Grammar Customization with the LinGO Grammar Matrix

MS/UW Symposium, January 22, 2010

Emily M. Bender

Joint work with: Scott Drellishak, Laurie Poulson, Antske Fokkens, Safiyyah Saleem, Michael Wayne Goodman, Kelly O'Hara, Joshua Hou, Daniel Mills, Dan Flickinger, Stephan Oepen

Overview

- Precision grammars
- Typology and typological variation
- LinGO Grammar Matrix
- Grammar customization
- Evaluation
- Future work

Precision grammars

- Map surface strings to syntactic and semantic representations, often bidirectionally
- Represent grammaticality
- Have been developed to broad coverage for a handful of languages in a handful of syntactic frameworks (Flickinger 2000, Siegel & Bender 2002, Müller & Kasper 2000)
- Can now parse efficiently (Oepen et al 2002)
- Scale more effectively than tree-bank derived grammars (in the sense of including new kinds of information)
- Can be made more robust with statistical lexical acquisition (Blunsom & Baldwin 2006) and other kinds of knowledge engineering/ML hybridization (Zhang & Kordoni 2008)
- ... but are expensive to build.

Human languages

- Vary along many dimensions, but not infinitely
- Can be seen as solving many of the same problems in different ways
- Just might share some core properties in common
- Can we leverage what's been learned in developing large-scale precision grammars for some languages to the development of grammars for others?

The Grammar Matrix

- Developed in the context of the DELPH-IN consortium (www.delph-in.net)
- Uses HPSG (Pollard and Sag 1994) and MRS (Copestake et al 2005)
- Core grammar originally abstracted from English Resource Grammar (Flickinger 2000) with reference to Jacy Japanese grammar (Siegel and Bender 2002)
- Aims to support both rapid initial development and long-term grammar build-out
- Promotes cross-grammar consistency in semantic representations
- Is also an exercise in exploration of potential universals
- http://www.delph-in.net/matrix

Customization System

- The Grammar Matrix core grammar is not itself a functioning grammar fragment

- Many phenomena are “widespread, but not universal” (Drellishak, 2009)

- Grammar customization is an approach to massively multilingual grammar code reuse

- Can the same analysis of e.g., SVO word order, split-ergativity, or “pro-drop” work in different languages?

- Web-based questionnaire elicits typological and lexical information, then outputs working “starter grammar”

Customization system: Current libraries

- Word order* (Bender & Flickinger 2005, Fokkens forthcoming)
- Morphotactics (O'Hara 2008)
- Case (+ direct-inverse marking) (Drellishak 2009)
- Agreement (person, number, gender) (Drellishak 2009)
- Tense and aspect (Poulson 2009)
- Sentential negation* (Bender & Flickinger 2005)
- Coordination (Drellishak & Bender 2005)
- Matrix yes-no questions* (Bender & Flickinger 2005)
- Argument optionality (pro-drop) (Saleem forthcoming)

Evaluation: Do the existing libraries scale to unseen languages?

- Testsuites developed by (then) non-Matrix developer on the basis of descriptive grammars to cover phenomena represented in Matrix libraries
- Starter grammars developed through customization system
- Coverage, semantic accuracy, and overgeneration measured

Evaluation

Language	Coverage raw	Coverage treebanked	Overgeneration	Spurious ambiguity	Average readings
Abkhaz	100%	94.4%	0%	2.8%	1.08
Chemehuevi	82.8%	75.9%	0%	3.4%	1.04
Hausa	42.1%	36.8%	6.7%	5.3%	1.31
Jingulu	100%	100%	0%	46.7%	2.00
Malayalam	89.7%	87.2%	2.8%	2.8%	1.09
Nkore-Kiga	78.6%	78.6%	11.5%	0%	1.00
West Greenlandic	93.9%	93.9%	0%	0%	1.00

Evaluation

Phenomenon	abk	hau	jig	kal	mal	nyn	ute
Negation	+	−	+	+	+	+	+/−
Yes−No Questions	−	−	+	+	+	+	−
Word Order	−	+/−	+	+	+	−	−
N/NP Coordination	+/−	+/−		−	+/−	+/−	+
S Coordination			+	−	−	+	+
V/VP Coordination		+/−			−	−	−
Determiners/Definiteness	−	−			+		
Tense/Aspect	+	+/−	+	+	+	+	+
Auxiliaries		+/−	+			+	
Morphology	+	+	+/−	+	+	+	+/−
Case			+	+	+/−		+
Verb Object Agreement	+		+	+		+	+
Verb Subject Agreement	+	+	+	+		+	+
Person	+	+	+	+	+	+	+
Number	+	+	+	+	+	+	+/−
Gender	+	+	+	+	+	+	+

Future work

- More libraries: Modifiers, embedded clauses of various types, wh-questions, information structure, ...

- Lexical acquisition

- MOM (Matrix-ODIN Mash-up): Can the customization system questionnaire be filled out automatically on the basis of information in ODIN (Lewis & Xia 2008)?