Automatically Identifying Computationally Relevant Typological Features

William D. Lewis[*]
Microsoft Research
Redmond, WA 98052-6399
wilewis@microsoft.com

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

Abstract

In this paper we explore the potential for identifying computationally relevant typological features from a multilingual corpus of language data built from readily available language data collected off the Web. Our work builds on previous structural projection work, where we extend the work of projection to building individual CFGs for approximately 100 languages. We then use the CFGs to discover the values of typological parameters such as word order, the presence or absence of definite and indefinite determiners, etc. Our methods have the potential of being extended to many more languages and parameters, and can have significant effects on current research focused on tool and resource development for low-density languages and grammar induction from raw corpora.

1 Introduction

There is much recent interest in NLP in "low-density" languages, languages that typically defy standard NLP methodologies due to the absence or paucity of relevant digital resources, such as treebanks, parallel corpora, machine readable lexicons and grammars. Even when resources such as raw or parallel corpora exist, they often cannot be found of sufficient size to allow the use of standard machine learning methods. In some recent grammar induction and MT work (Haghighi and Klein, 2006; Quirk et al., 2005) it has been shown that even a small amount of knowledge about a language, in the form of grammar fragments, treelets or prototypes, can go a long way in helping with the induction of a grammar from raw text or with alignment of parallel corpora.

In this paper we present a novel method for discovering knowledge about many of the world's languages by tapping readily available language data posted to the Web. Building upon our work on structural projections across interlinearized text (Xia and Lewis, 2007), we describe a means for automatically discovering a number of computationally salient typological features, such as the existence of particular constituents in a language (*e.g.,* definite or indefinite determiners) or the canonical order of constituents (*e.g.,* sentential word order, order of constituents in noun phrases). This knowledge can then be used for subsequent grammar and tool development work. We demonstrate that given even a very small sample of interlinearized data for a language, it is possible to discover computationally relevant information about the language, and because of the sheer volume and diversity of interlinear text on the Web, it is possible to do so for hundreds to thousands of the world's languages.

2 Background

2.1 Web-Based Interlinear Data as Resource

In linguistics, the practice of presenting language data in interlinear form has a long history, going back at least to the time of the structuralists. Interlinear Glossed Text, or IGT, is often used to present data and analysis on a language that the reader may not know much about, and is frequently included in scholarly linguistic documents. The canonical form, an example of which is shown in (1), consists of three lines: a line for the language in question (often a sentence, which we will refer to here as the *target sentence*), an English gloss line, and an English translation.

(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 reader will note that many word forms are shared between the gloss and translation lines, allowing for the alignment between these two lines as an intermediate step in the alignment between the translation and the target. We use this fact to facilitate projections from the parsed English data to the target language, and use the resulting grammars to discover the values of the typological parameters that are the focus of this paper.

We use ODIN, the Online Database of INterlinear text (http://www.csufresno.edu/odin), as our primary source of IGT data. ODIN is the result of an effort to collect and database snippets of IGT contained in scholarly documents posted to the Web (Lewis, 2006). At the time of this writing, ODIN contains 41,581 instances of interlinear data for 944 languages.

[*] The work described in this document was done while Lewis was faculty at the University of Washington.

2.2 The Structural Projection and CFG Extraction Algorithms

Our algorithm enriches the original IGT examples by building phrase structures over the English data and then projects these onto the target language data via word alignment. The enrichment process has three steps: (1) parse the English translation using an English parser, (2) align the target sentence and the English translation using the gloss line, and (3) project the phrase structures onto the target sentence. The specific details of the projection algorithm are described in (Xia and Lewis, 2007). Given the projected phrase structures on target sentences, we then designed algorithms to extract context-free grammars (CFGs) for each of the languages by reading off the context-free rules from the projected target phrase structure. Identical rules are collapsed, and a frequency of occurrence is associated with each rule. CFGs so generated provide the target grammars we use for work of typological discovery we describe here.

Since the gloss line provides a means of associating the English translation with the target language, the projections from the English translation effectively project "through" the gloss line. Any annotations associated the projected words, such as POS tags, can be associated with words and morphemes on the gloss line during the enrichment process and then can be projected onto the target. These tags are essential for answering some of the typological questions, and are generally not provided by the linguist. This is especially important for associated particular grammatical concepts, such as number or tense, with particular word categories, such as verb and noun.

3 The IGT and English Biases

The choice of the IGT as our source data type presents two causes for concern. First, IGT is typically used by linguists to illustrate linguistically interesting phenomena in a language. A linguist often carefully chooses examples from a language such that they are representative of the phenomena he or she wishes to discuss, and in no way can they be seen as being randomly sampled from a "corpus" of day-to-day usage for the language. It might be argued, then, that a corpus built over IGT suffers from this bias, what we call the *IGT bias*, and results generated from IGT will be somewhat skewed. Second, since we enrich IGT using a method of structural projection from parses made to English translations, the language structures and the grammars extracted from them might suffer from an English-centrism, what we call *English bias*: we cannot assume that all languages will have the same or similar grammatical features or constructions that English has, and by projecting structures from English, we bias the structures we generate to the English source. The degree to which we overcome these biases will demonstrate not only the success of our methodology, but also the viability of a corpus of IGT instances.

4 Experimental Design

4.1 The Typological Parameters

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.[1] One of the most well known and well studied typological types, or *parameters*[2], is that of word order, made famous by Joseph Greenberg (Greenberg, 1963). In this seminal work, Greenberg identified six possible orderings of Subjects, Objects, and Verbs in the world's languages, namely, SVO, SOV, VSO, VOS, OSV and OVS, and identified correlations between word order and other constituent orderings, such as the now well known tendency for SVO languages (*e.g.*, English, Spanish) to have prepositional ordering in adpositional phrases and SOV (*e.g.*, Japanese, Korean) to have postpositional.

We take inspiration from Greenberg's work, and that of succeeding typologists (*e.g.*(Comrie, 1989; Croft, 1990)). Using the linguistic typological literature as our base, we identified a set of typological parameters which we felt could have the most relevance to NLP, especially to tasks which might require prototype or structural bootstraps. All of the parameters we identified enumerate various constituent orderings, or the presence or absence of particular constituents. The complete list of typological parameters is shown in table 1. There are two major categories of parameters shown: (1) Constituent order parameters, which are broken down into (a) word order and (b) morpheme order, and (2) constituent existence. For each parameter, we enumerate the list of possible values (what typologists typically call *types*), which is generally a permutation of the possible orderings, constraining the set of possible answers to these values. The value *ndo* is reserved to indicate that a particular language exhibits *no dominant order* for the parameter in question, that is, there is no default or canonical order for the language. The value *nr*, or *not relevant*, indicates that a primary constituent of the parameter does not exist in the language and therefore no possible values for the parameter can exist. A good example of this can be seen for the DT+N parameter: in some languages, definite and indefinite determiners may not exist, therefore making the parameter irrelevant. In the specific case of determiners, we have the Def and Indef parameters, which describe the presence or absence of definite and/or indefinite determiners

[1] See (Croft, 1990) for a thorough discussion of linguistic typology and lists of possible types.

[2] The term *typological parameter* is in line with common usage within the field of linguistic typology.

for any given language. Since the parameters *Def* and *Indef* are strictly existence tests, their possible values are constrained simply to *Yes* or *No*.

4.2 Creating the Gold Standards

The gold standards were created by examining grammars and typological analyses for each language, and in some cases, consulting with native speakers or language experts. A principal target was the *World Atlas of Language Structures*, or WALS (Haspelmath et al., 2005), which contains a typology for hundreds of the world's languages. For each of the parameters shown in Table 1, a WALS # is provided. This was done for the convenience of the reader, and refers to the specific section numbers in WALS that can be consulted for a detailed explanation of the parameter. In some cases, WALS does not discuss a particular parameter we used, in which case a WALS section number is not provided (*i.e.*, it is *N/A*).

5 Finding the Answers

As discussed, a typology consists of a parameter and a list of possible types, essentially the values this parameter may hold. These values are usually not atomic, and can be decomposed into their permuted elements, which themselves are types. For instance, the word order parameter is constrained by the types *SVO*, *SOV*, etc., whose atoms are the types *S* for Subject, *V* for Verb, and *O* for Object. When we talk about the order of words in a language, we are not talking about the order of certain words, such as the constituents *The teacher*, *read*, and *the book* in the sentence *The teacher read the book*, but rather the order of the types that each of these words maps to, *S*, *V*, and *O*. Thus, examining individuals sentences of a language tell us little about the values for the typological parameters if the data is not annotated.

The structural projections built over IGT provide the annotations for specific phrases, words or morphemes in the target language, and, where necessary, the structural relationships between the annotations as expressed in a CFG. There are three broad classes of algorithms for this discovery process, which correspond directly to each of the basic categories of parameters shown in Table 1. For the word order parameters, we use an algorithm that directly examines the linear relationship of the relative types in the CFG. For the DT+N variable, for instance, we look for the relative order of the POS tags DT and N in the NP rules. For the WOrder variable, we look for the relative order NPs and Vs in the S (Sentence) and VP rules. If a language has a dominant rule of S → NP VP, it is highly likely that the language is SVO or SOV, and we can subsequently determine VO or OV by examining the VP rule: VP → V NP indicates VO and VP → NP V indicates OV.

Table 2: Functional Tags in the CFGs

Tag	Meaning	Parameters Affected
NP-SBJ	Subject NP	WOrder, V-OBJ
NP-OBJ	Object NP	WOrder, V-OBJ
NP-POSS	Possessive NP	Poss-N
NP-XOBJ	Oblique Object NP	VP-OBJ
PP-XOBJ	Oblique Object PP	VP-OBJ
DT1	the	DT-N, Def
DT2	a,an	DT-N, Indef
DT3	this, that	Dem-N, Def
DT4	all other determiners	Not used

Determining morpheme order is somewhat simplified in that the CFGs do not have to be consulted, but rather a grammar consisting of possible morpheme orders, which are derived from the tagged constituents on the gloss line. The source of the tags varies: POS tags, for instance, are generally not provided by the linguist, and thus must be projected onto the target line from the English translation. Other tags, such as *case*, *number*, and *tense/aspect* are generally represented by the linguist but with a finer granularity than we need. For example, the linguist will list the specific case, such as NOM for Nominative or ACC for Accusative, rather than just the label "case". We use a table from (Lewis, 2006) that has the top 80 morpheme tags used by linguists to map the specific values to the case, number, and tense/aspect tags that we need.

The existence parameters—in our study constrained to Definite and Indefinite determiners—require us to test the existence of particular POS annotations in the set of relevant CFG rules, and also to examine the specific mappings of words between the gloss and translation lines. For instance, if there are no DT tags in any of the CFG rules for NPs, it is unlikely the language has definite or indefinite determiners. This can specifically be confirmed by checking the transfer rules between *the* and *a* and constituents on the gloss line. If either or both *the* or *a* mostly map to NULL, then either or both may not exist in the language.

6 Experiments

We conducted two experiments to test the feasibility of our methods. For the first experiment, we built a gold standard for each of the typological parameters shown in Table 1 for ten languages, namely Welsh, German, Yaqui, Mandarin Chinese, Hebrew, Hungarian, Icelandic, Japanese, Russian, and Spanish. These languages were chosen for their typological diversity (*e.g.*, word order), for the number of IGT instances available (all had a minimum of fifty instances), and for the fact that some languages were low-density (*e.g.*, Welsh, Yaqui). For the second experiment, we examined the WOrder parameter for 97 languages. The gold standard for this experiment was copied directly from an electronic version of WALS.

Table 1: Computationally Salient Typological parameters (ndo=no dominant order, nr=not relevant)

Label	WALS #	**Description**	**Possible Values**
		Word Order	
WOrder	330	Order of Words in a sentence	SVO,SOV,VSO,VOS,OVS, OSV,ndo[3]
V+OBJ	342	Order of the Verb, Object and Oblique Object (e.g., PP)	VXO,VOX,OVX,OXV,XVO,XOV,ndo
DT+N	N/A	Order of Nouns and Determiners (*a*, *the*)	DT-N, N-DT, ndo, nr
Dem+N	358	Order of Nouns and Demonstrative Determiners (*this*, *that*)	Dem-N, N-Dem, ndo, nr
JJ+N	354	Order of Adjectives and Nouns	JJ-N, N-JJ, ndo
PRP$+N	N/A	Order of possessive pronouns and nouns	PRP$-N, N-PRP$, ndo, nr
Poss+N	350	Order of Possessive NPs and nouns	NP-Poss, NP-Poss, ndo, nr
P+NP	346	Order of Adpositions and Nouns	P-NP, NP-P, ndo
		Morpheme Order	
N+num	138	Order of Nouns and Number Inflections (Sing, Plur)	N-num, num-N, ndo
N+case	210	Order of Nouns and Case Inflections	N-case, case-N, ndo, nr
V+TA	282	Order of Verbs and Tense/Aspect Inflections	V-TA, TA-V, ndo, nr
		Existence Tests	
Def	154	Do definite determiners exist?	Yes, No
Indef	158	Do indefinite determiners exist?	Yes, No

Table 3: Experiment 1 Results (Accuracy)

	WOrder	VP	DT	Dem	JJ	PRP$	Poss	P	N	N	V	Def	Indef	Avg
		+OBJ	+N	+N	+N	+N	+N	+NP	+num	+case	+TA			
basic CFG	0.8	0.5	0.8	0.8	1.0	0.8	0.6	0.9	0.7	0.8	0.8	1.0	0.9	0.800
sum(CFG)	0.8	0.5	0.8	0.8	0.9	0.7	0.6	0.8	0.6	0.8	0.7	1.0	0.9	0.762
CFG w/ func	0.9	0.6	0.8	0.9	1.0	0.8	0.7	0.9	0.7	0.8	0.8	1.0	0.9	0.831
both	0.9	0.6	0.8	0.8	0.9	0.7	0.5	0.8	0.6	0.8	0.7	1.0	0.9	0.769

Since the number of IGT instances varied greatly, from a minimum of 1 (Halkomelem, Hatam, Palauan, Itelmen) to a maximum of 795 (Japanese), as shown in the first column of Table 4, we were able to examine specifically the correlation between the number of instances and our system's performance (at least for this parameter).

6.1 Experiment 1 - Results for 10 Languages, 14 Parameters

As described, the grammars for any given language consist of a CFG and associated frequencies. Our first intuition was that for any given word order parameter, the most frequent ordering, as expressed by the most frequent rule in which it appears, was likely the predominant pattern in the language. Thus, for Hungarian, the order of the DT+N parameter is DT-N since the most frequent rule, namely *NP* → *DT N*, occurs much more frequently than the one rule with the opposing order, by a factor of 33 to 1. Our second intuition was based on the assumption that noise could cause an anomalous ordering to appear in the most frequent rule of a targeted type, especially when the number of IGT examples was limited. We hypothesized that "summing" across a set of rules that contained the list of constituents we were interested in might give more accurate results, giving the predominant patterns a chance to reveal themselves in the summation process.

An examination of the types of rules in the CFGs and the parameter values we needed to populate led us to consider enriching the annotations on the English side. For instance, if a CFG contained the rule S → NP V, it is impossible for us to tell whether the NP is a subject or an object, a fact that is particularly relevant to the WOrder parameter. We enriched the annotations with functional tags, such as SBJ, OBJ, POSS, etc., which we assigned using heuristics based on our knowledge of English, and which could then be projected onto the target. The downside of such an approach is that it increases the granularity of the grammar rules, which then could weaken the generalizations that might be relevant to particular typological discoveries. However, summing across such rules might alleviate some of this problem. We also divided the English determiners into four groups in order to distinguish their different types, and projected the refined tags onto the target. The full set of functional tags we used are shown in Table 2, with the list of typological parameters that were affected by the inclusion of each.[4] The results for the experiment are shown in Table 3.

[4] It should be noted some "summations" were done to the CFGs in a preprocessing step, thus affecting all subsequent processing. All variants of NN (NN, NNS, NNP) were collapsed into N and all of VB (VB, VBD, VBZ, etc.) into V. Unaligned words and punctuation were also deleted and the affected rules collapsed.

Table 4: Confusion Matrix for the Word Order Types

Word	# of	System Prediction			
order	languages	SVO	SOV	VSO	VOS
SVO	46	32	8	0	6
SOV	39	2	33	0	4
VSO	11	2	2	3	4
VOS	1	0	0	0	1

Table 5: Word Order Accuracy for 97 languages

# of IGT instances	Average Accuracy
100+	100%
40-99	99%
10-39	79%
5-9	65%
3-4	44%
1-2	14%

6.2 Experiment 2 Results - Word Order for 97 Languages

The second experiment sought to assign values for the WOrder parameter for 97 languages. For this experiment, a CFG with functional tags was built for each language, and the WOrder algorithm was applied to each language's CFG. The confusion matrix in Table 4 shows the number of correct and incorrect assignments. SVO and SOV were assigned correctly most of the time, whereas VSO produced significant error. This is mostly due to the smaller sample sizes for VSO languages: of the 11 VSO languages in our survey, over half had sample sizes less than 10 IGT instances; of those with instance counts above 70 (two languages), the answer was correct.

6.3 Error Analysis

There are four main types of errors that affected our system's performance:

- Insufficient data – Accuracy of the parameters was affected by the amount of data available. For the WOrder parameter, for instance, the number of instances is a good predictor of the confidence of the value returned. The accuracy of the WOrder parameter drops off geometrically as the number of instances approaches zero, as shown in Table 5. However, even with as few as 4-8 instances, one can accurately predict WOrder's value more than half the time. For other parameters, the absence of crucial constituents (*e.g.*, Poss, PRP$) did not allow us to generate a value.
- Skewed or inaccurate data – Depending on the number of examples and source documents, results could be affected by the *IGT bias*. For instance, although Cantonese (YUH) is a strongly SVO language and ODIN contains 73 IGT instances for the language, our system determined that Cantonese was VOS. This resulted from a large number of skewed examples found in just one paper.
- Projection errors – In many cases, noise was introduced into the CFGs when the word aligner or projction algorithm made mistakes, potentially introducing unaligned constituents. These were subsequently collapsed out of the CFGs. The absent constituents sometimes led to spurious results when the CFGs were later examined.
- Free constituent order – Some languages have freer constituent order than others, making calculation of particular parametric values difficult. For example, Jingulu (JIG) and German (GER) alternate between SVO and SOV. In both cases, our grammars directed us to an order that was opposite our gold standard.

7 Discussion

7.1 Data

In examining Table 5, the reader might question why it is necessary to have 40 or more sentences of parsed language data in order to generalize the word order of a language with a high degree of confidence. After all, anyone could examine just one or two examples of parsed English data to discern that English is SVO, and be nearly certain to be right. There are several factors involved. First, a typological parameter like WOrder is meant to represent a *canonical* characteristic of the language; all languages exhibit varying degrees of flexibility in the ordering of constituents, and discovering the canonical order of constituents requires accumulating enough data for the pattern to emerge. Some languages might require more instances of data to reach a generalization than others precisely because they might have freer word order. English has a more rigid word order than most, and thus would require less data.

Second, the data we are relying on is somewhat skewed, resulting from the IGT bias. We have to collect sufficient amounts of data and from enough sources to counteract any linguist-based biases introduced into the data. It is also the case that not all examples are full sentences. A linguist might be exploring the structure of noun phrases for instance, and not provide full sentences.

Third, we are basing our analyses on projected structures. The word alignment and syntactic projections are not perfect. Consequently, the trees generated, and the rules read off of them, may be incomplete or inaccurate.

7.2 Relevance to NLP

Our efforts described here were inspired by some recent work on low-density languages (Yarowksy and Ngai, 2001; Maxwell and Hughes, 2006; Drabek and Yarowsky, 2006). Until fairly recently, almost all NLP work was done on just a dozen or so languages, with the vast majority of the world's 6,000 languages being ignored. This is understandable, since in order to do serious NLP work, a certain threshold of corpus size must be achieved. We provide a means for generating small, richly annotated corpora for hundreds of languages using freely available data found on the Web. These corpora can then be used to generate other electronic resources, such as annotated corpora and associated NLP tools.

The recent work of (Haghighi and Klein, 2006) and (Quirk et al., 2005) were also sources of inspiration. In the former case, the authors showed that it is possible to improve the results of grammar induction over raw corpora if one knows just a few facts about the target language. The "prototypes" they describe are very similar to the our constituent order parameters, and we see our work as an incremental step in applying grammar induction to raw corpora for a large number of languages.

Quirk et al 2005 demonstrates the success of using fragments of a target language's grammar, what they call "treelets", to improve performance in phrasal translation. They show that knowing even a little bit about the syntax of the target language can have significant effects on success of phrasal-based MT. Our parameters are in some ways similar to the treelets or grammar fragments built by Quirk and colleagues and thus might be applicable to phrasal-based MT for a larger number of languages.

Although the reader might question the utility of using enriched IGT for discovering the values of typological parameters, since the "one-off" nature of these discoveries might argue for using existing grammars (*e.g.*, WALS) over harvesting and enriching IGT. However, it is important to recognize that the parameters that we specify in this paper are only a sample of the potential parameters that might be recoverable from enriched IGT. Further, because we are effectively building PCFGs for the languages we target, it is possible to provide gradient values for various parameters, such as the degree of word order variability in a language (*e.g.*, SVO 90%, SOV 10%), the potential for which we not explicitly explored in this paper. In addition, IGT exists in one place, namely ODIN, for hundreds of languages, and the examples that are harvested are also readily available for review (not always the case for grammars).

8 Conclusion

We demonstrate a method for discovering interesting and computationally relevant typological features for hundreds of the world's languages automatically using freely available language data posted to the Web. We demonstrate that confidence increases as the number of data points increases, overcoming the IGT and English biases. Inspired by work that uses prototypes and grammar fragments, we see the work we describe here as being quite relevant to the growing body of work on languages whose digital footprint is much smaller than the ten or so majority languages of the world.

References

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

B. Comrie. 1989. *Language Universals and Linguistic Typology: Syntax and Morphology*. Blackwell, Oxford.

William Croft. 1990. *Typology and Universals*. Cambridge University Press, New York.

Elliott Franco Drabek and David Yarowsky. 2006. Induction of fine-grained part-of-speech taggers via classifier combination and crosslingual projection. In *Proceedings of COLING/ACL2006 Workshop on Frontiers in Linguistically Annotated Corpora*.

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.

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

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. 2005. *The World Atlas of Language Structures*. Oxford University Press, Oxford, England.

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.

Mike Maxwell and Baden Hughes. 2006. Frontiers in linguistic annotation for lower-density languages. In *Proceedings of COLING/ACL2006 Workshop on Frontiers in Linguistically Annotated Corpora*.

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

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*.

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.