Inferring Grammars from Interlinear Glossed Text: Extracting Typological and Lexical Properties for the Automatic Generation of HPSG Grammars

Kristen Howell

University of Washington

June 8, 2020

Inferring Grammars from IGT

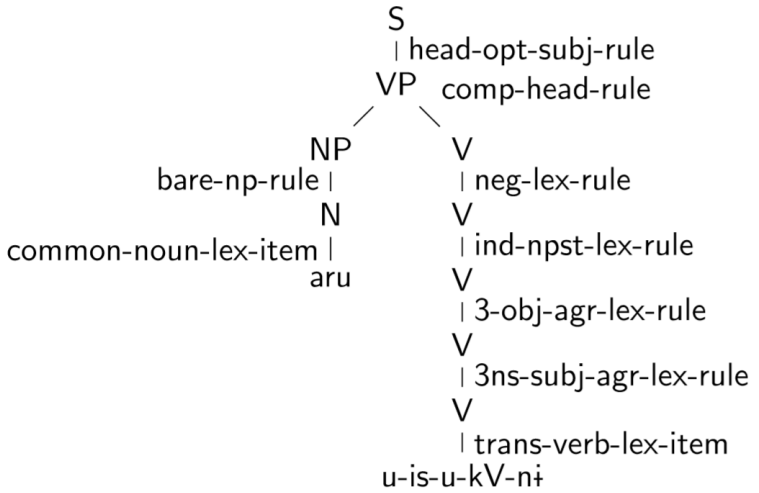
Chintang [ctn] (Bickel et al., 2013)

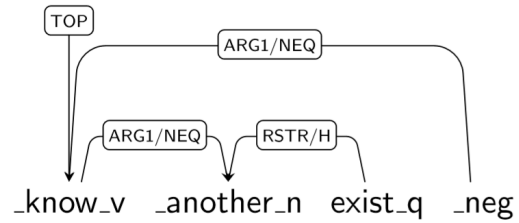
Umaa taktukVnɨ.
u-ma-a takt-u-kV-nɨ
3sPOSS-mother-ERG.A hold-3P-IND.NPST-NEG
'The mother does not hold it.'

Dodum tase.
dodum tas-e
talk reach-IND.NPST
'Enough was said.'

Tamakhu thua.
tamakhu thu-a
tobacco smoke-IMP
'Smoke some tobacco!'

Aru uisukVnɨ.
aru u-is-u-kV-nɨ
another 3nsS/A-know-3P-IND.NPST-NEG
'They did not know another [language].'





_know_v (ARG0 {SF *prop*, tense *npst*, aspect *ind*}, ARG1 {per *3*, num *ns* }, ARG2 {per *3*})

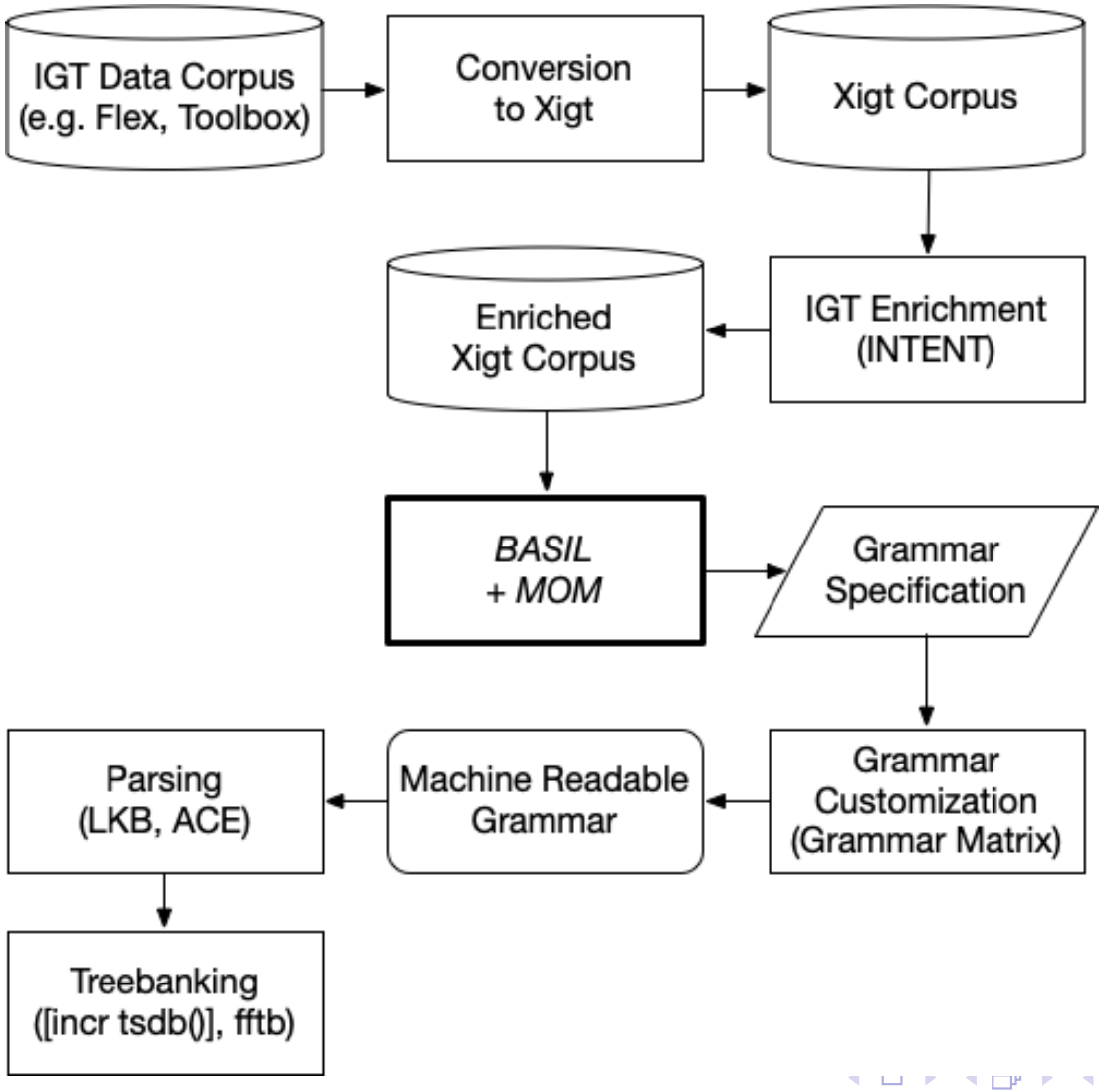
Research Question

Drawing on the linguistic information encoded in interlinear glossed text and stored syntactic analysis from a grammar customization toolkit, can machine-readable grammars can be automatically generated by inferring lexical, morphological and syntactic properties about the language from existing linguistic corpora?

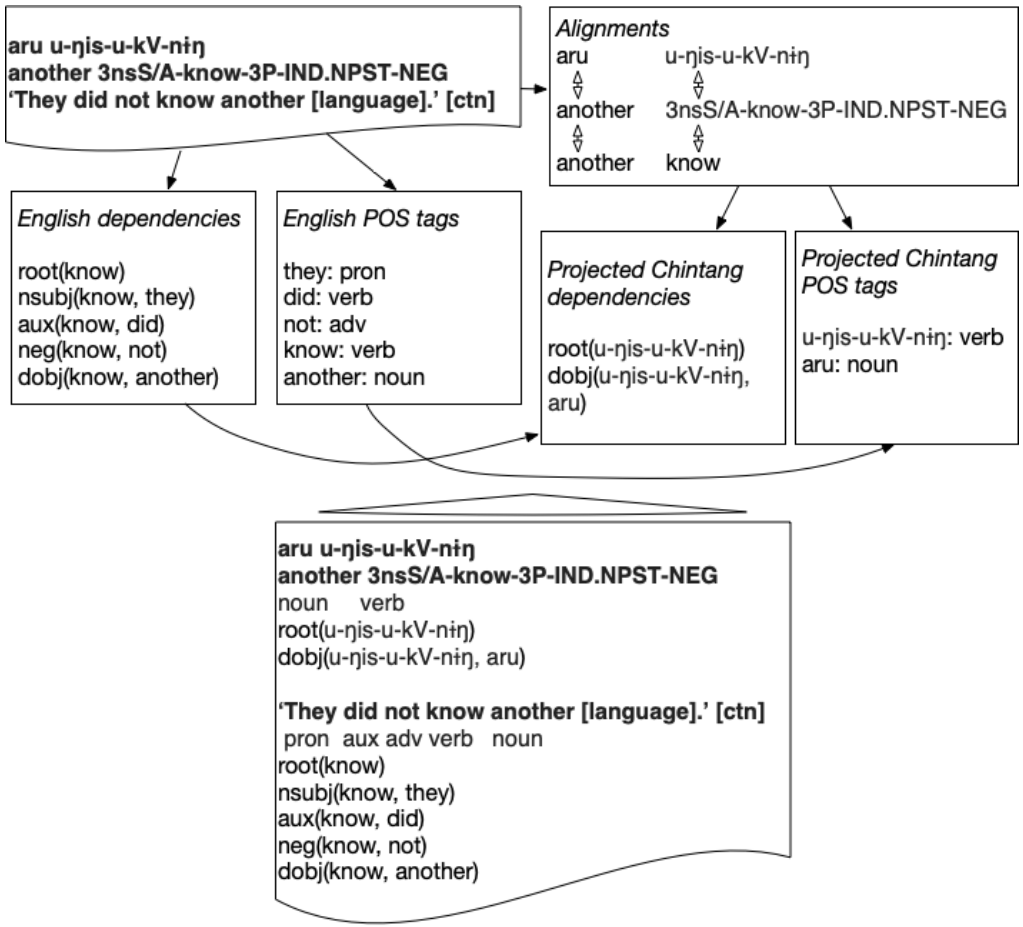
My Contributions

- BASIL: Building Analyses from Syntactic Inference in Low-resource languages
- Integrate existing inference modules into a single system
- Add modules for additional phenomena
- End to end testing on 9 development and 5 held-out languages

AGGREGATION Pipeline



Enriched Xigt



(Goodman et al., 2015; Georgi, 2016)

Grammar Specification

```
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
```

(Bender et al., 2010)

Grammar

```
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.
```

(Bender et al., 2002, 2010)

Previous AGGREGATION Grammar Inference

- Morphotactic inference with MOM (Wax, 2014; Zamaraeva, 2016; Zamaraeva et al., 2017)
  - nouns, verbs
- Syntactic inference
  - word order (Bender et al., 2013)
  - case system (Bender et al., 2013; Howell et al., 2017)
  - transitivity, case-frame (Bender et al., 2014; Zamaraeva et al., 2019)

Grammar Inference with BASIL

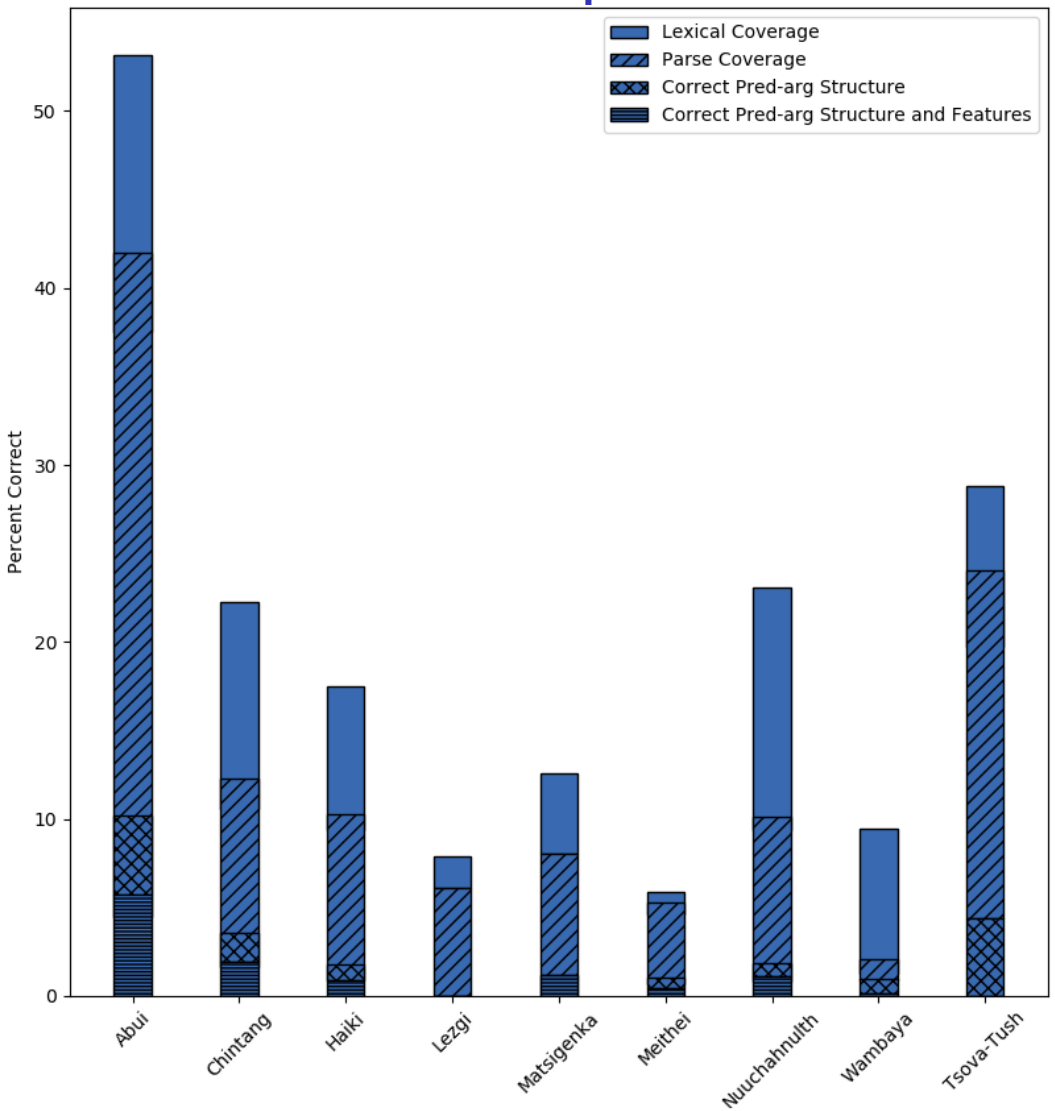
- Verbs (trans, intrans)
- Nouns (pronouns, other)
- Auxiliaries
- Case-marking adpositions
- Determiners
- Features: PNG, TAM
- Word order
- Case
- Argument optionality
- Sentential negation
- Coordination

Development Languages

Features inferred by BASIL

Phenomenon	# possible values	# 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	135	98

Data-Driven Development



Language [iso]	Ambiguity
Abui	2195
Chintang	5562
Haiki	161
Lezgi	10419
Matsigenka	2333
Meithei	3722
Nuuchahnulth	265
Wambaya	4
Tsova-Tush	3418

Held-out Languages

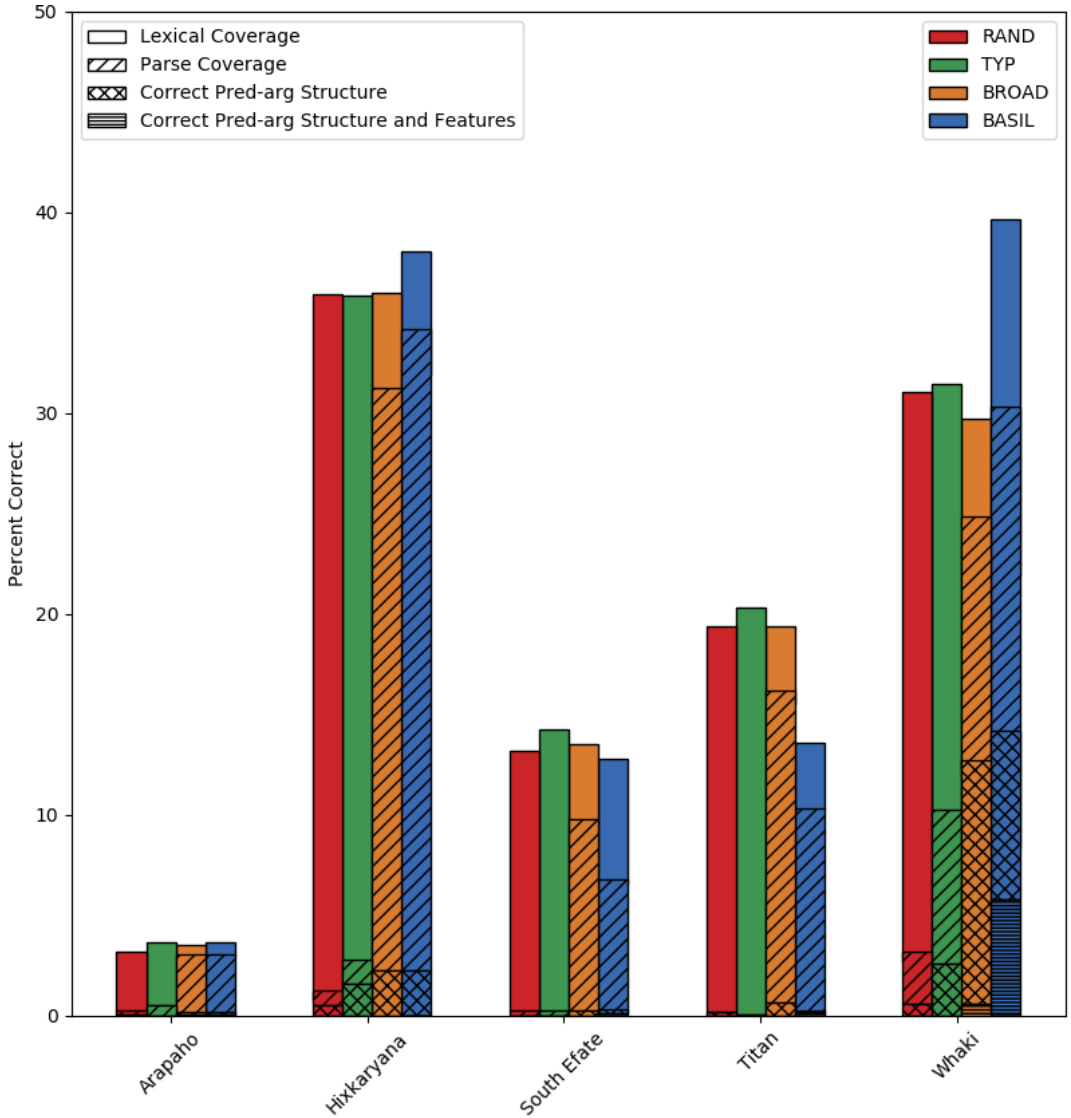
Evaluation Metrics

- **lexical coverage** – the proportion of sentences for which the grammar has a analysis for each word in the sentence
- **parse coverage** – the proportion of sentences the grammar parses
- **correct predicate-argument structure** – the proportion of sentences for which there is a parse with the correct predicate-argument structure
- **correct predicate-argument structure and semantic features** – the proportion of sentences for which there is a parse with the correct predicate-argument structure as well as the appropriate PNG and TAM features on those arguments
- **ambiguity** – the average number of results per sentence that parses

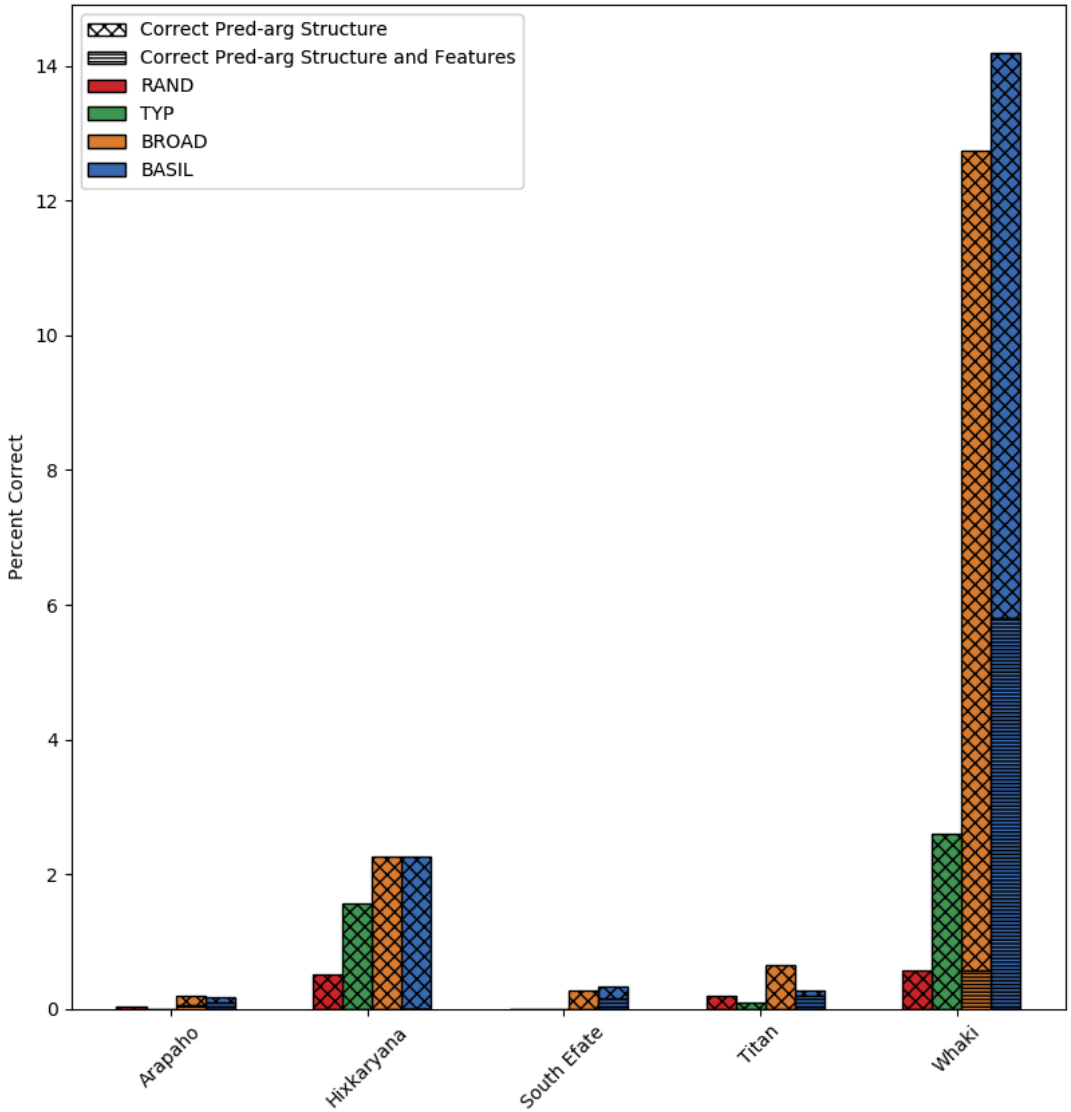
Baseline Systems

- Lexicon and morphological rules from MOM (Wax, 2014; Zamaraeva, 2016; Zamaraeva et al., 2017)
- Otherwise syntactically naive
  - BROAD-COV The grammar specifications expected to parse the most sentences
  - TYP The grammar specifications that are the typologically most common
  - RAND Grammar specifications chosen at random

Results


Lexical Coverage
Parse Coverage
Correct Pred-arg Structure
Correct Pred-arg Structure and Features
RAND
TYP
BROAD
BASIL
Percent Correct
0
10
20
30
40
50
Arapaho
Hixkaryana
South Efate
Titan
Whaki

Results


Correct Pred-arg Structure
Correct Pred-arg Structure and Features
RAND
TYP
BROAD
BASIL
Percent Correct
0
2
4
6
8
10
12
14
Arapaho
Hixkaryana
South Efate
Titan
Whaki

Ambiguity

Language	basil	broad-cov	typ	rand
Arapahoe [arp]	145	936	4	3
Hixkaryana [hix]	5642	15596	2	6
South Efate [erk]	126379	9759	2	4
Titan [ttv]	595	6201	2	1
Whaki [wbl]	10	26	1	2.5

Conclusion

- Developed a grammar inference system that starts with an IGT corpus and produces a machine-readable, HPSG grammar
- Developed algorithms based on a broad range of typologically diverse languages, doing end-to-end testing on 9
- Evaluated cross-linguistic generalizability on 5 previously unconsidered langauges
- Provide a starting point for broader coverage grammars

References I

Željko Agić, 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.

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

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. Research on Language & Computation, 8:1–50. ISSN 1570-7075.

Emily M. Bender, 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 Hawaii Press, Honolulu.

Emily M. Bender, Michael Wayne Goodman, Joshua Crowgey, and Fei Xia. August 2013. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 74–83, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2710.

Emily M. Bender, Joshua Crowgey, Michael Wayne Goodman, and Fei Xia. June 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, Maryland, USA, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-2206.

References II

Balthasar Bickel, Martin Gaenszle, Novel Kishore Rai, Vishnu Singh Rai, Elena Lieven, Sabine Stoll, G. Banjade, T. N. Bhatta, N Paudyal, J Pettigrew, and M Rai, I. P.and Rai. 2013. Durga. URL `https://corpus1.mpi.nl/qfs1/media-archive/dobes_data/ChintangPuma/Chintang/Narratives/Annotations/durga_exp.tbt`. Accessed: 2013.

Gosse Bouma, 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.

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

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

Luis Morgado da Costa, 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.

Östen Dahl. 1979. Typology of sentence negation. Linguistics, 17:79–106.

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

Matthew S Dryer. 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.

References III

Dan Flickinger. 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, CA.

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

Michael Wayne Goodman, 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.

Wenjuan Han, 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.

Lars Hellan. July 2010. From descriptive annotation to grammar specification. In Proceedings of the Fourth Linguistic Annotation Workshop, pages 172–176, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W10-1826.

Lars Hellan, 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, number 085, pages 435–439. Linköping University Electronic Press.

Kristen Howell, Emily M. Bender, Michael Lockwood, Fei Xia, and Olga Zamaraeva. 2017. Inferring case systems from IGT: Impacts and detection of variable glossing practices. ComputEL-2, pages 67–75.

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

References IV

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

Bevan Keeley Jones, Sharon Goldwater, and Mark Johnson. 2013. Modeling graph languages with grammars extracted via tree decompositions. In Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing, pages 54–62.

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

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

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

Ned Letcher and Timothy Baldwin. 2013. Constructing a phenomenal corpus: Towards detecting linguistic phenomena in precision grammars. In Workshop on High-level Methodologies for Grammar Engineering@ ESSLLI 2013, page 25.

Alfred Master and I General. 1946. The zero negative in dravidian. Transactions of the Philological Society, 45(1): 137–155.

Lev Michael, 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).

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

References V

Hiroshi Noji, 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.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Chris Manning, Dan Flickinger, and Thorsten Brants. 2002. The LinGO Redwoods treebank. Motivation and preliminary applications. Taipei, Taiwan, 2002.

Carl Pollard and Ivan A Sag. 1994. Head-Driven Phrase Structure Grammar. University of Chicago Press.

Chris Rogers. 2010. Fieldworks language explorer (flex) 3.0. Language Documentation & Conservation, 4.

Vee Satayamas and Asanee Kawtrakul. 2004. Wide-coverage grammar extraction from thai treebank. In Proceedings of Papillon 2004 Workshops on Multilingual Lexical Databases.

SIL International. Field Linguist's Toolbox. Lexicon and corpus management system with a parser and concordancer; URL: http://www-01.sil.org/computing/toolbox/documentation.htm, 2015.

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

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

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

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

References VI

Olga Zamaraeva. 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.

Olga Zamaraeva, 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.

Olga Zamaraeva, Kristen Howell, and Emily M Bender. 2019. 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.

Luke Zettlemoyer and Michael Collins. 2007. Online learning of relaxed ccg grammars for parsing to logical form. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 678–687.

Luke S Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In Proceedings of the Conference on Uncertainty in Artificial Intelligence.