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

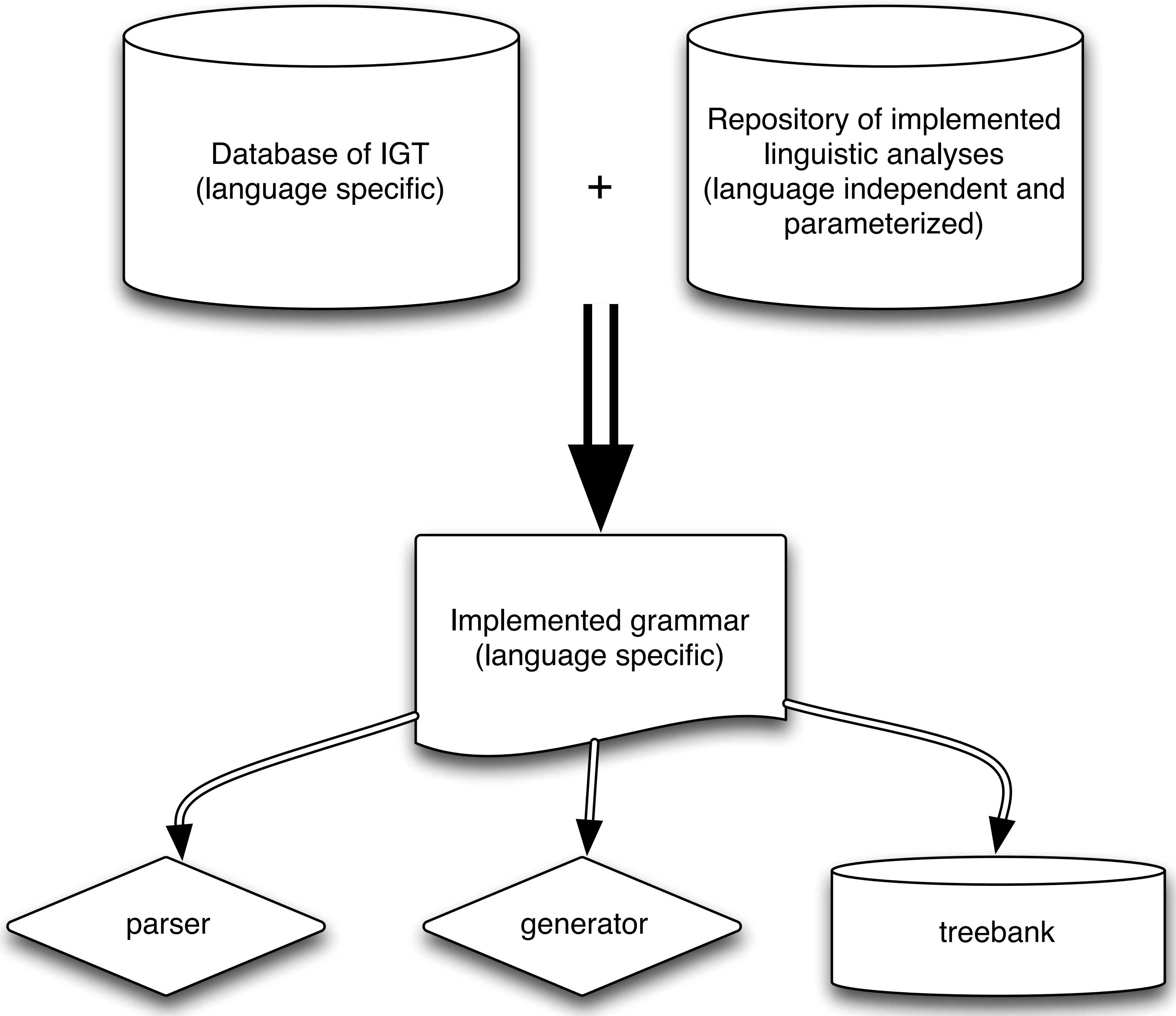
Overview

- 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



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

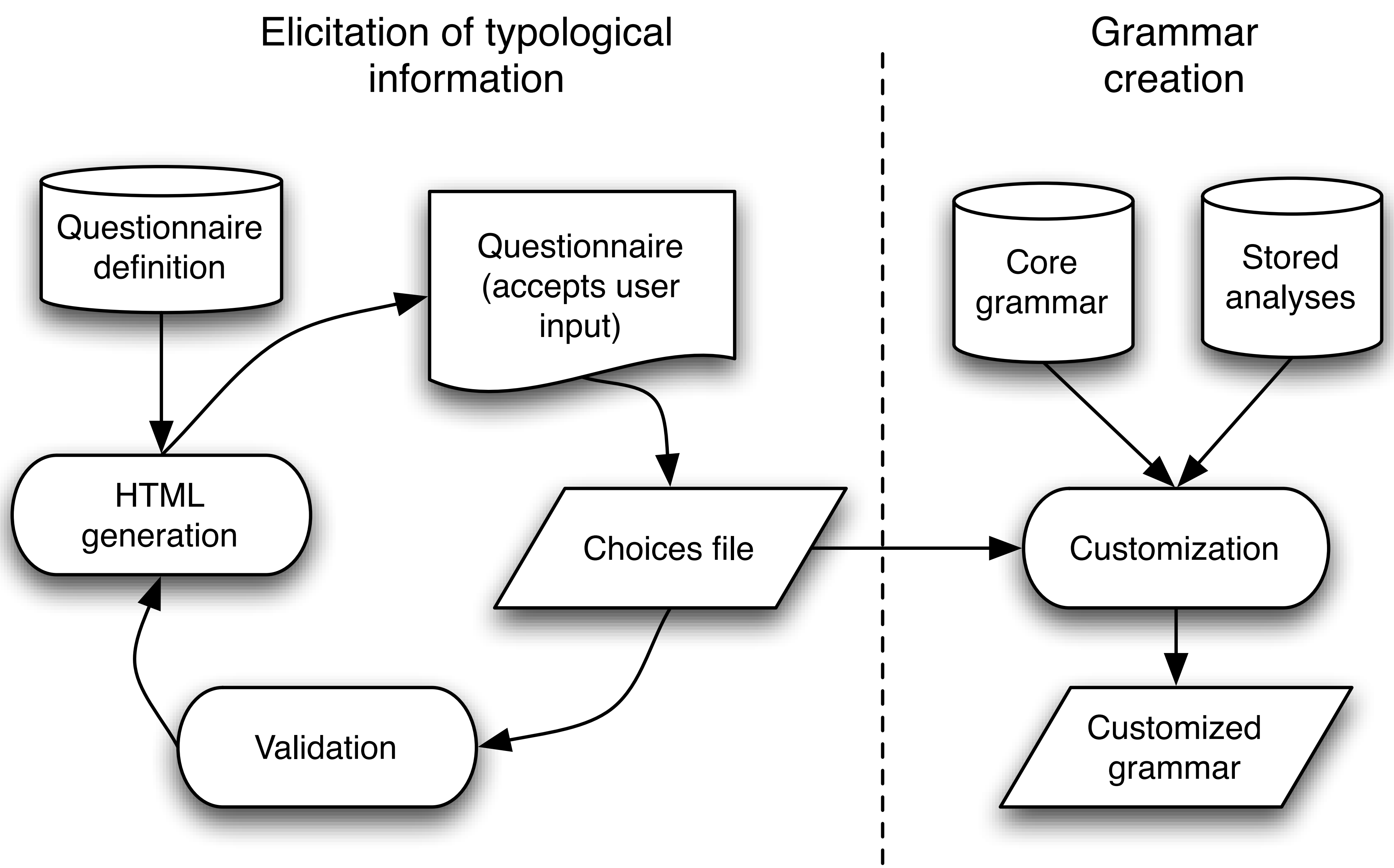
Sample hypothesized universals

- 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



(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

- Bender (2008) built a Matrix-based grammar for Wambaya based on description of Nordlinger (1998)
- 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]

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

section=
word-ord
has-dets
has-aux=

section=
  number
  number
  number

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

                        e
                        =pernum
                        e=1st
                        =pernum
                        e=2nd
                        =pernum
                        e=3rd
                        =topicality
                        e=topic
    scale4_feat1_name=pernum
    scale4_feat1_value=3rd
    scale4_feat2_name=topicality
    scale4_feat2_value=non-topic
scale-equal=direct

...
```

Customization system maps relatively simple 'choices' based description to working grammar fragment

Overview

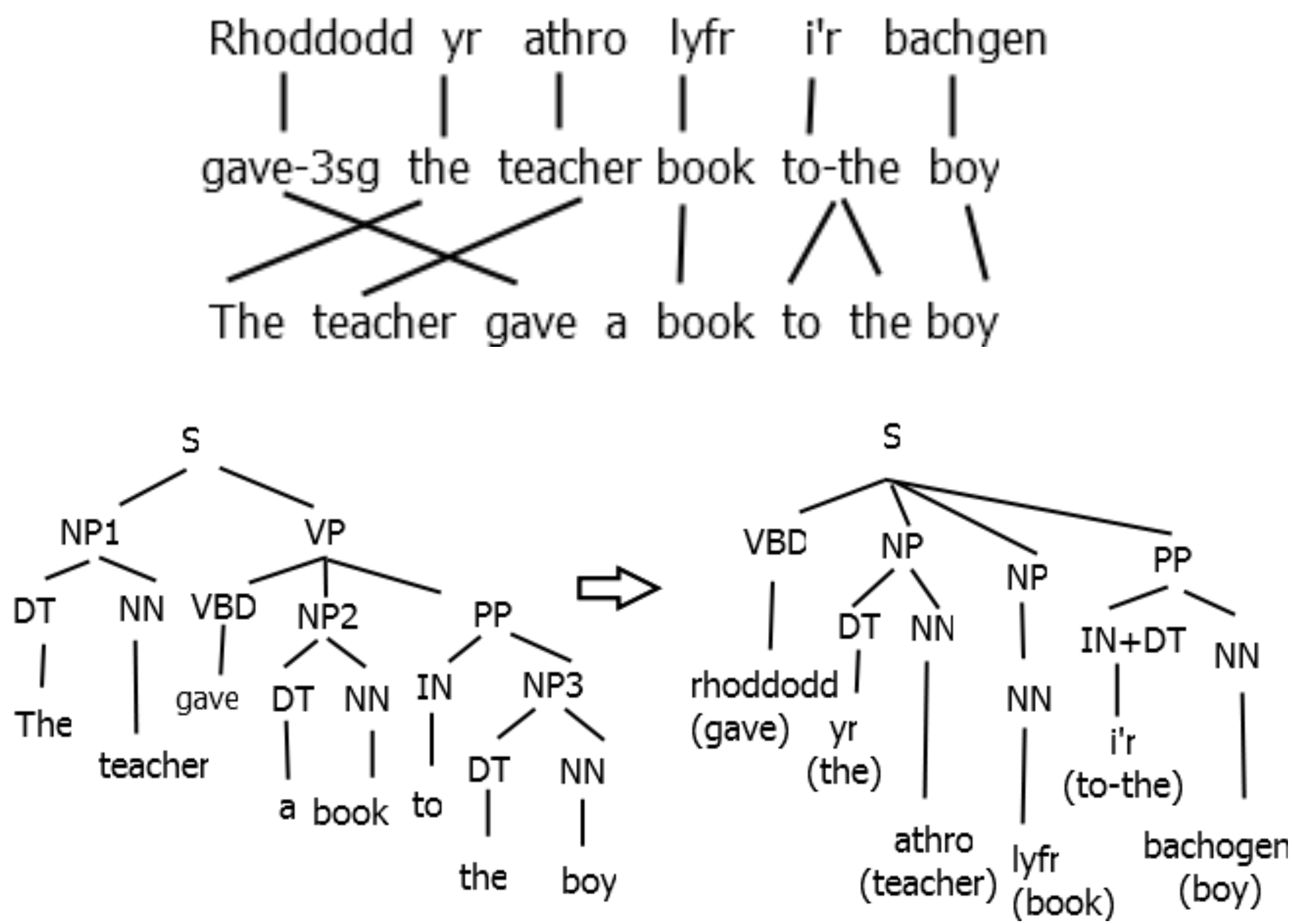
- AGGREGATION: Research goals
- The LinGO Grammar Matrix
- RiPLes
- Case study 1: Word order
- Case study 2: Case systems
- Conclusion & outlook

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)

	WOrder	VP +OBJ	DT +N	Dem +N	JJ +N	PRP$ +N	Poss +N	P +NP	N +num	N +case	V +TA	Def	Indef	Avg
basic CFG	0.8	0.5	0.8	0.8	1.0	0.8	0.6	0.9	0.7	0.8	0.8	1.0	0.9	0.800
sum(CFG)	0.8	0.5	0.8	0.8	0.9	0.7	0.6	0.8	0.6	0.8	0.7	1.0	0.9	0.762
CFG w/ func	0.9	0.6	0.8	0.9	1.0	0.8	0.7	0.9	0.7	0.8	0.8	1.0	0.9	0.831
both	0.9	0.6	0.8	0.8	0.9	0.7	0.5	0.8	0.6	0.8	0.7	1.0	0.9	0.769

Table 5: Word Order Accuracy for 97 languages

# of IGT instances	Average Accuracy
100+	100%
40-99	99%
10-39	79%
5-9	65%
3-4	44%
1-2	14%

(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

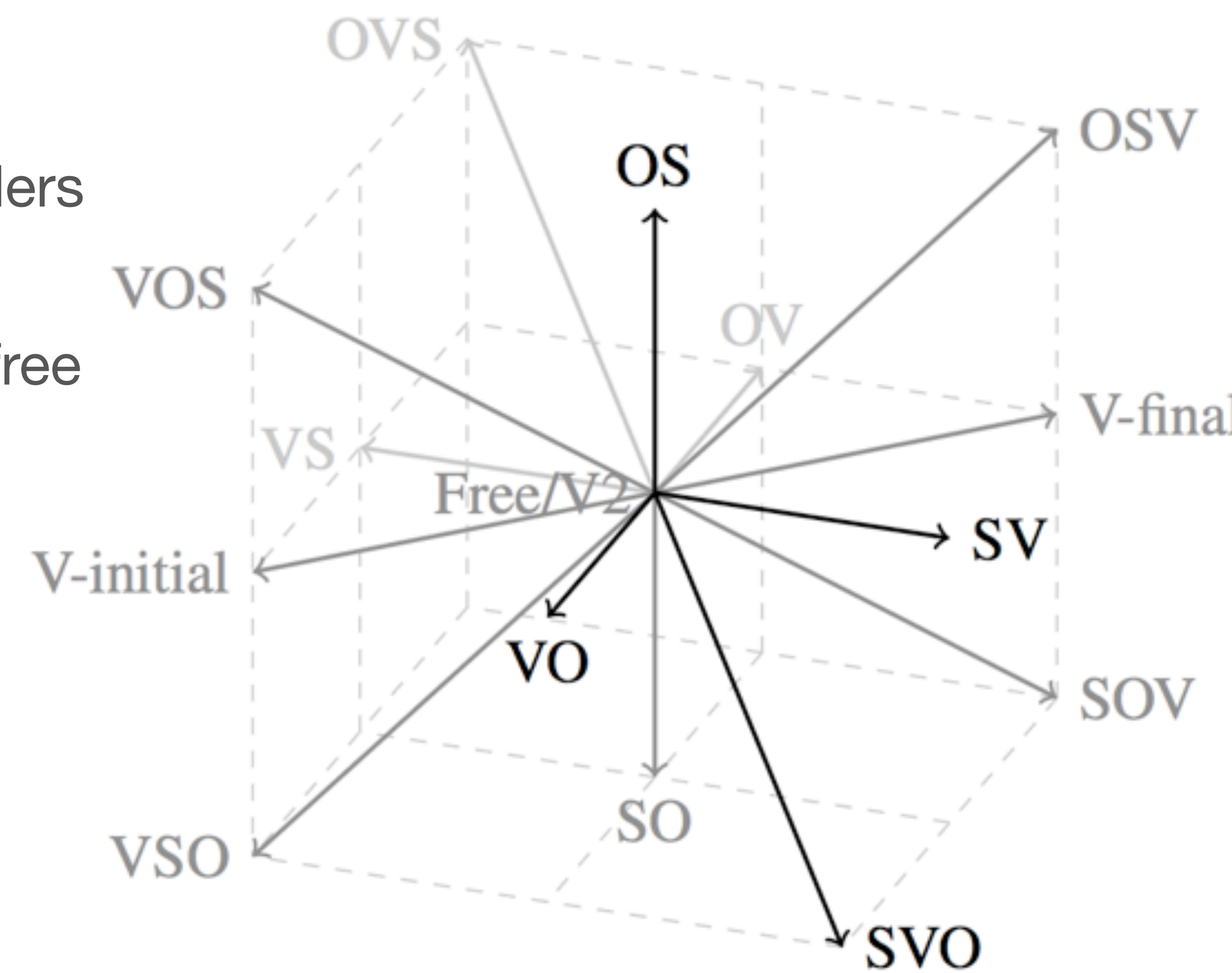
- More than a simple descriptive statement
- Affects phrase structure rules output by the system, but also interacts with other libraries (e.g., argument optionality)
- These phrase structure rules help model the mapping of syntactic to semantic arguments
- Underlying word order is not reflected in every sentence; testsuites won't have the same distribution as naturally occuring 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)

	DEV1	DEV2	TEST
Languages	10	10	11
Grammatical examples	16–359 (median: 91)	11–229 (median: 87)	48–216 (median: 76)
Language families	Indo-European (4), Niger-Congo (2), Afro-Asiatic, Japanese, Nadahup, Sino-Tibetan	Indo-European (3), Dravidian (2), Algic, Creole, Niger-Congo, Quechuan, Salishan	Indo-European (2), Afro-A Austro-Asiatic, Austronesi Arauan, Carib, Karvelian, N. Caucasian, Tai-Kadai, I

Results

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

Dataset	Inferred WO	Baseline
DEV1	0.900	0.200
DEV2	0.500	0.100
TEST	0.727	0.091

- 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)transitives

- 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:

Case sysem	Case grams present NOM ∨ ACC	ERG ∨ ABS
none		
nom-acc	✓	
erg-abs		✓
split-erg (conditioned on V)	✓	✓

- 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

Dataset	GRAM	SAO	Baseline
DEV1	0.900	0.700	0.400
DEV2	0.900	0.500	0.500
TEST	0.545	0.545	0.455

- 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

Bender, Emily M. 2007. Combining research and pedagogy in the development of a crosslinguistic grammar resource. In T. H. King and E. M. Bender (Eds.), *Proceedings of the GEAF 2007 Workshop*, Stanford, CA. CSLI Publications.

Bender, Emily M., Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation* 1–50. 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 J. Carroll, N. Oostdijk, and R. Sutcliffe (Eds.), *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, 8–14, Taipei, Taiwan.

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 S. Nordhoff and K.-L. G. Poggeman (Eds.), *Electronic Grammaticography*, 179–206. Honolulu: University of Hawaii Press.

Bickel, Balthasar, Bernard Comrie, and Martin Haspelmath. 2008. The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. Unpublished ms., Max Planck Institute for Evolutionary Anthropology and Department of Linguistics, University of Leipzig.

References

Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language & Computation* 3(4):281–332.

Drellishak, Scott. 2009. *Widespread But Not Universal: Improving the Typological Coverage of the Grammar Matrix*. PhD thesis, University of Washington.

Dryer, Matthew S. 2011. Order of subject, object and verb. In M. S. Dryer and M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library.

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.

Good, Jeff. 2004. The descriptive grammar as a (meta)database. In *Proceedings of the E-MELD Workshop 2004: Linguistic Databases and Best Practice*, Detroit, Michigan.

Lewis, William D., and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, 685–690, Hyderabad, India.

References

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

Xia, Fei, and William Lewis. 2009. Applying NLP technologies to the collection and enrichment of language data on the web to aid linguistic research. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*, 51–59, Athens, Greece, March. Association for Computational Linguistics.

Xia, Fei, and William D. Lewis. 2007. Multilingual structural projection across interlinearized text. In *NAACL-HLT 2007*, Rochester, NY.