Mining and Migrating Interlinear Glossed Text

William Lewis
California State University, Fresno

1.0 Introduction

The World Wide Web is rapidly becoming the primary source for disseminating data on the world's languages. Language researchers, linguists and language communities are regularly posting a variety of language data to the Web, including lexicons, teaching materials, language recordings and transcriptions, language descriptions, and grammars. Also posted in large quantities are scholarly papers on language, posted regularly by innovative online-only publications (e.g., Snippets), by traditional linguistic publications offering online editions, and by linguists themselves. Of significant potential utility to linguists is the language and linguistic data contained in these documents, specifically data presented in the form of Interlinear Glossed Text (IGT). Generally, IGT consists of a line of language data, often broken down by morpheme, a line of grammatical and gloss information aligned with the text in the first line, and a line representing the translation. An example is shown in (1). Variations to this basic form abound, but its most frequent instantiation is this basic three-line format.[1]

(1) yama +ajurnu jarti-ki+karn pikapikka-ka ngu-nngara.
leave +3pl.NS other-DAT+now children-DAT lie-OPT+FUT
'Some should be left for the other children.' [Simpson 1998]

Despite its ubiquitous presence, IGT on the Web can be difficult to find. Search engines are good at looking for text containing certain keywords using exact string matches, but are not quite as good at locating text of a certain structural form. Further, there is no way without significant manual labor to compare data drawn from multiple instances of IGT posted to the Web.

[1] For more information about varieties of IGT and what potentially can be described in IGT, see Bird and Bow 2003 and Zaeffer 2003, this workshop. Also see Peterson 2002:13.

Since IGT is designed for human consumption, it is also very difficult to process automatically. Alignment between the lines of data is implied, not explicitly represented. Further, deciphering the grammatical and gloss information contained in the second line (of the standard 3-line format) is subject to local interpretation, since the shape of the glosses can vary by author and by theoretical platform. This is further compounded by document format, since most IGT on the Web is embedded in PDF documents[2], a notoriously difficult file format to process and extract data from. IGT in PDFs is also rendered in a variety of fonts, most of which are difficult to extract and may in turn present problems for portability. Migrating IGT to best practice—where the data is stored in a consistent and portable format, and where relationships between data elements and their definitions are explicitly defined—can make IGT more useful and more accessible.

IGT is used to describe and enrich language data. As such it gives insight into the grammatical features of language that linguists feel are important. In other words, it provides information about what linguists talk about and how they talk about it. Such data can provide fuel for the development of a linguistic ontology. One such effort, the development of the Generalized Ontology of Linguistic Description (GOLD) (Farrar and Langendoen (to appear), Langendoen and Farrar 2003, this workshop) is directly benefiting from the distributional analysis of IGT.

IGT on the Web represents a very large corpus of enriched language data. It is broad in coverage, including data from a number of different languages and theoretical platforms. Coupled with an ontology, such a corpus could be useful for the development of tools for linguistic analysis, such as search engines, language comparison tools, and lexicon builders, among others. Successful implementation of such tools, coupled with the migration of IGT to best practice, provides a proof of concept which may supply the general impetus to migrate larger sets of data to best practice and to the Web.

[2] Initial searches for IGT on the web showed that approximately 95% were found in PDF documents, with the remainder found in HTML and DOC files.

This paper is divided into four sections: Section 1 is the introduction (this section). Section 2 discusses best practice and its applicability to IGT. The best practice recommendations presented in the section are minimal, and are only intended as evaluation metrics for exploring the utility of a best practice model for IGT. Section 3 discusses the processes involved with locating IGT on the Web, extracting data from it, and migrating it to best practice. This section also discusses some of the tools that have already been designed and tested for processing IGT. The issues involved with the resolution of the descriptive vocabulary used in IGT, although important to the process of converting IGT to best practice, have not been discussed in detail here. Much of this work is preliminary and still being evaluated. Section 4 is the conclusion.

2.0 Best Practice and IGT

This section will discuss the minimal encoding requirements for IGT as best practice and the contributions of the GOLD ontology to converted IGT. It does not elaborate to any detail on best practice recommendations, leaving that task to others. The best practice recommendations given in this section constitute a subset of those discussed in Bird and Simons (to appear).[3]

2.1 Descriptive Vocabulary and IGT

Whatever its eventual form, data formatted in best practice should minimally satisfy the following best practice recommendations:

(2) Best Practice Recommendations

Data in best practice should be

a) portable,
b) reusable,
c) encoded using a standard encoding format and formatting vocabulary,
d) renderable into human digestible form, and,
e) marked up using a standard descriptive vocabulary.

[3] Also see Simons 2003a and the EMELD proposal for more information.

XML (Extensible Markup Language), the standard for disseminating semantically enriched data on the Web, meets all of these criteria, with the exception of (e). (e) is problematic because XML says nothing about the semantics of a markup vocabulary—the "tagset" or "descriptive vocabulary"—leaving the semantics to the individual researcher. If one researcher chooses to mark a particular morpheme type as PERF, and another chooses PRF for the same, there is no way that an automated system will be able to tell that these two elements refer to the same concept. Opposing this is the problem of conflicting conceptual spaces for tags. For instance, the notion PAST has a wide variety of interpretations: it can refer to absolute past, to relative past, to immediate past, to remote past, etc. The tag PAST in and of itself tells us nothing about the type of past tense referred to: the conceptual space referred to by PAST for one researcher can intersect with, subsume, be subsumed by, or differ completely from that referred to by another.

Two obvious solutions exist: (i) define a standard vocabulary and recommend its use as part of best practice, or (ii) define a method for establishing the relationships between differing markup vocabularies. For (i), the problem is defining a standard that is comprehensive and flexible enough to satisfy the needs of most linguists. Since it is possible for the descriptive vocabulary used by one researcher to conflict with that of another (as with PAST), a comprehensive standard markup vocabulary may not, in fact, be possible. The second solution (ii) is the solution initially suggested by Lewis et al (2001), with significant refinement and greater detail in Farrar (2003) and Farrar and Langendoen (to appear). Farrar and Langendoen, in their development of GOLD, seek to define the specific semantics of markup, relating the instances of use to an ontological representation of the conceptual space defining linguistic vocabulary. The author proposes that (ii), although initially more difficult to implement, is the only one that insures long-term interpretability and portability of data.

The Web coupled with an ontology provides a unique solution to resolving the semantics of IGT. Rather than imposing a standard markup vocabulary, an ontology allows researchers to use their own vocabulary, as long as their vocabulary is consistent with the conceptual space of the ontology. Likewise, when analyzing marked-up text, the ontology, being available online and in real-time, can provide a translation of any “foreign” terms found to a locally interpretable and familiar vocabulary. The ontology on the Web as such acts as online markup interlingua. Given this, the best practice recommendations should be revised as follows (with the revisions in bold):

(3) Revised Best Practice Recommendations

Data in best practice should be

a) portable,
b) reusable,
c) encoded using a standard encoding format and formatting vocabulary,[4]
d) renderable into human digestible form,
e) marked up using a consistent, *documented* descriptive vocabulary, and,
f) readily accessible from any system.

Essentially, (f) recommends that data should be posted to the Web. This allows access to the data from any system, which in turn makes other resources, such as ontologies, automatically and immediately available. Naturally, granting access to a resource should be at the discretion of the author or provider, but once access is granted, the resource should be available from any system by anyone who has access privileges.

On the other hand, (e) (revised) recommends that the semantics of the descriptive vocabulary for any resource should be resolvable, which can be either within the resource itself (for instance, through the use of a glossary) or by reference. *By reference* can either mean that the semantics of tags are resolved by reference to a descriptive resource (such as an ontology), or that the document as a whole is classified with other resources that use the same descriptive vocabulary (and thus reference this class). The point is that the semantics of the descriptive vocabulary should be clear and resolvable by any user accessing the resource. The most explicit

[4] This is not a recommendation that a standard inflexible model be imposed on all language resources, rather that the formatting “language” observe a standard (such as XML). See Simons 2003b.

resolution, and thus the most useful to the development of automated tools, is one that references an ontology. An ontology enhances the versatility and interoperability of resources and its use is the suggested path for best practice.

2.2 Minimal Encoding Requirements for IGT

Essential to best practice is that data be formatted in a common encoding format, such as XML (best practice recommendation (c)). Bird and Bow (2003) (this workshop) describe a general purpose data model for representing IGT. They propose a four-level data model, with Text, Phrase, Word and Morpheme levels, and discuss an implementation using XML. A sample XML file using this model with the data from example (1) is shown in (4):[5]

[5] The XML implementation of the Bird and Bow model is currently under revision and may be somewhat different in their final paper.

(4) Example XML for Warumungu data from (1)

```
<it>
 <text>
  <level type="comments">From Simpson (1998)</level>
  <phrase>
   <level type="phrasal_translation" xml:lang="en">Some should be left for the other
children</level>
   <word start="1" end="4">
    <level type="ipa">…</level>
    <level type="orth" encoding="...">yama</level>
    <morph>
     <level type="morph">yama</level>
     <level type="gloss">leave</level>
    </morph>
   </word>
        …
   <word start="13" end="23">
    <level type="orth">jartikikarn</level>
    <morph>
     <level type="morph">jarti</level>
     <level type="gloss">other</level>
    </morph>
    <morph>
     <level type="morph">ki</level>
     <level type="gloss">DAT</level>
    </morph>
    <morph>
     <level type="morph">karn</level>
     <level type="gloss">now</level>
    </morph>
   </word>
        …
  </phrase>
 </text>
```

The implementation of the Bird and Bow model shown in (4) is quite good at representing Interlinear data, preserving the descriptive material contained in the human-readable IGT ((1)), while at the same time explicitly defining relationships between data elements, most especially relationships between morphemes and glosses in the first and second lines. Because it is in XML, it is inherently portable and reusable: using the XSL scripting language, one can render output in any one of a number of formats (satisfying recommendations (a), (b), and (d)). Being designed for the Web, XML also can easily be made available to any system (satisfying recommendation (f)).

The Bird and Bow model, however, says nothing about the resolution of the semantics of the glosses (recommendation (e)), something essential to the interoperability of resources. Enhancing this model to include semantic descriptions could greatly improve its utility. How this might be accomplished is discussed in the next section (section 2.3).

Also missing are some important distinctions commonly made in IGT. Linguists, for instance, often make a distinction between grammatical concepts, which I will call GRAMs, and glosses. We feel that a model should capture this distinction and make it explicit. Linguists also often note "types" of morphemes using various diacritics, and a model should explicitly capture this information as well. Further, language ID, authorship, and source should also be explicitly referenced in a model. These improvements are detailed below:

A. The distinction between GRAMs and glosses in IGT is typically made by using case, where upper case is used to mark GRAMs (e.g. DAT, NS), and lower case glosses.[6] In the model, another value for the *level* attribute for *morph* could accomplish the task of representing the GRAM/gloss distinction, as shown in (5). Explicitly encoding the distinction has implications for automated processing, since the grammatical notions have ontological relevance (at least in a linguistic ontology), whereas glosses do not. It is also possible for an entry to have both a GRAM and a gloss (for instance, a word can have a gloss such as *eat*, and a GRAM, such as the part of speech label *verb*)

(5) *gram* attribute for *level*

```
<morph>
  <level type="morph">ki</level>
  <level type="gram">DAT</level>
</morph>
```

B. Linguists regularly make a distinction between types of morphemes using markup diacritics that indicate the forms: "-" is often used to mark affixes, "=" or "+" are often used to mark clitics. This information should be explicitly encoded as well. It can be represented using a *type* attribute for *morpheme*:

[6] It is unclear whether this distinction generalizes to other language resource types, such as lexicons.

(6) *type* attribute for *morpheme*

```
<morph type=”clitic”>
  <level type="morph">karn</level>
  <level type="gloss">now</level>
</morph>
```

C. A language identifier attribute should be included at the *text* level. It should observe a standard classification system, such as Ethnologue language codes. An example is shown in (7):

(7) *languageID* attribute for *text*

```
<text languageID=”WRM”>
```

D. Additional attributes at the *text* level might include those for *author* and *source*. Author could refer to the original author/provider of the data, who transcribed it, etc. Source could refer to the source document, either a citation or a URL. The latter is useful for online searches.

2.3 Finding GOLD

Effective resolution of the semantics of GRAMs requires explicitly identifying what they mean. Ultimately, any GRAM in any document should resolve to some linguistic concept, wherever or however that concept is defined. Since the set of GRAMs within a document represents a kind of "linguistic dialect", it makes sense to resolve the dialect as a whole rather than resolve the individual GRAMs as they occur. In other words, what is needed is a glossary of the GRAMs used for the dialect. In the nomenclature of the Semantic Web (Berners-Lee, Hendler and Lassila 2001) such glossaries are called *terminology sets*. A terminology set is a list of terms used by a particular group. Each term is defined within the set, usually by reference to a concept within an ontology. For instance, in (8), the GRAMs CAUS, PASS, and PERF represent terms that are part of a terminology set, perhaps one for those who study Yaqui, or for Uto-Aztecanists in general, or even for the linguist Lilián Guerrero herself. In GOLD, these terms would resolve to the specific concepts of Causativizer, PassiveVoice, and PerfectiveAspect. Since the terminology set defines what each GRAM means, the body of marked up text can use the GRAMs themselves, as long as an explicit reference to the terminology set is made somewhere in the text.

(8) miik-tua-wa-k
give-CAUS-PASS-PERF
'to be made to give' Yaqui (Guerrero 2002)

3.0 Locating and Migrating IGT

This section will discuss the process involved with locating IGT on the Web, extracting relevant data from it, and migrating it to best practice. There are three basic steps to this process:

1. Search – locating potential sources of IGT on the Web.
2. Extraction – extracting relevant data from IGT and linking grammatical markup to the ontology.
3. Migration – converting the IGT that has been discovered to best practice.

Tools for 1 and 2 have been developed and are being tested and revised. Tools for 3 are still under development.

3.1 Locating and Recognizing IGT

Several approaches were taken to locate IGT resources. Core to each approach was a recognizer which was able to tell when IGT was found. It was assumed that scholarly linguistics documents would be in PDF form, so most of the effort in crawling the Web was to locate these documents. Once located, the documents were scanned for likely instances of IGT using the recognizer.

Locating potential sources of IGT on the Web proved to be quite difficult. Initially a general-purpose crawler was developed and used. The crawler was given a "seed" page, which it used as a starting point around which it built a "web" of pages linked to the seed page and subsequent pages. Even with a well-chosen seed page (such as a linguistics department web site, or a "document-rich" site at Linguistlist), the crawler quickly started crawling irrelevant spaces on the Web, wasting time searching through pages that were never likely to contain relevant documents. Constraining the search space proved difficult.

The crawler finally was abandoned in favor of a meta-crawler, which essentially used existing Web search engines to locate likely pages, performing "thin" crawls across these (using depth-constrained A* heuristics). Any likely PDF documents were downloaded for offline processing.

Offline processing consisted of PDF conversion, which used a publicly available PDF converter[7], and IGT recognition. The IGT recognizer used regular expressions (finite state grammatical rules) as "templates" to match IGT. A sample regular expression for a three line IGT instance is shown in (9). This regular expression matches any 3-line instances of IGT, such as (1). In (9), 9 refers to any number (and 9* to any string of numbers), X refers to any alphanumeric character, \t refers to any number of tabs or spaces (really this should be \t*, but was abbreviated to \t for convenience), \n refers to end of line, \' refers to any quote character, and material in parenthesis is optional.

[7] Both the Midas PDF converter and PDF2Text were tested. Midas was chosen since it had a higher conversion rate. Both are publicly available for download from http://www.tucows.com.

(9)

```
\t*(\()9*\)X*\n
\t*X*\n
\t*\‘X*\n
```

Since every instance of IGT is encoded across multiple lines, a computational device called a chart was used to store incomplete instances of IGT as they were being processed. Any rules that eventually failed to match text were removed from the chart and discarded. Those that did match, were kept as likely instances of IGT.

A test set of data was run through the recognizer. The test set consisted of 53 documents downloaded from the Web, comprising a representative sample of likely sources of IGT, including journal articles, theses, chapters from online texts, and miscellaneous papers from linguists’ home pages. Across the set of documents, there were 1499 possible instances of IGT. The recognizer recognized 897 instances accurately, giving a recall of 59.8%. There were 9 false positive errors (instances that were not IGT that were identified as IGT), giving a precision of 99%. See the table in (10).[8]

(10) Precision and Recall Stats for Recognition

A. Possible IGT	1499	
B. IGT discovered	897	
C. IGT errors (false pos)	9	
D. Recall	B/A =	59.8%
E. Precision	B/(B+C) =	99.0%

We are currently modifying the meta-crawler with an embedded PDF converter and IGT recognizer. This should improve performance with the added benefit of font preservation: existing PDF to text conversion tools tend to destroy font information.

We are also developing several specialized crawlers that will crawl online journals. Some tests have been performed. The results are still being analyzed, and are not presented here.

[8] Most of the IGT that was missed consisted of instances that lacked a gloss line or were not in the standard 3-line format. Since the regular expressions used by the recognizer were not designed to capture these data, the fact that these were missed is no surprise. Recall improves dramatically if these instances are excluded.

3.2 Extracting Grammatical Concepts from IGT

An essential part in the conversion of IGT is identifying what information on the gloss line (the second line in the standard model) is ontologically relevant. Fortunately, linguists are fairly consistent in separating glosses from GRAMs using case (as discussed earlier). We used this case distinction as a heuristic in the development of an IGT extractor, which gave the results in (11) (using the same test set of 53 documents).

(11) Number of GRAMs and glosses from trial run

Total number of GRAMs:	180
# of GRAMs correctly identified:	168
# of GRAMs incorrectly identifed:	4
# of unknowns:	8
Total number of glosses:	1287
# of glosses correctly identified:	1119
# of glosses incorrectly identified:	121
# of unknowns:	47

Making the assumption that linguists would tend to use the same terminology across the set of documents regardless of case, we then ran the extractor against the data again, except this time using the set of GRAMs identified in the first run as a lexicon to help identify additional GRAMs. The stats for this run are shown in (12) and show an improvement over (11).

(12) Number of GRAMs and glosses from second pass

Total number of GRAMs:	243
# of GRAMs correctly identified:	228
# of GRAMs incorrectly identifed:	7
# of unknowns:	8
Total number of glosses:	1224
# of glosses correctly identified:	1119
# of glosses incorrectly identified:	58
# of unknowns:	47

Currently, we are working to link GRAMs to concepts in the ontology. Much of the initial work is being done manually, but there are some distributional characteristics that have been helpful. For instance, there are some GRAMs that are used universally (such as NOM, ACC, 3SG, etc.), and thus can be identified and assigned automatically whenever they are encountered. Others are used locally, e.g., by particular researchers, or within particular language families or theoretical platforms. These are grouped into “clusters,” where identifying characteristics of the cluster can help determine what concept is referenced by a GRAM. Co-occurrence frequencies of GRAMs across documents help identify clusters, which in turn feed algorithms that use clusters to identify GRAMs. Each cluster ultimately has its own set of GRAMs, which then defines the terminology set for the cluster.

3.3 Migrating IGT

At this time, no mined IGT data has been migrated to best practice. Partly this is due to the fact that the format for best practice for IGT is still being worked out. Other issues include: identifying GRAMs that are not represented in the ontology, waiting for significant alterations to the ontology to be finished, problems with extracting and preserving font information from PDF documents (this has been particularly difficult), and problems with correctly aligning data across each line of IGT.

It is important to point out that mined IGT that has been converted to best practice cannot be updated to the source documents, but must be warehoused at our site. The intent is to keep the data as an index of IGT that exists on the Web, much like the indices that Google and Yahoo keep. Fair use dictates that all index entries will have appropriate citations.

4.0 Conclusion

IGT on the Web can be a useful source of information about linguistics, telling us about the field at the data level, which provides information helpful for the construction of a linguistic ontology. It also constitutes a vast resource of enriched language data, facilitating the development of tools for linguistic analysis that up until now has not been possible. Converting IGT to best practice makes the data more accessible, which in turn makes the development of tools possible. If such tools prove useful, then the work done here will only serve to encourage the wider adoption of best practice.

Converting IGT to best practice is not a substitute for encoding the data in best practice in the first place, however. Converting instances of IGT to best practice is not easy, and this difficulty may extend to the conversion of other legacy resources. For instance, the problems encountered with converting arcane file formats such as PDF, the fact that legacy sources may not clearly define descriptive vocabulary or clearly define the relationship between descriptive terms and data points, and the fact that a significant amount of manual intervention is required in a conversion process, may all be issues that would be relevant to the conversion of any legacy resource. As more and more linguistic data are posted to the Web, the adoption of a community agreed upon set of best practice standards becomes ever more critical.

5.0 Bibliography

Berners-Lee, T., J. Hendler and O. Lassila (2001) The Semantic Web. *Scientific American,* May 2001.

Bird, Steven and Cathy Bow. 2003. Interlinear Text Types. Paper to be presented at Workshop on Digitizing & Annotating Texts and Field Recordings, LSA Institute, Michigan State University, July 11th-13th, 2003. http://emeld.org/workshop/2003/papers03.html.

Bird, Steven and Gary Simons. (To appear). *Seven Dimensions of Portability for Language Documentation and Description*. Language 79. http://emeld.org/workshop/2003/birdandsimonspaper.pdf.

Farrar, Scott. 2003. Linguistics on the Semantic Web: Theory and Implementation. Unpublished dissertation. University of Arizona, Tucson.

Farrar, Scott and D. Terence Langendoen. (To appear). *An Ontology of Linguistic Annotation*. GLOT International. http://emeld.douglass.arizona.edu:8080/GOLD_glot.pdf.

Guerrero, Lilián. 2002. Macroroles and double object constructions in Yaqui. Unpublished ms., University at Buffalo.

Langendoen, D. Terence, and Scott Farrar. 2003. Markup and the GOLD Ontology. Paper to be presented at Workshop on Digitizing & Annotating Texts and Field Recordings, LSA Institute, Michigan State University, July 11th-13th, 2003. http://emeld.org/workshop/2003/papers03.html.

Lewis, William, Scott Farrar, and D. Terence Langendoen. 2001. Building a knowledge base of morphosyntactic terminology. Proceedings of the IRCS workshop on linguistic databases, ed. by Steven Bird, Peter Buneman, and Mark Liberman, 150–156. http://www.ldc.upenn.edu/annotation/database/.

Peterson, John. 2002. *Cross-linguistic Reference Grammar*. CIS-Bericht 02-130. Universität München, Centrum für Informations- und Sprachverarbeitung. http://www.cis.uni-muenchen.de/publications/reference_grammar.pdf.

Simons, Gary. 2003a. The Electronic Encoding of Text Resources: A Roadmap to Best Practice. Unpublished ms. http://emeld.org/workshop/2003/roadmaptext.html.

Simons, Gary. 2003b. Roadmap to Best Practice for Texts. Paper to be presented at Workshop on Digitizing & Annotating Texts and Field Recordings, LSA Institute, Michigan State University, July 11th-13th, 2003. http://emeld.org/workshop/2003/papers03.html.

Simpson, Jane. 1998. "Warumungu (Australian — Pama-Nyungan)." In the Handbook of Morphology, Spencer A & Zwicky A M, eds., 707-736. Oxford: Blackwell.

Zaefferer, Dietmar. A Unified Representation Format for Spoken and Sign Language Text. Paper to be presented at Workshop on Digitizing & Annotating Texts and Field Recordings, LSA Institute, Michigan State University, July 11th-13th, 2003. http://emeld.org/workshop/2003/papers03.html.