Parsing, Projecting & Prototypes: Repurposing Linguistic Data on the Web

William D. Lewis
Microsoft Research
Redmond, WA 98052
wilewis@microsoft.com

Fei Xia
University of Washington
Seattle, WA 98195
fxia@u.washington.edu

1 Introduction

Until very recently, most NLP tasks (*e.g.*, parsing, tagging, etc.) have been confined to a very limited number of languages, the so-called majority languages. Now, as the field moves into the era of developing tools for Resource Poor Languages (RPLs)—a vast majority of the world's 7,000 languages are resource poor—the discipline is confronted not only with the algorithmic challenges of limited data, but also the sheer difficulty of locating data in the first place. In this demo, we present a resource which taps the large body of linguistically annotated data on the Web, data which can be repurposed for NLP tasks. Because the field of linguistics has as its mandate the study of human language—in fact, the study of *all* human languages—and has wholeheartedly embraced the Web as a means for dissementating linguistic knowledge, the consequence is that a large quantity of analyzed language data can be found on the Web. In many cases, the data is richly annotated and exists for many languages for which there would otherwise be very limited annotated data. The resource, the Online Database of INterlinear text (ODIN), makes this data available and provides additional annotation and structure, making the resource useful to the Computational Linguistic audience.

In this paper, after a brief discussion of the previous work on ODIN, we report our recent work on extending ODIN by applying machine learning methods to the task of data extraction and language identification, and on using ODIN to "discover" linguistic knowledge. Then we outline a plan for the demo presentation.

2 Background and Previous work on ODIN

ODIN is a collection of Interlinear Glossed Text (IGT) harvested from scholarly documents. In this section, we describe the original ODIN system (Lewis, 2006), and the IGT enrichment algorithm (Xia and Lewis, 2007). These serve as the starting point for our current work, which will be discussed in the next section.

2.1 Interlinear Glossed Text (IGT)

In recent years, a large part of linguistic scholarly discourse has migrated to the Web, whether it be in the form of papers informally posted to scholars' websites, or electronic editions of highly respected journals. Included in many papers are snippets of language data that are included as part of this linguistic discourse. The language data is often represented as Interlinear Glossed Text (IGT), an example of which is shown in (1).

(1) Rhoddodd yr athro lyfr i'r bachgen ddoe
gave-3sg the teacher book to-the boy yesterday
"The teacher gave a book to the boy yesterday"
(Bailyn, 2001)

The canonical form of an IGT consists of three lines: a *language line* for the language in question, a *gloss line* that contains a word-by-word or morpheme-by-morpheme gloss, and a *translation line*, usually in English. The grammatical annotations such as *3sg* on the gloss line are called *grams*.

2.2 The Original ODIN System

ODIN was built in three steps. First, linguistic documents that may contain instances of IGT are harvested from the Web using metacrawls. Metacrawling involves throwing queries against an existing search engine, such as Google and Live Search.

Second, IGT instances in the retrieved documents are identified using regular expression "templates", effectively looking for text that resembles IGT. An example regex template is shown in (2), which matches any three-line instance (e.g., the IGT instance in (1)) such that the first line starts with an example number (e.g., *(1)*) and the third line starts with a quotation mark.

(2)
```
\s*\(\d+\).*\n
\s*.*\n
\s*\[‘’"].*\n
```

The third step is to determine the language of the language line in an IGT instance. Our original work in language ID relied on TextCat, an implementation of (Cavnar and Trenkle, 1994).

As of January 2008 (the time we started our current work), ODIN had 41,581 instances of IGT for 731 languages extracted from nearly 3,000 documents.[1]

2.3 Enriching IGT data

Since the language line in IGT data does not come with annotations (e.g., POS tags, phrase structures), Xia and Lewis (2007) proposed to enrich the original IGT

[1] For a thorough discussion about how ODIN was originally constructed, see (Lewis, 2006).

and then extract syntactic information (e.g., context-free rules) to bootstrap NLP tools such as POS taggers and parsers. The enrichment algorithm has three steps: (1) parse the English translation with an English parser, (2) align the language line and the English translation via the gloss line, and (3) project syntactic structure from English to the language line. The algorithm was tested on 538 IGTs from seven languages and the word alignment accuracy was 94.1% and projection accuracy (i.e., the percentage of correct links in the projected dependency structures) was 81.5%.

3 Our recent work

We extend the previous work in three areas: (1) improving IGT detection and language identification, (2) testing the usefulness of the enriched IGT by answering typological questions, and (3) enhancing ODIN's search facility by allowing structural and "construction" searches.[2]

3.1 IGT detection

The canonical form of IGT, as presented in Section 2.1, consists of three parts and each part is on a single line. However, many IGT instances, 53.6% of instances in ODIN, do not follow the canonical format for various reasons. For instance, some IGT instances are missing gloss or translation lines as they can be recovered from context (*e.g.*, other neighboring examples or the text surrounding the instance); other IGT instances have multiple translations or language lines (e.g., one part in the native script, and another in a latin transliteration).

Because of the irregular structure of IGT instances, the regular expression templates used in the original ODIN system performed poorly. We apply machine learning methods to the task. In particular, we treat the IGT detection task as a sequence labeling problem: we train a classifier to tag each line with a pre-defined tag set,[3] use the learner to tag new documents, and convert the best tag sequence into a span sequence. When trained on 41 documents (with 1573 IGT instances) and tested on 10 documents (with 447 instances), the F-score for *exact match* (i.e., two spans match iff they are identical) is 88.4%, and for *partial match* (i.e., two spans match iff they overlap) is 95.4%.[4] In comparison, the F-score of the RegEx approach on the same test set is 51.4% for exact match and 74.6% for partial match.

Table 1: The language distribution of the IGTs in ODIN

Range of IGT instances	# of languages	# of IGT instances	% of IGT instances
> 10000	3	36,691	19.39
1000-9999	37	97,158	51.34
100-999	122	40,260	21.27
10-99	326	12,822	6.78
1-9	838	2,313	1.22
total	1326	189,244	100

3.2 Language ID

The language ID task here is very different from a typical language ID task. For instance, the number of languages in ODIN is more than a thousand and could potentially reach several thousand as more data is added. Furthermore, for most languages in ODIN, our training data contains few to no instances of IGT. Because of these properties, applying existing language ID algorithms to the task does not produce satisfactory results.

As IGTs are part of a document, there are often various cues in the document (e.g., language names) that can help predict the language ID of the IGT instances. We designed a new algorithm that treats the language ID task as a pronoun resolution task, where IGT instances are "pronouns", language names are "antecedents", and finding the language name of an IGT is the same as linking a pronoun (i.e., the IGT) to its antecedent (i.e., the language name). The algorithm outperforms existing, general-purpose language identification algorithms significantly. The detail of the algorithm and experimental results is described in (Xia et al., 2009),

Running the new IGT detection on the original three thousand ODIN documents, the number of IGT instances increases from 41,581 to 189,244. We then ran the new language ID algorithm on the IGTs, and Table 1 shows the language distribution of the IGTs in ODIN according to the output of the algorithm. For instance, the third row says that 122 languages each have 100 to 999 IGT instances, and the 40,260 instances in this bin account for 21.27% of all instances in ODIN.

3.3 Answering typological questions

Linguistic typology is the study of the classification of languages, where a typology is an organization of languages by an enumerated list of logically possible types, most often identified by one or more structural features. One of the most well known and well studied typological types, or *parameters*, is that of canonical word order, made famous by Joseph Greenberg (Greenberg, 1963).

In (Lewis and Xia, 2008), we described a means for automatically discovering the answers to a number of computationally salient typological questions, such as the canonical order of constituents (e.g., sentential word order, order of constituents in noun phrases) or the exis-

[2] By constructions, we mean linguistically salient constructions, such as actives, passives, relative clauses, inverted word orders, etc., in particular those we feel would be of the most benefit to linguists and computational linguists alike.

[3] The tagset extends the standard BIO tagging scheme.

[4] The result is produced by a Maximum Entropy learner. The results by SVM and CRF learners are similar. The details were reported in (Xia and Lewis, 2008).

tence of particular constituents in a language (e.g., definite or indefinite determiners). In these experiments, we tested not only the potential of IGT to provide knowledge that could be useful to NLP, but also for IGT to overcome biases inherent to the opportunistic nature of its collection: (1) What we call the *IGT-bias*, that is, the bias produced by the fact that IGT examples are used by authors to demonstrate a particular fact about a language, causing the collection of IGT for a language to suffer from a potential lack of representativeness. (2) What we call the *English-bias*, an English-centrism in the examples brought on by the fact that most IGT examples provide a translation in English, which can potentially affect subsequent enrichment of IGT data, such as through structural projection. In one experiment, we automatically found the answer to the canonical word order question for about 100 languages, and the accuracy was 99% for all the languages with at least 40 IGT instances.[5] In another experiment, our system answered 13 typological questions for 10 languages with an accuracy of 90%. The discovered knowledge can then be used for subsequent grammar and tool development work.

The knowledge we capture in IGT instances—both the native annotations provided by the linguists themselves, as well as the answers to a variety of typological questions discovered in IGT—we use to populate *language profiles*. These profiles are a recent addition to the ODIN site, and are available for those languages where sufficient data exists. Following is an example profile:

```
<Profile>
  <language code="WBP">Warlpiri</language>
  <ontologyNamespace prefix="gold">
     http://linguistic-ontology.org/gold.owl#
  </ontologyNamespace>
  <feature="word_order"><value>SVO</value></feature>
  <feature="det_order"><value>DT-JJ-NN</value></feature>
  <feature="case">
     <value>gold:DativeCase</value>
     <value>gold:ErgativeCase</value>
     <value>gold:NominativeCase</value>
                    . . .
</Profile>
```

3.4 Enhancing ODIN's Value to Computational Linguistics: Search and Language Profiles

ODIN provides a variety of ways to search across its data, in particular, search by language name or code, language family, and even by annotations and their related concepts. Once data is discovered that fits the particular pattern that a user is interested in, he/she can either display the data (where sufficient citation information exists and where the data is relatively clean) or locate documents in which the data exists. Additional search facilities allow users to search across potentially linguistically salient structures and return results in the form of language profiles. Although language profiles are by no means complete—they are subject to the availability of data to fill in the answers within the profiles—they provide a summary of automatically available knowledge about that language as found in IGT (or enriched IGT).

4 The Demo Presentation

Our focus in this demonstration will be on the query features of ODIN. In addition, however, we will also give some background on how ODIN was built, show how we see the data in ODIN being used by both the linguistic and NLP communities, and present the kind of information available in language profiles. The following is our plan for the demo:

- Very brief discussion on the methods used to build ODIN (as discussed in Section 2.2, 3.1, and 3.2)
- An overview of the IGT enrichment algorithm (as discussed in Section 2.3).
- A presentation of ODIN's search facility and the results that can be returned, in particular language profiles (as discussed in Section 3.3-3.4). ODIN's current website is **http://uakari.ling.washington.edu/odin**. Users can also search ODIN using the OLAC[6] search interfaces at the LDC[7] and LinguistList.[8] Some search examples are given below.

4.1 Example 1: Search by Language Name

The opening screen for ODIN allows the user to search the ODIN database by clicking a specific language name in the left-hand frame, or by typing all or part of a name (finding closest matches). Once a language is selected, our search tool will list all the documents that have data for the language in question. The user can then click on any of those documents, and search tool will return the IGT instances found in those documents. Following linguistic custom and fair use restrictions, only instances of data that have citations are displayed. An example is shown in Figure 1. Search by language and name is by far the most popular search in ODIN, given the hundreds of queries executed per day.

4.2 Example 2: Search by Linguistic Constructions

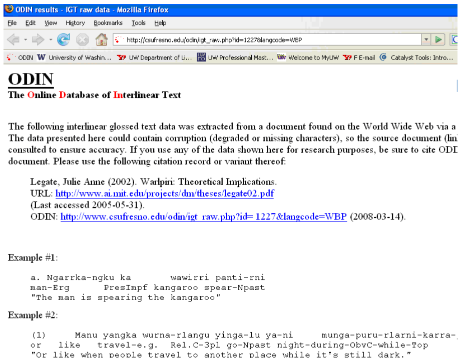
This type of query looks either at enriched data in the English translation, or at the projected structures in the target language data. Figure 2 shows the list of linguistic constructions that are currently covered.

[5] Some IGT instances are not sentences and therefore are not useful for answering this question. Further, those instances marked as ungrammatical (usually with an asterisk "*") are ignored for this and all the typological questions.

[6] Open Language Archives Community

[7] http://www.language-archives.org/tools/search/

[8] LinguistList has graciously offered to host ODIN, and it is being migrated to http://odin.linguistlist.org. Completion of this migration is expected sometime in April 2009.


ODIN results - IGT raw data - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://csufresno.edu/odin/igt_raw.php?id=1227&langcode=WBP
ODIN University of Washin… UW Department of Li… UW Professional Mast… Welcome to MyUW F E-mail Catalyst Tools: Intro…

ODIN
The Online Database of Interlinear Text

The following interlinear glossed text data was extracted from a document found on the World Wide Web via a
The data presented here could contain corruption (degraded or missing characters), so the source document (lin
consulted to ensure accuracy. If you use any of the data shown here for research purposes, be sure to cite ODI
document. Please use the following citation record or variant thereof:

Legate, Julie Anne (2002). Warlpiri: Theoretical Implications.
URL: http://www.ai.mit.edu/projects/dm/theses/legate02.pdf
(Last accessed 2005-05-31).
ODIN: http://www.csufresno.edu/odin/igt_raw.php?id= 1227&langcode=WBP (2008-03-14).

Example #1:

```
a. Ngarrka-ngku ka       wawirri panti-rni
man-Erg      PresImpf kangaroo spear-Npast
"The man is spearing the kangaroo"
```

Example #2:

```
(1)     Manu yangka wurna-rlangu yinga-lu ya-ni     munga-puru-rlarni-karra-
or   like   travel-e.g.  Rel.C-3pl go-Npast night-during-ObvC-while-Top
"Or like when people travel to another place while it's still dark."
```

Figure 1: IGT instances in a document

Suppose the user clicks on "*Word Order: VSO*", the search tool will retrieve all the languages in ODIN that have VSO order according to the PCFGs extracted from the projected phrase structures (Figure 3). The user can then click on the *Data* link for any language in the list to retrieve the IGT instances in that language.

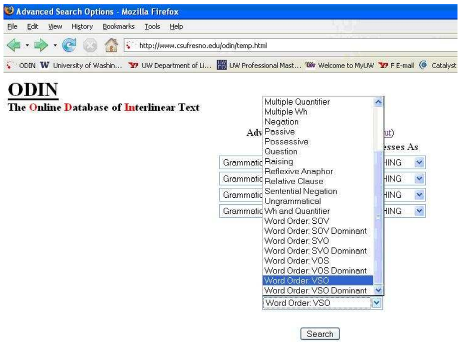
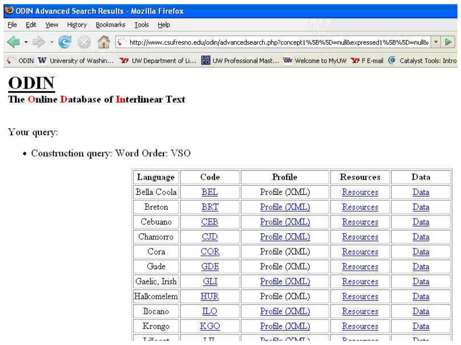


Figure 2: List of linguistic constructions that are currently supported

5 Conclusion

In this paper, we briefly discussed our work on improving the ODIN system, testing the usefulness of the ODIN data for linguistic study, and enhancing the search facility. While IGT data collected off the Web is inherently noisy, we show that even a sample size of 40 IGT instances is large enough to ensure 99% accuracy in predicting Word Order. In the future, we plan to continue our efforts to collect more data for ODIN, in order to make it a more useful resource to the linguistic and computational linguistic audiences. Likewise, we will further extend the search interface to allow more sophisticated queries that tap the full breadth of languages that exist in ODIN, and give users greater access to the enriched annotations and projected structues that can be found only in ODIN.


ODIN Advanced Search Results - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.csufresno.edu/odin/advancedsearch.php?concept1%5B%5D=null&expressed1%5B%5D=null&
ODIN University of Washin… UW Department of Li… UW Professional Mast… Welcome to MyUW F E-mail Catalyst Tools: Intro

ODIN
The Online Database of Interlinear Text

Your query:

- Construction query: Word Order: VSO

Language	Code	Profile	Resources	Data
Bella Coola	BEL	Profile (XML)	Resources	Data
Breton	BRT	Profile (XML)	Resources	Data
Cebuano	CEB	Profile (XML)	Resources	Data
Chamorro	CJD	Profile (XML)	Resources	Data
Cora	COR	Profile (XML)	Resources	Data
Gude	GDE	Profile (XML)	Resources	Data
Gaelic, Irish	GLI	Profile (XML)	Resources	Data
Halkomelem	HUR	Profile (XML)	Resources	Data
Ilocano	ILO	Profile (XML)	Resources	Data
Krongo	KGO	Profile (XML)	Resources	Data
Lillooet	LIL	Profile (XML)	Resources	Data

Figure 3: Languages in ODIN Determined to be VSO

References

John Frederick Bailyn. 2001. Inversion, Dislocation and Optionality in Russian. In Gerhild Zybatow, editor, *Current Issues in Formal Slavic Linguistics*.

W. B. Cavnar and J. M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, April.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, Cambridge, Massachusetts.

William D. Lewis and Fei Xia. 2008. Automatically Identifying Computationally Relevant Typological Features. In *Proceedings of The Third International Joint Conference on Natural Language Processing (IJCNLP)*, Hyderabad, January.

William D. Lewis. 2006. ODIN: A Model for Adapting and Enriching Legacy Infrastructure. In *Proceedings of the e-Humanities Workshop*, Amsterdam. Held in cooperation with e-Science 2006: 2nd IEEE International Conference on e-Science and Grid Computing.

Fei Xia and William D. Lewis. 2007. Multilingual structural projection across interlinearized text. In *Proceedings of the North American Association of Computational Linguistics (NAACL) conference*.

Fei Xia and William D. Lewis. 2008. Repurposing Theoretical Linguistic Data for Tool Development and Search. In *Proceedings of The Third International Joint Conference on Natural Language Processing (IJCNLP)*, Hyderabad, January.

Fei Xia, William D. Lewis, and Hoifung Poon. 2009. Language ID in the Context of Harvesting Language Data off the Web. In *Proceedings of The 12th Conference of the European Chapter of the Association of Computational Linguistics (EACL)*, Athens, Greece, April.