6

Enriching Language Data through Projected Structures

WILLIAM LEWIS, FEI XIA AND DAN JINGUJI

6.1 Introduction

This paper explores the potential for annotating and enriching data for minority or endangered languages via the alignment and projection of structure from annotated and parsed data for a resource-rich language such as English. The work presented here draws inspiration from the work of (Yarowksy and Ngai, 2001), who tested the methods for projecting linguistic annotations from one language to another, where the resulting projections could be used to train automated part-of-speech taggers and NP bracketers. However, unlike Yarowsky and Ngai, who sought to develop tools and resources for the 200+ major languages of the world, we seek to develop enriched, searchable resources for a larger number of the world's languages, most of which have no significant digital presence. We do this by tapping into the large 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 provide an automated means to search for them, facilitating a kind of structure-based, "construction" query. Further, the enriched data can be used to train and develop robust, statistically-based NLP tools, which can be used for the automated annotation and analysis of language data, especially that of resource-poor and computationally underrepresented languages.

Texas Linguistics Society 10.
Nicholas Gaylord, Stephen Hilderbrand, Heeyoung Lyu, Alexis Palmer and Elias Ponvert eds.

6.2 Background

6.2.1 The Problem

The field of linguistics stands on the edge of a precipice: half of the world's languages are expected to become extinct within the next 50 years (Krauss, 1992). Such "mass extinction" could have devastating consequences for a field that is so dependent on linguistic diversity. Despite noble efforts by a number of tool development and language preservation projects, the process of recording and analyzing linguistic data across the field has remained mostly unchanged, and there exist few to no substantive resources for most of the world's languages. Most of the knowledge base of linguistics persists in a nebulous, distributed form, existing as disconnected pieces of data, markup and analysis across thousands of books, manuscripts, technical reports and journal papers. A world where a linguist could ask a question about the world's languages—for instance, "What languages of North America are split-ergative?" or "Show me data for conditional constructions in the languages of Austronesia"—exists mostly in Science Fiction; the questions can be conceived, but cannot be answered without significant manual effort. One can also imagine a world where tools can be trained on constructions and annotation patterns observed in data for a large body of the world's languages, whose output could then be adapted to train other tools that could be used to enrich newly collected data, perhaps even to the point of generating hypotheses and analyses that could either be accepted or rejected by the linguist using the tool. The problem is unifying new and existing linguistic data into a central repository where automated means for locating and manipulating the data can be provided.

6.2.2 Tapping and Enriching the Existing Infrastructure

A central repository of linguistic data is the vision of the ODIN project (Lewis, to appear). ODIN, the Online Database of Interlinear Text, was developed to automatically locate, collect and house snippets of Interlinear Glossed Text (**IGT**) examples harvested from online scholarly linguistic papers. An example of IGT is shown in (1). A standard instance of IGT 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.[1]

At the time of this writing (2007/02/22), ODIN contains over 41,000 instances for over 700 languages found in 2,800 different documents. The initial purpose of the repository was to facilitate search, allowing linguists to find re-

[1] Although the gloss and translation lines could be encoded in languages other than English (*e.g.*, Spanish or German), we have found that the language of choice for IGT is most often English, even when the analysis presented in the surrounding document is in another language.

sources containing language data for hundreds of the world's languages. As it was originally implemented, linguists could use ODIN to search for data by language code or name, which they could then examine as it was extracted from papers or view it directly in the source documents themselves. More recent extensions to ODIN have included the facility to search by language family, by markup vocabulary (e.g., by markup tags such as 3SG, ERG, ACC, etc., normalized to a common vocabulary), and even by linguistically salient constructions (e.g., conditionals, imperatives, counterfactuals, passives, etc.). The construction query is a unique query facility in that it does not rely on the markup contained in the interlinear examples, but rather searches "enriched" content, where the enrichment of IGT is made possible through the use of statistical taggers and parsers applied to the English translation. Thus, a linguist could cast a query that looks for relative clause or raising constructions by looking for the tell-tale structural and content clues that indicate one of these constructions.

(1) Taro-wa John-ga kasiko-i-to omotta
Taro-TOP John-NOM smart-Pres-Comp think-past
"Taro thought that John was smart." Goro (2003)

It is this process of IGT enrichment that has led to the work discussed here, where we adapt the novel ideas presented in (Yarowksy and Ngai, 2001) for projecting structure across languages to the IGT data type. Since each interlinear example is essentially a snippet of aligned data between English and some source language, we can leverage the alignment provided within each example to project structure from the English onto the source. If a particular projection is successful and accurate, the resulting instance of enriched language data can itself be searched. Thus, the construction queries that currently can be cast only against the English data can instead be cast against the source language data. [2] The following list shows some example queries that a linguist might come up with. Note that query 5, in particular, taps in the unique structure of aligned and enriched IGT.

1. **Find examples of raising or control structures.** This query would look for specific raising and control verbs (*e.g.*, *seem*, *appear*, *ask*, etc.), and the structural clues indicating the presence of raising or control, namely where a noun phrase is the argument of both the raising/control

[2] We recognize that the process of projecting structures may not be successful and the resulting projected structures may not be accurate. Accuracy of the projected structures will be highly dependent on the quality, accuracy and "compatibility" of the English translations, as well as the robustness of the projection algorithm itself. This means that queries cast against projected structures will only be as precise as the projected structures are accurate. We also recognize that that many constructions that a linguist might be interested in may not exist in English, and thus will not be projectable. In other words, language specific constructions may not be recoverable nor searchable using the projection methodology we outline here.

verb and of the verb in a subordinate clause.

2. **Find examples of long distance anaphor binding.** By long distance, we mean cross clausal, where an anaphor that is "distant" from its antecedent is one that does not appear in the same clause. This query would require looking for a reflexive anaphor in a subordinate clause which does not have a possible antecedent in the same clause but does have a potential one in a matrix clause.
3. **Show examples where multiple Wh-words appear within the same clause.** This query requires looking for a clausal node which dominates multiple Wh-word descendants.
4. **Show examples where the matrix VP is marked for the past tense and the subordinate VP is marked for the present tense.** In this query, if cast against the source language data, the clausal boundaries must first be determined, and the annotations for past and present subordinate to two VPs must be searched for, or, if missing, must be inferred from the projections from the English parse (which we recognize could introduce error). The query specifically looks for a matrix VP whose tense is past which dominates another whose tense is present.
5. **Find examples where a noun phrase is headed by the translation of the English word *book* which is immediately followed by a verb whose English translation is *buy*.** Although this query is cast against the source language data, the English translation is used as a means to find the relevant elements in the source language data. In this case, the source equivalents for *buy* and *book* are discovered by following the alignment links that were created between English and the source. The structural part of the query, however, is cast only against the source language data (*i.e.*, the relevant NP and its head and the verb that follows).

Each of these queries cannot be satisfied through standard string based query engines, nor can they be satisfied by a corpus that has not been tagged and parsed. Likewise, casting these queries solely against an enriched English translation could miss novel or language-specific constructions that exist only in the source language data. For instance, structure senstive queries like 2-4, which are highly dependent on the phrasal, clausal, and dependency structures of the language data being queried, may not find relevant data points if only the English is queried; it is the structure of the source language data that is essential for these types of queries to succeed.

We feel that the language data in ODIN can be enriched in such a way to make the kinds of queries described above possible not only for one language but hundreds at a time. But beyond building a large, enriched repository of the world's languages, we see the work described here is a baby-step in the direction of developing robust tools for a large number of the world's

languages. We see the projected structures as facilitating the development part-of-speech taggers and parsers, specifically by providing "seed" structures and data for developing these tools across larger raw or aligned corpora, *a la* (Haghighi and Klein, 2006) . Likewise, the aligned IGT instances can be used for defining the *transfer rules* between English translations and source languages, which can be used to bootstrap transfer-based machine translation systems. Further, a large corpus of the world's languages can be used to some degree for large scale, statistically-based typological analyses.

6.3 The Enrichment Algorithm

Our algorithm enriches the original IGT examples by building syntactic structures over the English data and then projecting these onto the source language data via word and morpheme alignment. This is done in three steps:

1. Parse the English translation using off-the-shelf parsers.
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 and morpheme alignment.

6.3.1 Parsing English Sentences

Parsing is an important task in computational linguistics. Traditional parsers are rule-based and they require humans to manually craft grammars in some general formalisms. In the past decade, because of the availability of large-scale treebanks such as the English Penn Treebank (Marcus et al., 1994), there has been much progress in statistical parsing ((Magerman, 1995, Charniak, 1997, Ratnaparkhi, 1998, Collins, 1999)) and currently several high-quality English parsers are available to the public.

In this experiment, we used Charniak's English parser (Charniak, 1997), which was trained on the English Penn Treebank (Marcus et al., 1994). Figure 2(a) shows a phrase structure (in the Penn Treebank style) for the English translation in Example (1). Given a phrase structure, the standard method of creating the corresponding dependency structure is to use a head percolation table (Magerman, 1995). Our method is a variant of Magerman's algorithm, and the dependency structure for the English sentence is in Figure 3(a).

6.3.2 Word Alignment

The next step is to align the words in the source sentence with the words in the English translation. Word alignment is a common subtask for machine translation. Early work used bilingual dictionaries to obtain possible word alignments and then used heuristics to disambiguate. Since the pioneer research by Brown and his colleagues (Brown et al., 1993), most recent work

(*e.g.*, (Wu, 1994, Vogel et al., 1996, Melamed, 1999, Och and Ney, 2000)) uses machine learning methods to train word aligners automatically from a large amount of parallel data. Because many of the languages in ODIN are low-density languages with no on-line bilingual dictionaries or large parallel corpora, applying existing word alignment algorithms to the IGT instances directly would not yield satisfactory results.

We propose a new method for word alignment that uses the gloss line as a bridge between the source sentence and the English translation. To be more specific, we first align the source sentence and the gloss line, and then align the gloss line and the English translation (both are in English) with either automatically trained word aligners or heuristics. The composition of two alignments is an alignment between the source sentence and the English translation.

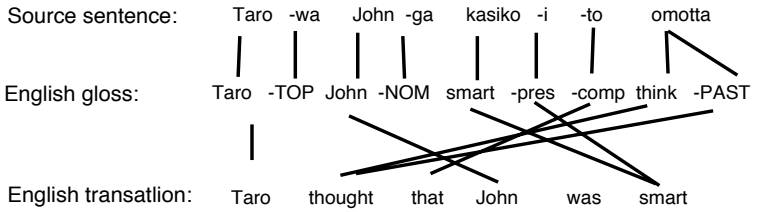

FIGURE 1 Aligning source sentence and English translation with the help of the gloss line

The process is illustrated in Figure 1. 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 would produce almost perfect alignment results on clean IGT data.[3] In contrast, the alignment between the gloss line and the English translation is more complicated as the alignment links can cross and a word on one side could align to multiple words on the other side. To align them, we could train a statistical word aligner with a parallel corpus formed by the second and third lines of all IGT instances; however, due to time constraints, for our preliminary experiments we implemented a word aligner that uses simple heuristics: we first ran an English morphological analyzer on both gloss and translation lines, and then linked two words if and only if they have the same root form. The aligner works reasonably well, as we shall discuss in Section 6.4.

[3] An IGT example is considered *clean* if it is not seriously corrupted when it is extracted from linguistic documents and stored in the ODIN database.

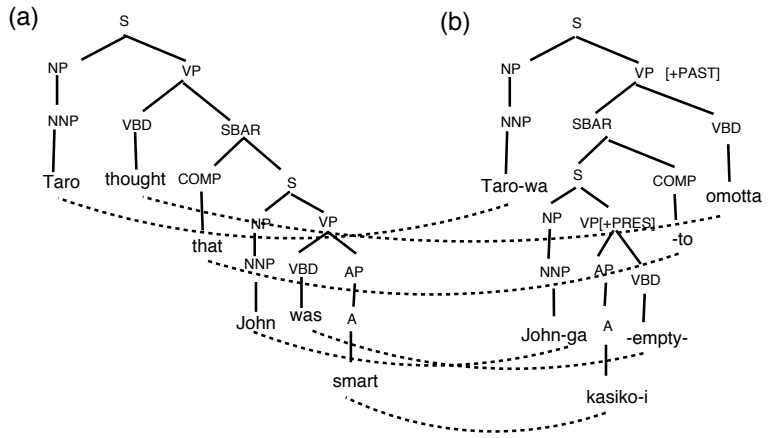

FIGURE 2 English and projected source phrase structures, and word alignment between them

6.3.3 Tree Projection

There have been some work on projection from one language onto another. Yarowksy and Ngai (2001) propose using aggressive generalization techniques and data filtering to prune out noise introduced by direct projection, and their experiments on projecting part-of-speech (POS) tags and base noun phrase bracketings from one language to another showed promising results.

Our goal is to project syntactic structures, not just POS tags or noun phrase bracketings. In this paper, a syntactic structure refers to either a phrase structure (PS) or a dependency structure (DS). Given the English PS and the word alignment between the source sentence and the English translation, conceptually, one could obtain the source PS by replacing English words in the English PS with the corresponding source words, reordering the resulting PS to get the same word order as in the source sentence. Figure 2(b) shows the source PS that could be derived from the English PS. However, in practice, such a straightforward algorithm might not perform well due to word alignment errors or mismatches between languages (e.g., translation divergence as defined in (Dorr, 1994)). We are currently working on a more sophisticated PS projection algorithm. [4]

[4]Note that PS projections, not DS projections, would be required in order to answer a few

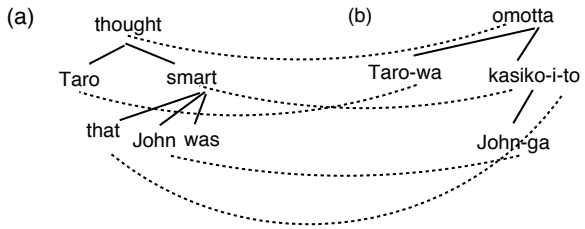

FIGURE 3 English and source dependency structures and word alignment between them

In contrast, dependency structure (DS) is much simpler than phrase structure, as illustrated in Figures 2 and 3. Our DS projection algorithm is similar to other DS projection algorithms such as (Hwa et al., 2002) and (Quirk et al., 2005) and it has four steps.

First, we make a copy of the English DS and remove all the unaligned English words from the DS. In cases where the unaligned word is an internal node in a DS, we adjust the DS such that a dependency relation is established between that word's parent and its children (so that there will not be any broken chains in the projected source DS).[5]

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 will keep only the copy that is closest to the root and remove all the other copies.[6]

In Step 4, we attach each unaligned source word to the DS with the heuristics described in (Quirk et al., 2005). Figure 3 shows the English DS converted from the English PS, and the source DS produced by the projection algorithm.

---

of the construction queries shown in Section 6.2.2. For more up-to-date background on the PS structural projection algorithms, we refer the reader to more recent work described in (Xia and Lewis, 2007).

[5]For further clarification, if the English word x depends on y, and y depends on z, but y is not aligned to any word in the source, we let x depend on z, and remove y from the English DS.

[6]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. We are using the dominating word to represent the whole set.

	Korean	German	Yaqui
# of IGT examples	53	57	69
# of source words	277	412	410
Average source sentence leng	5.23	7.23	5.94
# of English words	393	429	551
Average English sentence leng	7.41	7.53	7.99

TABLE 1 The size and average sentence length of the test data

6.4 Preliminary Results

We tested the feasibility of our approach on a small set of IGT examples for three languages: Korean, German, and Yaqui.[7] We chose these languages for several reasons. German and Korean were chosen because these languages are well-studied and have readily accessible resources that we could use to test the effectiveness and accuracy of our methods. Further, because German is typologically similar to English while Korean is not, we could use these two languages to test the differences in performance across typologically distinct languages. The third language, Yaqui, was chosen both because there was sufficient interlinear data for the language in ODIN and also because Yaqui, with fewer than 20,000 speakers, is a highly endangered language and serves as a demonstration of our methods for resource-poor and endangered languages.

For each of the languages, we randomly picked about 70 IGT examples from the ODIN database whose English translations had at least five tokens. The examples were manually checked and corrupted examples were thrown away. The remaining examples formed our test data. Table 1 shows the size and average sentence lengths of the test data. It is obvious that the source sentences in IGT examples are much shorter than naturally occurring text in newswire, but we believe that they contain useful information about the languages because they were specifically chosen by linguists to demonstrate linguistically interesting phenomena in the languages.

We ran our algorithm on the test data, and the system output was manually corrected by humans: the German and Yaqui data were each checked by two annotators (H_1 and H_2), and the disagreement between the annotators was adjudicated and a gold standard was created. The Korean data were corrected by H_1 only. Table 2 lists the results for the German data. The first column shows the inter-annotator agreement; the second and third columns list the accuracy of each annotator; the last column shows the performance of the

[7] We are in the process of extending our work to additional languages, namely Chamorro, Hausa, Irish, Malagasy, and Welsh, widening the degree of typological diversity in the languages we are studying. Work is ongoing, and we do not yet have results to report on these languages.

	H_1/H_2	H_1/Gold	H_2/Gold	Sys/Gold
English DS	96.34	99.30	96.95	89.80
Word alignment	96.35	98.98	97.99	91.21
Source DS	91.09	97.23	94.06	77.48

TABLE 2 The performance on the German data

	German	Yaqui	Korean
English DS	89.80	93.57	89.80
Word alignment	91.21	93.78	91.21
Source DS	77.48	79.85	77.48
w/ gold Eng DS	83.60	83.81	82.52
w/ gold word alignment	82.52	86.27	83.60
w/ both	88.69	90.20	88.69

TABLE 3 Performance and oracle results on the three languages

system output when compared with the gold standard.

We evaluated the results of the three major steps in our algorithm: the English DS derived from the PS produced by the English parser, the word alignment between the source and translation lines, and the projected source DS. We calculated precision, recall, and F-score of the dependency and word alignment links; the numbers in Tables 2 and 3 are F-scores.

We ran similar experiments on the Yaqui and Korean data. The results are in the first three rows of Table 3: the English parser and the word aligner work reasonably well with most F-scores well above 90%. The F-scores for the source DS are lower partly because the errors from early steps (English DS and word alignment) propagate to this step.

When we replaced the automatically created English DS and word alignment with the ones in gold standard, the F-measure of source DS increased greatly, as shown in the last three rows of Table 3. This result indicates that the improvement on the English parser and the word aligner will directly improve the quality of source DS.

Although the oracle results (the last row in Table 3) are much better than the current system performance (the third row), they are still far from perfect. In order to find out the causes of the remaining errors, we manually checked and classified the errors in the German data. Among the 43 errors, 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 caused by non-exact translations (e.g., the English translation is not an exact translation of the source sentence). Given that there are several ways we could improve word alignment and tree projection algorithms, we expect our system performance

on source DS to reach 90 percent after those improvements.

6.5 Future Directions

We see several future research directions based on the work we describe here. First, we plan to expand structural projections to a much larger sample of languages, and are currently applying the process to a small typologically diverse sample, which includes Chamorro, Hausa, Irish, Malagasy, and Welsh. Preliminary results are similar to those for German, Korean and Yaqui, suggesting that the underlying methodology is sound and that it can be applied without change to a much larger set of languages.

We also plan to improve the word alignment algorithms used on the gloss and translation lines. We are currently experimenting with statistical aligners such as GIZA++ (Och and Ney, 2003). They should outperform heuristic word aligners when the gloss and translation lines of IGT examples include many translation pairs which do not share orthographically similar forms. As an example, GIZA++ is particularly good at discovering gloss line grammatical annotations such as *2SG* that co-occur with words (in this case *you*) on the translation line.

Further, we intend to expand our projection algorithm to phrase structure. Although phrase structures are more difficult to project, preliminary analyses suggest that it is possible to do so. Once done, and given that we can generate enriched corpora of sufficient size, it will then be possible to train tools on the resulting enriched language data, tools such as statistical taggers, chunkers, and parsers. Because many of the languages in ODIN do not have significant digital resources, which is normally a requirement for developing such tools, these efforts could have lasting impacts on the field of linguistics. [8]

The potential for linguistic discovery across a syntactically enriched, multi-lingual corpus also opens the door to knowledge discovery that can have direct benefits to the field of computational linguistics. For instance, it has been shown that knowing a little about the possible orders of constituents in a language can significantly impact prototype-driven learning strategies used for grammar induction (Haghighi and Klein, 2006). Deriving syntactic constituency and the basic order of constituents from projected syntactic structures will be useful for applying these strategies to the languages in ODIN. Likewise, it has been shown that even a small sample of dependency annotated sentences can improve performance in statistical MT systems (Quirk and Corston-Oliver, 2006). We see the potential for producing a dependency structures for a large number of languages which we can

[8] We are aware that these tools will require testing against raw corpora for these languages. This fact, and the fact that the orthographic encoding in the testing corpus and in ODIN will need to coincide, will influence what languages we will eventually choose.

then use as "seeds" for applying these MT methodologies. In grammar engineering, Bender and colleagues (Bender et al., 2002, Bender and Flickinger, 2005), have shown that a small amount of typological information can facilitate the development of the core grammar for a language, allowing for deep parsing and language generation. We see mining strategies being applied to a large number of the languages in ODIN which we can use to supply much of the typological information they deem important.

6.6 Conclusion

In this paper we demonstrate a methodology for projecting structure from annotated English data onto source language data. Although the basic idea is not novel, having been demonstrated in (Yarowksy and Ngai, 2001), we provide several innovations. First, we tap into the existing linguistics infrastructure by enriching data that has been collected, analyzed and published on the Web by linguists. Second, because of the unique structure of IGT, we have been able to demonstrate the feasibility of an alignment and projection methodology that can be applied successfully to languages with very few digital resources. Third, because of the diversity of linguistic data that we have discovered, we are able to apply our algorithms *en masse* to data for a large number of the world's languages, and can do so with no changes to the underlying methodology. Fourth, the resulting enriched database can be searched using structurally sensitive search tools, and the enriched data themselves can act as "seeds" for tool development, *e.g.*, statistical taggers, parsers or even machine translation systems. The development of tools, in particular, is the greatest potential short-term benefit of the work we describe here: given the probable imminent death of thousands of world's languages, *any* language-specific tools that can be developed and used for collecting, enriching and analyzing language data can only help in the collection and preservation of the data that is so crucial to our field.

References

Bender, Emily M. and Dan Flickinger. 2005. Rapid prototyping of scalable grammars: Towards modularity in extensions to a language-independent core. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing IJCNLP-05 (Posters/Demos)*. Jeju Island, Korea.

Bender, Emily M., Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In J. Carroll, N. Oostdijk, and R. Sutcliffe, eds., *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14. Taipei, Taiwan.

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

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

Collins, Mike. 1999. *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

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

Goro, Takuya. 2003. Japanese disjunction and positive polarity.

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

Hwa, Rebecca, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, MI: Association for Computational Linguistics.

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

Lewis, William D. to appear. 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.

Magerman, David M. 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.

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

Melamed, Dan. 1999. Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics* 25(1):107–130.

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

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

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

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

Ratnaparkhi, Adwait. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania.

Vogel, Stefan, Hermann Ney, and Christoph Tillmann. 1996. HMM-Based Word Alignment in Statistical Machine Translation. In *Proc. of the 16th International Conference on Computational Linguistics (COLING 1996)*, pages 836–841.

Wu, Dekai. 1994. Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-1994)*.

Xia, Fei and William Lewis. 2007. Multilingual structural projection across interlinear text. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 452–459. Rochester, New York: Association for Computational Linguistics.

Yarowksy, David 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.