Towards Creating Precision Grammars from Interlinear Glossed Text

Emily M. Bender    Michael W. Goodman
Joshua Crowgey    Fei Xia
{ebender, goodmami, jcrowgey, fxia}@uw.edu

University of Washington

8 August 2013
Motivation:

- Many languages—an important kind of cultural heritage—are dying
- Language documentation takes a lot of time
- Linguists do the hard work and provide IGT, dictionaries, etc.
- Digital resources expand the accessibility and utility of documentation efforts (Nordhoff and Poggeman, 2012)
- Implemented grammars are beneficial for language documentation (Bender et al., 2012)
- We want to automatically create grammars based on existing descriptive resources (namely, IGT)
Example IGT from Shona (Niger-Congo, Zimbabwe)

(1) Ndakanga ndakatenga muchero
    ndi-aka-nga ndi-aka-teng-a mu-chero
    SBJ.1SG-RP-AUX SBJ.1SG-RP-buy-FV CL3-fruit

‘I had bought fruit.’ [sna] (Toews, 2009:34)
Background
The Grammar Matrix (Bender et al., 2002; 2010)

- Pairs a core grammar of near-universal types with a repository of implemented analyses
- Customization system transforms high-level description ("choices file") to an implemented HPSG (Pollard and Sag, 1994) grammar
- Customized grammars are ready for further hand-development
- Grammars can be used to parse and generate sentences, giving detailed derivation trees and semantic representations
- Front-end of the customization system is a linguist-friendly web-based questionnaire
Does your language have determiners (as independent words)? ○ yes ○ no
If so, what is the order of determiners with respect to nouns? ○ Noun-Det ○ Det-Noun

Does your language have auxiliary verbs? ○ yes ○ no
If so, please specify the following auxiliary properties:

**Word Order:** Does an auxiliary verb appear before or after its complement?
○ before
○ after

**Complements:** The complements of auxiliaries are:
○ saturated sentences
○ VPs, raising the subject
○ Vs, raising all of its arguments (argument composition)

**Figure:** The Grammar Matrix Questionnaire: Word Order
**Noun type 11:**

Type name: f-common-noun-non-human

Supertypes: non-pro (noun2), non-human (noun6), feminine (noun8)

Features:

- Name: person Value: 3rd

For nouns of this type, a determiner is: ◯ obligatory ◯ optional ◯ impossible

Stems:

- Spelling: mazzita, Predicate: _blutwurst_n_rel
- Spelling: ittra, Predicate: _letter_n_rel
- Spelling: universita, Predicate: _university_n_rel

---

**Figure:** The Grammar Matrix Questionnaire: Lexicon
**ODIN and RiPLes** (Lewis, 2006; Xia and Lewis, 2008)

- RiPLes parses the English line, and **projects** structure through the gloss line to the original language line

![Diagram](image-url)

**Figure**: Welsh IGT with alignment and projected syntactic structure
ODIN and RiPLeS (continued)

- Xia and Lewis (2008) did typological property inference from CFG rules extracted from projected structures
- **Question**: Can this process be adapted to customize Matrix grammars?
Methodology
Towards automatic grammar creation:

1. Word-order inference (of 10 word order types)
2. Case system inference (of 8 case system types)

Methodology overview:

- Obtain a corpus of IGT for a language
- Find observed (i.e. overt) patterns
- Analyze pattern distributions to infer underlying pattern/system

Data:

- Student-curated testsuites
- Avg 92 sentences per language (min: 11; max: 251)
- Clean and representative, but small

**Question:** The more voluminous/clean/representative the IGT, the better the model?
Word order

- Goal: Infer best word-order choice from projected structure
- Baseline: most frequent word-order (SOV) according to WALS (Haskelmath et al., 2008)
- For each IGT, get a projected parse from RiPLEs with functional and part-of-speech tags (SBJ, OBJ, VB)
- Extract observed binary word orders (S/V, O/V, S/O) as relative linear order
- Calculate observed word order coordinates on three axes: SV–VS; OV–VO; SO–OS
- Compare overall observed word-order to canonical word-orders types (SOV, OSV, SVO, OVS, VSO, VOS, V-initial, V-final, V2, Free)
- Select the closest canonical word-order by Euclidean distance
Figure: Three axes of basic word order and the positions of canonical word orders.
Word-order Results

| Dataset | # lgs | BASELINE | Inferred WO |
|---------|-------|----------|-------------|
| DEV1    | 10    | 0.200    | 0.900       |
| DEV2    | 10    | 0.100    | 0.500       |
| TEST    | 11    | 0.091    | 0.727       |

Table: Accuracy of word-order inference; BASELINE is ‘SOV’
Error Analysis:

- Noise (e.g. misalignments, non-standard IGT)
- Freer word orders (e.g. most-frequent vs unmarked)
- Unaligned elements (e.g. auxiliaries)
Case Systems—two approaches (and most-freq baseline):

**Case-gram presence (GRAM)**

- Look for case grams (NOM, ACC, ERG, ABS) on words
- Select system based on presence of certain grams

| Case system | Case grams present |
|-------------|--------------------|
|              | NOM ∨ ERG ∨ ACC    |
| none        | ✓                  |
| nom-acc     | ✓                  |
| erg-abs     | ✓                  |
| split-v     | ✓                  |

*Conditioned on V*

**Gram distribution (SAO)**

- Get gram lists for SBJ or OBJ
  - Transitive: $A_g, O_g$
  - Intransitive: $S_g$
  - Most frequent gram expected to be case-related

| Case system | Top grams |
|-------------|-----------|
| none        | $S_g=\neg A_g \neq O_g$, or $S_g \neq A_g \neq O_g$ and $S_g, A_g, O_g$ also present on the other argument types |
| nom-acc     | $S_g=\neg A_g \neq O_g$ |
| erg-abs     | $S_g \neq A_g \neq O_g$ |
| split-s     | $S_g \neq A_g \neq O_g$, and $A_g$ and $O_g$ both present on S list |
Case-system Results

| Dataset | # lgs | BASELINE | GRAM   | SAO   |
|---------|-------|----------|--------|-------|
| DEV1    | 10    | 0.400    | 0.900  | 0.700 |
| DEV2    | 10    | 0.500    | 0.900  | 0.500 |
| TEST    | 11    | 0.455    | 0.545  | 0.545 |

**Table:** Accuracy of case-marking inference; BASELINE is ‘none’
Error Analysis:

- **GRAM**: Non-standard case grams (e.g. “SBJ”)
- **SAO**: Unaligned elements (e.g. Japanese case markers)
- **SAO**: Top gram not for case (e.g. “3SG”)
- **Both**: Noise (e.g. erroneous annotation)
Conclusion
Summary:

- Language documentation is greatly facilitated with computational resources, including implemented grammars
- We show some first steps at inducing grammars from traditional kinds of resources
  - Inferring word order from projected syntax
  - Inferring case systems from case grams
- Initial results are promising, and informative
- ... but we’re still a long way from producing full grammars
Looking forward:

- Identify and account for noise
- Use larger data sets
- Analyze more phenomena
- Extrinsic evaluation techniques
Thank you!
Emily M. Bender, Scott Drelishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation*, pages 1–50. URL http://dx.doi.org/10.1007/s11168-010-9070-1, 10.1007/s11168-010-9070-1.

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

Emily M. Bender, 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.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors. 2008. *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. Http://wals.info.

William D. Lewis. 2006. ODIN: A model for adapting and enriching legacy infrastructure. In *Proceedings of the e-Humanities Workshop, Held in cooperation with e-Science*. Amsterdam.

Sebastian Nordhoff and Karl-Ludwig G. Poggeman, editors. 2012. *Electronic Grammaticography*. University of Hawaii Press, Honolulu.

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

Carmela Toews. 2009. The expression of tense and aspect in Shona. *Selected Proceedings of the 39th Annual Converence on African Linguistics*, pages 32–41.

Fei Xia and William D. Lewis. 2008. Repurposing theoretical linguistic data for tool development and search. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 529–536. Hyderabad, India.
Grammar Matrix choices file (Maltese):

section=word-order
word-order=free
has-dets=yes
noun-det-order=det-noun
has-aux=yes
aux-comp-order=before
aux-comp=v
multiple-aux=no
...
noun8_name=feminine
  noun8_feat1_name=gender
  noun8_feat1_value=fem
noun9_name=m-proper-noun
  noun9_supertypes=noun2, noun3, noun5, noun7
  noun9_feat1_name=person
  noun9_feat1_value=3rd
noun9_det=imp
  noun9_stem1_orth=Pawlu
  noun9_stem1_pred=_named_rel
  noun9_stem2_orth=Ganni
  noun9_stem2_pred=_name_rel
Grammar Matrix Libraries

- **Word Order**
  - SOV, OSV, SVO, OVS, VSO, VOS, V-initial, V-final, V2, Free
- **Number**
- **Person**
- **Gender**
- **Case** (and Direct-Inverse)
  - None, Nom-Acc, Erg-Abs, Tripartite
  - Split-S, Fluid-S, Split-V, Split-N, Focus
- **Tense, Aspect, and Mood**
- **Sentential Negation**
- **Coordination**
- **Yes/no questions**
- **Information structure**
- **Argument Optionality**
- **Lexicon and Morphology**
## Data distribution:

| Sets          | DEV1 (n=10)          | DEV2 (n=10)          | TEST (n=11)         |
|---------------|----------------------|----------------------|---------------------|
| Size range    | 16–251               | 11–229               | 14–216              |
| Size median   | 91                   | 87                   | 76                  |
| Families      | Indo-European (4),   | Indo-European (3),   | Indo-European (2),  |
|               | Niger-Congo (2),     | Dravidian (2),       | Afro-Asiatic,       |
|               | Afro-Asiatic,        | Algic, Creole,       | Austro-Asiatic,     |
|               | Japanese,            | Niger-Congo,         | Austronesian,       |
|               | Nadahup,             | Quechuan,            | Arauan, Carib,      |
|               | Sino-Tibetan         | Salishan             | Karvelian,          |
|               |                      |                      | N. Caucasian,       |
|               |                      |                      | Tai-Kadai, Isolate  |