Culture clash – Humanities research and computing: a case study of Interlinear Glossed Text (IGT)

Nick Thieberger
School of Languages and Linguistics, University of Melbourne, Parkville, Vic 3010, Australia
thien@unimelb.edu.au

Abstract

Interlinear glossed text (IGT) is an ideal representation of text for field linguists, but it is difficult to construct, and even more difficult to query as plain text. A major problem for IGT is the lack of agreed standards for its construction, underlying form, and presentation. Among field linguists technological standards are notoriously based on the use of specific tools so I developed a schema-based XML standard (EOPAS) for presentation of IGT. There is, however, a lack of connection between large-scale and well-funded language computation projects and the needs of linguists in language documentation that is symptomatic of a larger problem for humanities researchers. How can we articulate our needs and how can we obtain solutions to relatively trivial computational tasks that are beyond our abilities to implement by ourselves?

1 Introduction*

I am a linguist working on the description of previously undocumented languages (in central Vanuatu and in Western Australia). I also help run the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). My interest is in developing an efficient workflow for field linguists so that the material we produce is useful and accessible to use in our analysis, but also so that it can be archived and perhaps made available online. We are, in general, the only annotators of our corpus which is typically made up of a heterogenous set of texts collected opportunistically. While we may train speakers or have good intentions of training speakers to do this work, it is often the case that speakers do not want to do it, and the difficulty in using the tools described below is not a small impediment to their involvement. I am a corpus creator for South Efate, a language from Vanuatu of which I wrote a grammar and dictionary, and I want my collection of texts to be usable beyond my particular use of them, and with reference to the media from which they are transcribed.

This paper has three main points to make. The first is that Interlinear glossed text (IGT) is a problematic format that needs specialised tools, and those currently in use are not ideal for several reasons. The second is to outline the EOPAS system we built for viewing IGT together with media, using open-source tools and including what we propose as a standard schema for the underlying and archival representation of IGT. The third point is a general observation about the clash between the range of computational needs of humanities researchers and the higher-level problems addressed by computational linguists and the eResearch infrastructure more generally, leaving a middle ground in which little work appears able to be done. The contribution to the Australian National Corpus could be in the use of EOPAS or the steaming media server Annodex for presentation of text and media in languages or varieties other than standard English.

2 Interlinear Glossed Text (IGT)

Interlinear glossed text (IGT) is an ideal representation of text for linguists. Typically, it presents a line of text in another language, a morphemic line and a line with morphemic glosses, followed by a free translation as shown in Fig.1. Recent advances in access to digital media mean that the text can also be viewed together with the media it transcribes.

Ipitlak nmatu inru, rato elag Ep̃uf.
```
i=    pitlak nmatu i=    nru ra=      to   elag  Ep̃uf
3sRS= have   woman 3sRS= two 1d.exRS= stay above place
```
There are two women, they lived up at Bufa.
NT1-98004-A.mp3 210.34, 212.2

Fig. 1. Example of IGT

While a number of papers (cited below) have been written on the theoretical and underlying structure of IGT, we are yet to have a modern tool that allows a linguist to create well-formed IGT (in terms of the theoretical models discussed below). The process by which such text is created is by transcribing an oral account and then

* The work reported here was partly funded by Australian Research Council grants SR0566965, DP0450342 and DP0984419. Thanks to David Nash, Alexis Palmer and three anonymous reviewers for their comments on an earlier version of this paper.

making the textual transcript available in a form that can be opened by one of two tools, Toolbox or Fieldworks, both produced by the Summer Institute of Linguistics and provided for free download[1]. Toolbox is the product of just a couple of people's labour and does a wonderful job of automating the process of interlinearising texts. It takes some learning, but there are a number of tutorials available, as well as online resources and a sufficiently large userbase that can provide support. Creating these texts from transcriptions that are time-aligned with the primary media means that the IGT should also have links to playable media, at the level of the sentence (or utterance unit). In this way IGT texts can form a corpus that informs linguistic typology; creates archival forms of textual data; and perhaps represents languages in a distributed online museum.

A broader issue relates to the way in which Humanities researchers can benefit from the use of new methods in producing research outputs. While their needs are usually greater, in that they are typically not trained in or adept at using programming languages, the general community is likely to be interested in re-using the results of their work with primary data, in the form of historical records, recorded performances, or, in the case of linguistics, oral traditions from previously unrecorded languages.

Attempts to standardise IGT date at least to Lehmann (1983), who set out principles for aligning text and morphemic translations. Various software tools have allowed linguists to create IGT, including TRANSC(ript)[2] for CPM computers and IT[3] for DOS and Macintosh in the 1980s, followed by Shoebox and its successor Toolbox, ITE (Jacobson n.d.) and now Fieldworks (mentioned above) and TypeCraft[4]. These tools parse the text and some of them can populate the gloss line by selecting from an associated wordlist or dictionary. In some of these cases the relationship between a word and its gloss has not been explicit, rather it is the result of a visual alignment on the screen or page. This is how Toolbox presents its data, and as it is the most popular software for doing this work, most collections of text in IGT format are produced by Toolbox. A major problem with texts produced in this way is that any slippage in alignment between the text and morphemic lines results in loss of the encoding of the relationship between them. For a human reader the text and meaning is still readable, but for computational treatment of texts the slippage is a critical problem. While Toolbox exports to XML, it relies on the correct hierarchy being established by the user so that the XML output captures the internal relationships (sentence, word, morpheme, gloss, free gloss). No validation is provided by Toolbox[5] and as there is no published XML schema for IGT with any currency there is nothing to validate against. None of the just mentioned above produces IGT in a community-accepted schema-based standard format.

Theoretical modeling of IGT using XML includes Bow, Hughes & Bird (2003), Hughes, Bird, and Bow (2003), Hellmuth, Myers, and Nakhimovsky (2006), Schmidt (2003), Jacobson (2006), and Jacobson, Michailovsky, and Lowe (2001), Palmer and Erk (2007) for whom, in general, the solution is to encode relationships by inclusion within an XML element, something like that shown in Fig. 2.

Palmer and Erk (2007) provide a good summary of previous work and suggest a format for IGT that includes globally unique IDs rather than XML embedding for linking annotation layers (2007:179). This makes the links explicit rather than relying on an XML hierarchy between forms and glosses. A fragment of this model is given in Fig.3.

[1] http://www.sil.org/computing/toolbox/index.htm, http://www.sil.org/computing/fieldworks/index.html

[2] http://digitalhumanities.org/humanist/Archives/Virginia/v01/8803.1324.txt (see post of Sat, 19 Mar 88 20:42:37)

[3] http://sil.org/computing/catalog/show_software.asp?id=19

[4] For a detailed listing of annotation tools see http://annotation.exmaralda.org/index.php/Linguistic_ Annotation

[5] This is true of version 1.5.5.

```
<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>
```

Fig. 2., fragment of IGT encoded in XML

```
<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>
```

Fig. 3. Example of unique IDs in the Palmer and Erk (2007) model for IGT.

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>
           …
```

Fig. 4: Comparison of Toolbox and EOPAS XML.

The Text Encoding Initiative (TEI[6]) has no standard format for IGT, but presumably it could be developed within that framework. An implementation reported on by Canfield (2007) for the online delivery of Navajo texts uses TEI forms for text as the storage format. It is not clear if a standard schema is required or even desired in this approach.

A technical committee of the ISO (Ide 2006, Ide, Laurent, and de la Clergerie 2003) is planning a language annotation framework, for the design and implementation of linguistic resource formats and processes in order to facilitate the exchange of information between language processing modules. All of these models provide the basis for a tool that will encode IGT in a principled, robust form. But the tool does not yet exist.

3 Existing online uses of IGT

In a survey of online IGT, Lewis (2006) found a range of types, most of which are simply lines of text on a page with no underlying structure, often within pdf files. These have been collected and stored in the Online Database of Interlinear Text (ODIN[7]), a project that queries websites, looking for possible examples of IGT in legacy material by using clues like the 'grams' or grammatical glosses typically used in IGT, for example ERG and

[6] http://www.tei-c.org/

[7] http://www.csufresno.edu/odin/

ABS. Recognising that legacy text is never declared as being IGT, and so is unlikely to be easily located except with the kind of regular expression search employed by the ODIN project (Lewis 2006:[iii]), how much more efficient would this process be if we had a declared IGT type that could be located via normal search tools, eliminating the need for guesswork and manual handling of results?

Examples of the use of XML for IGT include Ballantyne's collection of Yapese texts[8] which are presented online. The LACITO[9] lab in Paris has an archive of texts encoded by Jacobson (Jacobson, Michailovsky and Lowe 2001, Jacobson 2006) who has built a system for presenting IGT using XML and associated media files. Extending the LACITO model I led a team to build EOPAS (Schroeter and Thieberger 2006) which is a system for uploading XML IGT from Toolbox, with media references, and a media file (in Ogg format) validating the XML against a schema[10], and presenting the corpus in an online data set, based on an XML database (eXist). The relationship of Toolbox fields used by EOPAS is shown in Fig.4. Functions provided for in the online representation include toggling a view of interlinear text and creating a keyword-in-context (concordance) view by clicking on any item in the morphemic line, as well as hearing or viewing media associated with chunks of text, called by time-offset from a streaming server.

4 What do we want?

If there were a known datatype 'IGT' that could be located via a central service, then research could target a body of texts knowing that, for example, a search over the contents of the *morphemic* line would only return morphemes of the language, and a search on the *gloss* line would only return a gloss. This is currently not possible.

At the moment, internet-based textual comparisons are necessarily made by hand from texts retrieved by hand (or via the semi-automated 'best-guess' approach that populates ODIN, above), for example, in exploring typological generalisations. Thus, a schema against which texts could be validated is a good first step towards automating the retrieval of objects of comparison. However, the goal of comparable texts is only part of a bigger picture. To encourage the development of corpora in IGT there need to be easy migration paths that allow various outputs. EOPAS takes an existing IGT and presents it online, perhaps together with media. This is an attractive goal in itself, but, once the text is structured in this way, it can be rendered in various formats for online or paper (rtf, pdf) publication and is in an ideal archival format. This, together with the access to their data that it provides to a linguist, should be reason enough for it to be taken up by field linguists.

The potential to create multiple outputs from a single, well-structured document is attractive to linguists, but only if the steps required are not too onerous – in fact, only if there are no obvious steps at all.

The bigger picture is the possibility of an online grammar (cf Thieberger 2009), in which specific datatypes have declared schemas and namespaces allowing them to be represented according to their specific characteristics. The 'IGT' type could be declared so that example sentences or complete texts are correctly rendered for onscreen presentation, including links to media, and providing a target for harvesting of IGT (as discussed earlier).

This service would allow a number of previously unavailable collections to be made public. With a schema and a declared namespace, the datatype 'IGT' could be harvested by a service that could then serve a distributed network of language museums, for example. As each corpus has a language code (ISO-639-3) it can be identified and geocoded, and thus located via a geographic search mechanism. These texts, if given persistent identifiers, could then provide a citable form of primary data. Subsequent research can then build on playable, searchable datasets.

Given the number of sources cited earlier in this paper, one could reasonably have expected that a useable system for the publication of IGT would have been developed by now, but this is not the case. There is a lack of connection between those for whom the theory of IGT provides a basis for theoretical papers, and those for whom an online system would be a useful tool. Thieberger (2007) suggests that linguists (and other humanities scholars) need simple tools now but that IT specialists don't find the creation of these tools to be challenging (or lucrative). However, without them linguists won't produce locatable and reusable corpora, which could be the basis for NLP efforts on a diverse range of languages. It is only in the outputs of tools that we find linguistic data in standard formats, that is, linguists are prepared to use these tools because there is an immediate benefit, and the fact that the data is structured in conformant XML is, in the main, irrelevant to them. Tools like Transcriber[11] and Elan[12] produce transcripts of media in XML files as the data that underlies their browser view, safely hidden from linguists who just want to use the tool and don't want to deal with XML themselves. Toolbox is capable of producing conformant XML, but it requires the user to adhere to predictable field naming and a hierarchy that can be rendered in the XML output. For the purposes of EOPAS we created a Toolbox template[13] that contained the requisite field

[8] http://www2.hawaii.edu/~ballanty/corpusintro.html
[9] Laboratoire de langues et civilisations à tradition orale (CNRS) http://lacito.vjf.cnrs.fr/archivage/index_fr.htm
[10] http://paradisec.org.au/eopas.xsd
[11] http://trans.sourceforge.net/en/presentation.php
[12] http://www.lat-mpi.eu/tools/elan
[13] http://wiki.arts.unimelb.edu.au/ethnoer/Main_Page#Toolbox_template_for_use_with_the_EOPAS_system

names and hierarchy for output to EOPAS-compliant XML.

5 Why can't we get it?

As we have seen, there are a number of proposals for data structures in a number of papers presented at conferences and published over the past decade, but there is not yet a working tool for the creation of IGT that builds in lexical lookup, media links, and a schema and namespace for validation of its outputs. It seems that a representational model of IGT does not satisfy computational linguists who want the underlying relationships in IGT to be the basis for the model, rather than the surface representation. As Bird (2009) notes, the field of natural language processing has little to offer language documentation efforts, as the example of the lack of services for IGT discussed here illustrates. There is a benefit for computational linguists in helping out here– only if field linguists get the right tools to allow them to make their data publicly available in digital format will computational linguists be able to use these data.

Further, there is a gulf between our needs as linguists and as humanities scholars, and the services available or planned at the Australian Federal government level. Over the past few years we have seen an increasing recognition of the need to support research using digital technologies, with large funding sources being channeled to projects with various acronyms (in Australia, ARROW[14] and FRODO ($12m), DART ($3.2m), AeRIC, NeAT, AAF, AREN ($88m), ANDS ($21m) some of which are part of NCRIS[15] ($542m)) which have resulted in some changes for the humanities scholar in terms of access to published research data.

On the other hand, there has been little or no support for creation of data within the research process, which would require humans to assess existing workflows and to suggest ways in which they could be improved, and to write necessary software for developing repositories or conversion algorithms to make existing data reusable. At the same time we are seeing universities withdrawing funding from IT support. Instead of more IT specialists who understand research needs we are getting fewer IT staff overall whose focus is on maintaining wires, disks and desktop computers. Humanities scholars cannot usually articulate their needs in terms comprehensible to programmers. We are typically not able to construct entity relationship diagrams in advance of the project implementation, but rather we prefer to develop in an iterative way, much to the frustration of the programmer. As Pitti notes, 'Although technologists who elect to participate in digital humanities projects may themselves have some background in the humanities, it will more often be the case that technologists have little training in humanities disciplines. … Carefully negotiated and apparently shared understandings will frequently be illusory, requiring further discussion and renegotiation.' (Pitti 2004:486)

The US report 'Our Cultural Commonwealth' discusses the nature of Humanities and Social Science (HASS) data as forming part of the public good, and details the distinctive needs and contributions that HASS researchers have for cyberinfrastructure. 'Extensive and reusable digital collections are at the core of the humanities and social science cyberinfrastructure. Scholars must be engaged in the development of these collections. […] The extensive digitization of cultural heritage materials is one of the most exciting developments in the humanities and social sciences in the past century.' (American Council of Learned Societies 2006:38).

6 Conclusion

In this paper I have used the example of a relatively simple computational problem, the presentation of several lines of text with internal relationships (IGT) and associated media, to illustrate the gulf between linguists (as humanities scholars) and Information Technology specialists. The problem of converting text from the output format of one tool to the input format of another is insurmountable for many linguists, but is too simple for a programmer to bother with. The example of IGT illustrates a problem that has theoretical interest for computational linguists, but has resulted in a gap between the working tool and the possibilities offered for online access and presentation of IGT. To get good IGT from the everyday practice of linguists involved in language documentation we need *tools* that can produce IGT in standard formats, and *converters* (like the one discussed by Margetts 2009 and available online[16]) to take data from transcription formats (like Transcriber and Elan) into formats that can be annotated to produce IGT. Most of all, we need good *advice* and *programming support* to enable us to make the best choices in the way that we create the data that we produce as part of our normal research.

While it has long been a part of the everyday practice of those in the physical sciences to use computing tools at a high level, it is only in the recent past that we have seen humanities computing emerge as a recognised area of interest. We need to establish a base of research materials and practices using new technologies, working collaboratively on complex data sets, such as, for example, large textual corpora (in the order of terabytes of textual material), digital versions of archival records, or audiovisual data over streaming servers, or real-time analysis of spoken voice interactions.

In order for these aims to be achieved, we need to establish work practices and appropriate data sets now. Data sets are being produced routinely in the course of our research, but usually there is no focus on conforming

[14] http://arrow.edu.au/

[15] http://ncris.innovation.gov.au '$542 million over 2005-2011 to provide researchers with major research facilities'

[16] Margetts, incidentally, is not a programmer by training but has produced a simple conversion tool here: http://linguisticsoftwareconverters.zong.mine.nu/

to standards of data structure, nor to the large problems of managing this data and storing it safely for later reuse. Much of this data is stored in analog form and so is becoming largely unusable due to the obsolescence of the machinery on which it was recorded, or the deterioration of the media itself.

Humanities scholars need guidance in how to create well-formed research outputs that will be reusable and potentially interoperate with similar types of data produced by other researchers. IGT is a good example of this kind of data.

References

American Council of Learned Societies. (2006). Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. New York: American Council of Learned Societies. http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf

Bird, Steven. (2009). Natural Language Processing and Linguistic Fieldwork. *Computational Linguistics* September 2009, Vol. 35, No. 3: 469–474. http://www.mitpressjournals.org/doi/pdf/10.1162/coli.35.3.469

Bow, C., Hughes B. & Bird, S. (2003). Towards a General Model of Interlinear Text. Proceedings of the EMELD Language Digitisation Project Conference 2003: Workshop on Digitizing and Annotating Texts and Field Recordings. Retrieved September 23, 2009, from http://emeld.org/workshop/2003/bowbadenbird-paper.html

Kip Canfield. (2007). The Navajo Language Literature Project: A Case Study in Client-side Design Patterns Using Asynchronous Requests. *Literary and Linguistic Computing*, Vol. 22, No. 4, 395-403.

Hellmuth, C., Myers, T. & Nakhimovsky, A. (2006). The Linguist's Toolbox and XML Technologies. Paper presented at the EMELD meeting, Retrieved September 23, 2009 from http://emeld.org/workshop/2006/papers/hellmuth.html

Hughes, B., Bird. S., & Bow, C. (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). Retrieved September 23, 2009, from http://repository.unimelb.edu.au/10187/1311

Ide, Nancy. (2006). Linguistic Annotation Framework ISO/TC 37/SC4 N311. http://www.tc37sc4.org/new_doc/ISO_TC_37_SC4_N311_Linguistic%20Annotation%20Framework.pdf

Ide, Nancy, M. Romary, Laurent, Eric de la Clergerie. (2003). Outline Of The International Standard Linguistic Annotation Framework http://www.aclweb.org/anthology-new/W/W03/W03-1901.pdf

Jacobson, M. (2006). The LACITO Archiving Project. *Ethnographic Eresearch Annotation Conference, University of Melbourne*, February 15-17, 2006

Jacobson, M. (n.d.). Interlinear text editor. Retrieved September 23, 2009, from http://michel.jacobson.free.fr/ITE/index_en.html

Jacobson, M., Michailovsky, B., & Lowe, J.B. (2001). Linguistic documents synchronizing sound and text. *Speech Communication 33* (1-2), 79-96.

Lehmann, Christian. (1983). Directions for interlinear morphemic translations. *Folia Linguistica* 16, 1982:193-224.

Lewis, William D. (2006) ODIN: A Model for Adapting and Enriching Legacy Infrastructure. In *Proceedings of the e-Humanities Workshop*: http://faculty.washington.edu/wlewis2/

Max Planck Institute for Psycholinguistics. (2006). *Tools: ELAN*. Retrieved September 23, 2009, from http://www.mpi.nl/tools/elan.html

Maeda, Kazuaki and Steven Bird. (2000). A Formal Framework for Interlinear Text. *Proceedings of the Workshop on Web-based Documentation and Description, Philadelphia*, USA; December 12-15, 2000.

Margetts, Andrew. 2009. Using Toolbox with Media Files. *Language Documentation & Conservation* 3(1): 51-86. http://hdl.handle.net/10125/4426

Open Language Archives Community (OLAC). (2006). Retrieved September 23, 2009, from http://www.language-archives.org

Palmer, Alexis & Katrin Erk. (2007). IGT-XML: an XML format for interlinearized glossed texts. Proceedings of the Linguistic Annotation Workshop (LAW-07), ACL07. Prague. http://whitepapers.zdnet.com/abstract.aspx?docid=889125

Pitti, Daniel V. (2004). Designing Sustainable Projects and Publications. In Schreibman, Susan, Ray Siemens, and John Unsworth. (eds.) 2004. *A companion to digital humanities*. Malden, MA ; Oxford : Blackwell.

Schmidt, T. (2003). Visualising Linguistic Annotation as Interlinear Text. *Arbeiten zur Mehrsprachigkeit. Working papers in multilingualism*. Series B. Hamburg: Univ. Hamburg. Viewed on September 23, 2009 http://www1.uni-hamburg.de/exmaralda/Daten/4D-Litertur/Paper_LREC.pdf

Schroeter, Ronald & Nicholas Thieberger. (2006). EOPAS, the EthnoER online representation of interlinear text. Barwick, Linda and Nicholas Thieberger. (eds.) 2006. *Sustainable Data from Digital Fieldwork* Sydney: Sydney University Press. 99-124. http://repository.unimelb.edu.au/10187/2137

Thieberger, Nick. (2007). Does Language Technology Offer Anything to Small Languages? *Proceedings of the Australasian Language Technology Workshop 2007,* http://www.alta.asn.au/events/altw2007/cdrom/pdf/ALTA2007_02.pdf

Thieberger, Nicholas. (2009). Steps toward a grammar embedded in data. Epps, Patricia and Alexandre Arkhipov. (eds*.) New Challenges in Typology: Transcending the Borders and Refining the Distinctions*. Berlin ; New York, NY : Mouton de Gruyter Mouton. 389-408. http://repository.unimelb.edu.au/10187/4864

Alexis Palmer: First, I entirely agree that there is a need for a better IGT tool, and as well that it's frustrating that this development has not yet occurred. As a community, I think we are close enough to an agreed-upon model for the data structure. And the type of models that have been discussed also I think provide a decent level of flexibility to meet varying presentational needs (EOPAS in particular). Now it's really a matter of implementation. As you point out, most field linguists are not prepared/equipped to implement complex software systems, but in fact neither are many computational linguists.

My primary comment is that it isn't entirely clear to me what the aim of the paper is... is it arguing for creation of a tool? Of a schema and articulated namespace? Of putting resources toward the problem in general? I'm also a little confused by references to various computer-y groups/fields... while I know you're not just conflating computational linguists, programmers, and IT professionals, it's not totally clear to me what distinctions are being made. I think it's important to make distinctions, though, because each group has a very different role to play in the humanities research ecosystem. (I suspect too some of my confusion comes simply from not knowing the audience/venue for the paper...)

Another slightly fuzzy point is whether the focus is meant to be on conversion and metadata-provisioning for legacy data or on production of new data, or both.

A third point: no question that presentation and underlying representation are different concerns that result in different priorities for working with IGT, but *both* rely on preserving relationships between lines of text. For us, developing a model for IGT was a necessary milestone in order to do further research involving IGT, not just an abstract theoretical interest. And the practical goal of that research is to facilitate more rapid production of IGT, reducing human effort. But this research is separate from the need for a better interlinearization tool. Our work could figure as an extension to such a tool, but it doesn't handle the basic functions handled by e.g. Toolbox or Fieldworks.

section 3 -- I like this term 'slippage in alignment' ... it's exactly the source of many of the problems we encountered in our own research. It might be worth pointing out that human readers are able to tolerate much more 'slippage' and still be able to understand the relationships between lines of text. The machine reader has extremely low tolerance to slippage.

figure 4 -- a little bit hard to read... alignment slippage... could just be because I printed from Google Docs.

section 5 -- 'there need to be easy migration paths that allow various outputs' YES!

section 6 -- I completely agree that we need tools and converters for IGT, but I would argue that such resources are broadly applicable and valuable... computational linguistics stands as much to benefit as language documentation... for both fields, such tools and converters (as well as the higher-quality and more readily-reusable data they'd help us to produce) have the potential to increase our output, improve our research, and help us to make our research more valuable to the world at large. So for me it doesn't really work to set comp.ling and doc.ling up in opposition to each other... it's not the case that we have all of these skills just at hand but don't care to share them. What's needed is a better way for researchers across the two fields to collaborate. My two cents.

Paper: 11
Title: Culture clash – Humanities research and computing: a case study of Interlinear Glossed Text (IGT)

---------- review 1 ----------

PAPER: 11
TITLE: Culture clash – Humanities research and computing: a case study of Interlinear Glossed Text (IGT)

OVERALL RATING: 1 (weak accept)
REVIEWER'S CONFIDENCE: 2 (medium)
THEMATIC FIT: 2 (Perhaps appropriate/relevant for this workshop)
CLARITY: 4 (The paper is fairly clear, but a few small points could be clarified)
TECHNICAL CORRECTNESS: 4 (Overall technically quite sound, but a few minor claims could do with more backing up or some theoretical work need further support)
ORIGINALITY: 2 (Fairly original)
REFERENCES / COMPARISON: 3 (Very good comparison to existing approaches)
QUALITY OF ENGLISH: 3 (The English is fine)

---------- REVIEW ----------

The paper is a manifesto of the need to get IT specialists working with linguists in order to build tools to facilitate linguists' work and also ensure the creation of archivable and usable data corpora. The paper takes the example of Interlinear Glossed Text to illustrate what needs to be done, and why it might not be of interest to researchers in Natural Language Processing. The author also points out that there are typically no resources for IT specialists to work with linguists, and, furthermore, it is difficult for the 2 communities to understand each other (linguists and IT), so that explaining requirements/etc is difficult.

The author presents the issues clearly and directly.

The paper is related to issues that have been raised at least in HCSNet (if not at ALTA) regarding the creation of a National Corpus. Of course, though, it is not totally related to ALTA itself. Yet, it might be good to have field linguists and computational linguists (the traditional attendees of the ALTA workshop) to talk to each other.

While reading the paper, I wondered if some of these issues should not be presented at ADCS (Australasian Document Computing Symposium) instead of ALTA - or may be INEX, where researchers are concerned with information retrieval on XML data. But if, as the author says, the solutions needed for the linguists are too simple for people in IT to be concerned with, may be that would not help either.
I also wondered if researchers in Digital Libraries might not be interested in the issues.

---------- review 2 ----------

PAPER: 11
TITLE: Culture clash – Humanities research and computing: a case study of Interlinear Glossed Text (IGT)

OVERALL RATING: 2 (accept)
REVIEWER'S CONFIDENCE: 2 (medium)
THEMATIC FIT: 3 (Definitely appropriate/relevant for this workshop)
CLARITY: 4 (The paper is fairly clear, but a few small points could be clarified)
TECHNICAL CORRECTNESS: 3 (Technically OK at the base, but for publications some fixes must be made)
ORIGINALITY: 2 (Fairly original)
REFERENCES / COMPARISON: 2 (Ok, but somewhat incomplete comparison to existing approaches)
QUALITY OF ENGLISH: 3 (The English is fine)

---------- REVIEW ----------

This paper presents the interesting problem field linguists face when they want to describe languages: the lack of standard to describe and represent linguistic data, and, more importantly, the lack of appropriate tools.
There also seems to be a need to bridge a gap between linguists and ICT researchers. I would have thought myself that computational linguists would have been the right kind of person for that… Also, I'm not sure that the tools we are talking about are that trivial to develop.

In terms of contribution to the ALTA workshop, as opposed to "traditional research papers" where the authors present their work, here, the author presents an account of some of the issues field linguists are facing, taking the IGT as an example. So, the paper is more like a position paper for me that could generate interesting discussions.
Maybe, it would be worth including in the conclusion some ways forward, some suggestions as what would be the first steps to more collaboration between linguists and IT specialists, and the development of useful tools.

---------- review 3 ----------

PAPER: 11
TITLE: Culture clash – Humanities research and computing: a case study of Interlinear Glossed Text (IGT)

OVERALL RATING: -2 (reject)
REVIEWER'S CONFIDENCE: 2 (medium)
THEMATIC FIT: 2 (Perhaps appropriate/relevant for this workshop)
CLARITY: 2 (There are quite a few problems with clarity, but something can be salvaged from it)
TECHNICAL CORRECTNESS: 1 (No, this is really flawed)
ORIGINALITY: 2 (Fairly original)
REFERENCES / COMPARISON: 2 (Ok, but somewhat incomplete comparison to existing approaches)
QUALITY OF ENGLISH: 2 (The English is OK, but there are some minor editorial issues and typos)

---------- REVIEW ----------

This paper calls for closer collaboration between computational linguists and field linguists with the aim to produce tools that support the production, storage and retrieval of language samples in a standardised format. The paper mentions a number of existing formats and tools but laments that none of them are technically sound enough to provide consistent data storage and retrieval capabilities. The author suggests that the main reasons for the lack of a technically sound, easy to use tool are the difficulty of communication between linguists (and other Humanities researcher) and computer scientists, the lack of interest of computer scientists in producing such a tool, the lack of interest of linguists in the underlying representation of their data, and the lack of public funding for computational support in linguistics.
I sympathise with the author's plea for help in constructing such tools and I can see one very good argument why the CL community should be interested in helping to create them: only if field linguists get the right tools to allow them to make their data publicly available in digital format will computational linguists be able to use these data. Unfortunately, this argument is not made clearly in the paper but only appears "between the lines" in the conclusion.
Instead, the problem (lack of a good tool) and the above mentioned reasons for this (lack of communication, interest and funding) are reiterated throughout the paper.
Beyond the possibility of using field linguists' data there seems to me little reason why the CL community should be interested in helping build such tools. Any reasons that might help attract CL researchers to this problem should be elaborated in the paper, if attracting the interest of the CL community is its aim.
My main concern with the paper is that it goes little beyond stating the problem. In order to make a serious contribution to solving the problem it would need to present a clearly structured and elaborate list of the requirements that a tool for recording and retrieving field linguists' data would need to meet. Section 5 makes a start by mentioning a few things that might be desirable, but it reads a bit like an ad hoc collection of thoughts. Being a field linguist himself who clearly has computational expertise, the author seems in the perfect position to compile and present a comprehensive list comprising both technical requirements regarding the representation of IGT and interface requirements regarding the usability of a potential new tool.
Another possibility for a contribution to a solution of the problem would be a comparison and evaluation of the existing XML formats listed in section 3 which could point someone interested in implementing a tool towards the representation format they should adopt.
In sections 2 and 3 the paper mentions a number of existing tools for the creation of IGT. As the main point of the paper is the lack of adequate tools, some more effort should be spent on explaining in some detail why the existing ones don't hit the mark to a reader who has never used them.

A few more technical points that might help improve the paper:
The paper would be a lot easier to read for an outsider, if it provided more more background information about the work practices of field linguists. Also, providing a clear statement of the problem and the contribution the paper makes to solving it as well as indicating the structure of the paper in the introduction would add clarity.
Sections 2 3 and 4 seem to be subsections of one larger section about the use of IGT and existing tools to date.

Clearer subheadings and separation of the content into three sections (the elements of IGT, tools that produce IGT, XML modelling of IGT) would greatly enhance readability.
Similarly, some content of section 5 seems to belong into section 6 (top of second column on page 4) and the last three paragraphs of section 6 into the introduction or conclusion
In places, the paper reads a bit like an informal opinion piece due to the use of informal language, shorthand (/ instead of 'or'), inclusion of biographical background of the author and others (Margetts) and switching between the first and third person.
Long sentences presenting too many propositions and long paragraphs ranging across different topics as well as the lack of connecting sentences between paragraphs make it hard to follow the thread of the argument.
I notice that the paper was submitted as a short paper. However, it exceeds the page limit by a whole page. If it had been submitted as a full paper, there would have been ample space to address the issues pointed out here.
(I tried not to give a score for TECHNICAL CORRECTNESS because the paper does not present a technical solution to a problem, but the review system did not allow this.)