Learning Grammar Specifications from IGT: A Case Study of Chintang

Emily M. Bender Joshua Crowgey Michael Wayne Goodman Fei Xia
Department of Linguistics
University of Washington
Seattle, WA 98195-4340 USA
{ebender,jcrowgey,goodmami,fxia}@uw.edu

Abstract

We present a case study of the methodology of using information extracted from interlinear glossed text (IGT) to create of actual working HPSG grammar fragments using the Grammar Matrix focusing on one language: Chintang. Though the results are barely measurable in terms of coverage over running text, they nonetheless provide a proof of concept. Our experience report reflects on the ways in which this task is non-trivial and on mismatches between the assumptions of the methodology and the realities of IGT as produced in a large-scale field project.

1 Introduction

We explore the possibility of learning precision grammar fragments from existing products of documentary linguistic work. A *precision grammar* is a grammar which encodes a sharp notion of grammaticality and furthermore relates strings to elaborate semantic representations. Such objects are of interest in the context of documentary linguistics because: (1) they are valuable tools in the exploration of linguistic hypotheses (especially regarding the interaction of various phenomena); (2) they facilitate the search for examples in corpora which are not yet understood; and (3) they can support the development of treebanks (see Bender et al., 2012a). However, they are expensive to build. The present work is carried out in the context of the AGGREGATION project,[1] which is exploring whether such grammars can be learned on the basis of data already collected and enriched through the work of descriptive linguists, specifically, collections of IGT (interlinear glossed text).

The grammars themselves are not likely targets for machine learning, especially in the absence of treebanks, which are not generally available for languages that are the focus of descriptive and documentary linguistics. Instead, we take advantage of the LinGO Grammar Matrix customization system (Bender et al., 2002; Bender et al., 2010) which maps from collections of statements of linguistic properties (encoded in *choices files*) to HPSG (Pollard and Sag, 1994) grammar fragments which in turn can be used to parse strings into semantic representations in the format of Minimal Recursion Semantics (MRS; Copestake et al., 2005) and conversely, to generate strings from MRS representations. The choices files are a much simpler representation than the grammars derived from them and therefore a more approachable learning target. Furthermore, using the Grammar Matrix customization system to produce the grammars results in much less noise in the automatically derived grammar code than would arise in a system learning grammars directly.

Here, we focus on a case study of Chintang, a Kiranti language of Nepal, described by the Chintang Language Research Project (CLRP) (Bickel et al., 2009). Where Lewis and Xia (2008) and Bender et al. (2013) apply similar methodologies to extract large scale properties for many languages, we focus on a case study of a single language, looking at both the large scale properties and the lexical details. This is important for two reasons: First, it gives us a chance to look in-depth at the possible sources of difficulty in extracting the large scale properties. Second, while large-scale properties are undoubtedly important, the bulk of the information specified in a precision grammar is far more fine-grained. In this case study we apply the methodology of Bender et al. (2013) to extract general word order and case properties and examine the sources of error affecting those results. We also explore extensions of those methodologies and that of Wax (2014) to extract lexical entries and specifications for morpho-

[1] http://depts.washington.edu/uwcl/aggregation/

Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 43–53,
Baltimore, Maryland, USA, 26 June 2014. ©2014 Association for Computational Linguistics

logical rules. Together with a few default specifications, this information is enough to allow us to define grammars through the Grammar Matrix customization system and thus evaluate the results in terms of parsing coverage, accuracy and ambiguity over running text. Chintang is particularly well-suited for this case study because it is an actual endangered language subject to active descriptive research, making the evaluation of our techniques realistic. Furthermore, the descriptive research on Chintang is fairly advanced, having produced both large corpora of high-quality IGT and sophisticated linguistic descriptions, making the evaluation and error analysis possible.

2 Related Work

This work can be understood as a task related to both grammar induction and grammar extraction, though it is distinct from both. It also connects with and extends previous work using interlinear glossed text to extract grammatical properties.

Grammar induction (Clark, 2001; Klein and Manning, 2002; Klein and Manning, 2004; Haghighi and Klein, 2006; Smith and Eisner, 2006; Snyder et al., 2009, inter alios) involves the learning of grammars from unlabeled sentences. Here, *unlabeled* means that the sentences are often POS tagged, but no syntactic structures for the sentences are available. Most of those studies choose probabilistic context-free grammars (PCFGs) or dependency grammars as the grammar framework, and estimate the probability of the context-free rules or dependency arcs from the data. These studies improve parsing performance significantly over some baselines such as the EM algorithm, but the induced grammars are very different from precision grammars with respect to content, quality, and grammar framework.

Grammar extraction, on the other hand, learns grammars (sets of rules) from treebanks. Here the idea is to use heuristics to convert the syntactic structures in a treebank into derivation trees conforming to a particular framework, and then extract grammars from those trees. This has been done in a wide range of grammar frameworks, including PCFG (e.g. Krotov et al., 1998), LTAG (e.g. Xia, 1999; Chen and Vijay-Shanker, 2000), LFG (e.g. Cahill et al., 2004), CCG (e.g. Hockenmaier and Steedman, 2002, 2007), and HPSG (e.g. Miyao et al., 2004; Cramer and Zhang, 2009). However, this approach is not applicable to work on endangered language documentation, as treebanks are not available for such languages.

```
word-order=v-final
has-dets=yes
noun-det-order=det-noun
...
case-marking=erg-abs
erg-abs-erg-case-name=erg
erg-abs-abs-case-name=abs
...
verb4_valence=erg-abs
  verb4_stem1_orth=sams-i-ne
  verb4_stem1_pred=_sams-i-ne_v_re
...
verb-pc3_inputs=verb-pc2
  verb-pc3_lrt1_name=2nd-person-subj
    verb-pc3_lrt1_feat1_name=pernum
    verb-pc3_lrt1_feat1_value=2nd
    verb-pc3_lrt1_feat1_head=subj
    verb-pc3_lrt1_lri1_inflecting=yes
    verb-pc3_lrt1_lri1_orth=a-
```

Figure 1: Excerpts from a choices file

A third line of research attempts to bootstrap NLP tools for resource-poor languages by taking advantage of IGT data and resources for resource-rich languages. The canonical form of an IGT instance includes a language line, a word-to-word or morpheme-to-morpheme gloss line, and a translation line (typically in a resource-rich language). The bootstrapping process starts with word alignment of the language line and translation line with the help of the gloss line. Then the translation line is parsed and the parse tree is projected to the language line using the alignments (Xia and Lewis, 2007). The projected trees can be used to answer linguistic questions such as word order (Lewis and Xia, 2008) or bootstrap parsers (Georgi et al., 2013). Our work extends this methodology to the construction of precision grammars.

3 Methodology

Our goal in this work is to automatically create *choices files* on the basis of IGT data. The choices files encode both general properties about the language we are trying to model as well as more specific information including lexical classes, lexical items within lexical classes and definitions of lexical rules. Lexical rule definitions can include both morphotactic information (ordering of affixes) as well as morphosyntactic information, though here our focus is on the former. Sample excerpts from a choices file are given in Fig 1. These choices files are then input into the Grammar Matrix customization system[2] which produces HPSG gram-

[2] SVN revision (for reproducibility): 27678.

mar fragments that meet the specifications in the choices files. The Grammar Matrix customization system provides analyses of a range of linguistic phenomena. Here, we focus on a few that we consider the most basic: major constituent word order, the general case system, case frames for specific verbs, case marking on nouns, and morphotactics for verbs. In §3.1 we describe the dataset we are working with. §3.2 describes the different approaches we take to building choices files on the basis of this dataset. §3.3 explains the metrics we will use to evaluate the resulting grammars in §4.

3.1 The Chintang Dataset

Chintang (ISO639-3: ctn) is a language spoken by about 5000 people in Nepal and believed to belong to the Eastern subgroup of the Kiranti languages, which in turn are argued to belong to the larger Tibeto-Burman family (Bickel et al., 2007; Schikowski et al., in press). Here we briefly summarize properties of the language that relate to the information we are attempting to automatically detect in the IGT, and in many cases make the problem interestingly difficult.

Schikowski et al. (in press) describe Chintang as exhibiting information-structurally constrained word order: All permutations of the major sentential constituents are expected to be valid, with the different orders subject to different felicity conditions. They state, however, that no detailed analysis of word order has yet been carried out, and so this description should be taken as preliminary.

In contrast, much detailed work has been done on the marking of arguments, both via agreement on the verb and via case marking of dependents (Bickel et al., 2010; Stoll and Bickel, 2012; Schikowski et al., in press). The case marking system can be understood as following an ergative-absolutive pattern, but with several variations from that theme. In an ergative-absolutive pattern, the sole argument of an intransitive verb (here called S) is marked the same as the most patient-like argument of a transitive verb (here called O) and differentiated from the most agent-like argument of a transitive verb (here called A). Most A arguments are marked with an overt case marker called ergative, while S and O arguments appear without a case marker. In most writing about the language, this unmarked case is called nominative; here we will use the term absolutive. Similarly, verbs agree with up to two arguments, and the agreement markers for S and O are generally shared and distinguished from those for A.

Divergences from the ergative-absolutive pattern include variable marking of ergative case on first and second person pronouns as well as valence alternations such as one that licenses occurrences of transitive verbs with two absolutive arguments (and S-style agreement with the A argument) when the O argument is of an indefinite quantity (Schikowski et al., in press). Furthermore, the language allows dropping of arguments (A, S, and O). Finally, there are of course valences beyond simple intransitive and transitive, as well as case frames even for two-argument verbs other than { ERG, ABS }. As a result of the combination of these facts, the actual occurrence of ergative-case-marked arguments in speech is relatively low: Examining a corpus of speech spoken to and around children, Stoll and Bickel (2012) find that only 11% of (semantically) transitive verb tokens have an overt, ergative-marked NP A argument. As discussed below, these properties make it difficult for automated methods to detect both the overall case system of the language and accurate information regarding the case frames of individual verbs.

The dataset we are using contains 9793 (8863 train, 930 test) IGT instances which come from the corpus of narratives and other speech collected, transcribed, translated and glossed by the CLRP.[3] An example is shown in Fig. 2. As can be seen in Fig. 2, the glossing in this dataset is extremely thorough. It is also supported by a detailed Toolbox lexicon that encodes not only alternative forms for each lemma as well as glosses in English and Nepali, but also valence frames for most verb entries which list the expected case marking on the arguments. Finally, note that morphosyntactic properties without a morphological reflex are systematically unglossed in the data, so that ABS never appears (nor does SG for singular nouns, etc.).

In our experiments, we abstract away from the problem of morphophonological analysis in order to focus on morphosyntax and lexical acquisition. Accordingly, our grammars target the second line of the IGT, which represents each form as a sequence of phonologically regularized morphemes.

3.2 Grammars

In this section, we describe the different means we use for extracting the different kinds of informa-

[3] http://www.spw.uzh.ch/clrp

unisaŋa	khatte	mo	kosi	moba
u-nisa-ŋa	khatt-e	mo	kosi-i	mo-pe
3sPOSS-younger.brother-ERG.A	take-IND.PST	DEM.DOWN	river-LOC	DEM.DOWN-LOC

'The younger brother took it to the river.' [ctn] (Bickel et al., 2013c)

Figure 2: Sample IGT

tion required to build the choices files (see Fig 1 above). We first describe our points of comparison (oracle, §3.2.1 and baseline, §3.2.2), and then consider different ways of detecting the large-scale properties (word order, §3.2.3; overall case system, §3.2.4). Next we turn to different ways of extracting two kinds of lexical information: the constraints on case (i.e. case frames of verbs and the case marking on nouns, §3.2.5) and verbal morphotactics (§3.2.6). Finally, we describe a small set of hand-coded 'choices' which are added to all choices files (except the oracle one) in order to create working grammars (§3.2.7).

The alternative approaches to extracting the various kinds of information can be cross-classified with each other, giving the set of choices files described in Table 1. The first column gives identifiers for the choices files. The second specifies how the lexicon was created, the third how the value for major constituent word order was determined, and the fourth how the values for case were determined, including the overall case system, the case frames, and the case values for nouns. These options are all described in more detail below.

3.2.1 Oracle choices file

As an upper-bound, we use the choices file developed in Bender et al., 2012b. This file includes hand-specified definitions of lexical rules for nouns and verbs as well as lexical entries created by importing lexical entries from the Toolbox lexicon developed by the CLRP. This lexicon, as noted above, lists valence frames for most verbal entries. As the Grammar Matrix customization system currently only provides for simple transitive and intransitive verbs, only two verb classes were defined: intransitives with the case frame { ABS } and transitives with the case frame { ERG, ABS }. In addition, there is one class of nouns. Finally, the choices file includes hand-coded lexical entries for pronouns. As an upper-bound, this choices file can be expected to represent high precision and moderate recall: verbs that don't fit the two classes defined aren't imported.

Note that the Grammar Matrix customization system does not currently support the definition of adjectives, adverbs, or other parts of speech outside of verb, noun, determiner, (certain) adpositions, conjunctions and auxiliaries. Thus while we expect each grammar to be able to parse at least some sentences in the corpus, to the extent that sentences tend to include words outside the classes noun, verb and determiner, we expect relatively low coverage, even from our upper-bound.

3.2.2 Baseline choices file

Our baseline choices file is designed to create a working grammar, without particular high-level information about Chintang, that focuses on coverage at the expense of precision. We hand-specified the (counter-factual) assertion that there is no case marking in Chintang, and in addition that Chintang allows free word order (on the grounds that this is the least constrained word order possibility). It also defines bare-bones classes of nouns, determiners and transitive verbs, and then populates the lexicon by using a variant of the methodology in Xia and Lewis 2007. In particular, we parse the translation line using the Charniak parser (Charniak, 1997) and then use the correspondences inherent in IGT to create a projected tree structure for the language line, following Xia and Lewis. An example of the result for Chintang is shown in Fig 3. The projected trees include part of speech tags for each word that can be aligned. For each such word tagged as noun, verb, or determiner, we create an instance in the corresponding lexical type. In this baseline grammar, all verbs are assumed to be transitive, but since all arguments can (optionally) be dropped, the grammar is expected to be able to cover intransitive sentences, even if the semantic representation is wrong.

Since this baseline choices file models Chintang as if it had no case marking, we expect it the resulting grammar to have relatively high recall in terms of the combination of nominal and verbal constituents. On the other hand, since it is building a full-form lexicon and Chintang is a morphologically complex language, we expect it to have relatively low lexical coverage on held-out data.

Choices file	Lexicon	Word order	Case
ORACLE	Manual	Manual	Manual
BASELINE	Fullform	Default	None
FF-AUTO-NONE	Fullform	Auto	None
FF-DEFAULT-GRAM	Fullform	Default	Auto (GRAM)
FF-AUTO-GRAM	Fullform	Auto	Auto (GRAM)
FF-DEFAULT-SAO*	Fullform	Default	Auto (SAO)
FF-AUTO-SAO*	Fullform	Auto	Auto (SAO)
MOM-DEFAULT-NONE	MOM	Default	None
MOM-AUTO-NONE	MOM	Auto	None
MOM-DEFAULT-GRAM*	MOM	Default	Auto (GRAM)
MOM-AUTO-GRAM*	MOM	Auto	Auto (GRAM)
MOM-DEFAULT-SAO*	MOM	Default	Auto (SAO)
MOM-AUTO-SAO*	MOM	Auto	Auto (SAO)

Table 1: Choices files generated

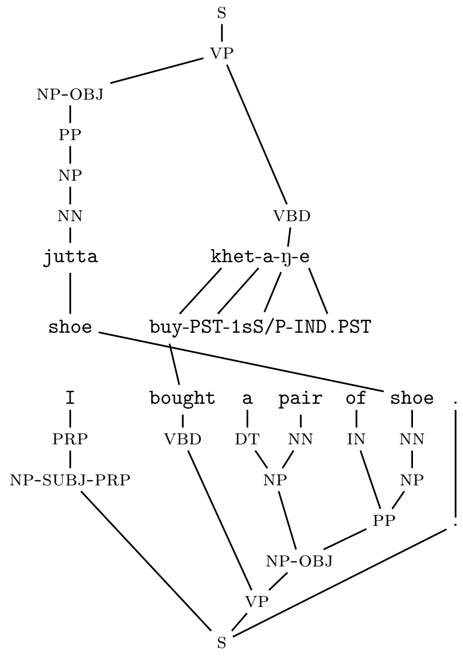

Figure 3: Projected tree structure (ex. from (Bickel et al., 2013d))

3.2.3 Word order

We applied the methodology of Bender et al. (2013) for determining major constituent order. For our dataset, the algorithm chose 'v-final', which matches what is in the ORACLE choices file, but is not necessarily correct. We created two versions of each of the other choices files, one with the default (baseline) answer of 'free word order' and one with this automatically supplied answer.

3.2.4 Case system

Similarly, we applied extended versions of the two methods for automatically discovering case systems from Bender et al. 2013: GRAM which looks for known case grams in glosses (not using projected trees) and SAO which extends the structure-projection methodology of Xia and Lewis (2007) to detect S, A and O arguments and then looks for the most frequent gram associated with each of these.[4] The GRAM method determines the case system of Chintang to be ergative-absolutive, while the SAO method indicates 'none' (no case). Specifying a case system in a choices file has no effect on the coverage or precision of the resulting grammar if the lexical items don't constrain case. Thus the case system choices only make sense in combination with the case frames choices (§3.2.5).

3.2.5 Case frames and case values

The HPSG analysis of case involves a feature CASE which is constrained by both verbs and nouns: Nouns constrain their own CASE value, while verbs constrain the CASE value of the arguments they select for.[5] In order to constrain verbs and nouns appropriately, we first need a range of possible case values. For choices files built based on the GRAM system, we consider case markers to be any of those included in the set of grams defined by the Leipzig Glossing Rules (Bickel et al., 2008): ABL, ABS, ACC, ALL, COM, DAT, ERG, GEN, INS, LOC, and OBL. For choices files built based on the SAO system, we consider as case markers only those grams (automatically) identified as marking S, A, or O. In the present study, that should only be ergative; as there is no marked case for absolutive, all other nouns were treated as absolutive (regardless of their actual case marking, since the SAO system has no way to detect other case grams).

[4] Our extensions involved making the system able to handle the situation where one or more of S, A and O are morphologically unmarked and therefore unreflected in the glosses.

[5] For the details of the analyses of case systems provided by the Grammar Matrix, see Drellishak 2009.

In choices files which specify case systems, we constrain the case value for nouns by creating one noun class for every case value, and then assigning the lexical entries for nouns to those lexical classes based on the grams in the gloss of the noun.[6]

Similarly, we create lexical classes for each case frame identified for transitive and intransitive verbs: We look for case grams on each argument of the verb, as determined by the function tags in the projected tree (e.g. NP-SUBJ-PRP in Fig 3).[7] For each case frame we identify, we create a lexical class, and we create lexical entries for verbs based on the case frames we extract for them. When the system identifies both an overt subject and an overt object, it considers the verb to be transitive and constrains the case of its two arguments based on the observed case values. If either argument is overt but not marked for case, the verb is constrained to select for the default case on that argument, according to the detected case marking system (i.e. ergative for transitive subjects and absolutive for transitive objects, in this instance). When there is an overt subject but no overt object, the verb is treated as intransitive and is constrained to select for a subject of the observed case (or the default case, here absolutive, if the overt subject bears no case marker). When there is an overt object but no subject, the verb is assumed to be transitive and the object's case assigned as with other transitives but the subject's case is constrained to the default (i.e. ergative, in this instance). Verbs with no overt arguments are not matched.

3.2.6 MOM choices file: Automatically extracted lemmas and lexical rules

The final refinement we try on our baseline is to apply the 'Matrix-ODIN Morphology' (MOM) methodology of Wax 2014. This methodology attempts to automatically identify affixes and create appropriate descriptions of lexical rules in a choices file to model those affixes. As a result, it also identifies stems. Thus we use the same basic choices as in the baseline choices file, but now populate the lexicon with stems rather than full-forms. Compared to BASELINE, this one should result in a grammar with better lexical coverage on held-out data, to the extent that the MOM system is able to correctly extract both stems and inflectional rules. We note that while the MOM system uses the same conceptual approach to alignment as that in the BASELINE, GRAM and SAO approaches, the implementation is separate, and so does not find exactly the same set of verbs.

3.2.7 Shared choices

The ORACLE choices file ran as-is. For the remaining choices files, we also needed to answer the questions about determiners (whether there are any, position with respect to the noun). Based on initial experiments, we chose 'yes' for the presence of determiners and 'det-noun' order. In an attempt to boost coverage generally, we also coded the choices that allow any argument to be dropped. While the determiner-related choices are specific to Chintang, the latter set of choices could be expected to boost coverage (at the cost of some precision) for any language.

3.2.8 Summary

Table 1 shows the 10 logical possibilities that arise from combining the methods discussed in this section, in addition to the ORACLE grammar and the BASELINE grammar. However, we test only a subset of these possibilities for the following reasons:[8] The SAO system chose no case as the case system for Chintang. As a result, this makes FF-DEFAULT-SAO and FF-AUTO-SAO the same as BASELINE and FF-AUTO-NONE, respectively. In future work, we aim to improve the SAO system but until it is effective enough to pick some case system for Chintang, these options do not require further testing. Secondly, while it is possible in principle to combine the output of the MOM system (which classifies verbs based on their morphological combinatoric potential) with the output of the system behind the GRAM choices files (which classifies verbs based on their case frames), doing so is non-trivial because these classifications are orthogonal, yet each verb must inherit from each dimension. We thus leave the exploration of MOM-DEFAULT-GRAM and MOM-AUTO-GRAM (and likewise MOM-DEFAULT-SAO and MOM-AUTO-SAO) for future work.

3.3 Evaluation

We evaluate the grammars generated by the choices files over both the data used to develop them ('training'; 8863 items) as well as data not included in the development process (held-out

[6] In future work, we plan to extend the MOM approach (§3.2.6) from verbs to nouns, but for now, the nouns are treated as full-form lexical entries across all choices files.

[7] While the GRAM method doesn't require the projected trees to determine the overall case system, we do need them here to find case frames for particular verbs.

[8] Untested choices files are marked with an * in the table.

'test' data; 930 items). We run both of these evaluations because we are actually testing two separate questions. The first is whether the grammars generated in this way can provide useful analytical tools to linguists. In this primary use-case, we expect a linguist to provide the system with all of their IGT and then use the generated grammars in order to gain insights into that same data. This does not amount to a case of testing on the training data because the annotations provided to the system (IGT) are not the same as those produced by the system (full parses, including semantic representations). However, we are still interested in also testing on held-out data in order to answer the second question: whether grammars generated in this way can also generalize to further texts.

We evaluate the grammars generated by the choices files we create in terms of *lexical coverage*, *parse coverage*, *parse accuracy* and *ambiguity*. Lexical coverage measures how many items consist only of word forms recognized by the grammar. Any item with unknown lexical items won't parse.[9] Parse coverage is the number of items that receive any analysis at all, where ambiguity is the number of different analyses each item receives. To measure parse accuracy, we examined the items that parse and determined which parses had semantic representations whose predicate-argument structures plausibly matched what was indicated in the gloss.

4 Results

Table 2 compares the lexical information encoded in each of the choices files in a quantitative fashion. The first thing to note is that the grammars vary widely in the size of their lexicons. The BASELINE/FF lexicons are expected to be larger than the others because they take each fully inflected form encountered as a separate lexical entry. On the other hand, the ORACLE choices file was built on the basis of the Toolbox lexicon (dictionary) from the CLRP and thus is effectively created on the basis of a much larger dataset. The GRAM choices files only contain verbs for which a case frame could be identified. If the projected tree was not interpretable by our extraction heuristics or if the example had no overt arguments, then the verb will not be extracted. The MOM choices files, on the other hand, only need to identify verbs in the string to be able to extract them, and should be able to generalize across different inflected forms of the same verb. This gives a number of verb entries intermediate between that for BASELINE/FF and the GRAM files. For nouns, there is less variation: the MOM files use the same data as the BASELINE, while the GRAM method faces as simpler problem than for verbs: it only needs to identify the case gram (if any) in a noun's gloss. The slightly larger numbers of nouns in the GRAM files v. the others can be explained by the same form being glossed in two different ways in the training data.

The remaining differences can be briefly explained as follows: The ORACLE choices file does not contain any entries for determiners. The others all contain the same 240 entries; one for any word aligned by the algorithm to a determiner in the English translation. Only the ORACLE and MOM choices files attempt to handle morphology, and so far MOM only does verbal morphology.

Table 3 presents the results of parsing training and test data with the various grammars, in absolute numbers and in percentages of the entire data set. The 'lexical coverage' columns indicate for how many items the grammars were able to recognize each constituent word form. The 'items parsed' columns show the number of items that received any analysis at all, while 'items correct' show the number of items that were judged (by one of the authors) to have a predicate-argument structure that plausibly reflects the gloss given in the IGT. The final column shows the average number of distinct analyses the grammars find for the items they parse at all.

The results are in fact barely measurable with these metrics (especially on the test data), but nonetheless speak to the differences between the grammars. Regarding lexical coverage, the ORACLE grammar does best on the test data set. This is because it is the only choices file not derived from the training data. Not surprisingly, the BASELINE grammar has the highest number of readings per item parsed, followed closely by FF-AUTO-NONE which adds only a minor constraint on word order.[10] On the other hand, comparing the number of items parsed to the number judged correct, except for the MOM choices files, the 'survival rate' was over 50% for all other tests.[11] This suggests

[9] There are methods for handling unknown lexical items (e.g. Adolphs et al., 2008) in more mature grammars of this type, but these are not applicable at this stage.

[10] It is in this relative lack of constraint that BASELINE mostly clearly forms a baseline to improve upon.

[11] The vast majority of the incorrect parses for the MOM

Choices file	# verb entries	# noun entries	# det entries	# verb affixes	# noun affixes
ORACLE	900	4751	0	160	24
BASELINE	3005	1719	240	0	0
FF-AUTO-NONE	3005	1719	240	0	0
FF-DEFAULT-GRAM	739	1724	240	0	0
FF-AUTO-GRAM	739	1724	240	0	0
MOM-DEFAULT-NONE	1177	1719	240	262	0
MOM-AUTO-NONE	1177	1719	240	262	0

Table 2: Amount of lexical information in each choices file

	Training Data (N = 8863)				Test Data (N = 930)			
	lexical	items	items	average	lexical	items	items	average
choices file	coverage (%)	parsed (%)	correct (%)	readings	coverage (%)	parsed (%)	correct (%)	readings
ORACLE	1165 (13)	174 (3.5)	132 (1.5)	2.17	116 (12.5)	20 (2.2)	10 (1.1)	1.35
BASELINE	1276 (14)	398 (7.9)	216 (2.4)	8.30	41 (4.4)	15 (1.6)	8 (0.9)	28.87
FF-AUTO-NONE	1276 (14)	354 (4.0)	196 (2.2)	7.12	41 (4.4)	13 (1.4)	7 (0.8)	13.92
FF-DEFAULT-GRAM	911 (10)	126 (1.4)	84 (0.9)	4.08	18 (1.9)	4 (0.4)	2 (0.2)	5.00
FF-AUTO-GRAM	911 (10)	120 (1.4)	82 (0.9)	3.84	18 (1.9)	4 (0.4)	2 (0.2)	5.00
MOM-DEFAULT-NONE	1102 (12)	814 (9.2)	52 (0.6)	6.04	39 (4.2)	16 (1.7)	3 (0.3)	10.81
MOM-AUTO-NONE	1102 (12)	753 (8.5)	49 (0.6)	4.20	39 (4.2)	10 (1.1)	3 (0.3)	9.20

Table 3: Results

that, despite the noise introduced by the automatic methods of lexical extraction, the precision grammar backbone provided by the Grammar Matrix can still provide high-quality parses.

For example, the BASELINE grammar produces six parses of the string in (1):

(1) din khiptukum
din khipt-u-kV-m
day count-3P-IND.NPST-1/2nsA
'(We) count days.' [ctn] (Bickel et al., 2013b)

Among these six is one which produces the semantic representation in (2). While this grammar does not yet capture any of the agreement morphology that indicates that the subject is first person plural, it does correctly link the 'day' to the semantic ARG2 of 'count'.

(2)
$$\left\langle h_1, \left| \begin{array}{l} h_3{:}\_\mathsf{din\_n\_day}(x_4), \\ h_5{:}\_\mathsf{exist\_q\_rel}(x_4, h_6, h_7), \\ h_6{:}\_\mathsf{khipt\text{-}u\text{-}kv\text{-}m\_v\_count}(e_2, x_9, x_4) \end{array} \right| \{ h_6 =_q h_3 \} \right\rangle$$

Finally, we note that the longest items we are able to parse consist of one verb and two NPs, each of which can have only up to two words (a determiner and a noun). Most of the examples that do parse consist of only one or two words, while the full data set ranges from items of length 1 to items of length 25 (average 4.5 words/item in training, 5 words/item in test). The Grammar Matrix already supports some longer sentences in the form of coordination, so one avenue for future work is to explore the automatic detection of coordination strategies. Otherwise, branching out to longer sentences will require additions to the Grammar Matrix allowing the specification of modifiers and a wider range of valence types for verbs.

5 Error Analysis

The opportunity to work closely with one language has allowed us to observe several ways in which the assumptions of the systems we are building on do not match what we find in the data. Here we briefly review some of those mismatches and reflects on what could be done to handle them.

The first observation concerns the non-glossing of zero-marked morphosyntactic features, such as absolutive case in Chintang. From the point of view of a consumer of IGT it is certainly desirable to have as much information as possible made explicit in the glossing. From the point of view of a project creating IGT in the context of on-going fieldwork, however, it is likely often difficult to reliably gloss zero morphemes and thus the decision to leave them systematically unglossed is quite sensible. Both the GRAM method and especially the SAO method for detecting case systems, which we extended to extracting case frames for particular verbs, are not yet fully robust to the possibility that certain case values are unmarked morphologically and thus not glossed in the data.

choices files involved analyses of words for 'yes', 'well', 'what' and the like as verbs. Note that one form of 'yes' is the copula, and such examples were accepted. Another source of incorrect parses for many grammars involves homophony between the focus particle and a verb meaning 'come'.

While we extended them to a certain extent in this work, there is still more to be done on this front.

A second observation concerns the glossing of proper names, as in (3):

(3)	pailego	ubhiyauti	paphuma
	paile-ko	u-bhiya	paphu-ma
	first-GEN	3A-marriage	a.clan.of.Rai.people-F

'His first marriage was with a Phuphu woman.' [ctn] (Bickel et al., 2013a)

We use statistical alignment between the translation line and the gloss line and between the gloss line and the language line in order to project information from the analysis of the translation line onto the language line. Glosses such as 'a.clan.of.Rai.people' tend to confuse this alignment process, though they are very informative to a human reader of the IGT. Error analysis of sentences for which we were unable to extract subject and object arguments at all suggested that many of the errors were caused by misalignments likely due to the aligner not being able to cope with this kind of glossing. Future work will explore how to train the aligner to function better in such cases.

In addition to properties of the glossing conventions, there are also properties of the language that proved challenging for our system. The first is the intricate nature of the case-marking system as discussed in §3.1. In particular, our system does not model any distinction between 1st and 2nd person pronouns and other nouns, such that when the pronouns appear without a case marker, they are taken to be in the unmarked case (i.e. absolutive), though this is not necessarily so. The second property of the language that our system found difficult is the optionality of arguments. We were able to adapt our case frame extraction strategy to handle dropped subjects, but dropped objects are more confounding: our system is unable so far to distinguish such verbs from intransitives. One possible way forward in this case is to draw more information from the English translation in the IGT: English tends not to drop arguments, and so when we find an object (especially a pronominal object) in the English translation that is not aligned to anything in the language line, we would have evidence that the verb in question may be transitive.

Finally, we looked closely at the items in the test data for which we had complete lexical analysis, but which still failed to parse. We did this both for the fullform and MOM-based lexicons. The goal here was to evaluate whether (a) our assignment of items to lexical categories was correct (and there was some other issue standing in the way of analyzing the item) or (b) we should have parsed a given item, but our system had misidentified the words in question in such a way that no syntactic analysis could be found. For the baseline system, we found that although some items had misidentified categories (specifically, pronouns and adverbs were sometimes misidentified as determiners), the two major obstacles to parsing came from multi-verb constructions or sentential fragments. Of the 26 unparsed items with lexical coverage, 10 contained multiple verbs and 12 were NP or interjectory fragments (eg: 'Yes, yes, yes.'). We observed a similar pattern among 23 unparsed items from the MOM-based lexicon. We can take two lessons from this assessment: (1) since much of our data comes from naturally occurring speech, it may be useful to rerun our tests with an NP fragment as a valid root symbol in our grammars; (2) proper identification of auxiliary verbs is an important next step for improving our system.

6 Conclusion

In this paper we have taken the first steps towards creating actual precision grammars by creating Grammar Matrix customization system choices files on the basis of automated analysis of IGT. Measured in terms of coverage over held-out data, the results are hardly impressive and might seem discouraging. However, we see in these initial forays rather a proof-of-concept. Moreover, the process of digging into the details of getting an IGT-to-grammar system working for one particular language has been a very rich source of information on the mismatches between the assumptions of systems built to handle high-level properties and the linguistic facts and glossing conventions of the kind of data they are meant to handle.

7 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. BCS-1160274. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

We would like to thank David Wax for his assistance in setting up the MOM system, Olga Zamaraeva for general discussion, and especially the CRLP for providing access to the Chintang data.

References

Peter Adolphs, Stephan Oepen, Ulrich Callmeier, Berthold Crysmann, Dan Flickinger, and Bernd Kiefer. 2008. Some fine points of hybrid natural language parsing. Marrakech, Morocco, May.

Emily M. Bender, 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 John Carroll, Nelleke Oostdijk, and Richard Sutcliffe, editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation*, pages 1–50. 10.1007/s11168-010-9070-1.

Emily M. Bender, Sumukh Ghodke, Timothy Baldwin, and Rebecca Dridan. 2012a. From database to treebank: Enhancing hypertext grammars with grammar engineering and treebank search. In Sebastian Nordhoff and Karl-Ludwig G. Poggeman, editors, *Electronic Grammaticography*, pages 179–206. University of Hawaii Press, Honolulu.

Emily M. Bender, Robert Schikowski, and Balthasar Bickel. 2012b. Deriving a lexicon for a precision grammar from language documentation resources: A case study of Chintang. In *Proceedings of COLING 2012*, pages 247–262, Mumbai, India, December. The COLING 2012 Organizing Committee.

Emily M. Bender, Michael Wayne Goodman, Joshua Crowgey, and Fei Xia. 2013. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 74–83, Sofia, Bulgaria, August. Association for Computational Linguistics.

Balthasar Bickel, Goma Banjade, Martin Gaenszle, Elena Lieven, Netra Paudyal, Ichchha Rai, Manoj Rai, Novel Kishor Rai, and Sabine Stoll. 2007. Free prefix ordering in Chintang. *Language*, 83(1):43–73.

Balthasar Bickel, Bernard Comrie, and Martin Haspelmath. 2008. The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. Max Planck Institute for Evolutionary Anthropology and Department of Linguistics, University of Leipzig.

Balthasar Bickel, Martin Gaenszle, Novel Kishore Rai, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Netra Paudyal, Judith Pettigrew, Ichchha P. Rai, Manoj Rai, Robert Schikowski, and Sabine Stoll. 2009. Audiovisual corpus of the chintang language, including a longitudinal corpus of language acquisition by six children, plus a trilingual dictionary, paradigm sets, grammar sketches, ethnographic descriptions, and photographs. *DOBES Archive*, `http://www.mpi.nl/DOBES`.

Balthasar Bickel, Manoj Rai, Netra P. Paudyal, Goma Banjade, Toya N. Bhatta, Martin Gaenszle, Elena Lieven, Ichchha Purna Rai, Novel Kishore Rai, and Sabine Stoll. 2010. The syntax of three-argument verbs in Chintang and Belhare (Southeastern Kiranti). In *Studies in Ditransitive Constructions: A Comparative Handbook*, pages 382–408. Mouton de Gruyter, Berlin.

Balthasar Bickel, Martin Gaenszle, Novel Kishore Rai, Vishnu Singh Rai, Elena Lieven, Sabine Stoll, G. Banjade, T. N. Bhatta, N Paudyal, J Pettigrew, and M Rai, I. P.and Rai. 2013a. Hatuwa. Accessed: 15 January 2013.

Balthasar Bickel, Martin Gaenszle, Novel Kishore Rai, Vishnu Singh Rai, Elena Lieven, Sabine Stoll, G. Banjade, T. N. Bhatta, N Paudyal, J Pettigrew, and M Rai, I. P.and Rai. 2013b. Khadak's daily life. Accessed: 15 January 2013.

Balthasar Bickel, Martin Gaenszle, Novel Kishore Rai, Vishnu Singh Rai, Elena Lieven, Sabine Stoll, G. Banjade, T. N. Bhatta, N Paudyal, J Pettigrew, and M Rai, I. P.and Rai. 2013c. Tale of a poor guy. Accessed: 15 January 2013.

Balthasar Bickel, Martin Gaenszle, Novel Kishore Rai, Vishnu Singh Rai, Elena Lieven, Sabine Stoll, G. Banjade, T. N. Bhatta, N Paudyal, J Pettigrew, and M Rai, I. P.and Rai. 2013d. Talk of kazi's trip. Accessed: 15 January 2013.

Aoife Cahill, Michael Burke, Ruth O'Donovan, Josef Van Genabith, and Andy Way. 2004. Long-distance dependency resolution in automatically acquired wide-coverage pcfg-based lfg approximations. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 319–326, Barcelona, Spain, July.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of AAAI-1997*.

John Chen and K. Vijay-Shanker. 2000. Automated Extraction of TAGs from the Penn Treebank. In *Proc. of the 6th International Workshop on Parsing Technologies (IWPT-2000), Italy*.

Alexander Clark. 2001. Unsupervised induction of stochastic context-free grammars using distributional clustering. In *Proc. of the 5th Conference on Computational Natural Language Learning (CoNLL-2001)*.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language & Computation*, 3(4):281–332.

Bart Cramer and Yi Zhang. 2009. Construction of a german hpsg grammar from a detailed treebank. In *Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks (GEAF 2009)*, pages 37–45, Suntec, Singapore.

Scott Drellishak. 2009. *Widespread But Not Universal: Improving the Typological Coverage of the Grammar Matrix*. Ph.D. thesis, University of Washington.

Ryan Georgi, Fei Xia, and William D. Lewis. 2013. Enhanced and portable dependency projection algorithms using interlinear glossed text. In *Proceedings of ACL 2013 (Volume 2: Short Papers)*, pages 306–311, Sofia, Bulgaria, August.

Aria Haghighi and Dan Klein. 2006. Prototype-driven grammar induction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, pages 881–888, Sydney, Australia, July. Association for Computational Linguistics.

Julia Hockenmaier and Mark Steedman. 2002. Acquiring compact lexicalized grammars from a cleaner treebank. In *Proc. of LREC-2002*, pages 1974–1981.

Julia Hockenmaier and Mark Steedman. 2007. Ccgbank: A corpus of ccg derivations and dependency structures extracted from the penn treebank. *Computational Linguistics*, 33(3):355–396.

Dan Klein and Christopher Manning. 2002. A general constituent context model for improved grammar induction. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, PA.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, Barcelona, Spain.

Alexander Krotov, Mark Hepple, Robert Gaizauskas, and Yorick Wilks. 1998. Compacting the Penn Treebank Grammar. In *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics (ACL-1998)*, Montreal, Quebec, Canada.

William D. Lewis and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 685–690, Hyderabad, India.

Yusuke Miyao, Takashi Ninomiya, and Junichi Tsujii. 2004. Corpus-oriented grammar development for acquiring a head-driven phrase structure grammar from the penn treebank. In *Proc. of the First International Joint Conference on Natural Language Processing (IJCNLP-2004)*, Hainan, China.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. The University of Chicago Press and CSLI Publications, Chicago, IL and Stanford, CA.

Robert Schikowski, Balthasar Bickel, and Netra Paudyal. in press. Flexible valency in Chintang. In B. Comrie and A. Malchukov, editors, *Valency Classes: A Comparative Handbook*. Mouton de Gruyter, Berlin.

Noah A. Smith and Jason Eisner. 2006. Annealing structural bias in multilingual weighted grammar induction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL/COLING 2006)*, pages 569–576, Sydney, Australia, July. Association for Computational Linguistics.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 73–81, August.

Sabine Stoll and Balthasar Bickel. 2012. How to measure frequency? Different ways of counting ergatives in Chintang (Tibeto-Burman, Nepal) and their implications. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek, editors, *Potentials of Language Documentation: Methods, Analyses, and Utilization*, pages 83–89. University of Hawai‘i Press, Manoa.

David Wax. 2014. Automated grammar engineering for verbal morphology. Master's thesis, University of Washington.

Fei Xia and William D. Lewis. 2007. Multilingual structural projection across interlinear text. In *Proc. of the Conference on Human Language Technologies (HLT/NAACL 2007)*, pages 452–459, Rochester, New York.

Fei Xia. 1999. Extracting Tree Adjoining Grammars from Bracketed Corpora. In *Proc. of 5th Natural Language Processing Pacific Rim Symposium (NLPRS-1999)*, Beijing, China.