Building Analyses from Syntactic Inference in Local Languages: An HPSG Grammar Inference System

Kristen Howell, University of Washington and LivePerson Inc., USA kphowell@uw.edu

Emily M. Bender, University of Washington, USA ebender@uw.edu

Abstract We present a grammar inference system that leverages linguistic knowledge recorded in the form of annotations in interlinear glossed text (IGT) and in a meta-grammar engineering system (the LinGO Grammar Matrix customization system) to automatically produce machine-readable HPSG grammars. Building on prior work to handle the inference of lexical classes, stems, affixes and position classes, and preliminary work on inferring case systems and word order, we introduce an integrated grammar inference system called BASIL that covers a wide range of fundamental linguistic phenomena. System development was guided by 27 genealogically and geographically diverse languages, and we test the system's cross-linguistic generalizability on an additional 5 held-out languages, using datasets provided by field linguists. Our system out-performs three baseline systems in increasing coverage while limiting ambiguity and producing richer semantic representations, while also producing richer representations than previous work in grammar inference.

1 Introduction

Machine-readable grammars for human languages that are grounded in theoretical syntactic formalisms can be useful tools in the context of endangered language documentation and revitalization. First, they support treebanking (Oepen et al., 2002), which in turn supports data exploration (Letcher and Baldwin, 2013; Bouma et al., 2015); and second, they facilitate the development of tools such as grammar checkers (da Costa et al., 2016) and automated tutors (Hellan et al., 2013). In spite of these advantages, the use of such grammars is hindered by the time-consuming process of developing them together with the need of a specific skillset required for grammar engineering, which is distinct from the skills involved in documentation itself. We are therefore motivated to investigate whether we can create machine-readable grammars automatically.[1] Endangered languages represent scenarios where the type of resources required for typical natural language processing techniques are scarce to non-existent. Furthermore, the output we are targeting goes well beyond simple labels or even structured representations, but rather must be a coherent and well-formed formal object — a grammar.

Fortunately, we have two rich sources of linguistic knowledge from which to work: The first is corpora of interlinear glossed text (IGT), annotated by field linguists during the process of documentation and analysis. Due to the efforts of field linguists and archivists, a number of archives (many of which we list in Appendix A) make IGT data publicly available. An example from Chintang [ISO 639-3: ctn] is shown in (1). Such annotations are linguistically rich, showing what grammatical information is marked morphologically and providing further information implicitly via a translation into a language of broader communication (in all examples we work with, this language of broader communication is English). Using the methodology of annotation projection, as applied to IGT (Xia and Lewis, 2007; Georgi, 2016), we can leverage parsers available for the translation language and project structural information such as part-of-speech (POS) tags and syntactic dependencies onto words in the target language.

(1) Aru unisokonɨŋ.
aru u-ŋis-u-kV-nɨŋ
another 3nss/A-know-3P-IND.NPST-NEG
'They did not know another [language].' [ctn] (Bickel et al., 2013a)

The second source of linguistic knowledge that we have in hand is the LinGO Grammar Matrix customization system (Bender et al., 2002, 2010; Zamaraeva et al., forthcoming), which maps from relatively

[1] This is similar in spirit to the work of Sarveswaran et al. (2019) who present an effort to create FSMs to provide computational benefits in the context of morphological analysis without requiring additional technical skillsets.

simple grammar specifications to full-fledged machine-readable grammars, couched in the framework of Head-driven Phrase Structure Grammar (HPSG; Pollard and Sag 1994; Müller et al. 2021), and compatible with DELPH-IN[2] processing tools. The Grammar Matrix customization system consists of a core grammar, hypothesized to be shared across languages, and a series of typologically-informed libraries of analyses of cross-linguistically variable phenomena.

Leveraging these sources, the question we investigate here is whether and how we can create machine-readable HPSG grammars for typologically diverse local[3] and/or endangered languages on the basis of corpora of IGT and the Grammar Matrix. In particular, we build on the open-source code base provided by the AGGREGATION project (Bender et al., 2014, inter alia) to produce the following contributions: (1) We integrate all existing inference modules into a single system to which (2) we add modules for additional grammatical phenomena and (3) where previous end-to-end testing treated only a single language, we use 27 diverse languages in development, doing end-to-end system testing on 9 of the 27, and then evaluate on 5 additional held-out languages not considered during system development.

We begin by situating our work on grammar inference against the broader background of automatic grammar generation in Section 2 and then provide background on the AGGREGATION project in Section 3. Section 4 describes our methodology for grammar inference, including lexical, morphological and syntactic aspects of an inferred grammar. In Section 5, we describe the languages we used in system development and how we use the DELPH-IN suite of software tools to evaluate the grammars we create by parsing and treebanking held-out data from each language. We use that same methodology for held-out languages to evaluate the generalizability of the system, finding that though the coverage of the grammars is still limited, the proposed methodology generally produces higher quality grammars than three baseline approaches. The languages we test on and the results of this evaluation are presented in Sections 6 and 7. Finally, Section 8 provides error analysis and discussion. We conclude in Section 9 with discussion of applications of grammars produced in this fashion.

2 Automatic Grammar Generation

Interest in creating machine-readable grammars is likely as old as the field of computational linguistics itself, with published work in *grammar engineering*—the process of creating machine-readable grammars by hand—going back at least as far as Zwicky et al. (1965) and continuing into the present day. Our work in grammar inference builds on grammar engineering work (in the form of the Grammar Matrix; Bender et al., 2002, 2010; Zamaraeva et al., forthcoming), but also fits into a tradition of work on *automatic grammar generation*, which is the development of systems that automatically create grammars on the basis of data. Within automatic grammar generation, we distinguish four broad categories of approaches, differentiated by the types of inputs they take: *grammar induction from strings*—automatic grammar generation based on text alone (§2.1); *grammar extraction*—automatic grammar generation based on treebanks (§2.2); *grammar induction from meaning representations*—automatic grammar generation based on strings paired with some form of semantic representation (§2.3); and *grammar inference*—automatic grammar generation based on text annotated with partial grammatical information but not full parse trees or logical forms (§2.4).

Just as these four approaches to grammar generation differ in their input, they also differ in the types of grammars they can produce. Grammar induction, if working from strings alone, will produce noisy representations that align only partially with structures created by linguists. Grammar extraction will produce grammars that provide the same kind of representations as given in the source treebank and similarly, grammar induction based on strings paired with semantic representations will produce grammars that can output those semantic representations. In each of these cases, the generated grammar will also typically include a parse selection model, based on observed patterns in the corpus. Grammar inference systems, by contrast, draw on both partial annotation in their input data and some external source of grammatical knowledge. For this reason, the inferred grammars can generate richer representations than those found in the input.

2.1 Grammar Induction from Strings

Often characterized as an incomplete data problem (see inter alia Klein and Manning, 2001), where the complete data would be a corpus of trees, *grammar induction* from surface strings seeks to produce grammars solely on the basis of text. Early grammar induction work focused on producing context-free grammars (CFGs), which involved two components: (1) identifying con-

[2] www.delph-in.net

[3] These are often called ‘low-resource languages’, but Bird (2022) argues that this label projects a number of Eurocentric beliefs onto these languages. Bird proposes describing languages as *standardized*, *local* and *contact* rather than *high* and *low resource*.

stituents and (2) identifying their categories (see Klein and Manning, 2001, 2002). Klein and Manning (2004) improved upon this work by inducing an unlabeled syntactic dependency grammar and combining it with the induced CFG for better performance parsing over English [eng], German [deu] and Mandarin [cmn]. This basic approach has informed work which further tuned the algorithm by preferring short vs. long dependencies and testing on additional languages, as in Smith and Eisner 2006. One shortcoming of these approaches is that they only take into account contiguous dependencies. Bod (2009) introduces an approach that allows discontiguous subtrees and thereby handles non-adjacent dependencies. Most recently, neural nets, such as BERT (Devlin et al., 2019), have proven effective in producing unlabeled dependency parses, as demonstrated by Hewitt and Manning (2019), although only parses and not a human-interpretable grammar have been generated. While unlabeled syntactic dependencies can be inferred from text and are useful for some tasks, they do not provide any information regarding the type of syntactic relationship between two constituents. Therefore, other methodologies of automatic grammar generation have focused on using inputs that are encoded with more linguistic information.

Still another strand of recent work seeks to improve grammar induction by using strings (still without linguistic labels) that are captions of still images (Shi et al., 2019; Zhao and Titov, 2020) or descriptions of videos (Zhang et al., 2021). These sources of grounding have been shown to improve recall of different constituent types, but the resulting parsers still produce quite impoverished and noisy representations.

2.2 Grammar Extraction

In contrast with the impoverished input used by grammar induction from surface strings, grammar extraction uses the syntactic information available in treebanks — collections of syntactic trees — to define grammars. Typically these grammars are produced by walking the trees in a treebank, collecting rules that could produce those structures and pruning to remove redundant rules (Krotov et al., 1998).

Because an extracted grammar is informed by the formalism and theory implicit in the tree structures in the input, it will produce trees with roughly the same amount of syntactic information as the formalism used to create the treebank. This can range from context-free grammars (CFG), as in Krotov et al. 1994, to grammar formalisms such as HPSG, as in Simov 2002. However, while the level of detail in the treebanked parses limits that of the resulting grammar, work has been done to extract a grammar in a different formalism than that represented in the input. Xia (1999), for example, proposed an algorithm to do additional bracketing on the Penn Treebank II-style trees (Marcus et al., 1994) in order to extract a Lexical Tree Adjoining Grammar (LTAG), which was more expressive than the CFG in the input. Similarly, Hockenmaier and Steedman (2007) present an approach to converting the Penn Treebank to Combinatory Categorial Grammar (CCG) representations, adding significant information, from which CCG grammars can then be extracted (e.g. Hockenmaier and Steedman, 2002; Clark and Curran, 2004). Neural networks have also been used to generate parse trees based on syntax trees in the training data. KERMIT (Zanzotto et al., 2020) generates syntactic parses of the same form as those in the training data and lends a great deal of interpretability to the underlying BERT (Devlin et al., 2019) model, although it does not produce a grammar or human-interpretable rules.

In principle, grammar extraction is possible for any language for which there is a treebank and recent work has leveraged the Universal Dependencies Treebank (Nivre et al., 2016), a collection of dependency treebanks for over 100 languages, to generate grammars for a wide range of languages (see inter alia Agić et al., 2016; Noji et al., 2016; Han et al., 2019). Our goals in this work, however, are to generate grammars for local languages,[4] many of which are not represented in the UD collection, and to produce syntactic and semantic representations which are richer than dependency parses.

2.3 Grammar Induction from Meaning Representations

In contrast with grammar extraction which relies on a treebank of syntactic parses, grammar induction from meaning representations relies on *sembanks*, typically pairing sentences with either semantic dependencies or logical forms. The types of semantic representations used in this work have ranged from formal query language (Kate et al., 2005; Kate and Mooney, 2006) to semantic dependencies from the Redwoods treebanks, which are based on Minimal Recursion Semantics (MRS; Copestake et al., 2005) as in Buys and Blunsom 2017 and Chen et al. 2018. The input is not always limited to meaning representations alone, and for example, previous work has also used additional input lexical templates to better handle morphological complexity (Kwiatkowski et al., 2011).

Due to the richness of semantic information in the input, grammars induced from text paired with semantic representations rather than text alone are capable of capturing much more detailed and meaningful semantic relations than the unlabeled syntactic dependency relations produced by grammars induced only from surface forms. Such semantic representations are still, however, constrained by what's available in the

[4] See footnote 3.

training data.

2.4 Grammar Inference

Grammar inference systems take as input a collection of text with partial grammatical annotations and use some external source of grammatical knowledge that is not specific to the language at hand to produce grammars that give richer representations than those produced by grammar induction without requiring a treebank. While these systems generally are not probabilistic and do not necessarily include a parse-selection model, as is common with induced or extracted grammars, they allow us to automatically generate formal linguistic grammars without a treebank.

To produce grammars in the Minimalist Grammar formalism (MG; Stabler, 1996) of the Minimalist Program (Chomsky, 1995), Indurkhya (2020) used a set of sentences annotated for part-of-speech (POS), agreement, predicate-argument structure and clause type (interrogative or declarative). This system inferred a lexicon for English on the basis of those annotations, pruned it with a set of Minimalist axioms, and combined it with a non-language-specific notion of merge (with internal and external subtypes) to create a machine-readable Minimalist Grammar.

Whereas Indurkhya used a custom annotation scheme for the input data, Hellan (2010) and Bender et al. (2014) leveraged the rich annotation already present in interlinear glossed text (IGT), illustrated in (1). IGT is a particularly rich source of data because it includes morpheme segmentation, glosses for each morpheme which encode morpho-syntactic information and a translation into a language with many NLP resources (frequently English). A particularly attractive fact about IGT data is that it is the format broadly used in linguistics to record data during collection and analysis, so IGT corpora exist for many languages that do not otherwise have very much written text.

Hellan (2010) and Hellan and Beermann (2011) inferred grammars using a combination of specially annotated IGT and the grammar engineering toolkit *TypeGram*. TypeGram is based on the DELPH-IN Joint Reference Formalism (Copestake, 2002a) which supports the development of typed feature structure grammars, typically within the HPSG framework. Hellan (2010) positioned TypeGram as a hybrid of HPSG and Lexical Functional Grammar (LFG; Kaplan and Bresnan, 1982). In addition to the annotations of typical IGT, their input data also included labels indicating syntactic properties such as valence patterns and constructions such as passive. The TypeGram resource included grammatical rules which are named by the same inventory of label types and thus could directly instantiate a grammar off of an appropriately annotated corpus. The authors illustrate their system with examples from Ga [gaa] and Kiswahili [swh].

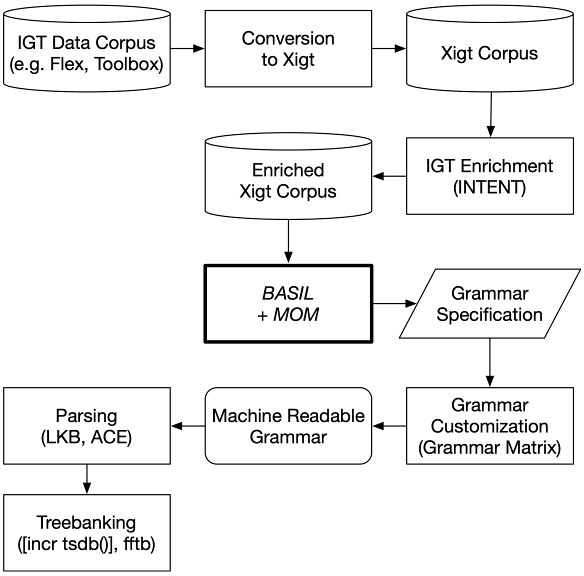


Figure 1: AGGREGATION Pipeline

Bender et al. (2014) also produced HPSG grammars in the DELPH-IN formalism on the basis of IGT data. However, they worked directly from the type of annotations typically produced by documentary linguistics projects, that is, IGT with thorough segmentation and glossing at the morpheme level, but no clause-level annotations. They inferred a lexicon, morphological rules and syntactic properties, and encoded this information in grammar specifications. Using the Grammar Matrix, which allows the user to define a grammar specification that selects from a typologically broad catalog of analyses for different syntactic phenomena and pairs these analyses with a core grammar used across languages, they generated grammars for Chintang [ctn] from their inferred specifications.

Our goal is to create precise syntactic grammars for languages without existing extensive NLP resources, using the rich annotated data that already exists for many of these languages. We build on the approach set forth by Bender et al. (2014), which we describe in detail in the following section. In addition, we extend the typological breadth of work on automatic grammar generation by focusing on languages which are far from the NLP mainstream.

3 The AGGREGATION Project

The AGGREGATION project (Bender et al., 2013, 2014; Howell et al., 2017; Zamaraeva et al., 2017, 2019a), describes its primary goal as providing the benefits of implemented, formal grammars to documentary linguists, without their having to invest time in developing those grammars by hand. Such grammars are useful for testing linguistic hypotheses against data (Bierwisch, 1963; Müller, 1999; Bender, 2008b; Fokkens, 2014; Müller, 2015) as well as building treebanks which are useful for discovering examples of phenomena in a language (Bender et al., 2012; Letcher and Baldwin, 2013; Bouma et al., 2015). The task of developing a grammar by hand is very time consuming and not likely to be taken up by field linguists already busy with the work of language documentation and description. However, the detailed analysis involved in annotating IGT data (another time consuming task that documentary linguists are doing anyway) provides a very rich starting point for producing these grammars automatically. Therefore, an end-to-end pipeline that begins with an IGT corpus and results in a machine-readable grammar has the potential to serve the language documentation community without requiring additional work on their end, either in the form of data curation or grammar engineering.[5] The AGGREGATION project has produced many key components towards this goal, as well as a rudimentary end-to-end pipeline (tested on Chintang in Bender et al. 2014 and Zamaraeva et al. 2019a). In this work, we build on those components to create a more robust and full-featured pipeline. In this section, we present the overall AGGREGATION pipeline as it is developed in our work, with reference to previous work.

In (2; repeated from 1) we present an example of interlinear glossed text (IGT) from the Chintang Language Research Project (CLRP; Bickel et al., 2013b). Based on the information encoded in this IGT and others in the corpus, our goal is a grammar that parses this sentence to produce an HPSG syntactic representation, like the one in Figure 2, and an MRS semantic representation, as in Figure 3.

(2) Aru unisokoniŋ.
aru u-ŋis-u-kV-nɨŋ
another 3nsS/A-know-3P-IND.NPST-NEG
'They did not know another [language].' [ctn] (Bickel et al., 2013a)

Inferring an implemented HPSG grammar directly from an IGT corpus would probably be prohibitively difficult, given the intricate nature of the target grammar. However, we have established a pipeline that leverages a number of existing resources to extract information from an IGT corpus and produce a customized grammar for that language. This pipeline, illustrated in Figure 1, expects as its starting point an IGT corpus, typically from Toolbox (SIL International, 2015) or FLEx

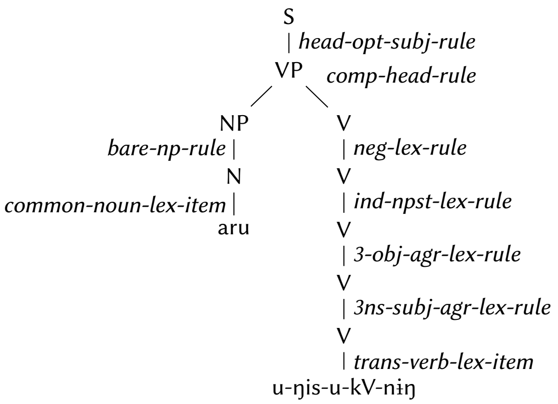
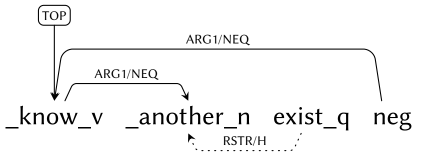


Figure 2: The parse tree for the sentence in (2), which was generated by an inferred grammar of Chintang and corresponds to the semantic representation in Figure 3



Key features on semantic variables:
_know_v (ARG0 {SF *prop,* TENSE *npst,* ASPECT *ind*}, ARG1 {PER *3rd*, NUM *ns*}, ARG2 {PER *3rd*})

Figure 3: A semantic representation for the sentence in (2), generated by an inferred grammar of Chintang

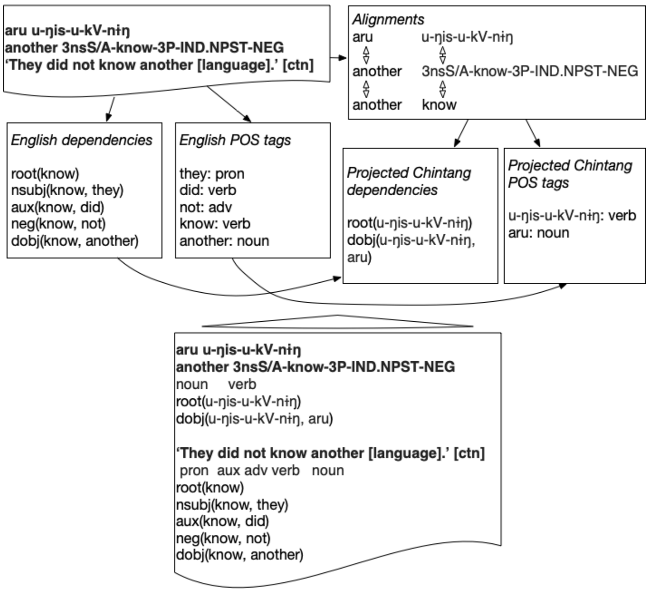


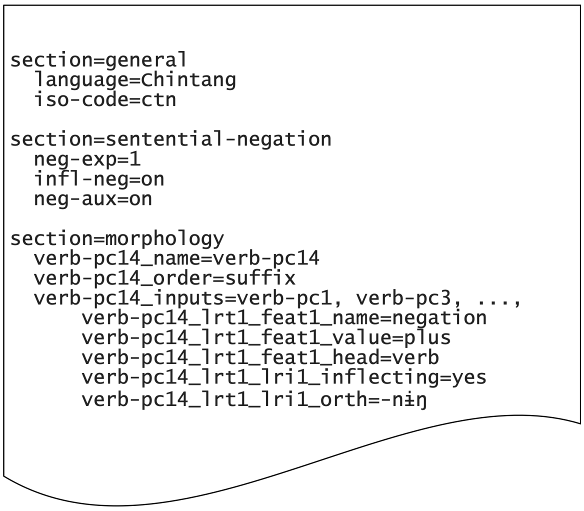
Figure 4: IGT Enriched with INTENT

[5] Ultimately, we hope to serve the communities whose languages are being documented, whether by outsider or insider linguists, by enabling further language technology. However, the immediate audience for implemented grammars remains linguists as opposed to language teachers and learners.

(also from SIL, see (Rogers, 2010)), that was collected by a field linguist, which we convert to an extensible and flexible XML-based format for IGT data called Xigt (Goodman et al., 2015). We then enrich the IGT using INTENT (Georgi, 2016), which projects syntactic dependencies and part-of-speech (POS) tags onto words in the language from a parse of the English translation, as shown in Figure 4.

The enriched corpus provides four key components that are necessary for grammar inference: morpheme segmentation, glossing, POS tags and syntactic dependencies, which can be seen in the final box in Figure 4. The morpheme segmentation and glossing are provided by the linguist in the source IGT and are necessary to extract a lexicon, infer the morphotactic system and associate morpho-syntactic and morpho-semantic information with the corresponding morphemes. POS tags are often provided in the source IGT, but if they are not, they can be acquired from INTENT. INTENT creates alignments between the English translation and the sentence by leveraging the one-to-one alignment between words of the sentence and words in the gloss line and noisy alignment between the gloss words (frequently English lemmas) and the English translation line. It then parses the English sentence and projects the POS and syntactic dependency tags from the English parse onto the aligned words in the source language. While this approach only provides an approximation, as POS and dependencies do not necessarily map across languages, it serves as a useful starting point for inference. Finally, the projected dependencies allow us to discriminate between arguments, modifiers and conjuncts and to identify different types of constituents in the sentence in order to infer syntactic properties.

Our grammar inference system uses these four components to produce a grammar specification file. As an example of our target output, Figure 5 illustrates some of the values we infer that are relevant to sentential negation in Chintang. Chintang expresses sentential negation with a verbal suffix -*nɨŋ*. We indicate that negation is expressed with a single morpheme by setting the negation exponence (`neg-exp`) to 1 in the grammar specification. In the morphology section of the grammar specification, we define one or more lexical rules for a morpheme with orthography *nɨŋ* and morpho-semantic feature `negation: plus`. This grammar specification can be input to the Grammar Matrix customization system (Bender et al., 2002, 2010), which uses stored syntactic analyses to produce customized grammars for languages based on the specification. The customized grammar generated by the Grammar Matrix for this specification will contain the appropriate lexical rule(s) to model negation (Crowgey, 2012), which are illustrated in Figure 6.

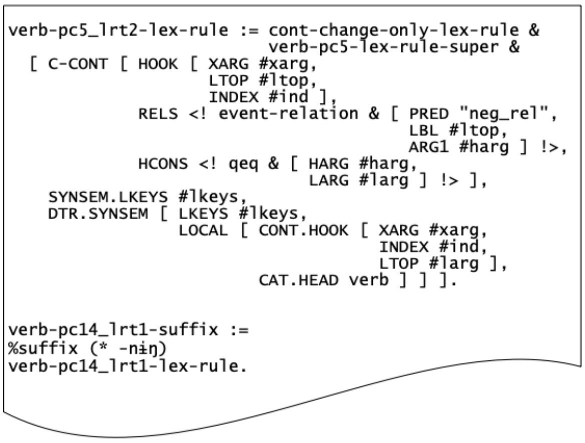


```
section=general
  language=Chintang
  iso-code=ctn

section=sentential-negation
  neg-exp=1
  infl-neg=on
  neg-aux=on

section=morphology
  verb-pc14_name=verb-pc14
  verb-pc14_order=suffix
  verb-pc14_inputs=verb-pc1, verb-pc3, ...,
      verb-pc14_lrt1_feat1_name=negation
      verb-pc14_lrt1_feat1_value=plus
      verb-pc14_lrt1_feat1_head=verb
      verb-pc14_lrt1_lri1_inflecting=yes
      verb-pc14_lrt1_lri1_orth=-nɨŋ
```

Figure 5: A portion of the grammar specification containing (some of) the relevant specifications for sentential negation in Chintang



```
verb-pc5_lrt2-lex-rule := cont-change-only-lex-rule &
                          verb-pc5-lex-rule-super &
  [ C-CONT [ HOOK [ XARG #xarg,
                    LTOP #ltop,
                    INDEX #ind ],
             RELS <! event-relation & [ PRED "neg_rel",
                                         LBL #ltop,
                                         ARG1 #harg ] !>,
             HCONS <! qeq & [ HARG #harg,
                              LARG #larg ] !> ],
    SYNSEM.LKEYS #lkeys,
    DTR.SYNSEM [ LKEYS #lkeys,
                 LOCAL [ CONT.HOOK [ XARG #xarg,
                                     INDEX #ind,
                                     LTOP #larg ],
                         CAT.HEAD verb ] ] ].


verb-pc14_lrt1-suffix :=
%suffix (* -nɨŋ)
verb-pc14_lrt1-lex-rule.
```

Figure 6: The relevant lexical rule for negation in the Chintang grammar, produced from the specification in Figure 5

The lexical rule in Figure 6 licenses the topmost V node in Figure 2 and introduces the neg predication in Figure 3. This rule is expressed in the DELPH-IN Joint Reference Formalism (called tdl; Copestake, 2002a), which can be used to implement HPSG-style typed feature structures. A grammar encoded in this way can be loaded into DELPH-IN processing tools like the LKB (Copestake, 2002b) and ACE (Crysmann and Packard, 2012) for parsing and [incr tsdb()] (Oepen, 2001) and FFTB (Packard, 2015) for treebanking.

Previous work in the AGGREGATION Project has produced grammar specifications that contain a lexicon of nouns and verbs, morphological rules and descriptions of the language's word order and case system as well as case frames for individual words. The lexicon and morphotactic rules are inferred using MOM (Wax, 2014; Zamaraeva, 2016), which we describe in Sections 4.2 and 4.3. These rules abstract away from morphophonology, so the inferred grammars are tested by parsing the morpheme-segmented line of the IGT. Inference algorithms for basic word order and case system were developed by Bender et al. (2013) and this inference together with lexical inference was used to generate grammars by Bender et al. (2014) and Zamaraeva et al. (2019a).

In this work, we present BASIL, an inference system that extends the number of phenomena that can be inferred by building on the existing morphotactic and syntactic inference systems. This system, also described in Howell 2020, infers additional lexical items including determiners, case-marking adpositions, coordinators and auxiliaries as well as properties including argument optionality, sentential negation and coordination. We also integrate syntactic and morphological inference to handle person, number and gender information on nouns, agreement between verbs and their arguments, and tense, aspect and mood contributed morphologically or by auxiliaries. Finally, whereas previous work has either evaluated the correctness of the grammar specifications on a variety of languages (Bender et al., 2013; Howell et al., 2017) or grammar performance on a single language (Bender et al., 2014; Zamaraeva et al., 2019a), we evaluate our system on grammar performance using 14 genealogically and geographically diverse languages.

4 Methodology: Inferring Grammar Specifications

This section focuses on our approach to inferring the grammar specifications illustrated in the previous section. We take as our starting point the system of Zamaraeva et al. (2019a) which integrates the morphological inference module (called MOM; Wax, 2014; Zamaraeva, 2016; Zamaraeva et al., 2017) and a module for inference of a few syntactic properties (Bender et al., 2014; Howell et al., 2017). To this integrated system we add extended inference for morphologically marked syntactic and semantic features, additional lexical classes and further syntactic properties to create BASIL, Building Analyses from Syntactic Inference in Local languages. BASIL takes an enriched (using INTENT; Georgi, 2016) corpus of the Xigt (Goodman et al., 2015) data type as input and produces a grammar specification file which can be input into the Grammar Matrix to generate a custom grammar for the language. This grammar specification (§4.1), often referred to as a 'choices file' in the Grammar Matrix literature, contains specifications for a lexicon (§4.2), a collection of morphological rules (§4.3), definitions of syntactico-semantic features (§4.4) and definitions of syntactic properties (§4.5) for the language at hand. During development, we used a set of 9 core languages to design and tune BASIL's algorithms and consulted an additional 18 languages that were illustrative of particular phenomena we wished to test (see §5.1). In this section, we describe each of BASIL's inference modules, including the typological range covered, what specifications the Grammar Matrix customization system requires, and how we infer appropriate specifications for a language based on IGT.[6]

4.1 The Grammar Specification

In this section, we give a brief quantitative overview of the space in which the inference system is operating. The grammar specification contains definitions for lexical items, morphological rules, syntactico-semantic features and syntactic rules. These take the form of features with either fixed or open-ended values, depending on the linguistic characteristics being defined. While a number of phenomena can be defined in the Grammar Matrix, BASIL focuses on a particular subset of lexical items and syntactic phenomena, which are modeled by 50 fixed features with 136 possible values in addition to a number of open-ended features, which allow the user to enter any value they like, rather than requiring them to choose from a menu. For some features, multiple values lead to similar coverage in the resulting grammars, so we simplify the system by focusing on a subset of the possible values. Other values are difficult to infer with sufficient accuracy from the available data or are so typologically rare that they are more likely to be inferred in error than correctly. For these reasons, BASIL targets only 99 of the 136 values, as summarized in Table 1.

While individual lexical entries and morphological rules have features that must be selected from a menu with a fixed set of values, the number of lexical items

[6] A more detailed description of these modules and the algorithms they use can be found in Howell 2020.

Phenomenon	number possible values	number targeted by inference
noun lexical entry	4	2
verb lexical entry	4	2
auxiliary lexical entry	6	4
adposition lexical entry	3	3
morphological rule	5	5
person	9	8
tense	2	1
word order	10	9
determiner order	4	4
auxiliary order	9	9
case system	9	3
argument optionality	18	15
sentential negation	41	23
coordination	12	11
total	136	99

Table 1: The number of possible values for the 50 features with a fixed value set in the grammar specification and those targeted by the inference system, broken down by syntactic category

and morphological rules defined by BASIL depends on the number of forms attested in the training corpus. Thus the size of the lexicon and morphology sections of the grammar specification varies depending on both the morphological complexity of the language and the diversity and number of samples in the training corpus. Similarly, many of the syntactico-semantic features supported by the Grammar Matrix allow the definition of unbounded numbers of possible values. For case, person, number, gender, tense, aspect and mood, we[7] compiled a list of 116 common values from the Leipzig Glossing Rules (Bickel et al., 2008), the ODIN corpus (Xia et al., 2016), Unimorph (Sylak-Glassman et al., 2015), the GOLD Ontology (GOLD, 2010) and our own observation, which the inference system can add to grammar specifications.

4.2 The Lexicon

The most accurate and fully detailed typological specification cannot produce a working grammar without a lexicon. At the same time, decent coverage over unseen texts for languages with any morphological complexity requires a lexicon built in terms of lexical entries for roots plus some model of morphological processes. The Grammar Matrix customization system elicits, as part of its input grammar specifications, descriptions of lexical classes and lexical rules. In this section, we describe lexical class specifications and how we infer them.

In brief, a lexical class is defined in terms of its part-of-speech, any further features specific to the class, and a set of lexical entries, which give the orthographic representations and semantic predicate symbols[8] for entries in that class. As an example, Figure 7 illustrates a lexical class for a type of common nouns in Meithei [mni].

```
section=lexicon
noun1_name=noun1
noun1_feat1_name=person
noun1_feat1_value=3rd
noun1_det=opt
noun1_stem1_orth=kekrú
noun1_stem1_pred=_blackberry_n_rel
noun1_stem2_orth=khoy
noun1_stem2_pred=_bee_n_rel
```

Figure 7: The definition of a common noun lexical class for Meithei

The Grammar Matrix customization system interface provides for nouns, intransitive verbs, transitive verbs, clausal complement verbs, auxiliaries, copulas, determiners, case-marking adpositions, and adjectives in its lexicon section. In addition, sections for particular syntactic phenomena allow for the definition of lexical entries for such items as conjunctions, subordinating conjunctions, complementizers, and negation adverbs. This classification of basic types of words brings with it a set of assumptions about what word classes exist in the world's languages, for example, that nouns and verbs are distinct cross-linguistically. We make no claims regarding the actual parts of speech of the lexical items MOM and BASIL infer, but attempt to model these words effectively in the resulting grammar. (For recent work showing that even languages with apparent category flexibility can be fruitfully analyzed in this way, see Crowgey's 2019 study of Lushootseed [lut].)

BASIL infers only a subset of the lexical categories supported by the Grammar Matrix, which are shown in Figure 8. In this section, we describe the process of extracting these definitions from the IGT corpus, with a focus on nouns and verbs and their subcategorization.

4.2.1 Noun and Verb Extraction

At the highest level of abstraction, lexical inference involves the definition of classes of words and the allocation of words to classes. In our system, the first pass classification of words involves parts of speech. The next level concerns inflection classes: which words

[7] This list comes from joint work with Olga Zamaraeva.

[8] We use the DELPH-IN convention for predicate symbols which includes a lemma followed by the part-of-speech (Flickinger et al., 2014a). For ease of evaluation in our current context, we use English glosses as the lemmas. For most applications, it is better to use lemmas from the language being modeled instead, as one cannot expect perfect word-level translational equivalence across languages.

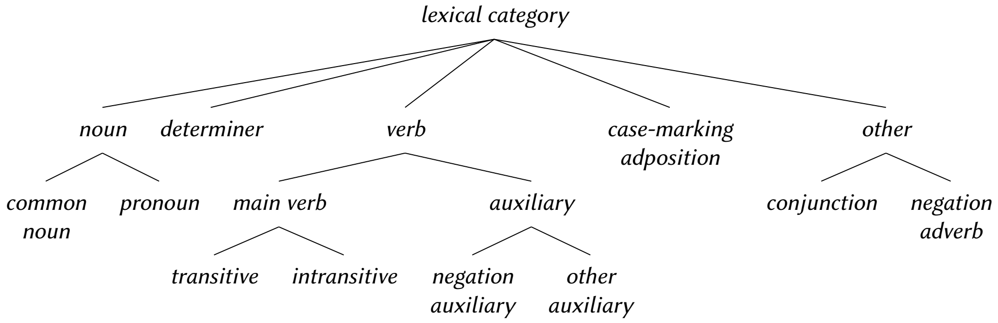


Figure 8: A taxonomy of the lexical categories that BASIL infers, organized according to the inference process

(within a part of speech) can be input to which lexical rules. To define these classes for nouns and verbs, we leverage the MOM morphological inference system. MOM identifies nouns and verbs based on their POS tags and uses a graph-based approach to identify and define inflection classes. (The morphotactic inference is further described in Section 4.3.)

4.2.2 Noun and Verb Subcategorization

In addition to defining lexical classes based on their morphotactic patterns, we must also group lexical entries based on their syntactic properties. In principle, this grouping can either be included in the input to MOM or performed on the output. Zamaraeva et al. (2019a) take the former approach to subcategorize verbs based on their valence properties by first inferring verbal case frame and including this information in MOM's input. MOM does not merge verbs with different valences, so the lexicon it produces includes separate classes for e.g. intransitive and transitive verbs, and those classes are further subcategorized based on their morphotactics.

To account for pronouns separately from common nouns and auxiliaries separately from verbs, we take the lexical classes in MOM's output and divide them based on their glosses: BASIL identifies nouns whose predication (in MOM's output) includes either an English pronoun or person, number, gender (PNG) or case features with no lemma and moves them into new lexical classes. BASIL constrains all common noun lexical classes to be third person, leaving number to the morphological analysis and inherent gender to future work (as shown in Figure 7 above). Pronoun lexical classes have more varied PNG and case values than common nouns, which BASIL accounts for by identifying any PNG and case glosses in MOM's output predication and specifying them as features on the pronoun's lexical entry.

Extracting auxiliaries from the verbal lexical classes and accounting for them in the grammar specification requires information regarding the auxiliary's syntactic distribution. For this reason, BASIL identifies auxiliaries from the source IGT rather than from MOM's lexicon, as we will describe in Section 4.5.1.

4.2.3 Additional Lexical Items

The Grammar Matrix does not support morphological inflection for determiners or adpositions, so it is not advantageous to infer these using MOM. Instead, BASIL extracts the full form orthographic representation and PNG and case features from the IGT. Where possible, we identify determiners from the POS tags, and if those are not available, BASIL looks for specific grams or lemmas in the gloss. Our grammars also support negation and coordination particles, which are described in their respective subsections of Section 4.5.

4.3 Morphotactics

The morphological component of a machine-readable grammar ultimately needs to account for which morphemes can co-occur and in which order, what the syntactic and semantic contributions of each morpheme are, and the morphophonological processes that relate the actual word forms to the collection of morphemes that make them up. The Grammar Matrix abstracts away from the morphophonology, assuming that the generated grammars will be interfaced with an external morphophonological analyzer (Bender and Good, 2005).[9] Accordingly, our inference system is only concerned with morpheme order, co-occurrence, and syntactico-semantic contributions.

The grammar specification files handle morpheme co-occurrence in terms of position classes (PCs), each of which specify what they can attach to (their 'input'),

[9] In brief, the idea is that morphophonological phenomena are best handled with different formal approaches than morpho-syntactic ones, so a parser using our grammars would be pipelined with bidirectional morphophonological analyzers. These latter map between surface realizations and morphophonolgically regularized sequences of morphemes, such as what is often found in the morpheme segmented line of IGT.

```
section=morphology
  noun-pc1_name=noun-pc1
  noun-pc1_order=suffix
  noun-pc1_inputs=noun1
     noun-pc1_lrt1_name=noun-pc1_lrt1
        noun-pc1_lrt1_feat1_name=case
        noun-pc1_lrt1_feat1_value=nom
        noun-pc1_lrt1_lri1_inflecting=yes
        noun-pc1_lrt1_lri1_orth=-pə
```

Figure 9: The definition of a position class for Lezgi

whether they are prefixes or suffixes, and which lexical rules they house. The lexical rules are defined in terms of lexical rule type (LRTs) which bear type constraints (feature/value pairs) and which in turn are instantiated by lexical rule instances (LRIs), which have specific affix spellings or are flagged as zero affixes (non-spelling-changing rules) (Goodman, 2013). An example of the specification for a position class in Lezgi [lez] is shown in Figure 9. Each PC must have at least one input (a lexical class or another PC) and a position (prefix or suffix)[10] and can be marked obligatory. Each PC must also have one or more LRTs, which can specify features on the word or on the arguments of the word. Each LRT must have one or more LRIs, which includes an orthographic form or a flag indicating that the rule involves no overt morpheme.

We use the MOM morphotactic inference system (Wax, 2014; Zamaraeva, 2016; Zamaraeva et al., 2017, 2019a) to infer the morphological rules. MOM infers a graph of the morphemes by collecting the affixes for each word with a noun or verb POS tag, creating a PC with an LRT which includes any features found in the gloss and an LRI with the appropriate orthographic representation and merging PCs that have overlapping inputs.[11]

While the morphotactic graph is essential for processing individual words, the morpho-syntactic or morpho-semantic features on those morphemes are key to producing the correct parse for larger phrases and sentences. MOM uses a feature dictionary comprising a large number of known glosses, grouped by their type, to map common grams to features. For example, the grams 'IPFV', 'IMPFV' and 'IMPERF' are all mapped to imperfective aspect. When MOM constructs the lexical rule types, it adds the features corresponding to any PNG, TAM or case grams to the lexical rule.

Non-inflecting lexical rules pose a particular challenge because they are not typically glossed as separate morphemes in IGT but rather indicated with a gram attached to the previous element with a ".", if they are indicated at all. MOM only creates non-inflecting rules for glosses it is able to map to PNG, case or TAM features, and only when such a gloss is found attached to the gloss for a stem. For example, if a noun is glossed as 'dog.NOM', MOM creates a non-inflecting lexical rule to add nominative case. All PCs which contain a non-inflecting LRI are made obligatory, so that forms without overt affixes do not end up only optionally bearing the features associated with that part of the paradigm.[12]

The result of morphological inference with MOM is a set of lexical rules grouped into position classes modeling their combinatorial potential. Within those position classes are lexical rule types that contribute features and in turn contain lexical rule instances, which either correspond to a particular orthography or are non-inflecting. Both the morphological rules in this section and the lexical entries in Section 4.2 contain morpho-syntactic features which interact with the syntactic inference in Section 4.5. The next section is concerned with how we define those features in the grammar specification, so that they will interact properly in the resulting grammars.

4.4 Syntactico-semantic Features

A great deal of semantic information is expressed morphologically in the form of person, number and gender (PNG) marking on nouns or agreement on verbs and tense, aspect and mood (TAM) inflection on verbs and auxiliaries. In order to model these features, the grammar specification must contain two types of definitions: First, the features and values themselves must be defined as belonging to the appropriate PNG or TAM category; and second, they must be associated with the appropriate lexical entries or morphological rules. The work of associating these features with the appropriate forms was described in Sections 4.2 and 4.3. When building the lexicon and morphological rules, MOM associates each feature value (e.g. perfective) with a type (e.g. aspect) according to their classifications in the GOLD Ontology (GOLD, 2010) and Unimorph (Sylak-Glassman et al., 2015). In this section we describe how BASIL uses these features and types to define more detailed type definitions for each PNG and TAM category, so the syntactic constraints contributed by these features can be used in the grammar and their semantic contributions will be reflected in the semantic representations.

[10] The Grammar Matrix does not handle circumfixes separately. These must be specified as individual prefixes and suffixes. Infixes are not explicitly handled; instead the Matrix assumes that a morphophonological analyzer regularizes these to prefixes or suffixes. See footnote 9.

[11] For more detail, see op cit.

[12] The addition of non-inflecting lexical rules to MOM, as well as the functionality of collecting the initial set of grams and adding features to lexical rules described in the preceding paragraph, is from unpublished work by Olga Zamaraeva.

4.4.1 Person

Generally speaking, person is a feature that marks the entities in an utterance with respect to discourse participants (Siewierska, 2004), where *first* is the speaker, *second* is the addressee and *third* is someone or something outside of the discourse context. Combinations of these persons, such as *first+second* ‘I and you’ and *first+third* ‘I and they’ are sometimes given special grammatical treatment and are often referred to as *inclusive* and *exclusive* (Cysouw, 2013). The Grammar Matrix’s library for person (Drellishak, 2009) provides a set of six options for person distinctions: first, second, third; first, second, third and fourth; first and non-first; second and non-second; third and non-third; and none. It also allows three options with regard to subtypes in the first person: none, inclusive vs. exclusive (along with the number categories in which this distinction applies) and other.

After collecting all of the person features from the lexical items and morphological rules, BASIL posits that the language contains first, second, third and fourth person if it found 4th person; first, second and third person if it found 3rd and either 1st or 2nd; and then first and non-first if it found 1st; second and non-second if it found 2nd; third and non-third if it found third; and otherwise none. BASIL then checks for inclusive and exclusive features and if it finds any, it defines an inclusive/exclusive distinction.

4.4.2 Number

Number indicates how many entities are being referred to. If a language marks number at all, this distinction can be as simple as singular vs. plural or may be more modular distinguishing dual (two), paucal (a few) and other numbers of entities (Corbett, 2000). The numbers distinguished by a language vary cross-linguistically and it is possible for these features to form a hierarchy (e.g. non-singular might subsume dual and plural). Thus, the Grammar Matrix allows number features to be freely added to the specification file, forming a hierarchy if desired (Drellishak, 2009). BASIL defines a number value for each of the numbers found in the morphology and lexicon. Currently, it defines each of these as sister types, rather than inferring a hierarchy of supertypes and subtypes, which we leave to future work.

4.4.3 Gender

Gender is another fairly open-ended category in the world’s languages. While some languages like Russian [rus] distinguish just masculine, feminine and neuter, Bantu languages such as Kiswahili [swh] distinguish a complex system of genders (Corbett, 1991). Linguists also vary in their annotation of gender features either using grams like M or MASC or using numerals for more complex systems. To accommodate this flexibility in the gender distinctions in language and linguists’ annotation preferences, the Grammar Matrix allows the addition of any number of genders by any name, and allows the specification of a hierarchy (e.g. to support agreement markers that are ambiguous between two or more gender values). As with number, BASIL defines a gender value for each of the genders found in the morphology and lexicon, but does not infer a hierarchy.

4.4.4 Tense, Aspect and Mood

Every language has some grammatical expression of time, which falls into the categories of tense, aspect and/or mood, and these features can be marked either morphologically on the verb, with an auxiliary or morphologically on an auxiliary, and a single utterance may include a combination of these expressions (Hopper, 1982).[13] For example, in the IGT from Matsigenka [mcb] in (3), the verb *oataira* is marked with regressive aspect (REG) and realis mood (REALIS), while the verb *oponiakara* is marked with perfective (PERF) aspect and realis mood (REALIS). Michael (2008) characterizes the regressive aspect as a subtype of perfective aspect that indicates motion back to a salient point of origin.

(3) ovashi oataira
ovashi o-a-t-a-i=ra
SO 3fS-go-EPC-REG-REALIS=SUB

oponiakara.
o-poni-ak-a=ra
3fS-come.from-PERF-REALIS.REFL=SUB

‘Then she went back to where she came from.’ [mcb] (Michael et al., 2013)

The TAM categories contain a number of possible values cross-linguistically and, as illustrated by the regressive and perfective aspects described by Michael, can form hierarchies. As with the number and gender libraries, the TAM library of the Grammar Matrix (Poulson, 2011) also allows the definition of any number of values for each of tense, aspect and mood and also allows the definition of hierarchies. BASIL defines each TAM feature as either tense, aspect or mood in the respective section of the grammar specification, leaving the inference of hierarchies to future work.

4.4.5 Summary

We described six categories of syntactico-semantic features: person, number, gender, tense, aspect and mood. These features are added to the specifications of lexical

[13] In NLP, the TimeML specification language (Pustejovsky et al., 2003) has been used in an effort to standardize such expressions of time, and has been made more cross-linguistically viable by efforts such as Zymla 2017.

entries or morphological rules according to the methodologies described in Sections 4.2 and 4.3 and defined as belonging to their respective categories. The result of these definitions is a grammar that produces semantic representations that contain this information and enforces agreement between heads and their arguments.

4.5 Syntactic Properties

In this section, we provide a high-level description of the algorithms used for inferring each of the syntactic phenomena accounted for in our grammars. Using the projected dependency tags provided by INTENT and typologically-informed heuristics, we make generalizations about distributional properties of the language and posit the appropriate definitions for that grammar specification for a range of syntactic phenomena. These include broad-brush, language-level properties (e.g. 'the case alignment is ergative-absolutive'), properties associated with specific constructions (e.g. 'this form can coordinate VPs in a monosyndetic pattern') and specific lexical items (e.g. 'negation is marked via an auxiliary with this orthography that combines with a VP and raises the subject').

4.5.1 Word Order and Auxiliaries

Languages vary in both their degree of word-order flexibility and, if only specific orders are allowed, which ones are (e.g. Dryer, 2013c). When linguists talk about the 'word order' of a language, they are frequently referring to the relative order of a verb and its arguments (subject, complement), but there are also cross-linguistic differences in the order of determiners (if present) with respect to their head nouns, adpositions with respect to NPs, and others. The 'word order' section of a Grammar Matrix grammar specification takes information about each of these (Bender et al., 2010).

We adopt the approach of Bender et al. (2013), which maps constituent word orders observed in the data to one of ten canonical word orders (SOV, SVO, OSV, OVS, VSO, VOS, v-initial, v-final, v2 and free). This approach identifies verbs based on their POS tags and their subjects and objects using projected dependency labels. Each observed order of verbs and subjects, verbs and objects and subjects and objects is counted to compute a three dimensional vector representing the respective order of verbs, subjects and objects in the language, which can be compared to the vector representations for each canonical word order. Following Bender et al., BASIL posits the canonical word order whose vector has the shortest euclidean distance from the observed language vector as the canonical word order for the language.

Also following Bender et al. (2013), we take a simpler approach to predict determiner-noun order. Collecting each noun and determiner pair from the projected dependencies, we count the number of observed determiners before vs. after the noun and posit whichever order is most common.

Whereas previous work did not account for auxiliaries, BASIL both identifies auxiliaries as lexical items and infers their syntactic properties. This includes identifying their position with respect to the main verb and inferring what type of constituent they attach to (a verb (V), verb phrase (VP) or sentence (S)), whether they attach before or after that constituent, and whether multiple auxiliaries are possible. We identify auxiliaries in the corpus as words that are either glossed with an English auxiliary or modal or glossed with only morphosyntactic or morpho-semantic features and no lemma. While collecting auxiliaries from the corpus we identify the main verb and its subject and object from the projected dependencies. We use these to discover whether the auxiliary occurs before or after the main verb and check for a subject intervening between an auxiliary and verb, which would indicate that the auxiliary takes an S complement instead of a VP, or an auxiliary intervening between a verb and its object, which would indicate that the auxiliary attaches to a V, rather than a VP. If no evidence for V or S attachment is found, BASIL defaults to VP attachment, as the argument-composition analysis that the Grammar Matrix uses to model auxiliaries with V complements is computationally very expensive (see Bender 2010) and we hypothesize that S attaching auxiliaries are typologically rare.

Because the MOM morphotactic inference system infers auxiliaries as verbs when constructing the lexicon, BASIL must reclassify these lexical items to give them the proper definitions to function as auxiliaries in the grammar. BASIL does this by finding any verbs in the MOM-generated lexicon that have the same lemma as those it identified as auxiliaries. For each, BASIL defines an auxiliary lexical class that is input to the same morphological position classes and contains the same features as the verb lexical class inferred by MOM. Because auxiliaries are often homophonous with main verbs, BASIL does not remove the main verb lexical entry.

In addition to the lemma, feature and morphological combinatorial information described above, the Grammar Matrix requires specifications for the semantic contribution of the auxiliary. When BASIL constructs the auxiliary lexical items from verb lexical items inferred by MOM, it specifies the auxiliary as semantically contentful and adds the predication value from the verb if the original verb's predication contains an English lemma (e.g. _should_v_rel), rather than containing only grams for syntactico-semantic features. BASIL also adds a negation predication if the auxiliary contributes negation (see Section 4.5.4 for negation inference).

Finally, the lexical entry includes a value for the case of its subject, which can be specified as a specific case, no case restrictions, or the case assigned by the verbal complement. With our development languages, we tested an algorithm in which BASIL checks for differences in the case on subjects in sentences with and without auxiliaries, and adds this constraint to the lexicon. We found that this inference is frequently confounded by other factors that can affect the subject's case, so we did not include this inference in BASIL and leave a more accurate algorithm to future work. Currently BASIL posits no case restrictions if A) the language does not have a case system or B) the auxiliary always occurs with a different case than the one inferred for the verb's case frame (this leads to some ambiguity, but avoids the loss in coverage that results from positing a case that was assigned due to other syntactic factors). Otherwise it posits that the auxiliary takes its case restrictions from the main verb.

After identifying the auxiliaries in the corpus, we allow for a post-hoc change to the main word order to account for second position clitic clusters. The Grammar Matrix supports an analysis set forth by Bender (2008c) of second position clitics/clitic clusters as auxiliaries in a V2 language, when those clitics express TAM and/or agreement features. Clitic clusters that contain PNG agreement and TAM information are identified during auxiliary inference and if they occur overwhelmingly as the second word of each sentence, BASIL posits V2 word order for the language to leverage this analysis.

4.5.2 Case System and Case Frame

A language which marks case has variations in the forms of the noun phrases correlated with their function in the sentence (Comrie, 1989; Dixon, 1994). A typical case system will involve both the case required of core arguments of typical verbs, as well as additional cases used when NPs function as modifiers (e.g. locative case) and sometimes selected for idiosyncratically by specific verbs. Case systems are differentiated according to the alignment they provide for the core arguments of intransitive and transitive verbs. The Grammar Matrix customization system's case library (Drellishak, 2009) provides nine overarching case systems (core argument case alignments) and facilitates defining any number of additional cases. The selection of the core case system enables default case frames for each verb type, but grammar specifications can also bypass these and define verb types which leave case underspecified or select for alternate case patterns.

To infer the overarching case system, we use an algorithm developed by Bender et al. (2013) and reimplemented to use an enriched Xigt corpus by Howell et al. (2017), which uses a simple heuristic based on the total counts of known case grams in the data. This approach only infers four case systems: nominative-accusative, ergative-absolutive, split-ergative and none. Because split-ergative requires information about the nature of the split, we map it to ergative-absolutive. In addition to inferring the overarching case system, we also collect any other case grams in the corpus and define these in the grammar specification, so that we can also handle verbs that require alternate case frames. Here we infer only intransitive and transitive verbs, leaving ditransitive (which are not currently supported by the Grammar Matrix) and clausal complement-taking verbs to future work.

To find the case frame of each intransitive and transitive verb in the corpus, BASIL uses the dependency parse of the English sentence to identify verbs that have zero or one direct object, skipping any that are passive or have an indirect object or clausal complement (following Zamaraeva et al. (2019a), such verbs will be excluded from the final grammar). We find the case of the subject and object in the gloss line and if no case gram is found in the gloss, we posit default case based on the overarching case system. In cases where the marked case doesn't match the default, we posit the attested case for that verb's arguments. Our approach is similar to that of Zamaraeva et al. (2019a), but differs in that we use projected dependency parses rather than phrase structure trees and that we account for verbal case frames that differ from the overarching system.

These constraints interact with the case features on noun-phrases when verbs unify with their arguments. Case features may be licensed by the morphological rules on nouns which were inferred by the morphological component described in Section 4.3, can be lexically specified (e.g. for pronouns, see Section 4.2.2) or can be indicated by the determiner or a case-marking adposition. If, for example, the feature specification [CASE acc] is associated with a lexical rule attaching an accusative case marker to a noun, or if [CASE acc] is in the lexical entry for a determiner or adposition, NPs or PPs built with these lexical entries or rules will be incompatible with argument positions that require [CASE nom].

Having described the inference algorithms and systems for phenomena such as morphotactics, word order and case, and the ways in which we refined, adapted and added to them, we now turn to the entirely new inference modules that we contribute in this paper, beginning with argument optionality.

4.5.3 Argument Optionality and Marking of Arguments on Verbs

Languages vary in the extent to which and under what conditions they allow dropped arguments: some languages allow core arguments of any verb to be dropped freely, while others are more restrictive if argument dropping is possible at all. These restrictions range from the specific verbs for which argument dropping is allowed, subject vs. non-subject arguments, specific syntactic contexts (e.g. only in certain tenses), or whether the verb is required to agree with overt vs. dropped arguments (Ackema et al., 2006; Dryer, 2013a). The Matsigenka example in (4) shows a verb with no overt arguments that is inflected for agreement with both the subject and object.[14]

(4) oogaigavakari
o-og-a-ig-av-ak-a=ri
3FS-eat-EPV-PL-TRNS-PERF-REALIS.REFL=3MO
'She ate them.' [mcb] (adapted from Michael et al., 2013)

The Grammar Matrix accounts for subject and object dropping as either lexically licensed (allowed for certain verbs) or possible for any verb (Saleem, 2010; Saleem and Bender, 2010). It also allows argument dropping to be constrained by agreement markers on the verb which can be optional, required or not allowed when the subject/object is overt, and similarly when the subject/object is dropped. Finally, specific syntactic contexts in which subject dropping is possible can be defined. Our inference focuses on determining whether argument dropping is permitted for subjects and objects in a language and leaves constraints on the context to future work. We infer whether agreement is required for dropped vs. overt arguments, which requires differentiating subject agreement markers and object agreement markers; however, we leave the integration of this inference with the morphological rules that license agreement to future work.

In order to identify whether subject and/or object dropping is possible in the language, BASIL begins by collecting all of the transitive and intransitive verbs[15] in the corpus together with their overt arguments, based on the projected dependencies as it did for case-frame inference (§4.5.2). Whereas the case-frame inference methodology determines if a verb is transitive based solely on the presence of an overt object in the English translation, here we account for the fact that some English verbs allow object dropping. If the corresponding verb in the English translation has a direct object, we assume that the verb is transitive. If no object is found, BASIL cross-references the verb's gloss with a list of English object-dropping verbs from the lexical entries in the English Resource Grammar (ERG v. 1214; Flickinger, 2000, 2011) of the type v_np*. If the verb is found in this list, BASIL posits that the verb is transitive and otherwise intransitive. Although the argument optionality of verbs does not necessarily map across languages, leveraging this list of English object-dropping verbs allows us to err on the side of positing transitivity, and we find that doing so improves the coverage of the resulting grammars.

Agreement with the subject or object can be marked either on the main verb or on an auxiliary. To determine whether a verbal complex has subject and/or object marking, BASIL identifies any auxiliaries associated with each verb and collects all agreement markers (across the verb and any auxiliaries), using a hand-compiled list of common agreement glosses. We compiled this list from the agreement glosses used by MapGloss (Lockwood, 2016) as well as observed glosses in the development data. Although agreement is not the only way arguments are marked on verbs (for example, in Hausa the verb's inflected form depends on whether or not an overt object is present, but this form does not include any PNG information (Newman, 2000)), it is the most common form and the easiest to identify. In addition to collecting all agreement markers, we use a heuristic to identify whether the agreement markers correspond to more than one argument: if the set of agreement glosses has multiple glosses of a particular category (e.g. person, number or gender), BASIL says that the verb is marked for more than one argument. This approach is particularly valuable when a single morpheme is used to mark two arguments. For example in (5) from Basque [eus], *dio* is glossed as 3ABS-3DAT.3ERG, containing three third person glosses, so BASIL counts three agreement glosses on that verb.

(5)

Eduk	neska	Toniri	aipatu	dio
Edu-k	neska	Toni-ri	aipatu	d-io
Edu-ERG	girl.ABS	Toni-DAT	mention	3ABS-3DAT.3ERG

'Edu has mentioned the girl to Toni.' [eus] (adapted from Xia et al., 2016)

We use the presence of agreement features on any verb in the set to detect argument marking on the main verb. Intransitive verbs with any agreement gloss are classified as having subject marking. The orthographies associated with these glosses are saved in a set of known subject markers. After all of the subject markers on intransitive verbs have been collected, BASIL looks at the transitive verbs. Transitive verbs with more than one agreement gloss (like that in (5)) are classified as having subject and object marking. Transitive verbs with only one agreement gloss which corresponds to the orthography of a known subject marker are classified as having subject marking and the remainder are classified as having object agreement. The set of known

[14] We analyze the pronominal clitics in Matsigenka as affixes, rather than independent words, following Inman (2015).

[15] Because BASIL does not infer ditransitive or clausal complement-taking verbs, it excludes them from consideration when inferring argument dropping.

subject glosses is included in the input to MOM. When deciding if a PNG gram should be identified with the subject or object, MOM consults this list and associates it with the subject if the verb is intransitive or the morpheme is in the set of subject morphemes and with the object otherwise.

BASIL's inference for argument optionality has two components: (1) inferring whether subjects and objects can be dropped, and (2) inferring whether argument marking on the verb is possible or even required when arguments are dropped or overt. The latter involves identifying argument markers in the form of agreement morphemes and discriminating between subject and object agreement markers. Our approach focuses on increasing the coverage of the inferred grammars, while future work to enforce or prohibit argument marking on verbs with overt versus dropped arguments would decrease ambiguity.

4.5.4 Sentential Negation

All human languages have a means of expressing sentential negation, but they vary in how many markers are used and whether those markers are independent words, bound morphemes (Östen Dahl, 1979; Dryer, 2005, 2013b; Miestamo, 2008) or a missing morpheme in the paradigm, such as the absence of a tense marker indicating negation in some south Dravidian languages (Master, 1946). Crowgey (2012) models sentential negation in the Grammar Matrix, allowing it to be marked with 0, 1 or 2 morphemes (calling these strategies *zero, simple* and *bipartite*), which can be bound morphemes, syntactic heads (auxiliaries) or uninflected particles (adverbs). The analyses provided by the Grammar Matrix ensure that there is only one negation predication in the semantics, regardless of the number and type of markers in the strategy. BASIL infers each of the possible combinations as described below.

We first identify sentences with sentential negation based on the English translation and then target the gloss line of the IGT to find negation morphemes, based on common glosses, such as 'NEG' and 'not'. BASIL considers glosses on affixes to be inflectional negation. We expect that zero-marked negation will be annotated with a negation gloss on a stem or on another morpheme and will therefore be modeled with a non-inflecting lexical rule as described in 4.3, so BASIL accounts for it using the morphological negation specification. If inflectional negation is detected, this is indicated in the sentential negation portion of the grammar specification which in turn enables a negation pseudo-feature which can be added to lexical rules. The distributional properties for negation affixes (including zero-negation) are inferred and specified by the morphological inference system in Section 4.3, which puts the negation pseudo-feature on the appropriate lexical rule. The Grammar Matrix customization system interprets this pseudo-feature and ensures that the resulting lexical rules carry negation semantics, as shown in Figure 6.

A root glossed as negation could be either an auxiliary or an adverb. The English dependency parse does not help us decide which, as it simply encodes facts about negation in English. Instead, we compare these negation words with the auxiliaries collected in Section 4.5.1. If auxiliary entries were inferred for orthographies glossed for negation, we treat them as such. Otherwise we define them as adverbs. The distributional properties of negation auxiliaries were inferred as part of auxiliary inference (§4.5.1), so there is no additional work to be done. In the case of negation adverbs, we use the same process as we did for auxiliaries to decide what type of constituent they attach to (VP or S) and whether they occur before of after that constituent.

After identifying instances of sentential negation in the corpus, BASIL compares the number of sentences that include one negation marker with those that include more than one negation marker. Although BASIL only looks at sentences with sentential negation, it does not distinguish between sentential and constituent negation markers, and can mistake a negated sentence with additional constituent negation as bipartite negation. However, we seek to avoid confounding from constituent negation co-occurring with sentential negation by taking the most common strategy (simple or bipartite) found in the corpus.

If simple negation is the most common, the Grammar Matrix lets us add all of the strategies we found (affix, auxiliary, and adverb) to the grammar specification. For bipartite negation, we can only specify one combination of markers, so if bipartite negation was the most common strategy found in the corpus, we add the two most common co-occuring types of negation markers (e.g. adverb and affix) to the grammar specification. While the Matrix only allows us to add one orthography for a negation adverb (so we use the most common), we are able to specify as many negation affixes and auxiliaries as we find in the corpus.

4.5.5 Coordination

Coordination is possible for a wide range of constituent types, called coordinands, and can be marked with either free or bound morphemes, called coordinators. Coordinators can attach to all (omnisyndetic), all but one (polysyndetic), one (monosyndetic) or none (asyndetic) of the coordinands (Drellishak, 2004; Haspelmath, 2007). The Grammar Matrix models all of these possibilities and allows us to define any number of strategies for nouns, noun phrases, verbs, verb phrases and sentences (Drellishak and Bender, 2005).

As with sentential negation, BASIL identifies IGT that exhibit coordination based on the English translation and then finds the coordinators first by looking for the word aligned by INTENT with the English coordinator and then, because alignment isn't always successful, by looking for the glosses 'COORD', 'CONJ', 'CCONJ' and 'and'. Then BASIL uses the projected dependencies to collect the dependent of each coordinator and these dependents are assumed to be the coordinands. As a fallback, if BASIL cannot find coordinands via projected dependencies, it looks for them by collecting the words that occur in between coordinators, although this approach is less successful for monosyndetic coordination. BASIL then compares the number of coordinators and coordinands to decide if the sentence exemplifies asyndetic, monosyndetic or omnisyndetic coordination. Differentiating between mono- and polysyndetic coordination is rather difficult as most examples in the corpora only have two coordinands, and the construction 'A and B' could be either mono- or polysyndetic. However, monosyndetic coordination can be used to model polysyndetic (e.g. [[A and B] and C]), so BASIL defaults to monosyndetic in cases that might be mono- or polysyndetic.

For each coordination strategy, we also identify the lexical category of the coordinand (noun or verb) and use heuristics to decide at what level the coordination takes place (word or phrase in the case of nouns and word, phrase or sentence for verbs). Because the Grammar Matrix allows any number of coordination strategies, we add each distinct coordination strategy that we detect in the corpus to the grammar specification.

4.6 Summary

In this section we described four types of inference that produce the necessary components of our inferred grammar specifications: lexical, morphotactic, morpho-syntactic/morpho-semantic and syntactic. For inference of noun and verb lexical classes and lexical entries, we rely primarily on the MOM morphotactic inference system, but make new contributions to lexical inference in the form of auxiliary, adposition and determiner inference as well as lexical types defined as part of syntactic inference such as negation adverbs or coordinators. We also leverage MOM to infer morphological rules for nouns and verbs, and build on the system by improving the detection of subject and object agreement, as described in Section 4.5.3, and adding the definitions of PNG and TAM features to the grammar specification, so that these syntactico-semantic features can be included in the semantic representations. We built on previous algorithms for inferring syntactic properties such as word order and case and added new algorithms for argument optionality, negation and coordination.

The scope of this inference spans a large number of feature-value pairs in the grammar specification, as we illustrate in Table 1, and testing the inference for all of these on real data would require a vast set of datasets from typologically diverse languages. At the same time, it is possible that specifications allowed by the Grammar Matrix or targeted by BASIL are not sufficient to correctly model some languages. In the following section, we describe our data-driven approach to development in which we considered corpora from a wide range of diverse languages and from a variety of data formats to develop and test the algorithms detailed in this section.

5 Development Languages

We developed the inference algorithms described in Section 4 using a data-driven approach in which we consulted the typological literature for each phenomenon and actively tested each algorithm on a diverse set of languages throughout implementation. In this section, we describe the languages and datasets we used during development (§5.1), phenomena that appear in our datasets, both targeted by BASIL and otherwise (§5.2) and BASIL's performance on the development datasets (§5.3).

5.1 Dev Languages and Datasets

In order to thoroughly test BASIL on the phenomena described in Section 4, it is necessary to use languages that are typologically varied, representing as many language families and geographic areas as possible. For development, we made use of 9 datasets for languages from 7 language families and 4 continents. In addition to these core development datasets, we tested individual phenomena using datasets from another 18 languages to span a total of 19 language families and 6 continents. These languages, their language families and details of the corpora are listed in Table 2. Their geographic distribution is shown in Figure 10, with development languages in red (1-9) and additional consulted languages in blue (10-27).[16] Held-out languages which we discuss in Section 6.3 are in green (28-32).

We selected the core development languages based on the size and quality of the dataset as well as for some of the syntactic phenomena exhibited by those languages. The majority of these corpora come from a FLEx or Toolbox corpus that was curated by a documentary linguist (or a group of linguists). To support the development and implementation of inference for specific syntactic and morpho-syntactic phenomena, we also consulted additional datasets for languages which represent those phenomena. These datasets not only contribute to the diversity of the languages we worked

[16] In most cases, these coordinates come from WALS (Dryer and Haspelmath, 2013). If information from WALS was not available, we consulted other sources, starting with descriptions of where the languages are spoken from the reference grammars we worked with.

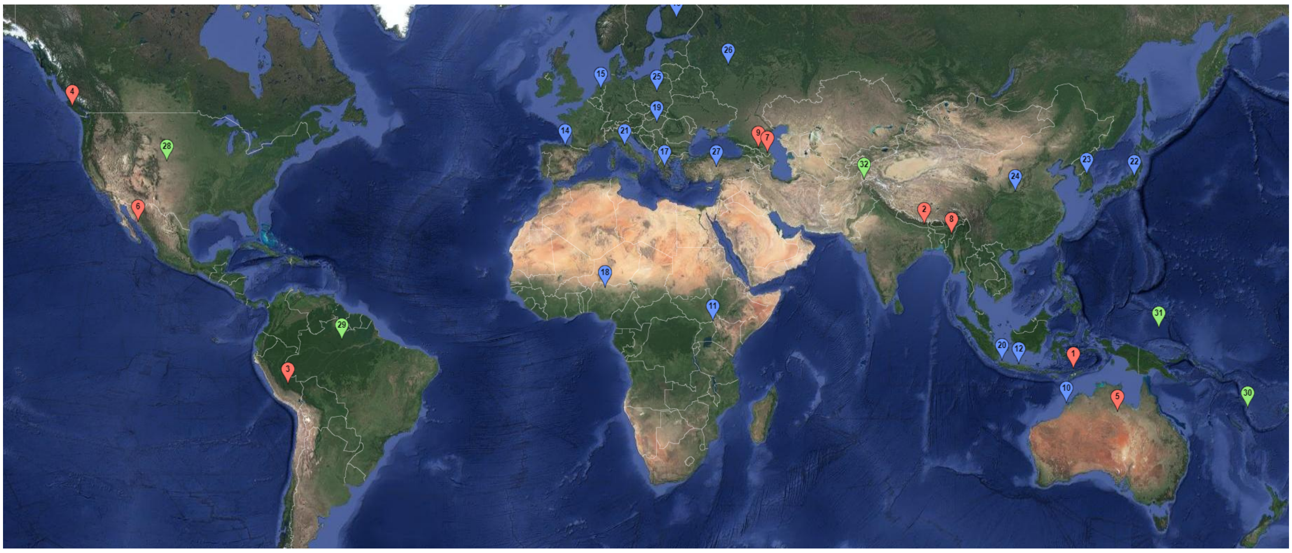


Figure 10: Map of the coordinates where languages used in the development are spoken

	Language	ISO 639-3	Family	Source Type	Number of IGT	POS tags in source
	Development					
1	Abui	abz	Trans-New Guinea	Toolbox	1568	yes
2	Chintang	ctn	Sino-Tibetan	Toolbox	9785	yes
3	Matsigenka	mcb	Arawakan	FLEx	349	yes
4	Nuuchahnulth	nuk	Wakashan	FLEx	641	no
5	Wambaya	wmb	Mirndi	Book	818	no
6	Haiki	yaq	Uto-Aztecan	FLEx	2235	yes
7	Lezgi	lez	Nakh-Daghestanian	FLEx	1168	yes
8	Meithei	mni	Sino-Tibetan	FLEx	955	yes
9	Tsova-Tush	bbl	Nakh-Daghestanian	FLEx	1601	yes
	Consulted					
10	Bardi	bcj	Nyulnyulan	Book	178	no
11	Ik	ikx	Eastern Sudanic	Book	201	no
12	Old Javanese	jav	Austronesian	Toolbox	308	no
13	Yup'ik	esu	Eskimo-Aleut	Book	217	no
14	Basque	eus	Basque	ODIN	1033	no
15	Dutch	nld	Indo-European	ODIN	3543	no
16	Finnish	fin	Uralic	ODIN	3123	no
17	Greek	ell	Indo-European	ODIN	2065	no
18	Hausa	hau	Afro-Asiatic	ODIN	2504	no
19	Hungarian	hun	Uralic	ODIN	2077	no
20	Indonesian	ind	Austronesian	ODIN	1699	no
21	Italian	ita	Indo-European	ODIN	3513	no
22	Japanese	jpn	Japonic	ODIN	6655	no
				Book	116	no
23	Korean	kor	Korean	ODIN	5383	no
24	Mandarin	cmn	Sino-Tibetan	ODIN	5045	no
25	Polish	pol	Indo-European	ODIN	2691	no
26	Russian	rus	Indo-European	ODIN	4161	no
27	Turkish	tur	Altaic	ODIN	2617	no

Table 2: Languages used in development

with, but also to the variety of source formats and dataset styles. A number of the datasets we consulted for individual phenomena (languages 14-27) come from the ODIN corpus (Xia et al., 2016), which is a collection of IGT scraped from academic papers. We also extracted four corpora from descriptive grammars, using the pipeline for extracting IGT from text and converting it to the Xigt data model developed by Xia et al. (2016). A full list of citations for the corpora and any descriptive resources we consulted are in Appendix C.

Later in this section, we describe BASIL's coverage over the development datasets. To contextualize that discussion, we begin with an overview of the languages and their respective datasets.

Abui [abz] is an Alor-Pantar language in the Trans-New Guinea language family. It has about 16,000 speakers and is primarily spoken on the Alor island of Indonesia (Kratochvíl, 2007). This dataset (Kratochvíl, 2019) comes from a Toolbox corpus which contains about 18,000 sentences from both elicitation and transcribed speech. As part of an ongoing documentation effort, the dataset is only partially glossed. We filtered the data based on the presence of full segmentation and glossing, and removed duplicates and examples marked as ungrammatical, to create a dataset of 1,500 sentences.

Chintang [ctn] is a Kiranti language of the Sino-Tibetan family spoken in Nepal with 4,000-5,000 speakers (Schikowski, 2013). The Toolbox dataset is quite large, coming from a long-term documentation effort (Bickel et al., 2013b). We use a fully segmented and glossed subset of the data containing almost 10,000 sentences. The type of language represented in the corpus is diverse, containing transcribed conversations, ritual language, narratives and a few other genres.

Haiki [yaq] is a Taracahitic language of the Uto-Aztecan family and is spoken by about 21,000 people in Mexico and the United States (Eberhard et al., 2019). There are multiple spellings of the name of this language, including Yaqui, which is the official name of the tribe in the United States and Mexico; however, Haiki is the correct spelling in the Pascua Yaqui orthography (Sanchez et al., 2015). The corpus (Harley, 2019) is quite large with almost 11,000 IGT, but as with most ongoing projects, is only partially annotated with interlinear glosses and part-of-speech tags. After filtering IGT with no glosses and removing ungrammatical examples and duplicates, we worked with a set of just over 2,000 IGT.

Lezgi [lez] belongs to the Lezgian subgroup of the Nakh-Daghestanian language family (Donet, 2014a). It is spoken by about 400,000 people (Eberhard et al., 2019), primarily in Daghestan and Azerbaijan (Donet, 2014a). The glossing and POS tagging in this corpus (Donet, 2014b) are fairly complete, resulting in a set of over 1,100 IGT after minor filtering and removing ungrammatical examples and duplicates.

Matsigenka [mcb] is a Maipurean language of the Arawakan family spoken in Peru by about 10,000 people (O'Hagan, 2018). The FLEx corpus (Michael et al., 2013) is made up of narratives that are fully segmented and glossed. Of the approximately 5,000 IGT in the corpus, some have English translations, while the vast majority of the translations are in Spanish. BASIL relies on computational resources for English, both through its dependency on the INTENT (Georgi, 2016) system (which parses the English translation of an IGT and projects the dependency parses onto the language) and through the list of English verbs referenced in Section 4.5.3, and thus BASIL requires IGT with English translations. From the full Matsigenka corpus, we[17] identified about 350 IGT with English translations.

Meithei [mni] is a Kuki-Chin-Naga language of the Sino-Tibetan language family. It is spoken predominantly in Manipur State, but has about 56 million speakers living across a wide region, including in China, India, Nepal and Myanmar (Chelliah, 2011). The FLEx corpus (Chelliah, 2019) contains about 1,800 IGT, but as part of an ongoing documentation effort, is only partially annotated. After filtering for fully-glossed IGT and removing duplicates and ungrammatical examples, the corpus has about 1,000 items. Compared to other corpora in our development set, this corpus contains a high proportion of complex sentences, which include subordinate clauses that are not covered by inference. Nevertheless, it is a strong example for how much typological information can be learned from a corpus, even when many of the sentences contain phenomena that are beyond the scope of the inference system.

Nuuchahnulth [nuk] is Southern Wakashan language of Vancouver Island in Canada and has only about 130 fluent speakers (Eberhard et al., 2019). The FLEx dataset (Inman, 2019b) was curated in connection with a dissertation on multi-predicate constructions and contains both transcribed narratives and elicitations, many of which target this construction. The dataset includes about 650 examples which are fully glossed and segmented. Inman's corpus does not include POS tags, which are required by MOM to build the lexicon of nouns and verbs. For many IGT, these are available from the projected part of speech tags from INTENT. However, because INTENT does not always successfully find an alignment (this can be particularly challenging for polysynthetic languages), we use an additional heuristic to identify verbs. Because single-word sentences are very common in this poly-synthetic language, we supplemented the projected POS tags by pre-

[17] Most of these were identified by previous research assistants on the AGGREGATION project and more were extracted by Angelina McMillan-Major.

processing the corpus to assign a verbal POS tag to the only word in any one-word IGT if the dependency parse for the translation was headed by a verb.

Wambaya [wmb] is a West Barkly language in the Mirndi family, which has about 60 speakers (Eberhard et al., 2019). The Wambaya dataset is distinct from our other development datasets as it was extracted from the examples in a descriptive grammar (Nordlinger, 1998). As such, it does not contain linguist-provided POS tags and the possibility of alignment errors in the interlinearization is higher, due to the process of extracting IGT from text. Nevertheless, this language illustrates a number of phenomena that guided our development and the use of a descriptive grammar allows us to explore the possibility of inferring grammars to accompany descriptive resources along the lines of Bouma et al. 2015.

Tsova-Tush [bbl], also referred to by the endonym Bats or Batsbi, is a Northeast Caucasian language of the Nakh subgroup of the Nakh-Daghestanian language family (Hauk and Harris, forthcoming). It is spoken in Georgia by about 2,500-3,200 people (*ibid.*). The corpus (Hauk, 2016–2019) contains elicitation and transcribed text and the glossing and part of speech tags are almost complete, including over 1,600 IGT after removing ungrammatical examples and duplicates.

5.2 Dev Language Phenomena

In this section we quantify the degree to which the inference system was tested by the development languages described above. In Section 4.1, we described the space of the inference task in terms of the number of features and values that BASIL is designed to add to the grammar specification to account for the phenomena it handles. We identified 50 features with a fixed set of values (listed in Table 1) totaling 136 possible values in the Grammar Matrix grammar specifications that are relevant to the phenomena targeted by BASIL. Our system is designed to infer 99 of those 136 values. When inferring grammar specifications for the 9 development languages, 37 of the 50 features and 71 of the 99 values were inferred by BASIL from the development data, as detailed in Table 3. We also reported in Section 4.1 that BASIL can identify 116 morpho-syntactic and morpho-semantic features from their glosses in the IGT. 66 of those 116 features are found in the development datasets (see Table 4).

While the development languages test a significant portion of the phenomena targeted by BASIL, they do not exhaustively test every facet. For this reason, we consulted an additional 18 languages (represented in blue in Figure 10) to test as many of the feature-value pairs as possible, in order to create a system that would generalize beyond the development languages.

The phenomena targeted by BASIL (§4) are only a subset of the phenomena necessary to fully model a language or to parse all of the sentences in the corpora. For this reason, understanding the types of sentences we do not expect to parse lays the groundwork for understanding what the inferred grammars should parse, but don't. A number of lexical types that BASIL does not infer will prevent the grammar from having lexical coverage over sentences that contain those types of words. These include but are not limited to adjectives, adverbs and 'particles' marking complementation, subordination, information structure, questions and possession. Because these words may be homophonous with words that BASIL does handle, sentences with these lexical types may have lexical coverage and the grammar might even produce one or more parses for them, but those parses will not be correct. In addition, there are phenomena whose analysis doesn't depend on particular lexical items, but rather phrase structure rules for specific configurations (e.g. asyndetic coordination) or lexical rules for particular types of inflection (e.g. imperatives), or both in combination (e.g. adverbial clauses where subordination is marked morphologically). If the inferred grammars don't cover a phenomenon, we don't expect the grammars to parse sentences including that phenomenon (correctly, or at all).

Some parses have the correct predicate-argument structure but lack some semantic features as a result of out-of-scope syntactic phenomena that contribute information to the semantic structure. As an example, yes/no questions and imperatives are traditionally modeled in the DELPH-IN formalism with the SF (sentential force) feature, which can have the values prop (proposition), ques (question) or comm (command) (Flickinger et al., 2014b). The inferred grammars for some languages parse questions and imperatives with the correct predicate-argument structure, but they do not use the appropriate prop or comm, so the correct features are not fully specified. With this context established, the next subsection presents the performance of the development grammars.

5.3 Coverage for Dev Languages

We evaluated system performance on the development languages using 10-fold cross validation. We assessed the inferred grammars by parsing sentences in their respective test folds, using five metrics: *lexical coverage* — the proportion of sentences for which the grammar has an analysis for each word; *parse coverage* — the proportion of sentences for which the grammar can produce a syntactic analysis; *correct predicate-argument structure* — the proportion of sentences the grammar parses, producing a semantic representation that includes appropriate predications and arguments for each semantic entity; *correct predicate-argument structure and semantic features* — the proportion of sentences for which

Phenomenon	# possible	# targeted by inference	# inferred from dev languages
noun lexical entry	4	2	2
verb lexical entry	4	2	2
auxiliary lexical entry	6	4	4
adposition lexical entry	3	3	3
morphological rule	5	5	5
person	9	8	4
tense	2	1	1
word order	10	9	6
determiner order	4	4	4
auxiliary order	9	9	7
case system	9	3	2
argument optionality	18	15	12
sentential negation	41	23	9
coordination	12	11	10
total	136	99	71

Table 3: The number of possible values for the closed set features to define phenomena in the grammar specification and, those targeted by the inference system and those attested in the development languages

Feature Category	# Found
Number	4
Gender	5
Case	21
Tense	6
Aspect	16
Mood	14
Total	66

Table 4: The number of morpho-syntactic features found in the development languages. (Person features are not included because the Grammar Matrix defines them automatically based on the overarching person system.)

the grammar produces the correct predicate-argument structure as well as the appropriate PNG and TAM features on those arguments and the correct sentential force; and *ambiguity* — the average number of results per sentence that parses. For details on how we operationalized these metrics, see Section 6.

Table 5 presents the results using these metrics for each of the development languages. Whereas calculating the lexical coverage, parse coverage and ambiguity are automated processes, calculating the correct predicate-argument structure and features requires manual inspection of the semantic representations (for a detailed description of these processes, see §6.1). For this reason, we provide results for correct predicate-argument structure and correct predicate-argument structure and features across all folds for languages with less than 1,000 IGT, but for those with more IGT, we provide these metrics only for the first fold.

The sentences for which the grammar produces a semantic representation with the correct predicate-argument structure and features are a subset of those for which the grammar produces a semantic representation with the correct predicate-argument structure. In turn, those are a subset of the sentences with parse coverage, which are a subset of those with lexical coverage. This is illustrated by the bar graph in Figure 11.

To contextualize this performance, remember that the datasets come from a wide range of sources. Transcribed speech and elicitations often include sentence fragments, which the grammar will not accept as sentences. For this reason, and because of the many out-of-scope phenomena described above, we do not expect the inferred grammars to parse a very large portion of the held-out sentences they are tested on. Instead, the most useful comparison to consider is the number of sentences that parsed with the correct predicate-argument structure or correct predicate-argument structure and features versus the number of sentences that parsed, but did not have the correct semantic representation.

Previously, little work has been done that evaluates inferred grammars on held-out test items. Hellan (2010) and Hellan and Beermann (2011) do not present any evaluation for their inference system and Indurkhya (2020) evaluates his grammars over the same sentences as were seen in the training set. However, Bender et al. (2014) and Zamaraeva et al. (2019a) evaluate inferred grammars over held-out portions of the Chintang dataset. Here we use the same dataset of Chintang as one of our development sets, so we use Zamaraeva et al. 2019a as a point of external comparison.

By creating lexical items for determiners, adpositions, coordinators and negation words, we doubled the

Language [iso]	Lexical Coverage (%)	Parse Coverage (%)	Correct Pred-Arg Structure (%)	Correct Pred-Arg Strugure and Features (%)	Ambiguity
Abui [abz]	53.19	41.96	10.19*	5.73*	2195
Chintang [ctn]	22.29	12.24	3.58*	1.94*	5562
Haiki [yaq]	17.49	10.29	1.79*	0.89*	161
Lezgi [lez]	7.88	6.08	0.00*	0.00*	10419
Matsigenka [mcb]	12.61	8.02	1.15	1.15	2333
Meithei [mni]	5.86	5.24	1.05	0.42	3722
Nuuchahnulth [nuk]	23.09	10.14	1.87	1.09	265
Wambaya [wmb]	9.41	2.08	0.98	0.12	4
Tsova-Tush [bbl]	28.79	24.05	4.35*	0.00*	3418

Table 5: Coverage and Ambiguity for Development Languages. Results are averages across 10 folds. * indicates results for only a single fold

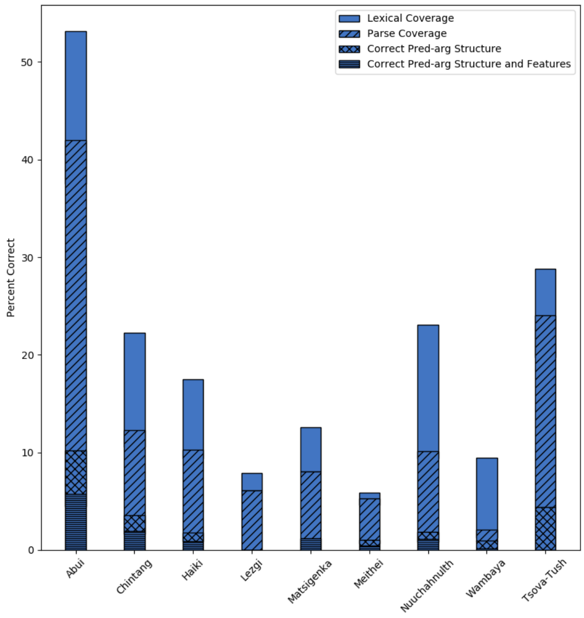


Figure 11: Lexical coverage, parse coverage, correct pred-arg structure and correct features by language for development languages

number of test items for which the inferred grammars can analyze each word, compared to Zamaraeva et al. (2019a) for Chintang. This is critical as the grammar has no chance at syntactic analysis if lexical analysis fails. Our lexical coverage averages 20% across the development languages. (Here and throughout, we use macro-averages weighting each language equally.)

The next thing to consider is what portion of the sentences for which the grammar can analyze each word can be analyzed syntactically. Zamaraeva et al.'s inferred Chintang grammar parsed 30% of the sentences it had lexical coverage for. Our inferred grammars have significantly closed that gap, parsing 84% of the Chintang sentences that had lexical coverage and 67% of the items with lexical coverage on average across all of the development languages.

The most important metric is the proportion of test items the grammar parses correctly. On the development languages, the number of sentences BASIL parses with the correct predicate-argument structure ranges from 0% to 10%. The number of sentences with correct predicate-argument structure for Chintang is more than double what it was for Zamaraeva et al. (2019a) and the introduction of semantic features increases the quality of these parses. BASIL has more spurious coverage than the system of Zamaraeva et al. (2019a), which correctly parsed 47% of its parsed sentences. BASIL produced parses with correct predicate-argument structure for only 19% of the Chintang sentences it parsed; however, for 9% of the sentences it parsed, BASIL also included the correct features in the semantic representation.

Finally, measuring ambiguity shows how many incorrect or redundant parses are produced by the grammar. Ideally, this should be minimal, as in Wambaya, for which our inferred grammars average four parses per sentence. However this average increases when there are multiple analyses for a morphological or syntactic phenomenon, some of which are valid and some of which are not. We go into this in more detail in Section 8.3 where we compare the ambiguity of the inferred grammars with baseline inference systems. At this stage, we simply note that there is an inherent tradeoff between coverage and ambiguity in inferred grammars, just as in hand-crafted grammars: Where sentences may seem unambiguous to humans, who have the benefit of context and world knowledge, computers are much better at finding alternative, often pragmatically odd, analyses. The more phenomena a grammar includes, the more such analyses are available.

5.4 Summary

In this section we described the languages and datasets that we used during development and assessed BASIL in terms of how it performs on them. We primarily used 9 development languages from 7 language families, but at times consulted others for a total of 27 languages from 19 families, in order to make BASIL as robust to cross-linguistic variation as possible. We showed that the 9 development languages tested most of the phenomena targeted by the inference system and performed well in terms of producing grammars that handle those phenomena correctly. With this performance at the end of development, we turn to evaluation on held-out languages to determine how well BASIL generalizes to previously unconsidered languages.

6 Evaluation Methodology

In Section 5, we present results for our development languages, where system development benefited from close error analysis. We use the same methodology to evaluate the system on held-out data from held-out languages. As above, we use the full end-to-end pipeline described in Section 3, with 10-fold cross-validation, and report the same five metrics from Section 5.3: lexical coverage, parse coverage, correct predicate-argument structure, correct predicate-argument structure and semantic features, and ambiguity. In this section, we describe how we measured these (§6.1), and present our baseline system (§6.2) and test languages (§6.3). The following sections (§§7–8) present our results and error analysis on the held-out languages.

6.1 Evaluation Metrics: Parsing and Treebanking

After inferring a grammar from the training data, we use the ACE parsing software (Crysmann and Packard, 2012) to parse each sentence in the test dataset (links to ACE and other software used for evaluation can be found in Appendix B). For each sentence, ACE outputs whether the grammar had a lexical analysis for each word in the sentence, from which we calculate *lexical coverage*. If each word has an analysis and the grammar accepts the sentence as grammatical, ACE returns a result which includes the syntactic parse trees and corresponding semantic representations (illustrated in Figures 12 and 13), and on this basis, we calculate *parse coverage*. In many cases the grammar contains *ambiguity*, returning multiple parses per sentence, and we report this as the average number of results for sentences that parse.

The process of finding the *correct predicate argument-structure* (and *semantic features*) is more involved. After parsing the test sentences with ACE, we use the Full Forest Treebanking software (FFTB; Packard, 2015) to examine the lexical and syntactic rules in the parse forest to identify any trees that represent an appropriate syntactic parse for the sentence. We then inspect the corresponding semantic structure by looking at the predicate-argument structure as well as the semantic features on each argument. Consider the syntactic and semantic representations in Figures 12 and 13 which were produced by an inferred grammar for the Matsigenka sentence in (6).

(6) Ikamagutakerotyo.
i-kamagu-t-ak-i=ro=tyo
3mS-look-EPC-PERF-REALIS=3fO=AFFECT
'He looked at it.' [mcb] (Michael et al., 2013)

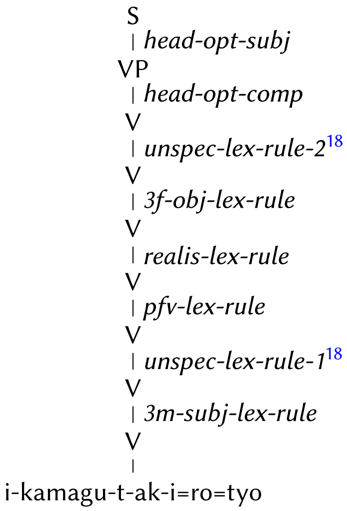


Figure 12: The syntax tree corresponding to the semantic representation in Figure 13

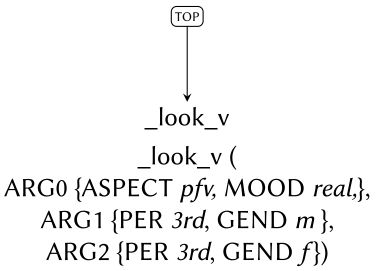


Figure 13: The best semantic representation produced by the inferred grammar for the sentence in (6)

[18] We use 'unspec' as a naming convention for lexical rules that do not add any morpho-syntactic or morpho-semantic features.

Sentence (6) has only one word[19] but includes three semantic arguments: an event and two entities. For this reason, the tree in Figure 12 contains a series of lexical rules (the nodes labeled as V) and two syntactic rules (object dropping, labeled by VP, and subject dropping, labeled by S).[20] The semantic dependency contains only one predicate, which is contributed by the verb *kamagu* 'look'. That predicate has three arguments. First is the event argument (ARG0), which is marked with perfective aspect and realis mood. Next there is the semantic argument (ARG1) corresponding to the unexpressed subject, which is marked with third person and masculine gender, and third is the semantic argument (ARG2) corresponding to the unexpressed object, marked with third person and feminine gender.

We consider the semantic representation in Figure 13 to have the correct predicate-argument structure because it contains all of the predications that should be in the semantic representation and no additional, incorrect predications, and because the predication has the correct arguments: an event and two entities. We consider the semantic features in Figure 13 to be correct because they reflect all of the semantic features that A) are in the IGT and B) the inference system targets: BASIL only targets PNG and TAM features, so those are the only ones we expect. The semantic representation does not reflect the affective meaning because BASIL does not extract stance features.[21]

Although using treebanking to check parses for correctness is an established practice (see inter alia Oepen et al., 2002; Flickinger et al., 2017), assessing the accuracy of semantic representations for languages that one doesn't speak fluently and isn't an expert on is a challenging task. For example, it can be hard to know if some locative dependents are core arguments of the verbs or if they are modifiers. Furthermore, glossing conventions vary from linguist to linguist and with limited familiarity with the datasets, one must make guesses as to implications of some grams and the ambiguous cases one might encounter are difficult to anticipate without first engaging with the data. Therefore, we established a practice of consulting both the gloss line and the translation line as the translation line might omit or add some semantic information compared to the gloss line, but the gloss line may be ambiguous with regards to which words are arguments of which and this can be learned from the translation.[22] After developing basic guidelines by discussing some specific examples from the development datasets, the authors of this paper independently treebanked one fold from each of the Abui and Chintang datasets. These folds contained approximately 100 parsed IGT each.

	Abui [abz]	Chintang [ctn]
Correct Parse	0.5714	0.7843
Matching Pred-Arg Structure	0.5714	0.7843
Matching Features	0.5714	0.5882
Exact Match MRS	0.5143	0.5882

Table 6: F1 scores for inter-annotator agreement on treebanked coverage for Abui and Chintang

Following the methodologies set forth by Dridan and Oepen (2011) for semantic evaluation and Bender et al. (2015) for inter-annotator agreement (IAA), with some adaptations to target our task-specific goals, we calculated IAA for the treebanked results of the two development sets, which we present in Table 6. Dridan and Oepen (2011) propose an Elementary Dependency Match (EDM) score calculated from multiple parts of the semantic representation. We used their EDM$_{na}$ metric for naming and argument identification, and added a metric for semantic features. Following Bender et al. (2015), and in light of the lack of chance-corrected metrics for such structures, we assess IAA for these metrics by calculating the F1 score for these metrics between the two annotators. These F1 scores are shown in Table 6 as Matching Pred-Arg Structure and Matching Features. To situate these measures we also present F1 scores for IAA for whether the parses for the item were considered to include one that was correct (Correct Parse) and whether the two semantic representations matched exactly (Exact Match MRS).

The F1 score for correct parse is the same for matching predicate-argument structure, which shows that when we agreed that there was a parse with an acceptable predicate-argument structure, we also agreed on what that predicate-argument structure should be.[23] Disagreements were often due to one author interpreting something as a modifier instead of an argument (the inferred grammars do not handle modifiers, so these parses would be rejected) or whether sentence fragments should be accepted or rejected, given an otherwise correct semantic representation.

The slightly lower F1 for Exact Match MRS for Abui is due to a slightly different but equally acceptable

[19] Although Michael et al. use an = to indicate two clitics (*=ro* and *=tyo*), BASIL analyzes them as affixes. We made this analytical choice because = in IGT frequently indicates less phonologically integrated affixes, rather than clitics in the sense of Zwicky and Pullum (1983).

[20] The treatment of these arguments as a dropped subject and object is consistent with Inman's (2015) analysis of pronoun incorporation in Matsigenka.

[21] The gloss AFFECT is not explicitly defined by Michael (2008), but from his discussion around such examples, we believe that this refers to stance. We assume that EPC marks an epenthetic consonant, and does not contribute any semantic feature.

[22] This is based on Bender's previous treebanking work in Bender 2008a, Bender et al. 2014 and Zamaraeva et al. 2019a.

[23] This does not necessarily mean that we chose the same syntactic parse, as spurious ambiguity may result in multiple syntactic structures producing the same semantic representation.

predication for the verb in one sentence: leave.for_v_rel vs. leave.for-or-step_v_rel, where the second represents two possible meanings of the verb. For Chintang the feature agreement is lower than predicate-argument structure agreement. For this language the grammars have a great deal of ambiguity in the lexical rules. In many cases, it was not possible to find a parse that had all of the correct features, and we chose parses with different subsets of correct and incorrect features.

After discussing our disagreements, we extended our definitions of correct parses. For all held-out languages a single author treebanked the results, according to the conventions decided through this process.

6.2 Baseline

The primary contribution of this paper is in inferring syntactic properties from IGT data and integrating these with lexical and morphological properties inferred by MOM (Wax, 2014; Zamaraeva, 2016; Zamaraeva et al., 2017). Therefore we compare our results to three baseline systems that are morphologically and lexically robust with respect to accounting for the training data, but are syntactically naive. Each of these use lexical entries and morphological rules from MOM for nouns and verbs. Although MOM extracts morpho-syntactic features for nouns and verbs and adds them to the lexicon and morphological rules, inference is required to define them appropriately in the grammar specification. Because a grammar specification with morpho-syntactic features on verbs and lexical entries with no definition of those features would not result in a working grammar we disable the feature extraction in MOM for all baselines.

Table 7 enumerates the syntactic specifications for our baseline systems. The first baseline (BROAD-COV) posits the specifications for each syntactic phenomenon we account for that we expect to result in the broadest coverage, given no specific knowledge of the language. The second baseline (TYP) posits the specifications that are typologically most common, according to the information available in WALS (Dryer and Haspelmath, 2013) and other typological resources. If a typologically-most-frequent choice could not be made, we select the specification at random if it is required by the Grammar Matrix, and omit it otherwise. Aside from specifications made at random (which are chosen with each run), the syntactic specifications under the BROAD-COV and TYP baselines are the same for all grammars, that is, they do not vary in response to the data presented. Finally the third baseline (RAND) selects a value for each specification at random. The baseline systems make a different random choice for each RC specification every time they are run, therefore the values in the baseline files for each fold of training data are different.

	BROAD-COV	TYP	RAND
word order	free	SOV	RC
has determiners	yes	yes	RC
noun-det order	RC	noun-det	RC
det required	optional	RC	RC
has auxiliaries	no	no	no
verb valence	trans	RC	RC
case frame	none	none	none
s coordination	asyndeton		RC
vp coordination	asyndeton		RC
np coordination	asyndeton		RC
n coordination	asyndeton		RC
subj-drop	all	all	RC
obj-drop	all		RC

Table 7: Grammar specifications for syntactic phenomena for three baseline systems. RC indicates a random choice

Language	ISO 639-3	Source	Number of of IGT	POS tags in source
Arapaho	arp	Toolbox	5000	yes
Hixkaryana	hix	Toolbox	5749	yes
South Efate	erk	Toolbox	1875	yes
Titan	ttv	Toolbox	1799	yes
Wakhi	wbl	FLEx	683	yes

Table 8: Source, number of IGT and presence of POS tags for the held-out datasets

6.3 Held-out Languages

To test how well BASIL generalizes to new languages, we acquired datasets for five additional languages, which we did not consider during development and which are genealogically and geographically varied from the development languages. These languages are listed in Table 8 and the locations where they are spoken are shown in green on the map in Figure 10.

We pre-processed each dataset by filtering out ungrammatical examples (examples marked with a *) and removing duplicates. For held-out evaluation, we selected only languages with POS tags in the original dataset. This information as well as the type of source dataset and the number of IGT after filtering are summarized in Figure 8. In this section, we provide a brief description of each language and dataset. For a full list of citations for datasets and descriptive resources referenced in this section, see Appendix C.

Arapaho [arp] is an Algonquian language of the Algic language family with only about 250 native speakers in the United States (Cowell and Moss Sr, 2011). The dataset we use is a 5,000 item subset of a ~60,000 IGT corpus (Cowell, 2018), randomly selected from fully-glossed examples. The corpus includes elicitations and transcribed conversations, among other genres.

Hixkaryana [hix] is a Cariban language in the Waiwai subgroup with about 1,200 speakers (Eberhard et al., 2019). After removing IGT with incomplete glosses, the corpus (Meira, 2020) contains almost 6,000 IGT.

South Efate [erk] is a Vanuatu language of the Austronesian language family, spoken by about 6,000 people on the Efate island in the Republic of Vanuatu (Thieberger, 2006b). From the 3,000 IGT corpus (Thieberger, 2006a), we use 1,900 fully glossed examples.

Titan [ttv] is also an Austronesian language, and while it and South Efate are both Oceanic, Titan is grouped as a language of the Admiralty Islands while South Efate is Central-Eastern Oceanic. The various dialects of Titan are spoken by approximately 3,500-4,500 people (Bowern, 2011). This corpus contains just under 1,800 IGT after filtering for glossing (Bowern, 2019). For this corpus, we obtain POS tags from the accompanying Toolbox lexicon. This introduces some noise, due to lexical ambiguity, but less than if we had used the projected POS tags from INTENT.

Wakhi [wbl] is an Iranian language of the Indo-European language family and is spoken primarily in Afghanistan and has a growing speaker population of about 17,000 (Eberhard et al., 2019). The dataset is small, containing only about 700 IGT after filtering (Kaufman et al., 2020). However, it is thoroughly glossed and is made up primarily of elicitations targeting specific syntactic phenomena.

7 Results

Using the methodology in Section 6, we performed ten-fold cross-validation on the evaluation languages for the BASIL inference system and the three baselines described in Section 6.2.[24] We show lexical coverage in Table 10, parse coverage in Table 11, coverage with correct predicate-argument structure in Table 12, coverage with correct predicate-argument structure and semantic features in Table 13 and ambiguity in Table 14.

For each language, we treebanked *n* folds such that the number of parsed sentences in *n* folds is greater than 100. The results for lexical coverage, parse coverage and ambiguity are averages across ten folds, while the results for coverage with correct predicate-argument structure and coverage with correct predicate-argument structure and features are averages across *n* folds where *n* is given in Table 9.

There is a great deal of variation in how well any of the systems did at inferring grammars that can parse held-out sentences for each language, as illustrated by the graph in Figure 14. Coverage for Arapaho was very low, at roughly 3% lexical coverage for each system and similar parse coverage for BASIL and BROAD-COV. Across all systems, Hixkaryana and Wakhi had significantly higher lexical and parse coverage, exceeding BASIL's performance on most of the development languages. South Efate and Titan fall between these two extremes. The correct coverage is more consistent across languages with Wakhi as an outlier. For Wakhi, BASIL achieves correct predicate-argument structure for 14.20% of the items in the test set and correct predicate-argument structure and features for 5.8% and the BROAD-COV baseline achieves 12.75% correct predicate-argument structure, while the remaining languages have much lower correct coverage across systems. Finally, the ambiguity (or average number of parses per parsed item) for these languages is quite low for Wakhi, on the order of tens, and extremely high for South Efate, on the order of 100,000. We provide more detail on the causes of ambiguity in the inferred South Efate grammar in Section 8.3.

Language	Tree-banked folds (*n*)	Parsed sentences in *n* folds	Total sentences in *n* folds
Arapaho [arp]	7	109	3500
Hixkaryana [hix]	1	198	575
South Efate [erk]	7	110	1504
Titan [ttv]	6	110	1080
Wakhi [wbl]	5	115	345

Table 9: Number of sentences treebanked across *n* folds for each held-out language

Overall, the systems performed best on Wakhi across the five metrics. Performance for Hixkaryana, South Efate and Titan was somewhat lower, with coverage for Arapaho being the lowest. In Sections 8.1 and 8.2, we explore sources of this variation, including characteristics of the languages and of the IGT datsets.

To understand the impact of *syntactic inference* on automatic grammar generation, we compare BASIL with three baselines that use the same morphotactic and lexical inference system as BASIL, but must specify the syntactic portions of the grammar specification through some other means. The BROAD-COV system uses the specifications that are expected to parse the most sentences, whether correctly or incorrectly. TYP uses the typologically most common specification and RAND uses a random choice (for details, see §6.2). Each of these baselines uses a random choice for at least one specification, where no clear determination could be made for broad coverage or typological frequency, so ten-fold cross validation (given that a new random choice is made when specifying the grammar for each fold) is important to reduce the effect of chance on the overall performance of each baseline.

Because the same morphotactic and lexical inference system was used for the baselines as for BASIL, the lexical coverage across systems is roughly compa-

[24] The code to reproduce these results is available at `https://git.ling.washington.edu/agg/repro/basil-2020`.

Language	BASIL	BROAD-COV	TYP	RAND
Arapaho [arp]	3.64	3.52	3.64	3.18
Hixkaryana [hix]	38.09	36.01	35.88	35.92
South Efate [erk]	12.80	13.55	14.29	13.17
Titan [ttv]	13.56	19.40	20.34	19.40
Wakhi [wbl]	39.68	29.72	31.48	31.04

Table 10: Lexical coverage for held-out languages as a percentage of the total number of test items across ten folds

Language	BASIL	BROAD-COV	TYP	RAND
Arapaho [arp]	3.04	3.06	0.50	0.26
Hixkaryana [hix]	34.18	31.28	2.80	1.25
South Efate [erk]	6.77	9.81	0.27	0.27
Titan [ttv]	10.34	16.18	0.06	0.17
Wakhi [wbl]	30.31	24.89	10.25	3.22

Table 11: Parse coverage for held-out languages as a percentage of the total number of test items across ten folds

Language	BASIL	BROAD-COV	TYP	RAND
Arapaho [arp]	0.17	0.20	0.00	0.03
Hixkaryana [hix]	2.26	2.26	1.57	0.52
South Efate [erk]	0.38	0.31	0.00	0.00
Titan [ttv][25]	0.28	0.65	0.09	0.19
Wakhi [wbl]	14.20	12.75	2.61	0.58

Table 12: Coverage with correct predicate-argument structure as a percentage of the total number of test items across *n* folds

Language	BASIL	BROAD-COV	TYP	RAND
Arapaho [arp]	0.09	0.06	0.00	0.00
Hixkaryana [hix]	0.00	0.00	0.00	0.00
South Efate [erk]	0.15	0.00	0.00	0.00
Titan [ttv]	0.19	0.00	0.00	0.00
Wakhi [wbl]	5.80	0.58	0.00	0.00

Table 13: Coverage with correct predicate-argument structure and semantic features as a percentage of the total number of test items across *n* folds

Language	BASIL	BROAD-COV	TYP	RAND
Arapaho [arp]	145	936	4	3
Hixkaryana [hix]	5642	15596	2	6
South Efate [erk]	126379	9759	2	4
Titan [ttv]	595	6201	2	1
Wakhi [wbl]	10	26	1	2.5

Table 14: Average number of results per parsed sentence for across ten folds

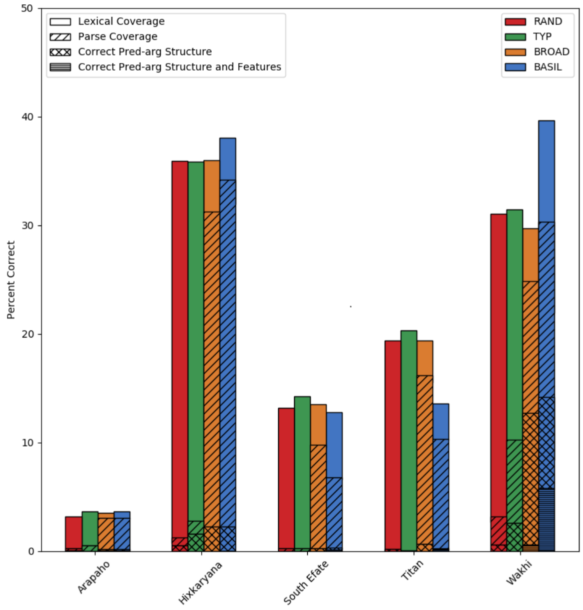


Figure 14: Lexical coverage, parse coverage, correct pred-arg structure and correct features by language for held-out languages

rable. For some languages, the baseline lexical coverage is lower because the baselines can only use POS tags to identify lexical items, while BASIL uses additional heuristics. For other languages, it is slightly higher because BASIL strategically excludes ditransitive and clausal complement-taking verbs (which it would not handle correctly) from the lexicon.[26] Additional variation in the lexical coverage across systems can be attributed to variations in the morphological graph: It is different for each baseline, because it is sensitive to verb valence assignments and these are done at random in each run for the TYP and RAND baselines.

A larger and more meaningful difference between the systems is seen in parse coverage. Here, the TYP and RAND baselines have much lower coverage than BASIL and BROAD-COV. While the TYP baseline has a better chance of using the correct value for each individual specification, it will not necessarily be correct for enough phenomena to produce a grammar that can parse simple sentences: For example, even if the order of verbs with respect to subjects and objects is correct, sentences with determiners won't parse if the determiner-noun order is incorrect. By design, the BROAD-COV system has the highest parse coverage, often outperforming BASIL; however, without syntactic in-

[25] For Titan we report a correct coverage that is higher than the parse coverage for the TYP and RAND baselines. This is possible because there were more parsed items per fold in the 6 folds we treebanked than in the remaining 4.

[26] BASIL cannot properly account for ditransitives as they are not currently supported by the Grammar Matrix. Clausal complement-taking verbs have also been left out of scope at this time.

ference this coverage could be spurious, so we must consider correct coverage (described in §6.1). Again, the typ and rand baselines under-perform the other systems, as there is a relatively low chance that their specifications will correctly model any given language. In terms of correct predicate-argument structure, basil outperforms broad-cov for South Efate and Wakhi, while broad-cov does better for Arapaho and Titan. They tie on Hixkaryana. As broad-cov is designed to maximize coverage, it specifies asyndetic coordination for each language, enabling it to parse sentences for languages where basil failed to infer this strategy. For correct predicate-argument structure and semantic features, basil outperforms all baselines, as they cannot posit semantic features. Only in rare cases did broad-cov have the 'correct features', because the semantic representation shouldn't include any features at all.

So far, we have shown that basil and broad-cov out-perform the other two baselines in parse coverage and correct predicate-argument structure, while basil out-performs all of the baselines in correct predicate-argument structure and semantic features, as illustrated in Figure 14. The last thing to consider is how much ambiguity each of the grammars contain. typ and rand produced grammars with very little ambiguity. These grammars only parsed simple sentences, so low ambiguity is not surprising. broad-cov was designed to maximize coverage, but this comes at the cost of increased ambiguity. For example, positing free word order for each language will ensure that all word orders will parse, but will also allow parses where the wrong constituents are identified as subjects and objects. As a result, the broad-cov baseline has significantly higher ambiguity than basil for all languages but South Efate.

While the results show a great deal of variation across the test languages, basil and broad-cov outperform the typ and rand baselines for most metrics. basil and broad-cov perform fairly comparably for a number of the metrics, but basil excels in two areas. First, basil generally has fewer parses per test item than broad-cov, suggesting that there is less spurious ambiguity in the inferred grammars than in that baseline. While typ and rand have even lower ambiguity scores, they also have such low coverage that this is not an advantage. Second, the semantic representations produced by basil are more correct in that they contain semantic features, resulting in higher scores for the correct predicate-argument structure and features metric.

8 Error Analysis

8.1 Out of Scope Phenomena

We begin our error analysis by establishing first what we do not expect basil's grammars to parse. Focusing on sentences where lexical coverage was achieved but the sentence did not parse or parsed incorrectly, we describe phenomena that are frequent in the test data but are beyond the scope of the current inference system.

basil currently handles a number of lexical types such as transitive and intransitive verbs, auxiliaries, nouns, determiners and case-marking adpositions, as well as phenomena including word order, case, argument optionality, sentential negation and coordination. However, it does not yet handle a number of very common phenomena such as adjectives, adverbs, ditransitive or clausal complement-taking verbs, content question words, possessives, etc. Therefore, sentences containing these lexical items will only have lexical coverage if a lexical item was inferred in error. At the same time, sentences that contain these syntactic phenomena will not parse at all or will not parse correctly.

In particular, frequent error types include: (i) verb valence, where basil posited intransitive or transitive entries for verbs which were actually ditransitive or clausal-complement taking; (ii) adnominal possession, where grammars produced by basil parsed but could not attribute the correct semantics to examples with possession; (iii) vocatives analyzed as subjects or objects; (iv) sentence linkers parsed as coordination; and (v) disfluency markers (e.g. *P* for 'pause') analyzed as verbs.

8.2 In Scope Phenomena

Whereas the previous section described common errors due to out of scope phenomena in the test data, this section focuses on errors due to basil failing to correctly infer phenomena that it was designed to handle. The sources of these errors range from the input data to problems with basil's inference algorithms or their implementation.

8.2.1 Wrong Part-of-Speech

Both basil and MOM rely on POS tags in the input to identify nouns and verbs. In some cases, the POS tag in the corpus may be incorrect. For example, in (7) the word *titko* is glossed as 'brazil.nut' but marked with a verbal POS tag. Such errors are not uncommon, as even the most careful human annotation is subject to error.

(7)	Tutko	yakahetxkoni.
	titko	y-akaha-yatxkoni
	Brazil.nut	rel-break-dpst2:col
	Vt	prs-Vt-tamn

'They were shelling Brazil nuts.' [hix] (adapted from Meira, 2020)

Because *titko* is glossed as a verb, the inferred grammar treats it semantically as an event instead of as a participant of the breaking/shelling event, resulting in an incorrect semantic representation.

8.2.2 Wrong Predication

We considered it an error anytime the predication associated with a word did not reflect the meaning in the gloss, even if the overall shape of the predicate-argument structure was correct. This can occur if MOM's heuristics for locating the root of a word fail in a particular case. For example, the IGT in (8) had spaces on both sides of the second hyphen. MOM guessed that the hyphen belonged to *neeni*, which in turn meant that *t* was the root, leading to a lexical entry with the predication _3.S_v_rel.

(8)	Nehe'	hinen	nihneenit.			
	nehe'	hinen	nih-	neeni	-	t
	this	man	PAST-	itis	-	3.S

'The man was the one.' [arp] (adapted from Cowell, 2018)

8.2.3 Missed Semantic Features

BASIL's greatest advantage over the baseline systems is its addition of semantic features to the grammars, but it still made some errors in feature inference. There is significant variation in the way linguists gloss syntactico-semantic features, and BASIL's most straight-forward source of error for semantic features was in not properly identifying all grams in the held-out corpora. BASIL uses a large dictionary of glosses, which it maps to 116 common PNG, TAM and case grams to identify morpho-syntactic and morpho-semantic features (see §4.1). Even so, the held-out corpora included grams that were not in this dictionary. In particular, this dictionary did not include any glosses for the pluperfect aspect 'PLPF', which appears in Wakhi, the immediate past 'IPST' or distant past 'DPST' used in Hixkaryana, or the narrative past 'NARRPAST' used in Arapaho. In addition, while the dictionary included 'D' as a gloss for dual number and quite a few person and number combinations (e.g. '3DU'), it did not contain '3D' which is used for third person, dual number in the South Efate corpus. This led to test items, which otherwise parsed correctly, not including all of the semantic features.

8.2.4 Auxiliaries

BASIL treats words that have only TAM and/or PNG agreement features as auxiliaries (see §4.5.1). The abundance of TAM auxiliaries in the held-out languages, such as the future tense auxiliary in (9), revealed a bug in our implementation of auxiliary inference. The clause in BASIL's code that infers where the auxiliary occurs (before or after its complement) assigns the wrong value. This caused some inferred grammars to require auxiliaries after their verbal complements instead of before. Though our development languages included auxiliaries, these freer word order languages (Wambaya and Nuuchahnulth) did not reveal this bug.

(9)	Tumrə	maz̨	jittu.
	tumrə	maz̨	jaw-tu
	FUT	1SG.OBL	eat-PLPF

'I will have had eaten.' [wbl] (adapted from Kaufman et al., 2020)

8.2.5 Coordination

Coordination inference, described in Section 4.5.5, errs on the side of positing VP coordination unless it finds explicit evidence of S coordination in the form of a projected subject dependency that intervenes between the coordinator and a verb in the coordinand. This algorithm may be too aggressive because dependency tag projection is not always successful. In addition to that, the algorithm does not consider cases where the subject is dropped or cases where there is no coordinator, because an asyndetic strategy is employed. Because the inference of S coordination relies on an overt coordinator, sentences like the one in (10) from Titan are taken by BASIL as evidence of VP coordination instead of S even though each coordinand has an overt subject. Thus asyndetic S coordination isn't added to the grammar and examples like this can't be parsed.

(10)	I	ani	pou	i	ani	ma.
	i	ani	pou	i	ani	ma
	3SG	eat	pig	3SG	eat	taro

'He ate the pig and he ate the taro.' [ttv] (adapted from Bowern, 2019)

In addition, examples of monosyndetic S coordination in Wakhi were misclassified as VP coordination because of failure to align the subjects between the English translation and the sentence. This prevented BASIL from inferring S coordination strategies and adding them to the grammar specifications. Because the BROAD-COV baseline posits asyndetic S coordination for all languages, that baseline was able to correctly parse sentences with asyndetic S coordination in Titan and Wakhi, giving it a boost in coverage over BASIL.

8.2.6 Case Frame

Finally, BASIL relies on the overt case markings on the subject and object (according to projected dependencies), to account for quirky case (§4.5.2). However, if no overt argument is found, the verb's case frame remains under-specified until it is merged with another instance of the same verb. Even though BASIL inferred the overarching nominative-accusative pattern for Wakhi, it found verbs in the training data with oblique subjects which were merged with verbs that did not have overt case marking on their subjects. Because of this, the inferred grammars for some of the Wakhi folds included a rather large transitive verb class with oblique case on the subject, resulting in a number of IGT with overtly marked nominative subjects in the test data that did not parse.

8.2.7 Summary

The majority of errors discussed in this section come from lexical inference. Beyond that, we identified three main sources of error in the syntactic specifications. One was a bug that resulted in auxiliaries having the wrong order with respect to their complements. Resolving this bug is trivial, while the errors in S coordination and case-frame inference require some re-designing of the algorithms. In particular, BASIL requires too much evidence to infer S coordination. As future work, we propose modifying the algorithm to rely less on projected dependencies and instead to leverage the dependency parse of the English translation to distinguish between VP and S coordination in the translation. The same redesign could be applied to N and NP coordination as well. The case frame inference algorithm may assign quirky case too readily and rather than merging lexical items with no case frame with those that have quirky case, should assign default case to those verbs unless a verb with the same orthography is found with quirky case in the corpus. Alternatively, better verb classes could be inferred with some re-tooling of the interaction between BASIL and MOM, so that case frame inference happens after morphotactic inference, similar to the pronoun and auxiliary inference methodologies in Section 4.2.2.

8.3 Ambiguity

BASIL's inferred grammars generally had less ambiguity than the BROAD-COV baseline for two intuitive reasons. First, the free word order, argument optionality and coordination specifications in BROAD-COV introduce a lot of ambiguity in the number of ways nouns and verbs can combine. Second, BASIL's specifications for case frame and agreement further constrain which arguments can be subjects and objects, even in freer word order languages. In spite of this, BASIL's grammars for South Efate have significantly more ambiguity than BROAD-COV's. To shed light on this, we present a specific example from the fourth test fold from South Efate.

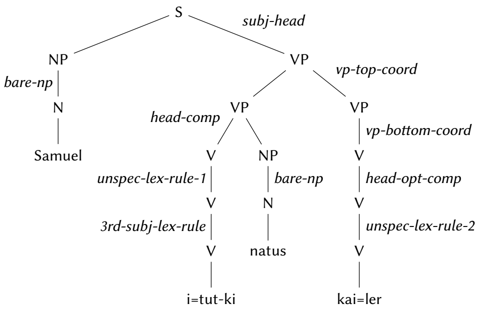


Figure 15: The parse tree generated by the BASIL and BROAD-COV grammars that corresponds with the semantic representation in Figure 16 for the sentence in (11)

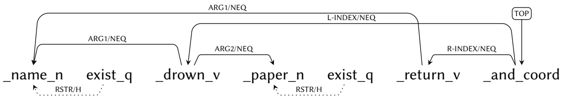


Figure 16: The best semantic representation generated by the BASIL and BROAD-COV grammars for the sentence in (11)

First of all, BASIL infers free word order, subject and object dropping and asyndetic coordination for VPs and NPs for this fold. Because of this, BASIL's inferred grammar is not less ambiguous than BROAD-COV in those areas. In order to understand why BASIL's grammar is even more ambiguous than BROAD-COV's, we explore the parse forest for the sentence in (11), which has asyndetic coordination, lexical ambiguity, morphological ambiguity and no overt case marking.

For this sentence, BASIL's grammar produces 2448 trees, while BROAD-COV's produces 19.[27] The best reading, produced by both grammars, is shown in the parse tree in Figure 15 and semantic representation in Figure 16.

(11) Samuel itutki natus kailer.
Samuel i=tut-ki natus kai=ler
Samuel 3S.RS1-drown-TR paper ES1-return
'Samuel threw in the paper and went back.' [erk] (Thieberger, 2006a)

We use the Full Forrest Treebanking software (FFTB; Packard, 2015) to efficiently investigate such large parse forests with discriminant-based tree selection (Carter, 1997). Figure 17 shows the choices among discriminants that we used to single out the tree in Figure 15 from the other 2447 trees in the parse forest.

The discriminants in Figure 17 are not ordered, and represent one of many paths in the decision space. The bottom 4 choices in the decision tree result in no difference in the semantic representation, yet combined they increase the ambiguity by a factor of 16. The *no-drop-lex-rule* is added by the Grammar Matrix's argument

[27] These numbers are estimates provided by FFTB based on the packed forest, as opposed to ACE, which we used for Table 14.

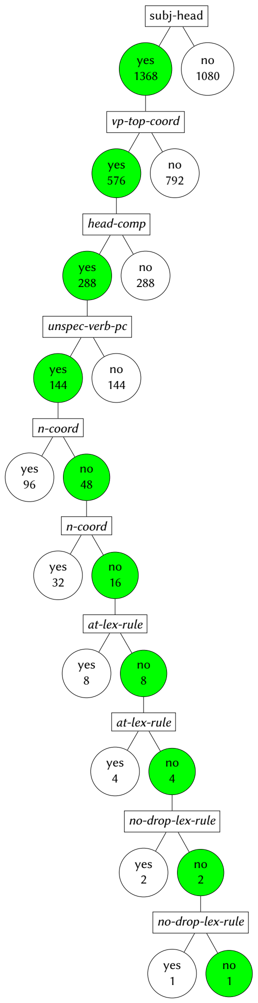


Figure 17: A decision tree illustrating the syntactic and lexical rules that discriminate between different parse trees produced by BASIL's grammar for the sentence in (11). The path in green shows the rules that we selected or excluded to identify the parse tree shown in Figure 15

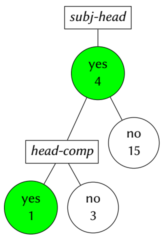


Figure 18: A decision tree illustrating the syntactic and lexical rules that discriminate between different parse trees produced by the BROAD-COV grammar for (11)

optionality library (Saleem, 2010; Saleem and Bender, 2010). This rule is intended to be further constrained by agreement restrictions for dropped arguments, but because BASIL does not add this information to the grammar, these optional, non-inflecting lexical rules add ambiguity for both verbs in (11). The two *at-lex-rule*s are added by the case library (Drellishak, 2009) for languages with case-marking adpositions. These rules apply to both nouns in (11) and because they apply optionally, each of these lexical rules and each of the words they apply to double the number of trees in the forest.[28]

In addition to these sources of ambiguity, there is an under-constrained noun coordination rule that applies optionally to each noun and can apply either before or after the bare-np rule, tripling the number of parse trees for each noun it can apply to. Because neither noun has an adjacent noun to attach to, these parses should not succeed, but they do as the result of a bug in the Grammar Matrix customization system.

All together the spurious case, coordination and argument optionality rules increase the number of possible trees by a factor of 144. Setting those aside, the number of possible trees looks much more reasonable. Additional ambiguity is added by two homophonous lexical rules for the *kai-* prefix: one adds first person agreement to the subject and the other (which produces the correct tree) does not add any features.[29]

The three choices at the top of the decision tree discriminate between trees in which *natus* is the object of *i=tut-ki* or *kai=ler* and indirectly, prevent *kai=ler* from being analyzed as a noun, coordinated with *natus*.

The decision tree for BROAD-COV to produce the parse shown in Figure 15 is shown in Figure 18. The lexical rules in the last four nodes in the tree in Figure 17 are not in the BROAD-COV grammar and therefore do not apply. Because ambiguity is a matter of combinatorics, the spurious lexical rules in BASIL's grammar inflate the ambiguity significantly. The same could be said for the sources of ambiguity in the BROAD-COV grammars for the other languages, where BASIL had less ambiguity.

Many of the sources of ambiguity in the South Efate grammars trace back to bugs in the Grammar Matrix customization system, rather than BASIL's inference. Furthermore, the high ambiguity for South Efate grammars was an outlier among the ambiguity in BASIL's grammars for the evaluation languages. This suggests that these sources of ambiguity, both from Matrix bugs and otherwise, are not particularly pervasive.

[28] The optionality of a non-inflecting lexical rule was a bug in the Grammar Matrix, and has since been addressed by (Conrad, 2021).

[29] The morpheme is glossed by the linguist as ES1. Thieberger (2006b) defines the ES abbreviation as "echo subject", and we assume that the 1 is a particular echo subject marker, but does not indicate first person, as there is no first person noun in the translation.

9 Conclusion

In this paper, we introduced BASIL — Building Analyses from Syntactic Inference in Local languages — a system for the automatic inference and generation of machine-readable grammars from IGT data. Leveraging the rich annotation in interlinear glossed text and syntactic information projected from parses of the English translation onto sentences in a local language, BASIL infers grammar specifications. These, in turn, can be input into the Grammar Matrix customization system to produce HPSG grammars.

BASIL utilizes an end-to-end pipeline that begins with an IGT corpus of a language and produces an HPSG grammar which can be loaded into parsing software to produce syntactic and semantic representations for strings in that language. Drawing on the linguistic information encoded in IGT text and generalizations about language from the typological literature, we designed algorithms that infer lexical and syntactic properties about a language and define these properties in a grammar specification. This grammar specification can be input into a grammar customization toolkit (the Grammar Matrix; Bender et al., 2002, 2010; Zamaraeva et al., forthcoming) to produce a machine-readable HPSG grammar for that language.

We built on previous work in grammar inference that produced both morphological (Wax, 2014; Zamaraeva, 2016; Zamaraeva et al., 2017) and syntactic (Bender et al., 2013, 2014; Howell et al., 2017; Zamaraeva et al., 2019a), specifications for a language. That work focused on lexical and morphotactic specifications for nouns and verbs, word order, case system and case frame for verbs. We integrated the existing modules into a single system which we scaled by adding inference for determiners, auxiliaries, case-marking adpositions, PNG and TAM features, argument optionality, negation and coordination.

The result is an inference system that identifies the overarching typological patterns for each of these phenomena and encodes that information in a grammar specification, which is then used to produce a grammar. As one of the goals of this work is to automatically infer grammars for a broad range of local and endangered languages, we developed inference algorithms using a data-driven process, testing our system on a genealogically and geographically diverse set of languages. During development, we consulted 27 languages from 19 language families, spread over 6 continents. We did end-to-end system testing on 9 of those 27 development languages.

In order to test the cross-linguistic generalizability of our inference system, we evaluated it using 5 languages from 4 language families that were not considered during development and did not come from any of the language families that we used in previous end-to-end testing. These languages were Arapaho, Hixkaryana, South Efate, Titan and Wakhi. We compared the performance of BASIL's inferred grammars with three baselines. The TYP baseline used the cross-linguistically most common specifications for each phenomenon (based on typological surveys), while RAND used random specifications. The low coverage of these baselines demonstrated that in order to produce a useful grammar, it is not sufficient to guess the right specifications for just some phenomena, but the specifications for a variety of interacting phenomena must be correct. The third baseline, BROAD-COV, was designed to parse as many sentences as possible in a language, and in spite of this, BASIL's overall coverage was comparable to BROAD-COV, while its grammars had less ambiguity for four of the five languages.

In addition to BASIL's parse coverage being higher than the TYP and RAND baselines and comparable with BROAD-COV, the semantic representations produced by BASIL's grammars were richer. In evaluation, we assessed not only the number of sentences that parsed, but the correctness of those parses in terms of the meaningfulness of their predications and the correctness of the argument relations for those predications. In this respect, BASIL and BROAD-COV performed comparably, outperforming the other two baselines by a large margin. However, BASIL's grammars also added semantic features for person, number, gender, tense, aspect and mood on the semantic predicates, resulting in even more detailed representations than those produced by the BROAD-COV grammars.

Because BASIL relies on the Grammar Matrix's typologically robust syntactic analyses to produce the grammars, BASIL can in principle be extended to account for phenomena as they are added to the Grammar Matrix. Recent work has added libraries for clausal complements (Zamaraeva et al., 2019b), adverbial clausal modifiers (Howell and Zamaraeva, 2018), nominalized clauses (Howell et al., 2018), adnominal possession (Nielsen, 2018; Nielsen and Bender, 2018) and constituent questions (Zamaraeva, 2021). Leveraging the analyses for these phenomena as well as others previously implemented in the Grammar Matrix, modules can be added to extend BASIL's scope.

Accounting for the characteristics of languages or datasets that have the most impact on system performance would enable better assessment of the system's weaknesses and ways to improve it. For this reason, we propose future work that systematically tests these factors by testing with different subsets of a single dataset with different sizes, genres, completeness of glossing or presence of part of speech tags. Upon identifying a threshold for these factors above which system performance stabilizes, it would then be possible to do more rigorous cross-linguistic testing to find language families or typological properties that BASIL struggles with.

Acknowledging that BASIL's grammars are currently limited to a certain number of phenomena and are subject to some degree of error, we turn to a brief discussion of possible uses for these grammars both now and after additional inference modules are added. The first of these is in accelerating the process of creating machine-readable grammars, as creating grammar specifications, especially for languages with complex morphology, can be quite tedious.

Machine readable grammars that are somewhat larger than those produced by BASIL have been used for a broad range of applications such as data exploration (Letcher and Baldwin, 2013; Bouma et al., 2015), grammar checkers (da Costa et al., 2016) and automatic tutors (Hellan et al., 2013). Accelerating the process of developing this type of grammar increases the number of grammars that can be used for these applications. At the current stage, inferred grammars could still be useful for data exploration as they can be used to search corpora for the phenomena they model. This type of data exploration could assist linguists in finding relevant examples of specific phenomena they wish to analyze (as in Zamaraeva et al. 2017), or it could be used to help teachers find varied examples to use in lessons. Once a sufficient number of phenomena are handled by grammar inference, machine-readable grammars inferred from descriptive grammars could accompany those descriptive resources as a tool for further investigating the language's syntax, as described by Bender et al. (2012) and Bouma et al. (2015). Our inferred grammars for Wambaya, which were based on IGT extracted from Nordlinger 1998, serve as proof of concept for this possibility. Finally, as inferred grammars help to streamline the process of grammar engineering, ultimately grammars that started with BASIL and were extended by hand could be used to produce grammar checkers along the lines of da Costa et al. 2016 and other educational tools in order to assist in the effort of language revitalization.

Finally, there is potential for a symbiotic relationship between BASIL and typological resources such as WALS (Dryer and Haspelmath, 2013), SAILS (Muysken et al., 2016) and others. In particular, previous work has found that a number of the Grammar Matrix's specifications map directly to WALS features (de Almeida et al., 2019). For languages where these features are encoded in WALS, this information can potentially be incorporated into the grammar inference pipeline to improve the accuracy of inference for some phenomena. On the other hand, for languages whose features have not been added to databases like WALS, BASIL could be used to automatically infer those features, if an IGT corpus (or a descriptive grammar from which IGT can be extracted) is available.

The primary contribution of this work is a grammar inference system that takes an IGT corpus as input and produces a machine-readable, HPSG grammar that can be used for parsing and generation. Although previous work has automatically generated grammars for English and other languages frequently studied in NLP contexts, BASIL focuses on producing language technology in the form of syntactically precise grammars for local and endangered languages. In light of this, we tested the system on a large number of genealogically and geographically diverse languages and verified its cross-linguistic generalizability. Although the grammars produced by BASIL are still relatively low-coverage over corpora containing the complexity and variety inherent to human language, they provide a valuable starting point for producing broader coverage grammars which can be used to assist data exploration and language documentation and revitalization.

Acknowledgements

This work builds on the contributions of many researchers who have been involved in the AGGREGATION Project over the years. In particular, we would like to thank Olga Zamaraeva for her many contributions to MOM, including multiple changes and updates to the system which this work relies on centrally, Fei Xia for conceptual discussions, and Angelina McMillan-Major and previous AGGREGATION RAs who downsampled the Matsigenka corpus. In addition, Alexis Palmer, Kristy K. Phillips and Olga Zamaraeva developed a tool for converting FLEx datasets to Xigt. Woodley Packard also provided a great deal of assistance to us using FFTB to do treebanking.

We are deeply indebted to the linguists who shared their datasets with us (all listed in Appendix C) and generously answered our questions, as well as to the speaker communities they have worked with.

In addition, we are grateful for the feedback on this paper that we received from Olga Zamaraeva, Woodley Packard, Angelina McMillan-Major, Gina-Anne Levow, Edith Aldridge and our anonymous reviewers.

This material is based upon work supported by the National Science Foundation under Grant No. BCS-1561833 (PI Bender).

References

Ackema, Peter, Patrick Brandt, Maaike Schoorlemmer, and Fred Weerman, editors. 2006. *Arguments and Agreement*. Oxford University Press, Oxford.

Acri, Andrea. 2018. Draft of an in-progress critical edition of chapter 3 of the Bhuvanakośa prepared for the 4th International Intensive Course in Old Javanese, Yogyakarta. 15–29 July 2018 (used with the author's permission).

Agić, Željko, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

de Almeida, Tifa, Youyun Zhang, Kristen Howell, and Emily M Bender. 2019. Feature comparison across typological resources. *Unpublished abstract, presented at TypNLP.*

Bender, Emily M. 2008a. Evaluating a crosslinguistic grammar resource: A case study of Wambaya. In *Proceedings of ACL-08: HLT*, pages 977–985, Columbus. Association for Computational Linguistics.

Bender, Emily M. 2008b. Grammar engineering for linguistic hypothesis testing. In *Proceedings of the Texas Linguistics Society X Conference: Computational Linguistics for Less-Studied Languages*, pages 16–36, Stanford. CSLI Publications.

Bender, Emily M. 2008c. Radical non-configurationality without shuffle operators: An analysis of wambaya. In *Proceedings of the International Conference on Head-Driven Phrase Structure Grammar*, pages 6–24, Stanford. CSLI Publications.

Bender, Emily M. 2010. Reweaving a grammar for Wambaya: A case study in grammar engineering for linguistic hypothesis testing. *Linguistic Issues in Language Technology*, 3(3):1–34.

Bender, Emily M, Joshua Crowgey, Michael Wayne Goodman, and Fei Xia. 2014. Learning grammar specifications from IGT: A case study of Chintang. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 43–53, Baltimore. Association for Computational Linguistics.

Bender, Emily M, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation*, 8(1):23–72. 10.1007/s11168-010-9070-1.

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 *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei.

Bender, Emily M, Dan Flickinger, Stephan Oepen, Woodley Packard, and Ann Copestake. 2015. Layers of interpretation: On grammar and compositionality. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 239–249, London. Association for Computational Linguistics.

Bender, Emily M, Sumukh Ghodke, Timothy Baldwin, and Rebecca Dridan. 2012. 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 Hawai'i Press, Honolulu.

Bender, Emily M and Jeff Good. 2005. Implementation for discovery: A bipartite lexicon to support morphological and syntactic analysis. In *Proceedings from the Panels of the Forty-First Meeting of the Chicago Linguistic Society*, pages 1–15.

Bender, Emily M, 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. Association for Computational Linguistics.

Bickel, Balthasar, 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.

Bickel, Balthasar, 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. Durga. https://corpus1.mpi.nl/qfs1/media-archive/dobes_data/ChintangPuma/Chintang/Narratives/Annotations/durga_exp.tbt Accessed: 2013.

Bickel, Balthasar, Sabine Stoll Stoll, Martin Gaenszle, Novel Kishor Rai, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Netra Prasad Paudyal, Judith Pettigrew, Ichchha Purna Rai, Manoj Rai, Taras Zakharko, and Robert Schikowski. 2013b. Audiovisual corpus of the chintang language, including a longitudinal corpus of language acquisition by six children, paradigm sets, grammar sketches, ethnographic descriptions, and photographs.

Bierwisch, Manfred. 1963. *Grammatik des deutschen Verbs*, volume II of *Studia Grammatica*. Akademie Verlag.

Bird, Steven. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Bod, Rens. 2009. From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science*, 33(5):752–793.

Booij, Geert Evert. 2002. *The morphology of Dutch*. Oxford University Press on Demand.

Bouma, Gosse, JM van Koppen, Frank Landsbergen, JEJM Odijk, Ton van der Wouden, and Matje van de Camp. 2015. Enriching a descriptive grammar with treebank queries. In *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, volume 14, pages 13–25.

Bowern, Claire. 2011. Sivisa Titan: sketch grammar, texts, vocabulary based on material collected by P. Josef Meier and Po Minis. *Oceanic Linguistics Special Publications*, 38:iii–466.

Bowern, Claire. 2012. *A grammar of Bardi*, volume 57. Walter de Gruyter.

Bowern, Claire. 2019. Titan materials. *Digital collection managed by PARADISEC [Open Access].* (Accessed January 2019).

Buys, Jan and Phil Blunsom. 2017. Robust incremental neural semantic graph parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1215–1226, Vancouver, Canada. Association for Computational Linguistics.

Carter, David. 1997. The TreeBanker: a tool for supervised training of parsed corpora. In *Computational Environments for Grammar Development and Linguistic Engineering*.

Chelliah, Shobhana Lakshmi. 2011. *A grammar of Meithei*, volume 17. Walter de Gruyter.

Chelliah, Shobhana Lakshmi. 2019. Meithei texts. Manipur Digital Resources in UNT Digital Library. University of North Texas Libraries. (Accessed August 2019).

Chen, Yufei, Weiwei Sun, and Xiaojun Wan. 2018. Accurate SHRG-based semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 408–418, Melbourne, Australia. Association for Computational Linguistics.

Chomsky, Noam. 1995. *The Minimalist Program*. MIT Press, Cambridge.

Clark, Stephen and James R. Curran. 2004. Parsing the WSJ using CCG and log-linear models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 103–110, Barcelona, Spain.

Comrie, Bernard. 1989. *Language Universals & Linguistic Typology*, second edition. University of Chicago, Chicago.

Conrad, Elizabeth. 2021. Tracing and reducing lexical ambiguity in automatically inferred grammars. Master's thesis, University of Washington.

Copestake, Ann. 2002a. Definitions of typed feature structures. In Stephan Oepen, Dan Flickinger, Junichi Tsujii, and Hans Uszkoreit, editors, *Collaborative Language Engineering*, pages 227–230. CSLI Publications, Stanford.

Copestake, Ann. 2002b. *Implementing typed feature structure grammars*. CSLI publications Stanford.

Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal Recursion Semantics: An introduction. *Research on Language and Computation*, 3(2-3):281–332.

Corbett, Greville G. 1991. Gender. *Cambridge: CUP.*

Corbett, Greville G. 2000. Number. *Cambridge: CUP.*

da Costa, Luis Morgado, Francis Bond, and Xiaoling He. 2016. Syntactic well-formedness diagnosis and error-based coaching in computer assisted language learning using machine translation. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 107–116.

Cowell, Andrew. 2018. Arapaho text database. Version 1, 2018. University of Colorado, Department of Linguistics (Accessed at https://github.com/Adamits/arapaho_library/tree/master/data February 2020).

Cowell, Andrew and Alonzo Moss Sr. 2011. *The Arapaho language*. University Press of Colorado.

Crowgey, Joshua. 2019. *Braiding Language (by Computer): Lushootseed Grammar Engineering*. Ph.D. thesis, University of Washington.

Crowgey, Joshua David. 2012. The syntactic exponence of sentential negation: A model for the LinGO Grammar Matrix. Master's thesis, University of Washington.

Crysmann, Berthold and Woodley Packard. 2012. Towards efficient HPSG generation for German, a nonconfigurational language. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 695–710.

Cysouw, Michael. 2013. Inclusive/exclusive distinction in independent pronouns. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Östen Dahl. 1979. Typology of sentence negation. *Linguistics*, 17(1-2):79–106.

Dedrick, John M and Eugene H Casad. 1999. *Sonora Yaqui Language Structures*. University of Arizona Press.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dixon, RMW. 1994. *Ergativity*. Cambridge University Press, Cambridge.

Donet, Charles. 2014a. The importance of verb salience in the followability of Lezgi oral narratives. Master's thesis, Dallas International University.

Donet, Charles. 2014b. Lezgi oral narratives. Dallas International University. Unpublished FieldWorks (FLEx) project. (Accessed August 2019).

Drellishak, Scott. 2004. A survey of coordination strategies in the world's languages. Master's thesis, University of Washington.

Drellishak, Scott. 2009. *Widespread but not universal: Improving the typological coverage of the Grammar Matrix*. Ph.D. thesis, University of Washington.

Drellishak, Scott and Emily M Bender. 2005. A coordination module for a crosslinguistic grammar resource. In *The Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar, Department of Informatics, University of Lisbon*, pages 108–128, Stanford. CSLI Publications.

Dridan, Rebecca and Stephan Oepen. 2011. Parser evaluation using elementary dependency matching. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 225–230.

Dryer, Matthew S. 2005. Negative morphemes. In Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors, *The World Atlas of Linguistic Structures (WALS)*, pages 454–457. Oxford University Press, Oxford.

Dryer, Matthew S. 2013a. Expression of pronominal subjects. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available at https://wals.info/chapter/101, Accessed 2022-05-04.

Dryer, Matthew S. 2013b. Negative morphemes. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available at https://wals.info/chapter/112, Accessed 2022-05-04.

Dryer, Matthew S. 2013c. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available at https://wals.info/chapter/81, Accessed 2022-05-04.

Dryer, Matthew S. and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available at https://wals.info/, Accessed 2022-05-04.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the World. Twenty-second edition. Available at http://www.ethnologue.com, Accessed 2022-05-04.

Flickinger, Dan. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6 (1) (Special Issue on Efficient Processing with HPSG):15 – 28.

Flickinger, Dan. 2011. Accuracy v. robustness in grammar engineering. In Emily M Bender and Jennifer E Arnold, editors, *Language from a Cognitive Perspective: Grammar, Usage and Processing*, pages 31–50. CSLI Publications, Stanford.

Flickinger, Dan, Emily M Bender, and Stephan Oepen. 2014a. ERG semantic documentation. Accessed on 2022-05-16.

Flickinger, Dan, Emily M Bender, and Stephan Oepen. 2014b. Towards an encyclopedia of compositional semantics: Documenting the interface of the English resource grammar. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 875–881, Reykjavik. European Language Resources Association (ELRA).

Flickinger, Dan, Stephan Oepen, and Emily M Bender. 2017. Sustainable development and refinement of complex linguistic annotations at scale. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 353–377. Springer Netherlands, Dordrecht.

Fokkens, Antske. 2014. *Enhancing Empirical Research for Linguistically Motivated Precision Grammars*. Ph.D. thesis, Department of Computational Linguistics, Universität des Saarlandes.

Georgi, Ryan. 2016. *From Aari to Zulu: Massively Multilingual Creation of Language Tools Using Interlinear Glossed Text*. Ph.D. thesis, University of Washington.

GOLD. 2010. General Ontology for Linguistic Description (GOLD). Bloomington, IN: Department of Linguistics (The LINGUIST List), Indiana University. Available at http://linguistics-ontology.org/, Accessed 2022-05-06.

Goodman, Michael Wayne. 2013. Generation of machine-readable morphological rules from human readable input. *Seattle: University of Washington Working Papers in Linguistics*, 30.

Goodman, Michael Wayne, Joshua Crowgey, Fei Xia, and Emily M Bender. 2015. Xigt: Extensible interlinear gloss text for natural language processing. *Language Resources and Evaluation*, 49 (2):455–485.

Han, Wenjuan, Ge Wang, Yong Jiang, and Kewei Tu. 2019. Multilingual grammar induction with continuous language identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5732–5737.

Harley, Heidi. 2019. Haiki text corpus. University of Arizona. Unpublished FieldWorks (FLEx) project. (Accessed August 2019).

Haspelmath, Martin. 2007. Coordination. In Timothy Shopen, editor, *Language typology and syntactic description*, volume 2. Cambridge University Press, Cambridge.

Hauk, Bryn. 2016–2019. Tsova-tush lexicon and texts. University of Hawai'i at Mānoa. Unpublished FieldWorks (FLEx) project. V2019.08.20. 2016–2019 (collection date).

Hauk, Bryn. 2020. *Deixis and reference tracking in Tsova-Tush*. Ph.D. thesis, University of Hawai'i at Mānoa.

Hauk, Bryn and Alice C. Harris. forthcoming. Batsbi. In Yuri Koryakov, Yury Lander, and Timur Maisak, editors, *The Caucasian languages: An international handbook*. De Gruyter Mouton.

Hellan, Lars. 2010. From descriptive annotation to grammar specification. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 172–176, Uppsala. Association for Computational Linguistics.

Hellan, Lars and Dorothee Beermann. 2011. Inducing grammars from IGT. In *Human Language Technology Challenges for Computer Science and Linguistics.*, volume 8287 of *LTC 2011. Lecture Notes in Computer Science*. Springer.

Hellan, Lars, Tore Bruland, Elias Aamot, and Mads H Sandøy. 2013. A grammar sparrer for Norwegian. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16*, 085, pages 435–439. Linköping University Electronic Press.

Hewitt, John and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis. Association for Computational Linguistics.

Hinds, John. 1986. *Japanese: Descriptive Grammar*. Routledge, New York.

Hockenmaier, Julia and Mark Steedman. 2002. Generative models for statistical parsing with Combinatory Categorial Grammar. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 335–342, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Hockenmaier, Julia and Mark Steedman. 2007. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Holton, David, Peter Mackridge, Irene Philippaki-Warburton, and Vassilios Spyropoulos. 2012. *Greek: A comprehensive grammar of the modern language*, 2nd edition. Routledge, London.

Hopper, Paul J. 1982. *Tense-aspect: Etween Semantics & Pragmatics: Containing the Contributions to a Symposium on Tense and Aspect, held at UCLA, May 1979*, volume 1. John Benjamins Publishing, Amsterdam/Philadelphia.

Howell, Kristen. 2020. *Inferring Grammars from Interlinear Glossed Text: Extracting Typological and Lexical Properties for the Automatic Generation of HPSG Grammars*. Ph.D. thesis, University of Washington.

Howell, Kristen, Emily M Bender, Michel Lockwood, Fei Xia, and Olga Zamaraeva. 2017. Inferring case systems from IGT: Impacts and detection of variable glossing practices. pages 67–75.

Howell, Kristen and Olga Zamaraeva. 2018. Clausal modifiers in the Grammar Matrix. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2939–2952.

Howell, Kristen, Olga Zamaraeva, and Emily M Bender. 2018. Nominalized clauses in the Grammar Matrix. In *Proceedings of the 25th International Conference on Head-Driven Phrase Structure Grammar, University of Tokyo*.

Indurkhya, Sagar. 2020. Inferring Minimalist grammars with an SMT-solver. In *Proceedings of the Society for Computation in Linguistics*, volume 3.

Inman, David. 2015. Pronoun incorporation in Matsigenka. Unpublished Manuscript, available at http://compling.hss.ntu.edu.sg/events/2015-hpsg/pdf/Inman.pdf, Accessed 2022-05-06.

Inman, David. 2019a. *Multi-predicate Constructions in Nuuchahnulth*. Ph.D. thesis, University of Washington.

Inman, David. 2019b. Nuuchahnulth texts. University of Washington. Unpublished FieldWorks (FLEx) project. (Accessed March 2019).

Kaplan, Ronald M and Joan Bresnan. 1982. Lexical-functional grammar: A formal system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations (MIT Press Series on Cognitive Theory and Mental Representation)*, pages 173–281. The MIT Press, Cambridge.

Kate, Rohit J and Raymond J Mooney. 2006. Using string-kernels for learning semantic parsers. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 913–920. Association for Computational Linguistics.

Kate, Rohit J, Yuk Wah Wong, and Raymond J Mooney. 2005. Learning to transform natural to formal languages. In *Proceedings of the 1st AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pages 1062–1068.

Kaufman, Daniel, Husniya Khujamyorova, and Ross Perlin. 2020. Wakhi texts. *Digital collection managed by KRATYLOS*. Uploaded from www.elalliance.org, Wakhi. In Finkel, R. and Kaufman, D., Kratylos: Unified Linguistic Corpora from Diverse Data Sources. Uploaded April 28, 2020 and retrieved from https://www.cs.uky.edu/ raphael/ela/ on May 20 2020.

Kenesei, István, Robert M Vago, and Anna Fenyvesi. 2002. *Hungarian*, 1st edition. Routledge, London.

Klein, Dan and Christopher D Manning. 2001. Natural language grammar induction using a constituent-context model. In *Advances in neural information processing systems 14*, pages 35–42.

Klein, Dan and Christopher D Manning. 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Klein, Dan and Christopher D 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-04)*, pages 478–485, Barcelona, Spain.

Kornfilt, Jaklin. 1997. *Turkish*. Routledge, London.

Kratochvíl, František. 2007. *A grammar of Abui*. LOT, Utrecht.

Kratochvíl, František. 2019. Abui Corpus. Electronic Database: Unpublished toolbox project (accessed March 2019). Nanyang Technological University, Singapore.

Krotov, Alexander, Robert Gaizauskas, and Yorick Wilks. 1994. Acquiring a stochastic context-free grammar from the Penn Treebank. In *Proceedings of the Irish Conference on NLP, Dublin*.

Krotov, Alexander, Mark Hepple, Robert Gaizauskas, and Yorick Wilks. 1998. Compacting the Penn Treebank Grammar. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL-1998)*, pages 699–703, Montreal.

Kwiatkowski, Tom, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2011. Lexical generalization in CCG grammar induction for semantic parsing. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1512–1523. Association for Computational Linguistics.

Letcher, Ned and Timothy Baldwin. 2013. Constructing a phenomenal corpus: Towards detecting linguistic phenomena in precision grammars. In *Proceedings of the Workshop on High-level Methodologies for Grammar Engineering at ESSLLI 2013*, pages 25–36.

Li, Charles N and Sandra A Thompson. 1989. *Mandarin Chinese: A functional reference grammar*. University of California Press, Berkeley/Los Angeles.

Lockwood, Michael. 2016. Automated gloss mapping for inferring grammatical properties. Master's thesis, University of Washington.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119. Association for Computational Linguistics.

Master, Alfred. 1946. The zero negative in Dravidian. *Transactions of the Philological Society*, 45(1):137–155.

Meira, Sérgio. 2020. Hixkaryana lexicon and texts. Unpublished Toolbox project. (Accessed March 2020).

Michael, Lev, Christine Beier, Zachary O'Hagan, (compilers), Haroldo Vargas, José Vargas, and (authors). 2013. Matsigenka text corpus (version june 2013; FLEx database and LaTeX interlinear output).

Michael, Lev David. 2008. *Nanti evidential practice: Language, knowledge, and social action in an Amazonian society*. Ph.D. thesis, University of Texas Austin.

Miestamo, Matti. 2008. *Standard negation: The negation of declarative verbal main clauses in a typological perspective*, volume 31. Walter de Gruyter, Berlin.

Miyaoka, Osahito. 2012. *A Grammar of Central Alaskan Yupik (CAY)*, volume 58. Walter de Gruyter, Berlin.

Monachesi, Paola. 1996. *A grammar of Italian clitics*. ITK Dissertations Series 1996-1.

Müller, Stefan. 1999. *Deutsche Syntax deklarativ. Head-Driven Phrase Structure Grammar für das Deutsche (Linguistische Arbeiten 394)*. Max Niemeyer, Tübingen.

Müller, Stefan. 2015. The CoreGram project: Theoretical linguistics, theory development and verification. *Journal of Language Modelling*, 3(1):21–86.

Müller, Stefan, Anne Abeillé, Robert D. Borsley, and Jean-Pierre Koenig, editors. 2021. *Head-Driven Phrase Structure Grammar: The handbook (Empirically Oriented Theoretical Morphology and Syntax 9)*. Language Science Press, Berlin. https://doi.org/10.5281/zenodo.5543318.

Muysken, Pieter, Harald Hammarström, Olga Krasnoukhova, Neele Müller, Joshua Birchall, Simon van de Kerke, Loretta O'Connor, Swintha Danielsen, Rik van Gijn, and George Saad, editors. 2016. *South American Indigenous Language Structures (SAILS) Online*. Max Planck Institute for the Science of Human History. Available at https://sails.clld.org, Accessed 2022-05-04.

Newman, Paul. 2000. *The Hausa language: An encyclopedic reference grammar*. Yale University Press, New Haven.

Nielsen, Elizabeth and Emily M Bender. 2018. Modeling adnominal possession in multilingual grammar engineering. In *Proceedings of the 25th International Conference on Head-Driven Phrase Structure Grammar, University of Tokyo*, pages 140–153, Stanford. CSLI Publications.

Nielsen, Elizabeth K. 2018. Modeling adnominal possession in the LinGO Grammar Matrix. Master's thesis, University of Washington.

Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Noji, Hiroshi, Yusuke Miyao, and Mark Johnson. 2016. Using left-corner parsing to encode universal structural constraints in grammar induction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 33–43.

Nordlinger, Rachel. 1998. *A Grammar of Wambaya, Northern Australia*. Pacific Linguistics, Canberra.

Oepen, Stephan. 2001. [incr tsdb()] — Competence and performance laboratory. User manual. Technical report, Computational Linguistics — Saarland University, Saarbrücken.

Oepen, Stephan, Kristina Toutanova, Stuart Shieber, Chris Manning, Dan Flickinger, and Thorsten Brants. 2002. The LinGO Redwoods treebank. Motivation and preliminary applications. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei.

O'Hagan, Zachary. 2018. The syntax of Matsigenka object-marking. *Berkeley Papers in Formal Linguistics*, 1(1).

Packard, Woodley. 2015. Full Forest Treebanking. Master's thesis, University of Washington.

Pollard, Carl and Ivan A Sag. 1994. *Head-Driven Phrase Structure Grammar (Studies in Contemporary Linguistics)*. University of Chicago Press, Chicago.

Poulson, Laurie. 2011. Meta-modeling of tense and aspect in a cross-linguistic grammar engineering platform. *University of Washington Working Papers in Linguistics (UWWPL)*, 28.

Pustejovsky, James, José M Castaño, Robert Ingria, Roser Saurí, Robert J Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the IWCS-5 Fifth International Workshop on Computational Semantics.*

Rogers, Chris. 2010. Fieldworks language explorer (FLEx) 3.0. *Language Documentation & Conservation*, 4:78–84.

Saleem, Safiyyah. 2010. Argument optionality: A new library for the Grammar Matrix customization system. Master's thesis, University of Washington.

Saleem, Safiyyah and Emily M Bender. 2010. Argument optionality in the LinGO Grammar Matrix. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1068–1076. Association for Computational Linguistics.

Sanchez, Jose, Alex Trueman, Maria Florez Leyva, Santos Leyva Alvarez, Mercedes Tubino Blanco, Hyun-Kyoung Jung, Louise St. Amour, and Heidi Harley. 2015. *An Introduction to Hiaki Grammar.* University of Arizona Press, Tucson.

Sarveswaran, Kengatharaiyer, Gihan Dias, and Miriam Butt. 2019. Using meta-morph rules to develop morphological analysers: A case study concerning Tamil. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 76–86, Dresden. Association for Computational Linguistics.

Schikowski, Robert. 2013. *Object-conditioned differential marking in Chintang and Nepali.* Ph.D. thesis, University of Zurich.

Schrock, Terrill B. 2014. *A Grammar of Ik (Icé-tód): Northeast Uganda's Last Thriving Kuliak Language.* LOT, Utrecht.

Shi, Haoyue, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1842–1861, Florence. Association for Computational Linguistics.

Siegel, Melanie, Emily M Bender, and Francis Bond. 2016. *Jacy: An implemented grammar of Japanese.* CSLI Publications, Stanford.

Siewierska. 2004. *Person.* Cambridge University Press, Cambridge.

SIL International. 2015. Field Linguist's Toolbox. Lexicon and corpus management system with a parser and concordancer; Available at https://software.sil.org/fieldworks/download/, Accessed 2022-05-04.

Simov, Kiril. 2002. Grammar extraction and refinement from an HPSG corpus. In *Proceedings of the ESSLLI Workshop on Machine Learning Approaches in Computational Linguistics*, pages 38–55.

Smith, Noah A 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. Association for Computational Linguistics.

Sneddon, James Neil, K Alexander Adelaar, Dwi N Djenar, and Michael Ewing. 2012. *Indonesian: A comprehensive grammar.* Routledge, Oxfordshire.

Sohn, Ho-Min. 1994. *Korean: A Descriptive Grammar.* Routledge, London/New York.

Stabler, Edward. 1996. Derivational minimalism. In *International Conference on Logical Aspects of Computational Linguistics*, pages 68–95, Berlin/Heidelberg. Springer.

Sulkala, Helena and Merja Karjalainen. 1992. *Finnish.* Routledge, London/New York.

Sylak-Glassman, John, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680.

Thieberger, Nick. 2006a. Dictionary and texts in South Efate. *Digital collection managed by PARADISEC [Open Access].* (Accessed March 2019).

Thieberger, Nick. 2006b. *A grammar of South Efate: an Oceanic language of Vanuatu*, volume 33. University of Hawai'i Press, Honolulu.

de Urbina, Jon Ortiz. 1989. *Parameters in the grammar of Basque: A GB approach to Basque syntax.* Foris, Dordrecht/Providence.

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

Xia, Fei. 1999. Extracting tree adjoining grammars from bracketed corpora. In *Proceedings of 5th Natural Language Processing Pacific Rim Symposium (NLPRS-1999)*, Beijing.

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

Xia, Fei, William D. Lewis, Michael Wayne Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey, and Emily M Bender. 2016. Enriching a massively multilingual database of interlinear glossed text. *Language Resources and Evaluation*, 50:321–349.

Zamaraeva, Olga. 2016. Inferring morphotactics from interlinear glossed text: combining clustering and precision grammars. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 141–150.

Zamaraeva, Olga. 2021. *Assembling Syntax: Modeling Constituent Questions in a Grammar Engineering Framework*. Ph.D. thesis, University of Washington.

Zamaraeva, Olga, Kristen Howell, and Emily M Bender. 2019a. Handling cross-cutting properties in automatic inference of lexical classes: A case study of Chintang. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, volume 1 Papers, pages 28–38, Honolulu, Hawai'i.

Zamaraeva, Olga, Kristen Howell, and Emily M Bender. 2019b. Modeling clausal complementation for a grammar engineering resource. In *Proceedings of the Society for Computation in Linguistics*, volume 2, page Article 6.

Zamaraeva, Olga, František Kratochvíl, Emily M Bender, Fei Xia, and Kristen Howell. 2017. Computational support for finding word classes: A case study of Abui. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 130–140.

Zamaraeva, Olga, TJ Trimble, Kristen Howell, Michael Wayne Goodman, Antske Fokkens, Guy Emerson, Chris Curtis, and Emily M Bender. forthcoming. 20 years of the Grammar Matrix: Cross-linguistic hypothesis testing of increasingly complex interactions. *Journal of Language Modeling*.

Zanzotto, Fabio Massimo, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. 2020. KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267.

Zhang, Songyang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, and Jiebo Luo. 2021. Video-aided unsupervised grammar induction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1513–1524, Online. Association for Computational Linguistics.

Zhao, Yanpeng and Ivan Titov. 2020. Visually grounded compound PCFGs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4369–4379, Online. Association for Computational Linguistics.

Zwicky, Arnold, Joyce Friedman, Barbara C. Hall, and D.E. Walker. 1965. The MITRE syntactic analysis procedure for transformational grammars. In *Proceedings Fall Joint Computer Conference*, volume 67, Pt 1, pages 317–326.

Zwicky, Arnold M and Geoffrey K Pullum. 1983. Cliticization vs. inflection: English n't. *Language*, 59(3):502–513.

Zymla, Mark-Matthias. 2017. Comprehensive annotation of cross-linguistic variation in tense and aspect categories. In *IWCS 2017-12th International Conference on Computational Semantics-Long papers*.

A Data Repositories

Alaskan Native Languages Archive (ANLA)
https://www.uaf.edu/anla/

Archive of Indigenous Languages in Latin America (AILLA)
http://www.ailla.utexas.org/site/welcome.html

Endangered Languages Archive (ELAR)
http://elar.soas.ac.uk/

Kaipuleohone
https://scholarspace.manoa.hawaii.edu/handle/10125/4250

Kratylos
https://www.kratylos.org/~kratylos/home.cgi

Multi-CAST
https://multicast.aspra.uni-bamberg.de/

ODIN
http://depts.washington.edu/uwcl/odin/

Pacific and Regional Archive for Digital Sources (PARADISEC)
http://www.paradisec.org.au/

B Code and Project Repositories

ACE
http://sweaglesw.org/linguistics/ace/

AGGREGATION, BASIL
https://git.ling.washington.edu/agg

DELPH-IN
www.delph-in.net

INTENT
https://github.com/rgeorgi/INTENT2

FFTB
http://moin.delph-in.net/FftbTop

Grammar Matrix
http://matrix.ling.washington.edu/index.html

MOM
https://git.ling.washington.edu/agg/mom

Xigt
https://github.com/xigt/xigt

C Languages, Corpora and Descriptive Resources

The languages and corpora used for this research are listed in the table below, together with any descriptive resources we consulted during BASIL's development and evaluation.

	Language	iso	Corpus	Descriptive Resource
	Development			
1	Abui	abz	Kratochvíl 2019	Kratochvíl 2007
2	Chintang	ctn	Bickel et al. 2013b	Schikowski 2013
3	Matsigenka	mcb	Michael et al. 2013	Michael 2008
4	Nuuchahnulth	nuk	Inman 2019b	Inman 2019a
5	Wambaya	wmb	Nordlinger 1998	Nordlinger 1998
6	Haiki	yaq	Harley 2019	Sanchez et al. 2015 Dedrick and Casad 1999
7	Lezgi	lez	Donet 2014b	Donet 2014a
8	Meithei	mni	Chelliah 2019	Chelliah 2011
9	Tsova-Tush	bbl	Hauk 2016–2019	Hauk and Harris forthcoming Hauk 2020
	Consulted			
10	Bardi	bcj	Bowern 2012	Bowern 2012
11	Ik	ikx	Schrock 2014	Schrock 2014
12	Old Javanese	jav	Acri 2018	
13	Yup'ik	esu	Miyaoka 2012	Miyaoka 2012
14	Basque	eus	Xia et al. 2016	de Urbina 1989
15	Dutch	nld	Xia et al. 2016	Booij 2002
16	Finnish	fin	Xia et al. 2016	Sulkala and Karjalainen 1992
17	Greek	ell	Xia et al. 2016	Holton et al. 2012
18	Hausa	hau	Xia et al. 2016	Newman 2000
19	Hungarian	hun	Xia et al. 2016	Kenesei et al. 2002
20	Indonesian	ind	Xia et al. 2016	Sneddon et al. 2012
21	Italian	ita	Xia et al. 2016	Monachesi 1996
22	Japanese	jpn	Siegel et al. 2016 Xia et al. 2016	Siegel et al. 2016 Hinds 1986
23	Korean	kor	Xia et al. 2016	Sohn 1994
24	Mandarin	cmn	Xia et al. 2016	Li and Thompson 1989
25	Polish	pol	Xia et al. 2016	
26	Russian	rus	Xia et al. 2016	
27	Turkish	tur	Xia et al. 2016	Kornfilt 1997
	Held Out			
28	Arapaho	arp	Cowell 2018	Cowell and Moss Sr 2011
29	Hixkaryana	hix	Meira 2020	
30	South Efate	erk	Thieberger 2006a	Thieberger 2006b
31	Titan	ttv	Bowern 2019	Bowern 2011
32	Wakhi	wbl	Kaufman et al. 2020	