Towards Automatic Detection of Morphosyntactic Systems from IGT

Exploring Data from Language Documentation
ZAS Berlin
10 May 2013

Emily M. Bender, Fei Xia, Joshua Crowgey and Michael Wayne Goodman
University of Washington
Overview

• AGGREGATION: Research goals

• The LinGO Grammar Matrix

• RiPLEs

• Case study 1: Word order

• Case study 2: Case systems

• Conclusion & outlook
Overview

• AGGREGATION: Research goals

• The LinGO Grammar Matrix

• RiPLes

• Case study 1: Word order

• Case study 2: Case systems

• Conclusion & outlook
AGGREGATION: Research goals

• Precision implemented grammars are a kind of structured annotation over linguistic data (cf. Good 2004, Bender et al 2012).

• They map surface strings to semantic representations and vice-versa.

• They can be used in the development of grammar checkers and treebanks, making them useful for language documentation and revitalization (Bender et al 2012)

• But they are expensive to build.

• The AGGREGATION project asks whether existing products of documentary linguistic research (IGT collections) can be used to boot-strap the development of precision implemented grammars.
Combining linguistic knowledge

Database of IGT (language specific) + Repository of implemented linguistic analyses (language independent and parameterized)

Implemented grammar (language specific)

parser

generator

treebank
Overview

- AGGREGATION: Research goals
- The LinGO Grammar Matrix
- RiPLes
- Case study 1: Word order
- Case study 2: Case systems
- Conclusion & outlook
LinGO Grammar Matrix: Goals and History

• Developed in the context of the DELPH-IN Consortium (http://www.delph-in.net)

• Compatible with open-source tools for parsing, generation, treebanking, parse ranking, machine translation and more

• Implements analyses in Head-driven Phrase Structure Grammar (Pollard & Sag 1994) with semantic representations in Minimal Recursion Semantics (MRS; Copestake et al 2005)

• Package what has been learned in 20+ person-years of development of the English Resource Grammar (Flickinger 2000) for easy reuse in grammars for other languages
Sample hypothesized universals

- Words and phrases combine to make larger phrases.
- The semantics of a phrase is determined by the words in the phrase and how they are put together.
- Some rules for phrases add semantics (but some don’t).
- Most phrases have an identifiable head daughter.
- Heads determine which arguments they require and how they combine semantically with those arguments.
- Modifiers determine which kinds of heads they can modify, and how they combine semantically with those heads.
- No lexical or syntactic rule can remove semantic information.
Cross-linguistic variation doesn’t preclude all grammar code sharing

• Many grammatical properties which vary cross-linguistically vary within a fairly well-understood range

• Hypothesis: Analyses can be developed for e.g., SOV word order which will work across SOV languages, regardless of language family or other typological properties

• ‘Libraries’ of analyses of ‘wide-spread but not universal’ (Drellishak 2009) properties facilitate rapid development of precision grammars

• ... while also constituting typological hypotheses
Grammar customization

Elicitation of typological information

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

Grammar creation

- Choices file
- Core grammar
- Stored analyses
- Customization
- Customized grammar

(Bender et al 2010)
Cross-linguistically robust

- Used in the development of small grammars for >80 genealogically diverse languages, plus several larger grammar fragments

- Systematically evaluated on 7 languages from 7 (non-IE) language families (Bender et al 2010)

- Core grammar and libraries both refined as evidence from new languages falsifies hypothesized universals and/or exposes new options
The Grammar Matrix and documentary linguistics


- 804 IGT instances in Nordlinger 1998 used as development data

- Grammar tested on narrative (held out test data), of which 76% received analyses matching the translation
  
  - The original descriptive work represents ~20x more effort

  - But the grammar engineering still took an expert grammar engineer 5.5 person weeks

  - Can we speed that up?
Sample choices file: Umatilla Sahaptin [uma]

section=general
language=Umatilla Sahaptin
iso-code=uma

section=word-order
word-order=vso
has-dets=no
has-aux=no

section=number
  number1_name=sg
  number2_name=du
  number3_name=pl

section=person
person=1-2-3
first-person=incl-excl
incl-excl-number=du, pl

section=gender

section=case
case-marking=nom-acc
  nom-acc-nom-case-name=nom
  nom-acc-acc-case-name=obj

section=direct-inverse
  scale1_feat1_name=pernum
  scale1_feat1_value=1st
  scale2_feat1_name=pernum
  scale2_feat1_value=2nd
  scale3_feat1_name=pernum
  scale3_feat1_value=3rd
  scale3_feat2_name=topicality
  scale3_feat2_value=topic
  scale4_feat1_name=pernum
  scale4_feat1_value=3rd
  scale4_feat2_name=topicality
  scale4_feat2_value=non-topic
  scale-equal=direct
...

...
Sample choices file: Umatilla Sahaptin [uma]

Customization system maps relatively simple ‘choices’ based description to working grammar fragment
Overview

• AGGREGATION: Research goals

• The LinGO Grammar Matrix

• RiPLeS

• Case study 1: Word order

• Case study 2: Case systems

• Conclusion & outlook
RiPLes: Goals

• RiPLes: Information engineering and synthesis for Resource Poor Languages

• Support rapid development of NLP resources for RPLs by bootstrapping through IGT

• Support cross-linguistic study through creating ‘language profiles’ based on IGT analysis

(Xia & Lewis 2007, Lewis & Xia 2008)
RiPLes: IGT projection methodology

(Xia & Lewis 2009)
RiPLEs: Results

Table 3: Experiment 1 Results (Accuracy)

<table>
<thead>
<tr>
<th></th>
<th>WOrder</th>
<th>VP +OBJ</th>
<th>DT +N</th>
<th>Dem +N</th>
<th>JJ +N</th>
<th>PRPS +N</th>
<th>Poss +N</th>
<th>P +NP</th>
<th>N +num</th>
<th>N +case</th>
<th>V +TA</th>
<th>Def</th>
<th>Indef</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic CFG</td>
<td>0.8</td>
<td>0.5</td>
<td>0.8</td>
<td>0.8</td>
<td>1.0</td>
<td>0.8</td>
<td>0.6</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>1.0</td>
<td>0.9</td>
<td>0.800</td>
</tr>
<tr>
<td>sum(CFG)</td>
<td>0.8</td>
<td>0.5</td>
<td>0.8</td>
<td>0.8</td>
<td>0.9</td>
<td>0.7</td>
<td>0.6</td>
<td>0.8</td>
<td>0.6</td>
<td>0.8</td>
<td>0.7</td>
<td>1.0</td>
<td>0.9</td>
<td>0.762</td>
</tr>
<tr>
<td>CFG w/ func</td>
<td>0.9</td>
<td>0.6</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>0.8</td>
<td>0.7</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>1.0</td>
<td>0.9</td>
<td>0.831</td>
</tr>
<tr>
<td>both</td>
<td>0.9</td>
<td>0.6</td>
<td>0.8</td>
<td>0.8</td>
<td>0.9</td>
<td>0.7</td>
<td>0.5</td>
<td>0.8</td>
<td>0.6</td>
<td>0.8</td>
<td>0.7</td>
<td>1.0</td>
<td>0.9</td>
<td>0.769</td>
</tr>
</tbody>
</table>

Table 5: Word Order Accuracy for 97 languages

<table>
<thead>
<tr>
<th># of IGT instances</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>100+</td>
<td>100%</td>
</tr>
<tr>
<td>40-99</td>
<td>99%</td>
</tr>
<tr>
<td>10-39</td>
<td>79%</td>
</tr>
<tr>
<td>5-9</td>
<td>65%</td>
</tr>
<tr>
<td>3-4</td>
<td>44%</td>
</tr>
<tr>
<td>1-2</td>
<td>14%</td>
</tr>
</tbody>
</table>

(Lewis & Xia 2008)
Overview

• AGGREGATION: Research goals

• The LinGO Grammar Matrix

• RiPLEs

• Case study 1: Word order

• Case study 2: Case systems

• Conclusion & outlook
Word order options

- Lewis & Xia 2008, Dryer 2011 (WALS)
  - SOV
  - SVO
  - OSV
  - OVS
  - VSO
  - VOS
  - no dominant order

- Grammar Matrix
  - SOV
  - SVO
  - OSV
  - OVS
  - VSO
  - VOS
  - Free (pragmatically determined)
  - V-final
  - V-initial
  - V2
Word order in the Grammar Matrix

• More than a simple descriptive statement

• Affects phrase structure rules output by the system, but also interacts with other libraries (e.g., argument optionality)

• These phrase structure rules help model the mapping of syntactic to semantic arguments

• Underlying word order is not reflected in every sentence; testsuites won’t have the same distribution as naturally occurring corpora

• Matrix users advised to choose fixed word order if deviations from that order can be attributed to specific syntactic constructions
Methodology

• Parse English translation and project the parsed structure onto the language line (per RiPLeS)

• Add -SBJ and -OBJ function tags to the English parse trees (by heuristic), and project these too

• *Observed word orders:* counts of the 10 patterns SOV, SVO, OSV, OVS, VSO, VOS, SV, VS, OV, and VO in the source language trees

• Decompose SOV, SVO, OSV, OVS, VSO, VOS into order of S/O, S/V and O/V
Methodology

- SOV, SVO, OSV, OVS, VSO, VOS

- Measure Euclidean distance to positions of canonical word orders

- In a separate step, distinguish free from V2
Dev and test data

- 31 testsuite + choices file pairs, developed in Linguistics 567 at UW (Bender 2007)

<table>
<thead>
<tr>
<th>Languages</th>
<th>DEV1</th>
<th>DEV2</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammatical examples</td>
<td>16–359 (median: 91)</td>
<td>Indo-European (3), Dravidian (2), Algic, Creole, Niger-Congo, Quechuan, Salishan</td>
<td></td>
</tr>
<tr>
<td>Language families</td>
<td></td>
<td></td>
<td>Indo-European (2), Afro-Austro-Asiatic, Austronesian, Arauan, Carib, Karvelian, N. Caucasian, Tai-Kadai, I</td>
</tr>
</tbody>
</table>
Results

• Compare to most-frequent-type (SOV, Dryer 2011)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Inferred WO</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEV1</td>
<td>0.900</td>
<td>0.200</td>
</tr>
<tr>
<td>DEV2</td>
<td>0.500</td>
<td>0.100</td>
</tr>
<tr>
<td>TEST</td>
<td>0.727</td>
<td>0.091</td>
</tr>
</tbody>
</table>

• Sources of error:
  
  • Testsuite bias
  
  • Misalignment in projections
Overview

- AGGREGATION: Research goals
- The LinGO Grammar Matrix
- RiPLes
- Case study 1: Word order
- **Case study 2: Case systems**
- Conclusion & outlook
Case system options in the Grammar Matrix:
Case marking on core arguments of (in)transitivs

- None
- Nominative-accusative
- Ergative-absolutive
- Tripartite
- Split-S
- Fluid-S
- Split conditioned on features of the arguments
- Split conditions on features of the V
- Focus-case (Austronesian-style)

- The choice among these options makes further features available on the lexicon page, including case frames
- There is always the option to define more cases and case frames
Two methods

- **GRAM:** Assume Leipzig Glossing Rules-compliance (Bickel et al 2008)

- Search gloss line for case grams, and assign system as follows:

<table>
<thead>
<tr>
<th>Case system</th>
<th>Case grams present</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NOM ∨ ACC    ERG ∨ 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>

- **SAO:** Use RiPLes to identify S, A, and O arguments

- Collect most frequent gram for each

- Compare most frequent grams across S/A/O to determine case system
Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GRAM</th>
<th>SAO</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEV1</td>
<td>0.900</td>
<td>0.700</td>
<td>0.400</td>
</tr>
<tr>
<td>DEV2</td>
<td>0.900</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>TEST</td>
<td>0.545</td>
<td>0.545</td>
<td>0.455</td>
</tr>
</tbody>
</table>

- GRAM confused by non-NOM/ACC style glossing
- SAO confused by testsuite bias (spurious most-frequent elements)
- SAO confused by alignment errors (e.g. case marking adpositions)
Overview

• AGGREGATION: Research goals

• The LinGO Grammar Matrix

• RiPLes

• Case study 1: Word order

• Case study 2: Case systems

• Conclusion & outlook
Summary

- First steps towards our long-term goal: Automatically create working grammar fragments from IGT, by taking advantage of
  - Grammar Matrix customization system’s mapping of relatively simple language description files to working grammars
  - Linguistic analysis encoded in IGT
  - RiPLes methodology for further enriching IGT

- Resulting grammars are of interest for testing the Grammar Matrix as a set of typological hypotheses

- And potentially for field grammarians (when built-out) as they can support the creation of treebanks and exploration of corpora for unanalyzed phenomena
Opportunities for collaboration

• We are interested in collections of IGT from field projects with detailed glosses, paired with ‘choices’ files

• We would gladly advise linguists in creating choices files for their languages
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.
References


References


