Partially Automated Grammar Engineering and Chintang

Emily M. Bender
University of Washington/Aarhus Universitet

Seminar für Allgemeine Sprachwissenschaft
Universität Zürich

10 July 2014
Acknowledgments

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

• Collaborators: Balthasar Bickel, Joshua Crowgey, Michael Goodman, Robert Schikowski, Fei Xia, Olga Zamarova

• The Chintang and Puma Documentation Project!
High-level Goals

• Create precision, implemented, morphosyntactic grammars in the context of language description

  • Both grammars and eventually treebanks (Bender et al 2012)

• Speed up the development of precision grammars, though:

  • Grammar code re-use (Grammar Matrix; Bender et al 2002, 2010)

  • Leveraging analyses encoded by field linguists, in Toolbox lexica and in IGT
Precision Grammars

• Map (near-)surface strings to semantic representations, and vice versa
  • Syntactic structures mostly a means to this end

• Encode a sharp notion of grammaticality

• Require models *ou tout se tient*

• In our case, use HPSG (Pollard & Sag 1994) and MRS (Copestake et al 2005)
Precision Grammars: Example

\[
\left\langle h_1, \right.
\begin{align*}
&h_3 : \text{din.n.day}(x_4), \\
&h_5 : \text{exist.q.rel}(x_4, h_6, h_7), \\
&h_6 : \text{khipt-u-kv-m.v.count}(e_2, x_9, x_4) \\
\{& h_6 = _q h_3 \} \right. \\
\right. \\
\Rightarrow \text{demo}
\]
Precision Grammars in Language Documentation: Wambaya [wmb]
Precision Grammars in Language Documentation: Wambaya [wmb]

- Wambaya [wmb] is a Mirndi language spoken in the West Barkly Tablelands region (Northern Territory) of Australia (Nordlinger 1998; Green & Nordlinger 2004)
Wambaya [wmb] is a Mirndi language spoken in the West Barkly Tablelands region (Northern Territory) of Australia (Nordlinger & Green 2004).
Precision Grammars in Language Documentation: Wambaya [wmb]

- Wambaya [wmb] is a Mirndi language spoken in the West Barkly Tablelands region (Northern Territory) of Australia (Nordlinger 1998; Green & Nordlinger 2004)
Precision Grammars in Language Documentation: Wambaya [wmb]

- Wambaya [wmb] is a Mirndi language spoken in the West Barkly Tablelands region (Northern Territory) of Australia (Nordlinger 1998; Green & Nordlinger 2004)

- Developed by hand (with jump start from the Grammar Matrix) on the basis of the analyses in Nordlinger 1998 (incl. IGT)
Precision Grammars in Language Documentation: Wambaya [wmb]

• Wambaya [wmb] is a Mirndi language spoken in the West Barkly Tablelands region (Northern Territory) of Australia (Nordlinger 1998; Green & Nordlinger 2004)

• Developed by hand (with jump start from the Grammar Matrix) on the basis of the analyses in Nordlinger 1998 (incl. IGT)

• Development set: 794 examples from Nordlinger 1998
Precision Grammars in Language Documentation: Wambaya [wmb]

• Wambaya [wmb] is a Mirndi language spoken in the West Barkly Tablelands region (Northern Territory) of Australia (Nordlinger 1998; Green & Nordlinger 2004)

• Developed by hand (with jump start from the Grammar Matrix) on the basis of the analyses in Nordlinger 1998 (incl. IGT)

• Development set: 794 examples from Nordlinger 1998

• Held-out test data: Narrative from Nordlinger 1998
Wambaya grammar scope

- Word order (2nd position auxiliary, discontinuous noun phrases)
- Argument optionality
- Linking of syntactic to semantic arguments
- Case (split ergativity)
- Agreement (verb-subject, verb-object, adj-noun)
- Lexical adverbs, including manner, time, location, and negation
- Derived event modifiers
- Derived nominal modifiers
- Lexical adjectives (demonstratives, possessives, numerals, others)
- Subordinate clauses (clausal complements, purposives, simultaneous and prior events)
- Verbless clauses: adjective, nouns, and adverbs functioning as predicates
- Illocutionary force: imperatives, declaratives and interrogatives, including wh-questions
- Coordination: of clauses and noun phrases
- Inalienable possession construction
- Secondary predicates
- Causatives of verbs and adjectives
Wambaya grammar development

(Bender 2008)
Wambaya grammar development

>95% of the work is Nordlinger’s original field work

(Bender 2008)
Wambaya grammar evaluation

- Held out test data “The two Eaglehawks”

- 72 sentences (orig text: 92, removed 20 seen sentences)

- Run twice: before and after adding lexical entries and adjusting morphophonology only

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>parsed</th>
<th>unparsed</th>
<th>average ambiguity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existing vocab</td>
<td>50%</td>
<td>8%</td>
<td>42%</td>
<td>10.62</td>
</tr>
<tr>
<td>w/added vocab</td>
<td>76%</td>
<td>8%</td>
<td>14%</td>
<td>12.56</td>
</tr>
</tbody>
</table>
Grammar Matrix

• Facilitate grammar development & grammar interoperability through code-reuse

• Combine breadth of typological analysis with depth of grammar implementation (“computational linguistic typology”)

• Core grammar: Information hypothesized to be cross-linguistically useful

• Libraries: Sets of analyses of cross-linguistically variable phenomena

• Accessed through a web-based questionnaire

• => Demo Grammar Matrix questionnaire
Grammar Matrix

Elicitation of typological information

- Questionnaire definition
- Questionnaire (accepts user input)
- HTML generation
- Validation
- Choices file

Grammar creation

- Core grammar
- Stored analyses
- Customization
- Customized grammar
Morphophonology and Morphosyntax

• Amenable to different kinds of formal analysis

• Should be kept separate (but linked) (Bender & Good, 2005)

• Grammar Matrix derived grammars assume the existence of morphophonological analyzers, and target the morpheme-segmented line of IGT:
Morphophonology and Morphosyntax

• Amenable to different kinds of formal analysis

• Should be kept separate (but linked) (Bender & Good, 2005)

• Grammar Matrix derived grammars assume the existence of morphophonological analyzers, and target the morpheme-segmented line of IGT:

\[
\begin{array}{cccccc}
\text{unisaa} & \text{khatte} & \text{mo} & \text{kosi} & \text{moba} \\
\text{u-nisa-a} & \text{khatt-e} & \text{mo} & \text{kosi-i} & \text{mo-pe} \\
3\text{POSS-yngr.bro-ERG.A} & \text{take-IND.PST} & \text{DEM.DOWN} & \text{river-LOC} & \text{DEM.DOWN-LOC}
\end{array}
\]

The younger brother took it to the river. [ctn]  

(Bickel et al., 2013c)
Morphophonology and Morphosyntax

• Amenable to different kinds of formal analysis

• Should be kept separate (but linked) (Bender & Good, 2005)

• Grammar Matrix derived grammars assume the existence of morphophonological analyzers, and target the morpheme-segmented line of IGT:

\[
\begin{array}{cccccc}
\text{unisaa} & \text{khatte} & \text{mo} & \text{kosi} & \text{moba} \\
\text{u-nisa-a} & \text{khatt-e} & \text{mo} & \text{kosi-i} & \text{mo-pe} \\
3\text{sPOSS-yngr.bro-ERG.A} & \text{take-IND.PST} & \text{DEM.DOWN} & \text{river-LOC} & \text{DEM.DOWN-LOC} \\
\end{array}
\]

The younger brother took it to the river. [ctn] (Bickel et al., 2013c)
Chintang as a case study language: Why?

- Morphologically rich
- Active descriptive/documentation work
- Relatively mature
- Treasure trove of IGT!
Chintang as a case study language: Why?

• Morphologically rich

• Active descriptive/documentation work

• Relatively mature descriptive materials

• Treasure trove of IGT!
Defining Morphosyntactic rules for Chintang (Bender, Schikowski & Bickel 2012)

- Done collaboratively with Robert Schikowski, on the basis of Schikowski 2012

- 160 verbal lexical rules grouped into 54 position classes
  - Including duplicated prefix/suffix rules to handle the multi-root verbs

- 24 nominal lexical rules grouped into 6 position classes

- Elsewhere in the choices file, define values for case, tense/aspect/mood, person/number/gender
  - => Partial modeling of the syntactico-semantic effects of 131/160 rules
Questionnaire Snippet

1st-person-object (verb-pc2)

Verb Position Class 2:
- Position Class Name: 1st-person-object
- Obligatorily occurs: [ ]
- Appears as a prefix or suffix: prefix
- Possible inputs: neg-prefix (verb-pc1)

Morphotactic Constraints:
- Add a Require constraint
- Add a Forbid constraint

Lexical Rule Types that appear in this Position Class:

1nseO (verb-pc2_lrt1)

Lexical Rule Type 1:
- Name: 1nseO
- Supertypes: [ ]

Features:
- Name: pernum
  - Value: 1du_excl, 1pl_excl
  - Specified on: the object NP

Morphotactic Constraints:
- Add a Require constraint
- Add a Forbid constraint

Lexical Rule Instances:

=> demo?
## Implementation of Morphology

<table>
<thead>
<tr>
<th>Feature</th>
<th># Rules</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERNUM</td>
<td>103</td>
<td>Person and/or number of verb’s dependent or of noun</td>
</tr>
<tr>
<td>CASE</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>FORM</td>
<td>35</td>
<td>Form (finite/non-finite) of verb</td>
</tr>
<tr>
<td>TENSE</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>ASPECT</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>MOOD</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>NEGATION</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>
Reuse of Resources I: Import from Toolbox Lexicon
(Bender, Schikowski & Bickel 2012)

• Define lexical types in Grammar Matrix customization system

• Use new ‘Import Toolbox lexicon’ subpage to map Toolbox lexical entries to types
  
  • Toolbox tags are user-defined; Import page allows user-specified tag-value pairs

  • Lexical types, import types, import tag-value constraints are all unbounded
Chintang Toolbox lexicon

\lex kond
\id 179
\psrev v
\val A-ERG P-NOM V-a(A).o(P)
\ge search; look.for
\dt 22/Feb/2011
Chintang Toolbox lexicon

\lex kond
\id 179
\psrev v
\val A-ERG P-NOM V-a(A).o(P)
\ge search; look for
\dt 22/Feb/2011
Grammar Matrix lexical entry: source code

khad := trans-verb-lex &
  [ STEM < "kond" >,
    SYNSEM.LKEYS.KEYREL.PRED "_search;look.for_v_rel" ].

trans-verb-lex := erg-abs-transitive-verb-lex.

erg-abs-transitive-verb-lex := transitive-verb-lex &
  [ ARG-ST < [ LOCAL.CAT.HEAD noun &
      [ CASE erg ] ],
    [ LOCAL.CAT.HEAD noun &
      [ CASE abs ] ] > ].
Final import types: entries imported

- Common nouns: 4,741
- Native Chintang intransitive verbs: 282
- Borrowed Nepali intransitive verbs: 142
- Native Chintang transitive verbs: 285
- Borrowed Nepali transitive verbs: 190
- Total: 5,640/9,034 (62%) of all entries; 899/1,440 (62%) of the verbs
- Most frequent unimported POS tags: adverbs (866), adjectives (515), interjections (377), affixes (286)
Evaluation procedure

- Define choices file, including: lexical types, lexical rules (as above), import classes, supporting grammatical information, 10 lexical entries for pronouns, 1 lexical entry for an auxiliary

- Create grammar using Grammar Matrix customization system

- Load grammar into LKB grammar development environment (Copestake 2002) with [incr tsdb()] (Oepen 2001) for test suite management

- Refine choices file based on parsing results over three sample narratives (2,906 word tokens)

- Test lexical coverage of resulting grammar on four held-out narratives

- Evaluate precision/recall of morphosyntactic and morphosemantic features
Test data narratives

• 1,453 total word tokens

• Varied genres: biographical monologue, Pear Story (Chafe 1980), traditional story, recipe

• Example test item (target is second line):

  thupro wassace  uyuwakte  pho
  thupro wassak-ce u-yuŋ-a-yakt-e  pho
  many  bird-NS  3NSS/A-live-PST-IPFV-IND.PST REP

  ‘There lived many birds.’ [ctn] story_rabbit.005
## Lexical coverage results

<table>
<thead>
<tr>
<th>Narrative</th>
<th>total type</th>
<th>total token</th>
<th># analyzed type</th>
<th># analyzed token</th>
<th>% analyzed type</th>
<th>% analyzed token</th>
<th>avg ambiguity type</th>
<th>avg ambiguity token</th>
</tr>
</thead>
<tbody>
<tr>
<td>Durga_Exp</td>
<td>206</td>
<td>489</td>
<td>120</td>
<td>265</td>
<td>58</td>
<td>54</td>
<td>1.24</td>
<td>1.14</td>
</tr>
<tr>
<td>choku_yakkheng</td>
<td>152</td>
<td>331</td>
<td>89</td>
<td>184</td>
<td>59</td>
<td>56</td>
<td>1.26</td>
<td>1.20</td>
</tr>
<tr>
<td>pear_6-1</td>
<td>206</td>
<td>433</td>
<td>105</td>
<td>203</td>
<td>56</td>
<td>51</td>
<td>1.20</td>
<td>1.62</td>
</tr>
<tr>
<td>story_rabbit</td>
<td>85</td>
<td>200</td>
<td>43</td>
<td>69</td>
<td>51</td>
<td>35</td>
<td>1.37</td>
<td>1.23</td>
</tr>
<tr>
<td>All</td>
<td>568</td>
<td>1453</td>
<td>324</td>
<td>721</td>
<td>57</td>
<td>50</td>
<td>1.40</td>
<td>1.27</td>
</tr>
</tbody>
</table>
Morphosyntactic/morphosemantic features: precision/recall

- 10 randomly selected word types from each of four test narratives

- Compared morphosyntactic and morphosemantic features in grammar’s analysis to gloss provided in Chintang corpus

- In case of ambiguity, select the best match from the grammar

<table>
<thead>
<tr>
<th>Narrative</th>
<th>Total gold attributes</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Durga_Exp</td>
<td>16</td>
<td>.48</td>
<td>.88</td>
<td>.62</td>
</tr>
<tr>
<td>choku_yakkheng</td>
<td>21</td>
<td>.57</td>
<td>.62</td>
<td>.59</td>
</tr>
<tr>
<td>pear_6-1</td>
<td>31</td>
<td>.71</td>
<td>.92</td>
<td>.80</td>
</tr>
<tr>
<td>story_rabbit</td>
<td>29</td>
<td>.62</td>
<td>.83</td>
<td>.71</td>
</tr>
<tr>
<td>Total</td>
<td>97</td>
<td>.61</td>
<td>.82</td>
<td>.70</td>
</tr>
</tbody>
</table>
Error analysis: features

• Glosses in Chintang corpus represent a systematic policy to not gloss zero morphs

• Grammar Matrix-derived grammar includes rules for zeroes
  • Default singular number: 19 errors
  • Default nominative case: 15 errors
  • Default [FORM fin]: 9 errors
  • Total errors in precision attributable to default values: 52 errors

• 18 recall errors are primarily due to cases where the intended lexical root is not available
Error analysis: Lack of coverage

- Randomly selected 10 unanalyzed word forms from each of the four narratives

- Lack of import class: 73%
- Not in Toolbox lexicon: 13%
- Affixes not implemented: 10%
- Other grammar error: 5%
AGGREGATION Project: Overview

- choices file
- Grammar Matrix
- precision grammar
- syntactico-semantic analyses

Diagram flow:
- Text flows into the Grammar Matrix.
- The Grammar Matrix generates precision grammar.
- Precision grammar leads to syntactico-semantic analyses.
AGGREGATION Project: Overview
Choices files

word-order=v-final
has-dets=yes
noun-det-order=det-noun
...
case-marking=erg-abs
erg-abs-erg-case-name=erg
erg-abs-abs-case-name=abs
...
verb4_valence=erg-abs
  verb4_stem1_orth=sams-i-ne
  verb4_stem1_pred=_sams-i-ne_v_re
...
verb-pc3_inputs=verb-pc2
  verb-pc3_lrt1_name=2nd-person-subj
    verb-pc3_lrt1_feat1_name=pernum
    verb-pc3_lrt1_feat1_value=2nd
    verb-pc3_lrt1_feat1_head=subj
    verb-pc3_lrt1_lrl1_inflecting=yes
    verb-pc3_lrt1_lrl1_orth=a-
Choices files

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

} Word order
Choices files

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

} Word order

} Case system
Choices files

```Private

word-order=v-final
has-dets=yes
noun-det-order=det-noun

... case-marking=erg-abs
erg-abs-erg-case-name=erg
erg-abs-abs-case-name=abs

... verb4_valence=erg-abs
   verb4_stem1_orth=sams-i-ne
   verb4_stem1_pred=_sams-i-ne_v_re

... verb-pc3_inputs=verb-pc2
   verb-pc3_lrt1_name=2nd-person-subj
      verb-pc3_lrt1_feat1_name=pernum
      verb-pc3_lrt1_feat1_value=2nd
      verb-pc3_lrt1_feat1_head=subj
      verb-pc3_lrt1_lrl1_inflecting=yes
      verb-pc3_lrt1_lrl1_orth=a-

```
Choices files

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

} Word order
} Case system
} Lex types & entries
} Lex rules
Projected structure
Projected structure

Charniak parser (Charniak 1997)
Projected structure

RiPLEs (Xia & Lewis 2007)

Charniak parser (Charniak 1997)
Learning General Choices from IGT: Word order
(Bender et al 2013)
Learning General Choices from IGT: Word order (Bender et al 2013)

• Count S-V and O-V orders in projected trees
Learning General Choices from IGT: Word order
(Bender et al 2013)

• Count S-V and O-V orders in projected trees

• Calculate observed word order on three axes (SV-VS, OV-VO, SO-OS)
Learning General Choices from IGT: Word order (Bender et al 2013)

- Count S-V and O-V orders in projected trees

- Calculate observed word order on three axes (SV-VS, OV-VO, SO-OS)

- Compare to canonical points
Learning General Choices from IGT: Word order (Bender et al 2013)

- Count S-V and O-V orders in projected trees
- Calculate observed word order on three axes (SV-VS, OV-VO, SO-OS)
- Compare to canonical points
Learning General Choices from IGT: Word order (Bender et al 2013)

- Count S-V and O-V orders in projected trees
- Calculate observed word order on three axes (SV-VS, OV-VO, SO-OS)
- Compare to canonical points
- Answer for Chintang: v-final
Learning General Choices from IGT: Case System (Bender et al 2013)

<table>
<thead>
<tr>
<th>Case system</th>
<th>Case grams present</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>NOM \ checkmark</td>
</tr>
<tr>
<td>nom-acc</td>
<td>NOM \ checkmark, ACC \ checkmark</td>
</tr>
<tr>
<td>erg-abs</td>
<td>ERG \ checkmark, ABS \ checkmark</td>
</tr>
<tr>
<td>split-erg</td>
<td>ERG \ checkmark, ABS \ checkmark</td>
</tr>
<tr>
<td>(conditioned on V)</td>
<td>NOM \ checkmark, ACC \ checkmark</td>
</tr>
</tbody>
</table>


Learning General Choices from IGT: Case System (Bender et al 2013)

<table>
<thead>
<tr>
<th>Case system</th>
<th>Case grams present</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nom-acc</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>erg-abs</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>split-erg (conditioned on V)</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

• “Gram” method
Learning General Choices from IGT: Case System (Bender et al. 2013)

<table>
<thead>
<tr>
<th>Case system</th>
<th>Case grams present</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NOM  V ACC  ERG  V ABS</td>
</tr>
<tr>
<td>none</td>
<td></td>
</tr>
<tr>
<td>nom-acc</td>
<td>✓</td>
</tr>
<tr>
<td>erg-abs</td>
<td>✓</td>
</tr>
<tr>
<td>split-erg</td>
<td>✓</td>
</tr>
<tr>
<td>(conditioned on V)</td>
<td>✓</td>
</tr>
</tbody>
</table>

- "Gram" method
- Results for Chintang: erg-abs
### Learning General Choices from IGT: Case System
(Bender et al 2013)

<table>
<thead>
<tr>
<th>Case system</th>
<th>Top grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>$S_g = A_g = O_g$, or $S_g \neq A_g \neq O_g$ and $S_g$, $A_g$, $O_g$ also present on the other argument types</td>
</tr>
<tr>
<td>nom-acc</td>
<td>$S_g = A_g$, $S_g \neq O_g$</td>
</tr>
<tr>
<td>erg-abs</td>
<td>$S_g = O_g$, $S_g \neq A_g$</td>
</tr>
<tr>
<td>tripartite</td>
<td>$S_g \neq A_g \neq O_g$, and $S_g$, $A_g$, $O_g$ absent from others</td>
</tr>
<tr>
<td>split-s</td>
<td>$S_g \neq A_g \neq O_g$, and $A_g$ and $O_g$ both present on $S$ list</td>
</tr>
</tbody>
</table>
Learning General Choices from IGT: Case System  
(Bender et al 2013)

<table>
<thead>
<tr>
<th>Case system</th>
<th>Top grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>$S_g = A_g = O_g$, or $S_g \neq A_g \neq O_g$ and $S_g$, $A_g$, $O_g$ also present on the other argument types</td>
</tr>
<tr>
<td>nom-acc</td>
<td>$S_g = A_g$, $S_g \neq O_g$</td>
</tr>
<tr>
<td>erg-abs</td>
<td>$S_g = O_g$, $S_g \neq A_g$</td>
</tr>
<tr>
<td>tripartite</td>
<td>$S_g \neq A_g \neq O_g$, and $S_g$, $A_g$, $O_g$ absent from others</td>
</tr>
<tr>
<td>split-s</td>
<td>$S_g \neq A_g \neq O_g$, and $A_g$ and $O_g$ both present on $S$ list</td>
</tr>
</tbody>
</table>

• “SAO” method
Learning General Choices from IGT: Case System (Bender et al 2013)

<table>
<thead>
<tr>
<th>Case system</th>
<th>Top grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>( S_g = A_g = 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</td>
</tr>
<tr>
<td>nom-acc</td>
<td>( S_g = A_g ), ( S_g \neq O_g )</td>
</tr>
<tr>
<td>erg-abs</td>
<td>( S_g = O_g ), ( S_g \neq A_g )</td>
</tr>
<tr>
<td>tripartite</td>
<td>( S_g \neq A_g \neq O_g ), and ( S_g ), ( A_g ), ( O_g ) absent from others</td>
</tr>
<tr>
<td>split-s</td>
<td>( S_g \neq A_g \neq O_g ), and ( A_g ) and ( O_g ) both present on ( S ) list</td>
</tr>
</tbody>
</table>

- “SAO” method

- Results for Chintang: None :(
Learning Case Frames from IGT
(Bender et al 2014)

- For each IGT with overt subject and/or object,
- use grams to find case assigned to subject and object NPs
- no case gram => use default case value for that position

```plaintext
# find 1-arg verbs (not dropped subj)
@S < (@VP < @VB=verb !< @OBJ) < @SBJ=sbj !< @OBJ;
# find 1-arg verbs for OSV or VSO langs
@S < @VB=verb < @SBJ=sbj !< @OBJ;

# find 2-arg verbs, both args overt
@S < (@VP < @VB=verb < @OBJ=objc) < @SBJ=sbj;
# 2-arg verbs for OSV or VSO
@S < @VB=verb < @SBJ=sbj < @OBJ=objc;

# find 2-arg verbs w/dropped sbj
@S < (@VP < @VB=verb < @OBJ=objc) !< @SBJ=sbj;
# 2-arg verbs for OSV or VSO
@S < @VB=verb !< @SBJ=sbj < @OBJ=objc;
```
## Putting It All Together: Interim Results
(Bender et al 2014)

<table>
<thead>
<tr>
<th>Choices file</th>
<th># verb entries</th>
<th># noun entries</th>
<th># det entries</th>
<th># verb affixes</th>
<th># noun affixes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORACLE</td>
<td>900</td>
<td>4751</td>
<td>0</td>
<td>160</td>
<td>24</td>
</tr>
<tr>
<td>BASELINE</td>
<td>3005</td>
<td>1719</td>
<td>240</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FF-AUTO-NONE</td>
<td>3005</td>
<td>1719</td>
<td>240</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FF-DEFAULT-GRAM</td>
<td>739</td>
<td>1724</td>
<td>240</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FF-AUTO-GRAM</td>
<td>739</td>
<td>1724</td>
<td>240</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MOM-DEFAULT-NONE</td>
<td>1177</td>
<td>1719</td>
<td>240</td>
<td>262</td>
<td>0</td>
</tr>
<tr>
<td>MOM-AUTO-NONE</td>
<td>1177</td>
<td>1719</td>
<td>240</td>
<td>262</td>
<td>0</td>
</tr>
</tbody>
</table>
### Putting It All Together: Interim Results
(Bender et al 2014)

<table>
<thead>
<tr>
<th>choices file</th>
<th>Training Data (N = 8863)</th>
<th>Test Data (N = 930)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lexical coverage (%)</td>
<td>items parsed (%)</td>
</tr>
<tr>
<td>ORACLE</td>
<td>1165 (13)</td>
<td>174 (3.5)</td>
</tr>
<tr>
<td>BASELINE</td>
<td>1276 (14)</td>
<td>398 (7.9)</td>
</tr>
<tr>
<td>FF-AUTO-NONE</td>
<td>1276 (14)</td>
<td>354 (4.0)</td>
</tr>
<tr>
<td>FF-DEFAULT-GRAM</td>
<td>911 (10)</td>
<td>126 (1.4)</td>
</tr>
<tr>
<td>FF-AUTO-GRAM</td>
<td>911 (10)</td>
<td>120 (1.4)</td>
</tr>
<tr>
<td>MOM-DEFAULT-NONE</td>
<td>1102 (12)</td>
<td>814 (9.2)</td>
</tr>
<tr>
<td>MOM-AUTO-NONE</td>
<td>1102 (12)</td>
<td>753 (8.5)</td>
</tr>
</tbody>
</table>
Next Steps

• Combine MOM and gram approaches

• Refine search patterns to capture more lexical entries in gram approach

• Extend MOM to noun morphology

• Extend approach to learning other morphosyntactic properties (both lexically specific, and language-level)

• Extend Grammar Matrix customization system to handle more phenomena

• Test on more languages
References


