inib'-al-il tu`
E3S-early-ABS-ABS DEM

'The child came out early that morning (they say)' [2]

The format of the IGT in this example is typical of the presentation of individual examples in the linguistics literature. The raw, unannotated text (1) is associated with three layers of annotation, shown in (2). The first annotation layer shows the same text with each word segmented into its constituent morphemes. The next layer, the gloss layer, is a combination of English translations of the Q'anjob'al lemmas and tags representing the linguistic information encoded by affixes on the lemmas. The third layer is an English translation.

IGT formats vary more widely in language documentation, where IGT is typically the product of linguistic analysis of texts transcribed from audio or audiovisual recordings. A broad survey of formats for interlinear texts (Bow et al., 2003) found variation in the number of rows, the type of analysis found in each row, as well as the level of granularity of analysis in each row.[3]

[2]KEY: ABS=abstract, COM=completive, CL=classifier, DEM=demonstrative, E=ergative, EV=evidential, S=singular, 3=third person

[3]Hughes et al (2003) also discuss variation in presentational factors, which we choose not to encode in our XML format.

Tools using IGT Shoebox/Toolbox[4] (*Shoebox* in following text) is a system that is widely used in documentary linguistics for storing and managing language data. It provides facilities for lexicon management as well as text interlinearization.

Figure 1 shows one sentence of Q'anjob'al IGT in the Shoebox output format.[5] Shoebox exports texts as plain text files. The different annotation layers are marked by labels at the beginning of the line. For example, in Figure 1 the label `\tx` marks the original text and the line starting with `\dm` contains its morphological segmentation.

One important test case for any XML format for IGT is whether it can represent existing IGT data. As Shoebox is a widely used tool, we take the Shoebox data format as a representative case study. Specifically, in Section 5 we show how texts from two different languages, interlinearized using Shoebox and represented in the Shoebox output format, can be encoded in IGT-XML.

In this paper we focus on the question of representation rather than format transformation. Each system managing IGT data will have different output formats, requiring different techniques for transforming the data to XML. The aim of this paper is simply to describe and demonstrate the IGT-XML format; a detailed automatic transformation method mapping other formats to IGT-XML is beyond the scope of this paper and will be addressed separately.

3 Previous work

This section discusses previous work on representation formats and specifically XML formats for interlinear text.

The BHB model: four levels of interlinear text. Building on Bow et al.'s (2003) analysis of different IGT formats used in the literature, Hughes et al. (2003) propose a four-level hierarchical model for representing interlinear text. The four levels encode elements common to most instances of IGT: *text*, *phrase*, *word*, and *morpheme*. One *text* may consist of several individual *phrases*. A *phrase* consists of one or more *words*, each of which consists

[4]`http://www.sil.org/computing/catalog/show_software.asp?id=79`

[5]Data from B'alam Mateo-Toledo, p.c.

```
\ref txt080_p2.002
\tx Exx  a    yet             junxa          tyempohal,  ayin ti'  xiwil+
\dm exxx a    y-   et         jun      - xa  tyempo -al, ayin ti   xiwil+
\ge INTJ ENF  E3-  de/cuando  ART/uno  - ya  tiempo -ABS yo   DEM  muchos
\cp intj part pref- sr        num      - adv s      -suf pro  part adv

\tes Eee en otro tiempo yo vi
```

Figure 1: Shoebox output: Q'anjob'al


```
<resource>
<interlinear_text>
 <item type="title">Example</item>
 <phrases>
   <phrase>
    <item type="gls">The child came out
    early that morning (they say)</item>
    <words>
      <word>
        <item type="txt">ek'elteq</item>
        <morphemes>
          <morph>
            <item type="txt">ek'</item>
            <item type="gls">pass</item>
          </morph>
          <morph>
            <item type="txt">el</item>
            <item type="gls">DIR</item>
          </morph>
          <morph>
            <item type="txt">teq</item>
            <item type="gls">DIR</item>
          </morph>
        </morphemes>
      </word>
    </words>
   </phrase>
 </phrases>
</interlinear_text>
</resource>
```


Figure 2: BHB IGT representation format: Q'anjob'al

of one or more *morphemes*. To make this more concrete, the example in (1) shows a single phrase (or a one-phrase text). The three annotation layers in (2) are situated at different levels in the hierarchy: The first and second annotation layers are both situated at the morpheme level, showing a separation of the original phrase into its constituent morphemes and a morpheme-by-morpheme gloss, respectively. The third annotation layer, the translation, is again situated at the phrase level, like the original text in (1).

The BHB model was originally developed in the context of the EMELD project,[6] which has focused on advancing the state of technologies, data representation formats, and methodologies for digital language documentation.

The BHB XML format. Figure 2 shows an example of the BHB XML format, which articulates the four nested levels of structure of the BHB model. It directly expresses the hierarchy of annotation levels in a nested XML structure, in which, for example, `<morph>` elements representing morphemes are embedded in `<word>` elements representing the corresponding words. The model maintains the link between the source text morpheme and the morpheme-level gloss annotation by embedding both as `<item>` elements within the `<morph>` and distinguishing the two by an attribute called `type`.

While this representation provides the needed link between morphemes and their glosses, it is rather inflexible because it is not modular: To add an additional annotation layer at the word level, one would need to access and change the representation of each word of each phrase. In this way, the BHB XML format is not ideally suited for an extensible annotation that would need to add additional layers of linguistic information in a flexible way.

[6] http://linguistlist.org/emeld

4 IGT-XML

In this section we propose a new XML representation for IGT, IGT-XML. Like the BHB XML format, it is based on the BHB four-level model, but it modularizes annotation levels. Linking between annotation levels is achieved via unique IDs.

The IGT-XML format.

Figure 3 illustrates the new IGT-XML format, showing a representation of the Q'anjob'al example of Figure 1, mostly restricted to a single word, *tyempohal*, for simplicity.

The IGT-XML format contains (at least) three main components:

- a *plaintext* component comprising phrases as well as the individual words making up each phrase, encased in the `<phrases>` XML element,
- a *morpheme* component giving a morphological analysis of the source text, encased in the `<morphemes>` XML element, and
- a *gloss* component including glosses at both the phrase and the word level.

Further annotation layers can be added by extending the format with additional components beyond these three, which describe the core four levels of interlinear text.

Within the `<phrases>` block, each individual phrase is encased in a `<phrase>` element, which includes the plain text within the `<plaintext>` element as well as each individual word of the plain text in a `<word>` element. Each `<phrase>` and each `<word>` has a globally unique ID, assigned in an `id` attribute. We choose to give explicit IDs to words, rather than rely on character offsets, to avoid possible problems with character encodings and mis-represented special characters.

The morphemes in the `<morphemes>` block are again organized by `<phrase>`. Each `<phrase>` in the `<morphemes>` block refers to the corresponding phrase in the `<phrases>` block by that phrase's unique ID.

Each individual morpheme, represented by a `<morph>` element, refers to the unique ID corresponding to the word of which it is a part. The linear order of morphemes belonging to the same word is reflected in the order in which `<morph>` elements appear, as well as in the running `id` of the morphemes. Morphemes have `id` attributes of their own such that further annotation levels can refer to the morphological segmentation of the source text, as is the case for the morpheme-by-morpheme gloss in the example in (2).

Whole-sentence glosses are collected in the `<translations>` block, while word-by-word glosses reside in the `<gloss>` block. Again, glosses are organized by `<phrase>`, linked to the original phrases by `idref` attributes. The glosses in `<gloss>` refer to individual morphemes, hence their `idref` attributes point to `id` attributes of the `<morphemes>` block.

Metadata information in the file header

As suggested in Figure 3, IGT-XML is easily extended with metadata for each text. We adopt the OLAC metadata set which uses the fifteen elements defined in the Dublin Core metadata standard (Bird and Simons, 2003a; Bird and Simons, 2001). These elements provide a framework for specifying key information such as annotators, format, and language of the text. In addition, the OLAC standard incorporates a number of qualifiers specific to the language-resource community, such as discourse types (story, conversation, etc.) and linguistic data types (lexicon, language description, primary text, etc.), and a process for adopting further extensions.

In addition to the metadata block at the head of the document, it would be possible to intersperse additional metadata blocks throughout the document, if for example we wanted to indicate change of speaker from one phrase to another in recorded conversation.

Discussion

Feature overview. The IGT-XML format we have presented groups annotation into blocks in a modular fashion. Each block represents an annotation layer. The format uses globally unique IDs (via `id` and `idref` attributes) rather than XML embedding for linking annotation layers. In particular, `<morph>` and `<word>` annotation is kept separate, such that additional layers of annotation at the word and morpheme levels can be added modularly without interfering with each other.

In its minimal form, the format has three blocks,

```
<text id="T1" lg="kjb" source_id="txt080_p2" title="Pixanej">
<metadata idref="T1">
  <!-- incorporate OLAC metadata standard -->
</metadata>
<body>
<phrases>
  <phrase id="T1.P2" source_id="txt080_p2.002">
    <plaintext>Exx a yet junxa tyempohal, ayin ti’ xiwil+</plaintext>
    <word id="T1.P2.W5" text="tyempohal"/>
  </phrase>
</phrases>
<morphemes source_layer="\dm">
  <phrase idref="T1.P2">
    <morph idref="T1.P2.W5" id="T1.P2.W5.M1" text="tyempo"/>
    <morph idref="T1.P2.W5" id="T1.P2.W5.M2" text="al">
      <type l="suf"/>
    </morph>
  </phrase>
</morphemes>
<gloss source_layer="\ge">
  <phrase idref="T1.P2">
    <gls idref="T1.P2.W5.M1" text="tiempo"/>
    <gls idref="T1.P2.W5.M2" text="ABS"/>
  </phrase>
</gloss>
<translations>
  <phrase idref="T1.P2">
    <trans id="T1.P2.Tr1" lg="en">Eee en otro tiempo yo vi</trans>
  </phrase>
</translations>
</body>
</text>
```

Figure 3: IGT-XML representation format: Q’anjob’al

for phrases, morphemes, and glosses, but it is extensible by further blocks, for example for POS-tags. It is also possible to have different types of annotation at the same linguistic level, for example manually created as well as automatically assigned POS-tags.

Mildly standoff annotation. The IGT-XML format keeps the plain text separate from all levels of annotation. However, it is not standoff in the strict sense of having all annotation levels refer to the plain text only and never to one another. The reason for this is that there is no clear "basic" level to which all other annotation could refer.

One obvious candidate is the plain text, but the morpheme-by-morpheme gloss refers not to words, but to the morpheme segmentation of the source text, as can be seen in example (2). This makes the morpheme-segmented source text another candidate for the basic level, but it is not guaranteed that this level of annotation will always be available. At the start of the annotation process the documentary linguist likely has a transcription and a translation, but he or she may or may not have determined the morphotactics of the language or even how to identify word boundaries.

So, in order (a) not to commit the annotator to one single order of annotation, or the presence of any particular annotation level besides the plain text, and (b) to allow annotation to refer to each of the levels identified in the BHB model – text, phrase, word, and morpheme –, we allow annotation levels to refer to each other via unique IDs.

Requirements for IGT formats. Given the nature of language documentation projects and IGT data, an IGT representation format should (1) support long term archiving of language data (Bird and Simons, 2003b), which requires platform-independent encoding, and it should (2) support a range of formats. IGT data from different sources may show differences in format and in what is annotated (Bow et al., 2003), and may be produced using different software systems. (3) It should be possible to add or exchange layers of annotation in a modular fashion. This is important because linguistic analysis in language documentation, which typically targets languages that are not well-studied, is often tentative and subject to change. This will also become increasingly important with the use of automation to aid and speed up language documentation: Automation techniques will typically target individual annotation layers, and it is desirable to be able to exchange automatic analysis tools freely.

Point (1), platform independence, is achieved by almost any XML format, since XML formats are plain text-based and mostly human-readable. Point (2), the coverage of IGT formats in all variants, can be achieved by adoption of the BHB model. Flexibility and modularity (point (3)) are the main motivations in the introduction of IGT-XML.

Beyond word-level annotation. For now the annotation focus in language documentation projects is mostly on the word level, especially on morphology and POS-tags. For annotation at the syntactic level, it is an open question what the features of a universally applicable annotation format should be. At the moment, TIGER XML (Mengel and Lezius, 2000), with its capability to represent discontinuous constituents, and constituent as well as dependency information, seems like a good candidate. Syntactic information could be represented in a separate top-level XML element, linking tree terminals to `<word>` elements by their ID attributes.

5 Data

An important goal of this research is to develop an XML format which will be viable for use in the broadest possible range of language documentation contexts. To that end, the format needs to stretch and morph with the needs and desires of the individual user. This section discusses some issues arising from actual use of the format. The points are illustrated with pieces of the XML representation rather than complete XML documents.

IGT-XML has been used to encode portions of texts from the Mayan language Q'anjob'al and the Mixe-Zoquean language Soteapanec (more commonly known as Sierra Popoluca). Q'anjob'al is spoken primarily in the northwestern regions of Guatemala, and Soteapanec is spoken in the southern part of the state of Veracruz, Mexico. Both texts come from ongoing documentation efforts, and both were first interlinearized using Shoebox.

5.1 Q'anjob'al

Figure 1 shows a Q'anjob'al sentence in the Shoebox export format. The annotation comprises original text (`\tx` level), morphological analysis (`\dm`), morpheme gloss (`\ge`), and parts of speech (`\cp`). The Q'anjob'al texts we received preserve links between Shoebox annotation layers only through typographical alignment. The IGT-XML representation makes these links explicit through global IDs using `id` and `idref` attributes. It also splits off punctuation, treating punctuation marks as separate words:

```
<word id="T1.P2.W5" text="tyempohal"/>
<word id="T1.P2.W6" text=","/>
<word id="T1.P2.W7" text="ayin"/>
```

In the part of speech annotation level (line `\cp`), the annotator has additionally marked prefixes and suffixes, using the labels `pref-` and `-suf`, respectively. In the IGT-XML, we have incorporated this information in the `<morphemes>` level as *type* information on a morpheme. Figure 3 shows an example of this, extended below:

```
<morph idref="T1.P2.W5" id="T1.P2.W5.M1"
       text="tyempo"/>
<morph idref="T1.P2.W5" id="T1.P2.W5.M2"
       text="al">
  <type l="suf"/>
</morph>
<morph idref="T1.P2.W6" id="T1.P2.W6.M1"
       text=",">
  <type l="punct"/>
</morph>
```

By encoding morpheme type as a `<type>` element embedded in the `<morph>`, we can allow a single morpheme to bear more than one type label. For example, an annotator may want to mark a single morpheme as being an inflectional morpheme which appears in a suffixal position. This would be indicated by associating multiple `<type>` elements with a single `<morph>` element, differentiating the `<type>` elements through use of the label (`l`) attribute, as shown in the constructed example below.

```
<morph idref="T3.P1.W3" id="T3.P1.W3.M2"
       text="al">
  <type l="suf"/>
  <type l="infl"/>
</morph>
```

Furthermore, as the type label is specified in an attribute value, each documentation project can specify its own list of possible labels.

```
\ref Jovenes 002
\t Weenyi  woony=jaych@@x+tyam
\mb weenyi  woonyi=jay.ty@@xi+tam
\gs algunos  varon+HPL

\t yo7om@7yyajpa+m
\mb 0+yoomo.7@7y-yaj-pa+m
\gs 3ABS+casar  con  mujer-3PL-INC+ALR

\f Algunos nin*os se casan.
```

Figure 4: Shoebox output: Soteapanec

5.2 Soteapanec

Figure 4 shows the Shoebox output for a Soteapanec phrase.[7] In the notation chosen in this project, the characters '7' and '@' refer to phonemes (glottal stop and mid high unrounded vowel, respectively), while '-', '+', '>', '=' and '.' all mark morpheme boundaries. Clitic boundaries are marked by '+', inflectional boundaries by '-', derivational boundaries by '>' or '.', and compounds are indicated with '='. The four different morpheme boundaries translate to morpheme types in the IGT-XML, which are encoded as in the Q'anjob'al case:

```
<morph idref="T1.P2.W1" id="T1.P2.W1.M1"
       text="weenyi"/>
<morph idref="T1.P2.W2" id="T1.P2.W2.M1"
       text="woonyi=jay">
  <type l="compound"/>
</morph>
<morph idref="T1.P2.W2" id="T1.P2.W2.M2"
       text="ty@@xi"/>
<morph idref="T1.P2.W2" id="T1.P2.W2.M3"
       text="tam">
  <type l="suf"/>
</morph>
```

The encoding of the compound represents one of many choices to be made by users of IGT-XML. We have chosen to present the compound *woonyi=jay* as a single morpheme, in line with the linguist's choice to notate the compounds this way in the text. An alternative would be to break the compound into two separate morphemes, each marked as a compound via the `l` attribute of the `<type>` element.

A similar choice exists with respect to the representation of other derivational morphology, both at the level of morphological segmentation and at the level of the plaintext. In this case, the plaintext of the Soteapanec includes boundary markers. IGT-XML

[7] Data from (Franco and de Jong Boudreault, 2005).

can accommodate this type of text as well as it can a truly plain text.

In this Shoebox output, there is no typographical alignment between annotation levels. So the manual transformation to IGT-XML had to rely on counting morphemes. However there are frequent mismatches between the number of morphemes in the morphological level (\mb) and the gloss level (\gs). The second group of lines in Figure 4 shows an example: There are six morphemes on the \mb level, but seven on the \gs level. We envision that automatic transformation to IGT-XML will flag such cases as mismatched, thus functioning as error detection for the annotation. Even in the manual transformation process, we have marked mismatches at the gloss level to facilitate adjudication by the annotator.

```
<morph idref="T1.P2.W2.M4"><gls text="HPL"
   flag="mismatch" flagsrc="amp"
   flagdate="031507"/>
</morph>
```

We also include the source and date of the flag, attributes which could easily be obtained automatically.

This section provides only a sample of the issues encountered using IGT-XML. One of our next steps is to work on automatic transformations from Shoebox data formats to IGT-XML, a stage at which many of these challenges will necessarily be addressed.

6 Conclusion

In this paper we have introduced a new XML format for representing language documentation data, IGT-XML. At the heart of the model is a representation of interlinearized glossed text (IGT). Building on the BHB model (Hughes et al., 2003), IGT-XML represents original text, its translation, a morphological analysis of the original text, and a morpheme-by-morpheme gloss. Different annotation layers are represented separately in a modular fashion, allowing for flexible annotation of individual layers as well as the extension by further annotation layers. Layers are linked explicitly via globally unique IDs, using `id` and `idref` attributes.

One main aim in the design of the IGT-XML format is to facilitate the (semi-)automatic annotation of language documentation data. In fact, our next step will be to explore the use of computational tools for speeding up and extending the annotation of less-studied languages. This connection of documentary and computational linguistics has the potential to be very useful to documentary linguists. It also represents an interesting opportunity for the use of semi-supervised machine learning techniques like active learning on a novel application.

Acknowledgments

We would like to thank Lynda de Jong Boudreault and B'alam Mateo-Toledo for sharing with us data collected in their documentation efforts.

References

Steven Bird and Gary Simons. 2001. The OLAC metadata set and controlled vocabularies. In *Proceedings of ACL Workshop on Sharing Tools and Resources for Research and Education*, pages 7–18, Toulouse.

Steven Bird and Gary Simons. 2003a. Extending Dublin Core Metadata to support the description and discovery of language resources. *Computing and the Humanities*, 37:375–388.

Steven Bird and Gary Simons. 2003b. Seven dimensions of portability for language documentation and description. *Language*, 79(3):557–582.

Catherine Bow, Baden Hughes, and Steven Bird. 2003. Towards a general model of interlinear text. In *Proceedings of EMELD Workshop 2003: Digitizing and Annotating Texts and Field Recordings*, LSA Institute: Lansing MI, USA.

Julia Albino Franco and Lynda de Jong Boudreault. 2005. Jovenes. Unpublished annotated text. University of Texas at Austin.

Baden Hughes, Steven Bird, and Catherine Bow. 2003. Encoding and presenting interlinear text using XML technologies. In Alistair Knott and Dominique Estival, editors, *Proceedings of the Australasian Language Technology Workshop*, pages 105–113.

Jonas Kuhn and B'alam Mateo-Toledo. 2004. Applying computational linguistic techniques in a documentary project for Q'anjob'al (Mayan, Guatemala). In *Proceedings of LREC 2004*, Lisbon, Portugal.

Andreas Mengel and Wolfgang Lezius. 2000. An XML-based encoding format for syntactically annotated corpora. In *Proceedings of LREC 2000*, Athens, Greece.