Multilingual Structural Projection across Interlinearized Text

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

William D. Lewis
Department of Linguistics
University of Washington
Seattle, WA 98195
wlewis2@u.washington.edu

Abstract

This paper explores the potential for annotating and enriching data for low-density languages via the alignment and projection of syntactic structure from parsed data for resource-rich languages such as English. We seek to develop enriched resources for a large number of the world's languages, most of which have no significant digital presence. We do this by tapping the body of Web-based linguistic data, most of which exists in small, analyzed chunks embedded in scholarly papers, journal articles, Web pages, and other online documents. By harvesting and enriching these data, we can provide the means for knowledge discovery across the resulting corpus that can lead to building computational resources such as grammars and transfer rules, which, in turn, can be used as bootstraps for building additional tools and resources for the languages represented.[1]

1 Introduction

Developing natural language applications is generally dependent on the availability of annotated corpora. Building annotated resources, however, is a significantly time consuming process involving considerable human effort. Although a number of projects have been undertaken to develop annotated resources for non-English languages, e.g., treebanks, the development of these resources has been no small feat, and to date have been limited to a very small number of the world's languages (e.g., Chinese, German, Arabic, Korean, etc.). Some notable efforts have been undertaken to develop automated means for creating annotated corpora through the projection of annotations (Yarowksy and Ngai, 2001; Xi and Hwa, 2005). The resulting methods, however, can only be applied to a small number of language pairs due mostly to the need for sizeable parallel corpora. Unfortunately, most languages do not have parallel corpora of sufficient size, making these methods inapplicable for the vast majority of the world's languages.

We describe a method for bootstrapping resource creation by tapping the wealth of multilingual data on the Web that has been created by linguists. Of particular note is the linguistic presentation format of "interlinearized text", a common format used for presenting language data and analysis relevant to a particular argument or investigation. Since interlinear examples consist of orthographically or phonetically encoded language data aligned with an English translation, the "database" of interlinear examples found on the Web, when taken together, constitute a significant multilingual, parallel corpus covering hundreds to thousands of the world's languages.

We do not propose that a database of interlinear text alone is sufficient to create NLP resources and tools, but rather that it may act as a means for more rapidly developing such tools using less data. We contend that such a resource allows one to develop computational artifacts, such as grammars and transfer rules, which can be used as "seed" knowledge for building larger resources. In particular, knowing a little about the structure of a language can help in developing annotated corpora and tools, since a little knowledge can go a long way in inducing accurate structure and annotations (Haghighi and Klein, 2006).

Of particular relevance to MT is the issue of struc-

[1] We would like to thank Dan Jinguji for creating the word alignment and source dependency structure gold standards. Our thanks also go to three anonymous reviewers for their helpful comments and suggestions.

tural divergence (Dorr, 1994). Many MT models implicitly make the so-called direct correspondence assumption (DCA) as defined in (Hwa et al., 2002). However, to what extent that assumption holds is tested only on a small number of language pairs using hand aligned data (Fox, 2002; Hwa et al., 2002; Wellington et al., 2006). A larger sample of typologically diverse language data can help test the assumption for hundreds of languages.

We contend that the knowledge garnered from structural projections applied to interlinear text can bootstrap the development of resources and tools across parallel corpora, where such corpora could be of smaller size and the resulting tools more robust, opening the door to the development of tools and resources for a larger number of the world's languages. Given the imminent death of half of the world's 6,000 languages (Krauss, 1992), the development of *any* language specific tools for a larger percentage of the world's languages than is currently possible can aid in both their documentation and preservation.

2 Background

The practice of presenting language data in interlinear form has a long history in the field of linguistics, going back at least to the time of the structuralists (see (Swanton, 1912) for early examples). The modern form of interlinear data presentation started to gel in the mid-1960s, resulting in the canonical three line form shown in Ex (1), which we will refer to as Interlinear Glossed Text, or IGT. The canonical form consists of three lines: a line for the language in question (often a sentence, which we will refer to here as the *source sentence*), an English gloss line, and an English translation.[2]

(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)

Although IGT is usually embedded in linguistics documents as part of a larger analysis, in and of itself it contains analysis and interesting information about the source language. In particular, the gloss line, which is word and morpheme aligned with the source, contains word and morpheme translations for the source language data, and can even contain grammatically salient annotations (e.g., 3sg for Third Person Singular). Further, the reader will note that many words are shared between the gloss and translation lines, allowing for the alignment between these two lines as a intermediate step in the alignment between the translation and the source.

An effort is underway to collect these interlinear snippets into an online searchable database, the primary purpose of which is to help linguists find analyzed data for languages they are interested in. We use this resource, called ODIN, the Online Database of INterlinear text (Lewis, 2006)[3], as our primary data source. At the time of this writing, ODIN contains 36,439 instances of interlinear data for 725 of the world's languages.

3 The Enrichment Algorithm

Our algorithm enriches the original IGT examples by building syntactic structures over the English data and then projects these onto the source language data via word alignment. The term *syntactic structure* in this paper refers to both phrase structure (PS) and dependency structure (DS). The enrichment process has three steps:

1. Parse the English translation using an off-the-shelf parser.
2. Align the source sentence and English translation with the help of the gloss line.
3. Project the English syntactic structures to obtain the source syntactic structures using word alignment.

3.1 Parsing English sentences

There are many English parsers available to the public, and in this experiment we used Charniak's parser (Charniak, 1997), which was trained on the English Penn Treebank (Marcus et al., 1994). Figure 1(a) shows a parse tree (in the Penn Treebank style) for the English translation in Ex (1). Given a parse tree, we use a head percolation table (Magerman, 1995) to create the corresponding dependency structure. Figure 2(a) shows the dependency structure derived from the parse tree in Figure 1(a).

3.2 Word alignment

Because most of the 700+ languages in ODIN are low-density languages with no on-line bilingual dictionaries or large parallel corpora, aligning the source sentence and its English translation *directly* would not work well. To take advantage of the unique layout of IGT examples, we propose using the gloss line as a bridge between the other two lines; that is, we first align the source sentence and the gloss line, and then align the gloss line and the English translation. The process is illustrated in Figure 3.

[2] As pointed out by a reviewer, there is a long tradition in the classical languages for using interlinear translations. So, too, in other literature bases. Our focus here is strictly limited to IGT, the interlinear form used in the field of linguistics.

[3] The url of ODIN is http://www.csufresno.edu/odin

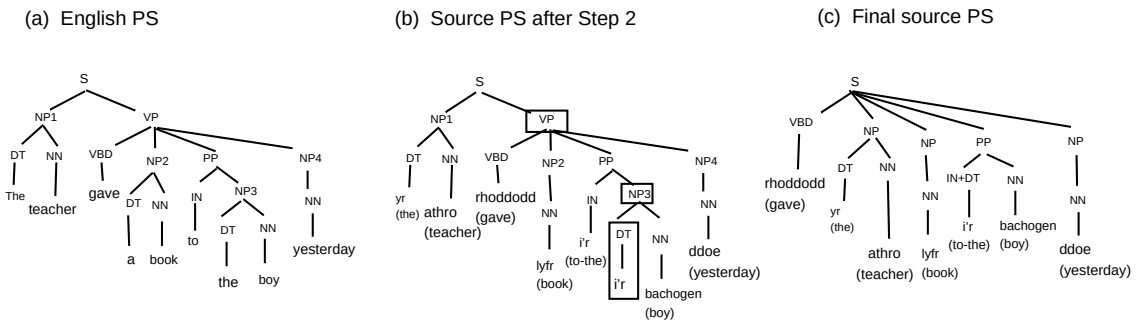

Figure 1: English PS produced by Charniak's parser, and source PS projected from the English PS

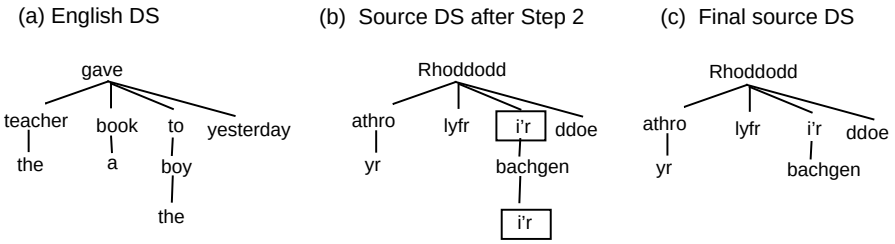

Figure 2: English DS derived from English PS, and source DS projected from the English DS

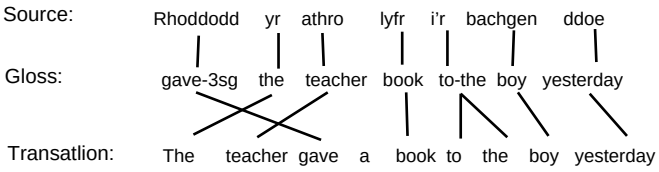

Figure 3: Aligning source sentence and English translation with help of the gloss line

The alignment between the source sentence and the gloss line is trivial and our preliminary experiments showed that simply using whitespace and dashes as delimiters, and assuming a one-to-one alignment produces almost perfect results. In contrast, the alignment between the gloss line and the English translation is more complicated since alignment links can cross and words on one side can link to zero or more words on the other side. We built two aligners for this stage, as described below.

3.2.1 Statistical word aligner

We create a parallel corpus by using the gloss lines and the translation lines of all the IGT examples for all the languages in the ODIN database. We then train IBM models (Brown et al., 1993) using the GIZA++ package (Och and Ney, 2000). In addition to the common practice of lowercasing words and combining word alignments from both directions, we adopt the following strategies to improve word alignment:

Breaking words into morphemes: Since a multi-morpheme word in a gloss line often corresponds to multiple words in the translation line, we split each word on the gloss line into morphemes using the standard IGT morpheme delimiters (*e.g.*, "-"). For instance, the seven words in the gloss line of Ex (1) become nine morphemes.

Adding (x,x) pairs: If a word x appears in the gloss and the translation lines of the same IGT example, it is highly likely that the two copies of the same word should be aligned to each other. To help GIZA++ recognize this property, we first identify and collect all such words and then add single word pairs (x,x) to the training data. For instance, from Ex (1), we would add a sentence pair for each morpheme (excepting *-3sg* which does not appear in the translation line).

3.2.2 Heuristic word aligner

Our second word aligner is based on the assumption that if two words (one on the gloss line, the other on the translation line) have the same root form, they are likely to be aligned to one other.We built a simple English morphological analyzer and ran it on the two lines, and then linked the words with the same

root form.[4]

3.3 Tree projection

We designed two projection algorithms: one which projects PS and the other which projects DS, both from the English to the source language.[5]

3.3.1 Projecting dependency structure

Our projection algorithm is similar to the projection algorithms described in (Hwa et al., 2002) and (Quirk et al., 2005). It has four steps: First, we copy the English DS, and remove all the unaligned English words from the DS.[6] Second, we replace each English word in the DS with the corresponding source words. If an English word x aligns to several source words, we will make several copies of the node for x, one copy for each such source word. The copies will all be siblings in the DS.

If a source word aligns to multiple English words, after Step 2 the source word will have several copies in the resulting DS. In the third step, we keep only the copy that is closest to the root and remove all the other copies.[7] In Step 4, we attach unaligned source words to the DS using the heuristics described in (Quirk et al., 2005). Figure 2 shows the English DS, the source DS after Step 2, and the final DS.

3.3.2 Projecting phrase structure

The algorithm for projecting PS also has four steps, the first two being the same as those for projecting DS. In the third step, we start from the root of the current source PS and for each node x with more than one child, we reorder each pair of x's children until they are in the correct order as dictated by the source sentence. Let y_i and y_j be two children of x, and their spans be $S_i = [a_i, b_i]$ and $S_j = [a_j, b_j]$. When we reorder y_i and y_j, there are four possible scenarios:

(1) S_i and S_j don't overlap: we put y_i before y_j if $a_i < a_j$ or the opposity if $a_i > a_j$.

(2) S_i is a strict subset of S_j: we remove y_j from the PS and *promote* its children: y_j's children will become children of y_j's parent.

(3) S_j is a strict subset of S_i: we remove y_i and promote its children.

(4) S_i and S_j overlap but neither is a strict subset of the other: we remove both y_i and y_j and promote their children. If both y_i and y_j are leaf nodes with the same span, we will merge the two nodes.[8]

The last step is to insert unaligned source words into the source PS. For each unaligned source word x, we will find its closest left and right neighbors that are aligned to some English words, and then attach x to the lowest common ancestor of the two neighbors. Figure 1 shows the English PS, the source PS after Step 2, and the final source PS. The three boxes in 1(b) mark the nodes that are removed in Step 3.

4 Experiments

We tested the feasibility of our approach on a small set of IGT examples for seven languages: German (GER), Korean (KKN), Hausa (HUA), Malagasy (MEX), Welsh (WLS), Irish (GLI), and Yaqui (YAQ). This set of languages was chosen because of its typological diversity: GER and HUA are SVO languages, KKN and YAQ are SOV, GLI and WLS are VSO, and MEX is VOS. In addition, while German and Korean are well-studied and have readily accessible resources that we could use to test the effectiveness and accuracy of our methods, Yaqui, with about 16,000 speakers, is a highly endangered language and serves as a demonstration of our methods for resource-poor and endangered languages.

4.1 Creating the gold standard for the test set

The number of IGT examples in ODIN varies greatly across the seven languages, ranging from less than one hundred for Welsh to over fifteen hundred for German. For each language, we randomly picked 50-150 IGT examples from the available examples whose English translations had at least five words.[9] The examples were manually checked and corrupted examples were thrown away. The remaining examples

[4] When a word is repeated in both the gloss and translation, the individual occurrences are aligned individually in left-to-right order.

[5] The DS projection algorithm as described does not guarantee that the yield of the resulting source DS has the same word order as the source sentence; however, if needed, the algorithm can be easily modified (by making its Step 3 similar to the Step 3 of the PS projection algorithm) to ensure the correct word order.

[6] Every time we remove an internal node x from a DS, we make x's children depend on x's parent directly.

[7] The heuristic is not as arbitrary as it sounds because very often when a source word aligns to multiple English words, one of the English words dominates the rest in the DS (e.g., the node for *to* in Figure 2(a) dominates the node for *the*). We are using the dominating word to represent the whole set.

[8] We will keep one copy and merge the POS tag of the words. For instance, the tag IN+DT in Figure 1(c) was created when two copies of *i'r* in Figure 1(b) were merged.

[9] We skipped examples with very short English translations because they are unlikely to contain much in the way of syntactic structures.

Table 1: The size and average sentence length of the test data

	GER	KKN	HUA	MEX	WLS	GLI	YAQ	Total
# of IGT examples	104	103	77	87	53	46	68	538
# of src words	739	526	441	498	313	252	404	3173
Ave src sent leng	7.11	5.11	5.73	5.72	5.91	5.48	5.94	5.90
# of Eng words	711	735	520	646	329	278	544	3823
Ave Eng sent leng	7.41	7.14	6.75	7.43	6.21	6.04	8.01	7.11
# of speakers	128M	78M	39M	9.4M	580K	260K	16K	255.3M

formed our test data. Table 1 shows the size and average sentence lengths of the test data by language.[10] The languages are sorted by number of speakers (as derived from the Ethnologue (Gordon, 2005)).

We ran our algorithm on the test data, and the system produced the following: an English PS, English DS, word alignment, projected source PS, and projected source DS. We asked human annotators to manually check the output and correct the English DS, word alignments and projected DS structures where necessary.[11] [12] In order to calculate inter-annotator agreement, the Yaqui data and half of the German data were each checked by two annotators, and the disagreement between the annotators was adjudicated and a gold standard was created. The inter-annotator agreement (a.k.a. the F-measure of dependency or alignment links) on English DS, gloss-translation alignment, and projected source DS are 96.34%, 96.35%, and 91.09%, respectively. The rest of the data were annotated by one annotator.

4.2 Word alignment results

We tested our word aligners on 70% (374 examples) of the whole test set (538 examples), while reserving the remaining 30% for future use.

4.2.1 Statistical word aligner

As indicated earlier, the ODIN database contains 36,439 IGT examples. We removed duplicates[13] and examples with missing lines, and used the remaining 28,902 examples for GIZA++ training.[14] Table 2 shows the statistics of the training data with all words lowercased. Tables 3–5 show the performance of the word aligner under three settings:

Table 2: The training data for GIZA++

# of sentences	28,902
# of words in gloss lines	174,765
# of morphemes in gloss lines	251,465
# of words in translation lines	217,022
Size of gloss word vocabulary	16360
Size of gloss morpheme vocabulary	14050
Size of translation word vocabulary	14029

Table 3: The word alignment performance when gloss words are not split into morphemes

	Precision	Recall	F-measure
Gloss → trans	0.674	0.689	0.681
Trans → gloss	0.721	0.823	0.769
Intersection	0.948	0.620	0.750
Union	0.590	0.892	0.711
Refined	0.846	0.780	**0.812**

(1): Not splitting words in the gloss lines into morphemes.

(2): Splitting words in gloss lines into morphemes.

(3): Doing (2) plus adding (x,x) sentence pairs into the training data, where x is a word that appears in both the gloss and translation lines of the same IGT example.

For each setting, we trained in both directions and combined the two alignments by taking the intersection, union, and refined as defined in (Och and Ney, 2000). The best F-score for each setting is in boldface. From the tables, it is clear that the third setting works the best, and combining the alignments

[10] There are three reasons why the sentences are so short. First, since is IGT used to present particular linguistically salient morphological or syntactic material, sentences in IGT are only as long as needed for the given exposé. Second, space constraints often dictate using shorter examples (i.e., they must fit on one line). Third, the IGT extraction algorithm currently used in ODIN does not search for the less common multi-line (*i.e.*, greater than three line) examples.

[11] The English PS and source PS were not corrected; without a thorough linguistic study of the source languages, it is impossible to devise appropriate *gold standards* for their phrase structures.

[12] The DS structures for the English and source language in the gold standard can be non-isomorphic.

[13] Duplicates are common since it is standard practice in linguistics to copy and cite language examples from other papers.

[14] Interestingly, although the IGT examples in the training data come from hundreds of languages in the ODIN, IBM Model 4 performs significantly better than Models 1 and 2 (by at least two percent points for F-measure); therefore, all the GIZA++ results reported in the paper are based on Model 4.

Table 4: The word alignment performance when gloss words are split into morphemes

	Precision	Recall	F-measure
Gloss → trans	0.746	0.889	0.811
Trans → gloss	0.797	0.863	0.829
Intersection	0.958	0.811	0.878
Union	0.659	0.941	0.775
Refined	0.918	0.900	**0.909**

Table 5: The word alignment performance when (x,x) pairs are added

	Precision	Recall	F-measure
Gloss → trans	0.759	0.922	0.833
Trans → gloss	0.801	0.924	0.858
Intersection	0.956	0.885	**0.919**
Union	0.666	0.961	0.787
Refined	0.908	0.921	0.915

from both directions works better than either direction alone.[15]

4.2.2 Heuristic word aligner

The word aligner has two settings. In the first one, the aligner aligns two words if and only if they have the same *orthographic form*. In the second, it aligns two words if and only if they have the same *root* form.[16] The results are shown in the first and second rows of Table 6.

We experimented with various methods of combining the two aligners, and the best one is an augmented heuristic word aligner which links two words if and only if they have the same root form or they are *good* translations of each other according to the translation model built by GIZA++.[17] The result is shown in the last row of Table 6. We used this aligner for the structural projection experiment.

[15]For languages with hundreds of IGT examples, one may wonder whether training GIZA++ with the data for that language alone would outperform the system trained with IGT examples from all the languages in ODIN. To answer this question, we ran three experiments on the German data (for which there are 1757 IGT examples in ODIN after removing duplicates): (a) trained on the (gloss, translation) pairs for all IGT data, (b) trained on the (gloss, translation) pairs of the German data alone, and (c) trained on the (source, translation) pairs of the German data. The test was run against 58 IGT examples, a subset of the German test data in Table 1. It turns out that (a) performs much better than (b) and (c), which justifies the approach we proposed in Section 3.2. For instance, the F-measures for the *refined* alignment for (a)-(c) are 92.5%, 90.2%, and 85.6%, respectively.

[16]For the second setting, we wrote a 90-line Perl application that finds the root for each English word by using a dozen regular expression patterns combined with a list of 163 irregular verbs with their inflected forms.

Table 6: The performance of heuristic word aligner

	Precision	Recall	F-measure
No morphing	0.983	0.742	0.846
With morphing	0.983	0.854	0.914
Augmented aligner	0.981	0.881	**0.928**

4.3 Projection results

We evaluated the results of the major steps in our algorithm: the English DS derived from the parse trees produced by the English parser, the word alignment between the gloss and translation lines, and the projected source DS. We calculated precision, recall, and F-score of dependency links and word alignment links. The F-scores are shown in Table 7.[18]

Both the English parser and the word aligner work reasonably well with most F-scores well above 90%. The F-scores for dependency links in the source DS are lower partly due to errors in early parts of the process (e.g., English DS and word alignment), which propagate to this step. When we replace the automatically generated English DS and word alignment with the ones in gold standard, the F-measure of source DS increases significantly, as shown in Table 8.

In order to identify the causes of the remaining errors in the *oracle* results, we manually checked and classified one third of the errors in the German data. Among the 43 errors in the source DS, 26 (60.5%) are due to language divergence (e.g., head switching), eight (18.6%) are errors made by the projection heuristics, and nine (20.9%) are due to non-exact translations such as the one shown in Ex (2). Because language divergence can reveal interesting typological distinctions between languages, the first type of error may, in fact, identify examples that could be of great value to linguists and computational linguists.

(2) der Antrag des oder der Dozenten
the petition of-the.SG or of-the.PL docent.MSC
"the petition of the docent." (Daniels, 2001)

5 Discussion

5.1 The IGT bias and Knowledge discovery from enriched data

From the enriched data, various kinds of linguistic information can be extracted, such as grammars and

[17]We treat a word pair, (e,f), as a good translation if and only if both $P(e|f)$ and $P(f|e)$ are high.

[18]The *Total* word alignment F-measure is higher than 0.928 as mentioned in Table 6 because the test set used here is the superset of the one used in that section.

Table 7: The system performance on the seven languages

	GER	KKN	HUA	MEX	WLS	GLI	YAQ	Total
English DS	94.25	89.78	96.15	95.51	91.49	93.53	93.57	93.48
Word alignment	94.91	94.20	94.71	94.26	95.65	88.11	93.64	94.03
Source DS	78.14	82.16	84.71	84.22	84.39	78.17	79.36	81.45

Table 8: The F-measure of source dependency links with perfect English DS and/or word alignment

	GER	KKN	HUA	MEX	WLS	GLI	YAQ	Total
With gold Eng DS	82.21	87.67	88.46	85.23	91.72	80.16	83.81	85.42
With gold alignment	85.77	86.15	86.07	88.44	84.98	82.40	86.27	86.00
With both	91.21	91.67	89.82	89.65	94.25	85.77	90.68	90.64

Table 9: Extracted CFGs and evidence of word order

	HUA	MEX	GLI	YAQ
Word order	SVO	VOS	VSO	SOV
# of rule types	102	129	86	115
# of rule tokens	384	466	202	295

transfer rules. We extracted CFGs for the seven languages by reading off the context-free rules from the projected source PS. The numbers of rule types and rule tokens for four of the languages are listed in Table 9.

It is important to note that IGT data is somewhat biased: examples tend to be short and are selected for the purposes of a particular rhetorical context. They, therefore, deviate from the "normal" usage that one might normally expect to find in a corpus of language data. As such, one might question whether the information extracted from IGT would also be skewed due to these biases.

To test the usefulness of the data for answering typological questions, we wrote a simple tool that predicted the canonical word order (e.g., SOV, SVO) of a language using simple heuristics. It was able to produce the correct answers for all seven languages in our sample.[19] We suspect that the number of IGT instances and their diversity (i.e., from multiple documents) is crucial to overcoming the IGT bias, and feel that the same heuristics could be applied to a much larger sample of languages, and adapted to additional typological parameters beyond word order (e.g., orders of heads and modifiers in phrasal constituents). We leave this to future work.

[19] Our code simply went through all the rules in the extracted CFGs and checked the position of the verb with respect to its subject and object. The -SBJ and -OBJ function tags were added to the English parse trees by simple heuristics and were carried over to the source PS via the projection algorithm. GER was the only problematic case, since there is disagreement about German's underlying word order, being either SVO or SOV. Our heuristics returned SOV.

Given syntactically enriched data, it is also possible to search for patterns that are linguistically interesting. For instance, we wrote a piece of code that automatically identified examples with crossing dependencies (i.e., the ones whose DS have crossing links). One such example from the Yaqui data is in Ex (3), where the coordinated noun phrase *kow-ta into mis-ta* "the pig and the cat" is separated by the verb *bwuise-k* "grasp". Note that the crossing dependencies can only be discovered in the Yaqui data and not in the English since none exist in the English.

(3) inepo kow-ta bwuise-k into mis-ta
1SG pig-NNOM.SG grasp-PST and cat-NNOM.SG
"I caught the pig and the cat." (Martínez Fabián, 2006)

So far, we have examined linguistically interesting information in the source. In the future, we plan to examine structures in both the source and English. For instance, we plan to extract transfer rules from the aligned source and English structures and also calculate head/modifier crossings between languages similar to those described in (Fox, 2002).

5.2 Tools and Resource Building

The information that we discover about a language can help with the development of tools for the language. The order of constituents, for instance, can be used to inform prototype-driven learning strategies (Haghighi and Klein, 2006), which can then be applied to raw corpora. It is also possible that small samples of data showing the alignment interactions between source language structures and those of English can provide essential bootstrap information for informing machine translation systems (cf (Quirk and Corston-Oliver, 2006)).

Proof of the utility of an enriched corpus built over ODIN will depend crucially on its evaluation, and we feel that an important part of our future work will be the development of parsers that have been trained on projected structures. These parsers can be evaluated against human built corpora such as treebanks (obviously, only for those languages that have treebanks).

Proof will also come from linguists who will be able to use the corpus to search for constructions of interest (*e.g.*, passives, relative clauses, etc.), and will likely be able to do so using standard tools such as tgrep.[20] Crucially, linguists would be able to conduct such searches over a very large number of languages.

6 Conclusion

In this paper we demonstrate a methodology for projecting structure from annotated English data onto source language data. Because each IGT instance provides an English translation and an intermediary gloss line, we are able to project full syntactic structures from the automatically parsed translation. The fact that our basic methodology and code were applied to a typologically diverse sample of seven languages *without modification* suggests the potential for application to a much larger sample, perhaps numbering into the hundreds of languages. The resulting enriched structures could be of great importance to the fields of linguistics and computational linguistics. For the former, search facilities could be built over the data that would allow linguists to find syntactically marked up data for a large variety of languages, and could even accommodate cross-linguistic comparisons and analyses. For the latter, we could automatically discern grammars and transfer rules from the aligned and marked up data, where these computational artifacts could act as bootstraps for the development of additional tools and resources.

References

John Frederick Bailyn. 2001. Inversion, dislocation and optionality in russian. In Gerhild Zybatow, editor, *Current Issues in Formal Slavic Linguistics*.

Peter Brown, Vincent Pietra, Stephen Pietra, and Robert Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Eugene Charniak. 1997. Statistical Parsing with a Context-Free Grammar and Word Statistics. In *Proc. of AAAI-1997*.

Michael W. Daniels. 2001. On a type-based analysis of feature neutrality and the coordination of unlikes. In *Proceedings of the 8th International HPSG Conference*. CSLI Publications.

Bonnie J. Dorr. 1994. Machine translation divergences: a formal description and proposed solution. *Computational Linguistics*, 20(4):597–635.

[20] This kind of search is reminiscent of Resnik's Linguists Search Engine (http://lse.umiacs.umd.edu), which allows structural search across text found on the Web.

Heidi Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of EMNLP 2002*, Philadelphia, Pennsylvania.

Raymond G. Gordon, editor. 2005. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, fifteenth edition.

Aria Haghighi and Dan Klein. 2006. Protoype-driven sequence models. In *Proceedings of HLT-NAACL*, New York City, NY.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, Pennsylvania.

Michael Krauss. 1992. The World's Languages in Crisis. *Language*, 68(1):4–10.

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.

David M. Magerman. 1995. Statistical Decision-Tree Models for Parsing. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-1995)*, Cambridge, Massachusetts, USA.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, et al. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In *Proc of ARPA Speech and Natural Language Workshop*.

Constantino Martínez Fabián. 2006. *Yaqui Coordination*. Ph.D. thesis, University of Arizona.

Franz-Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *the 38th Annual Conference of the Association for Computational Linguistics (ACL-2000)*, pages 440–447.

Chris Quirk and Simon Corston-Oliver. 2006. The impact of parse quality on syntactically-informed statistical machine translation. In *Proceedings of EMNLP 2006*.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency tree translation: Syntactically informed phrasal smt. In *Proceedings of ACL 2005*.

John R. Swanton. 1912. Haida songs. In Franz Boas, editor, *Publications of the American Ethnological Society, Volume III*. E. J. Brill.

Benjamin Wellington, Sonjia Waxmonsky, and I. Dan Melamed. 2006. Empirical lower bounds on the complexity of translation equivalence. In *Proceedings of ACL 2006*.

Chenhai Xi and Rebecca Hwa. 2005. A backoff model for bootstrapping resources for non-English languages. In *Proceedings of HLT-EMNLP*, pages 851–858, Vancouver, British Columbia, Canada.

David Yarowksy and Grace Ngai. 2001. Inducing Multilingual POS taggers and NP Bracketers via robust projection across aligned corpora. In *Proceedings of NAACL-2001*, pages 377–404.