The availability of language-specific computational tools such as parsers can greatly benefit linguistic research, but the development of such tools often depends on the availability of hand annotated data. In this work we explore whether a resource created for another purpose, spanning hundreds of otherwise resource-poor languages, can be used instead. We draw inspiration from Yarowsky and Ngai (2001), who tested methods for projecting linguistic annotations from one language to another. We seek to extend their methodology to a broader set of languages by tapping into the large body of linguistic data on the Web, most of which exists in small, analyzed chunks embedded in various scholarly online documents. Specifically, we exploit Interlinear Glossed Text (IGT) (example in (1)), since it is commonplace online, highly multi-lingual (with hundreds of languages over tens of thousands of instances (Lewis 2006)), and often used to document languages that otherwise have little or no significant digital presence.

(1) Maria tajkaim ya'a-su-kai am=bwa-ka.
Maria tortillas make-TERM-SUB 3SG.NNOM.PL=eat-PST
'After finishing making tortillas, Maria ate them.' (Martínez-Fabián 2006)

The typical IGT instance contains three lines: a line in the source language (Yaqui in (1)), an English gloss line, and an English translation. We parse the translation using the Charniak parser (Charniak 2005), and convert the resulting phrase structure into a dependency structure (DS).[1] We then project the DS onto the source using word-alignment across the three lines of IGT.

We tested our methodology on three languages, German, Korean, and Yaqui. German and Korean were chosen for their typological differences and because of the availability of other resources for evaluation purposes, and Yaqui, because it's highly endangered. Our results are illustrated in the following table. Evaluation was made against a manually created "gold" standard for each language.

	(1) # of IGT examples	(2) avg English sentence length	(3) accuracy English DS	(4) accuracy word alignment	(5) accuracy source DS
Korean	53	7.41	89.80	91.21	77.48
German	57	7.53	93.01	94.67	77.14
Yaqui	69	7.99	93.57	93.78	79.85

Column (3) shows the accuracy (F-measure) of the English DS generated from the Charniak parse. Column (4) shows the accuracy of alignment between the English and source sentences. Column (5) shows the accuracy of the DS projected onto the source, demonstrating the overall success of our methods. Note that the results are in the high 70s for all languages, despite typological differences, suggesting that the methodology can be applied to a much larger set of languages without changes to the underlying approach. We are currently testing the methodology against several dozen additional languages and are also testing the usefulness of the resulting enriched data for the development of NLP and structure-based query tools. It is the development of tools, especially for resource-poor and endangered languages, that could be the most significant consequence of the work we describe here, given the probably imminent death of so many of the world's languages (Krauss 1992).

[1] Dependency parses project more readily than traditional phrase-structure parses.