Site Update: University of Washington

Emily M. Bender

University of Washington

DELPH-IN Summit, 2 July 2010, Paris

Outline

UNIVERSITY OF WASHINGTON

Come and stay a while...

The Grammar Matrix Customization System is evolving from a place to download something to a place to stay and work

- More phenomena available through libraries
- $\longrightarrow$ it takes longer to input analyses
- $\longrightarrow$ more options to experiment with

UNIVERSITY OF WASHINGTON

Come and stay a while...

- Test by generation (Bender et al. 2010, ACL demo)
- Work in progress: Allowing users more flexibility in defining type hierarchies
  - Lexical entry and lexical rule types
  - Cross-classification
  - Fewer redundant constraints

The longer users stay, the bigger the grammars

UNIVERSITY OF WASHINGTON

Size of customized grammars (Ling 567, 2010)

Language	Family	Lg-specific types	Matrix+head types	Lex rules	Phrasal rules	Choices
ERG	Germanic	3654	N/A	71	226	N/A
Breton	Celtic	220	413+510	57	49	1692
Cherokee	Iroquoian	182	413+510	95	27	985
French	Romance	137	413+510	29	22	740
Jamamadí	Arauan	188	413+510	87	11	1151
Lushootseed	Salish	95	413+510	20	8	391
Nishnaabemwin	Algonquian	289	413+510	124	50	1754
Pashto	Iranian	234	413+510	86	19	1839
Pali	Indo-Aryan	237	413+510	92	55	1310
Russian	Slavic	190	413+510	56	35	993
Shona	Bantu	136	413+510	51	9	591
Vietnamese	Austro-Asiatic	105	413+510	2	26	362
Average		182.9	413+510	63.5	28.3	1073.5

UNIVERSITY OF WASHINGTON

Outline

Development of analyses

- Improvements to morphotactics (see Mike's talk and Goodman and Bender 2010 at MFG)
- Argument optionality (Saleem and Bender 2010, COLING)
- Information structure (see Emily and Sanghoun's talks on Tuesday)
- Improvements to word order (see Antske's talk later today)
- And more to come...

UNIVERSITY OF WASHINGTON

Outline

Evaluation

Current evaluation focus

How well do the analyses encoded in the core grammar + libraries scale across typologically and genetically diverse natural languages?

- Chose seven languages from seven different language families
- Created test suites
  - based on descriptive grammars
  - covering just those phenomena purported to be covered in the libraries
- Once test suites were frozen, created choices files based on descriptive grammars and test suites

UNIVERSITY OF WASHINGTON

The Map

(Created with Google Maps, using KML data from WALS Haspelmath et al. 2008.)

UNIVERSITY OF WASHINGTON

The Numbers

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

UNIVERSITY OF WASHINGTON

Qualitative 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 Subject Agreement	+	+	+	+		+	+
Verb Object Agreement	+		+	+		+	+
Person	+	+	+	+	+	+	+
Number	+	+	+	+	+	+	+/–
Gender	+	+	+	+	+	+	+

Outline

UNIVERSITY OF WASHINGTON

Matrix-ODIN Mash-up (MOM)

- The ODIN project is collecting IGT from pdfs published on the web (Lewis, 2006).
- Fei Xia and colleagues (e.g., Lewis and Xia 2008) have begun automatically creating typological profiles of languages based on this data.
- Could we get information rich enough to automatically fillout the Matrix customization system's questionnaire?
- Initial steps: Extracting verbal morphology 'choices' (David Wax, MA student)

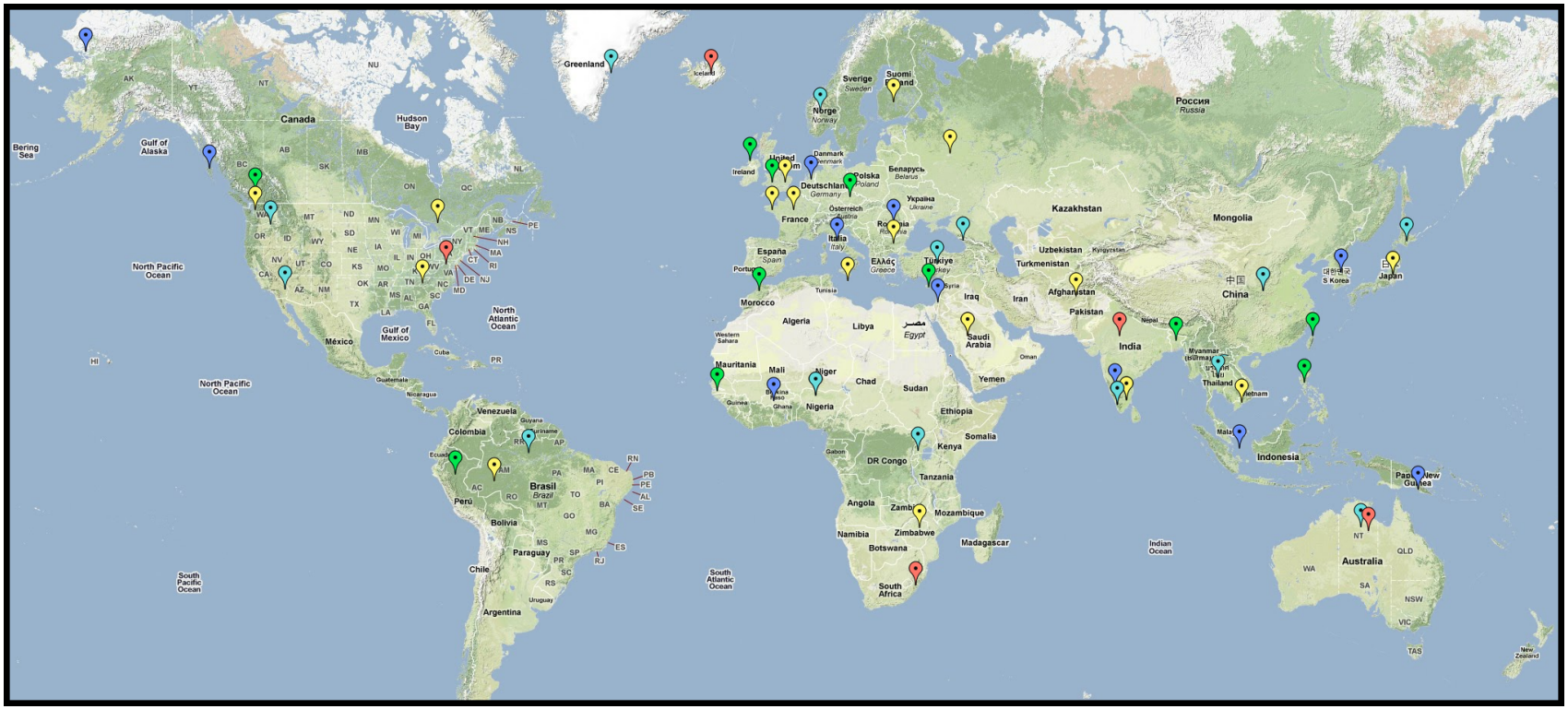
UNIVERSITY OF WASHINGTON

Translation mismatches

- Big picture goal: MT for many language pairs built on customized Matrix grammars and PanLex (Mausam et al., 2009)
- To what extent can we use simple pred-to-pred mapping for transfer rules? (Nygaard et al., 2006)
- Measuring and classifying translation divergence (Dorr, 1994) across a sample of translation pairs in PanLex (Francesca Gola's MA thesis)

UNIVERSITY OF WASHINGTON

Just for fun: Customized languages map



Map image courtesy of Google Maps, location data courtesy of WALS (Haspelmath et al., 2008).

Outline

Visibility

- All three major NLP explicitly included grammar engineering this year.
- We made a point of doing our part to show that grammar engineering merits such attention.
  - With some success! (Drellishak, 2010; Bender et al., 2010; Saleem and Bender, 2010)
- Also, tutorials: LREC, HPSG

UNIVERSITY OF WASHINGTON

Bibliography I

Bender, E. M., Drellishak, S., Fokkens, A., Goodman, M. W., Mills, D. P., Poulson, L., and Saleem, S. (2010). Grammar prototyping and testing with the lingo grammar matrix customization system. In *Proceedings of ACL 2010, Software Demonstrations*.

Dorr, B. J. (1994). Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Drellishak, S. (2010). Testing a grammar customization system with sahaptin. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 254–262, Los Angeles, California. Association for Computational Linguistics.

Goodman, M. W. and Bender, E. M. (2010). What's in a word? refining the morphotactic infrastructure in the lingo grammar matrix customization system. Poster presented at the Workshop on Morphology and Formal Grammar at HPSG 2010.

Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B., editors (2008). *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. http://wals.info.

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

Lewis, W. D. and Xia, F. (2008). Automatically identifying computationally relevant typological features. In *Proceedings of The Third International Joint Conference on Natural Language Processing (IJCNLP)*.

Mausam, Soderland, S., Etzioni, O., Weld, D., Skinner, M., and Bilmes, J. (2009). Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 262–270, Suntec, Singapore. Association for Computational Linguistics.

UNIVERSITY OF WASHINGTON

Bibliography II

Nygaard, L., Lønning, J. T., Nordgård, T., and Oepen, S. (2006). Using a bi-lingual dictionary in lexical transfer. In *Proceedings of the 11th Conference of the European Association for Machine Translation*, Oslo, Norway.

Saleem, S. and Bender, E. M. (2010). Argument optionality in the LinGO Grammar Matrix. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China. COLING 2010 Organizing Committee.

UNIVERSITY OF WASHINGTON