Crosslinguistic Resources for the Rapid Development of Precision Computational Grammars

Emily M. Bender

Dept of Linguistics, University of Washington

UW CSE Colloquium

May 17, 2005

Overview

- *Why build precision grammars?*
- *Hurdles to robust processing with precision grammars*
- *The Grammar Matrix*
  - *Crosslinguistic core*
  - *Modeling variation with 'modules'*
- *Future work: software support for endangered language documentation*

Background:
Why build precision grammars?

Natural language syntax & semantics

- *Constituent structure*
- *Mapping of linear string to predicate-argument structure (word order, case, agreement)*
- *Long distance dependencies*
  - What did Kim think Pat said Chris saw?
- *Idioms, collocations*

Formal/‘Generative’ Grammars

- *Characterize a set of strings (phrases and sentences)*
- *These strings should correspond to those that native speakers find acceptable*
- *Assign one or more syntactic structures to each string*
- *Assign one or more semantic structures to each string*

Formal/‘Generative Grammars

- *No complete generative grammar has ever been written for any language*

Precision Computational Grammars

- *Knowledge engineering of formal grammars, for:*
- *Parsing: assigning syntactic structure and semantic representation to strings*
- *Generation: assigning surface strings to semantic representations*

Why build precision grammars?

- *Linguistic hypothesis testing*
  - *Test interacting analyses for consistency*
  - *Test grammar against test suites and naturally occuring text*
  - *More precise language documentation*

Why build precision grammars?

- *'Deep' NLP/NLU*
  - *Automated customer service response*
  - *Machine translation (symbolic, hybrid)*
  - *Speech prostheses*
  - *Hybrid Q&A systems*

Why build precision grammars?

- *'Deep' NLP/NLU*
  - *Human-computer dialog/collaboration*
  - *Machine mediated human-human interaction*
  - *Better treebanks*

Background:
Hurdles to robust processing with precision grammars

Hurdles

- *Efficient processing* *(Oepen et al 2002)*
- *Ambiguity resolution* *(Baldridge & Osborn 2003, Toutanova et al 2005, Riezler et al 2002)*
- *Domain portability* *(Baldwin et al 2005)*
- *Lexical acquisition* *(Baldwin & Bond 2003, Baldwin 2005)*
- *Extragrammatical/ungrammatical input* *(Baldwin et al 2005)*
- *Scaling to many languages*

The LinGO Grammar Matrix

The Grammar Matrix: Overview

- *Motivation*
- *HPSG*
- *Semantic representations*
- *Cross-linguistic core*
- *Modules*

Matrix: Motivation

- *English Resource Grammar:*
  - *140,000 lines of code (25,000 exclusive of lexicon)*
  - *~3000 types*
  - *16+ person-years of effort*
- *Much of that is useful in other languages*
- *Reduces the cost of developing new grammars*

Matrix: Motivation

- *Hypothesis testing (monolingual and cross-linguistic)*
  - *Interdependencies between analyses*
  - *Adequacy of analyses for naturally occurring text*

Matrix: Motivation

- *Promote consistent semantic representations*
  - *Reuse downstream technology in NLU applications while changing languages*
  - *Transfer-based (symbolic or stochastic MT)*

The Grammar Matrix: Overview

- *Motivation*
- *HPSG*
- *Semantic representations*
- *Cross-linguistic core*
- *Modules*

HPSG

- *Head-Driven Phrase Structure Grammar (Pollard & Sag 1994)*
- *Mildly-context sensitive (Joshi et al 1991)*
- *Typed feature-structures*
- *Declarative, order-independent, constraint-based formalism*

An HPSG consists of

- *A collection of feature-structure descriptions for phrase structure rules and lexical entries*
- *Organized into a type hierarchy, with supertypes encoding appropriate features and constraints inherited by subtypes*
- *All rules and entries contain both syntactic and semantic information*

An HPSG is used

- *By a parser to assign structures and semantic representations to strings*
- *By a generator to assign structures and strings to semantic representations*
- *Rules, entries, and structures are DAGs, with type name labeling the nodes*
- *Constraints on rules and entries are combined via unification*

Example rule type

$$\textit{head-subj-phrase}: \begin{bmatrix} \textit{binary-headed-phrase}\ \& \\ \textit{head-compositional} \\ \text{SUBJ} & \langle\ \rangle \\ \text{COMPS} & \boxed{1} \\ \text{HEAD-DTR} & \begin{bmatrix} \text{SUBJ} & \langle \boxed{2} \rangle \\ \text{COMPS} & \boxed{1} \end{bmatrix} \\ \text{NON-HEAD-DTR} & \boxed{2} \end{bmatrix}$$

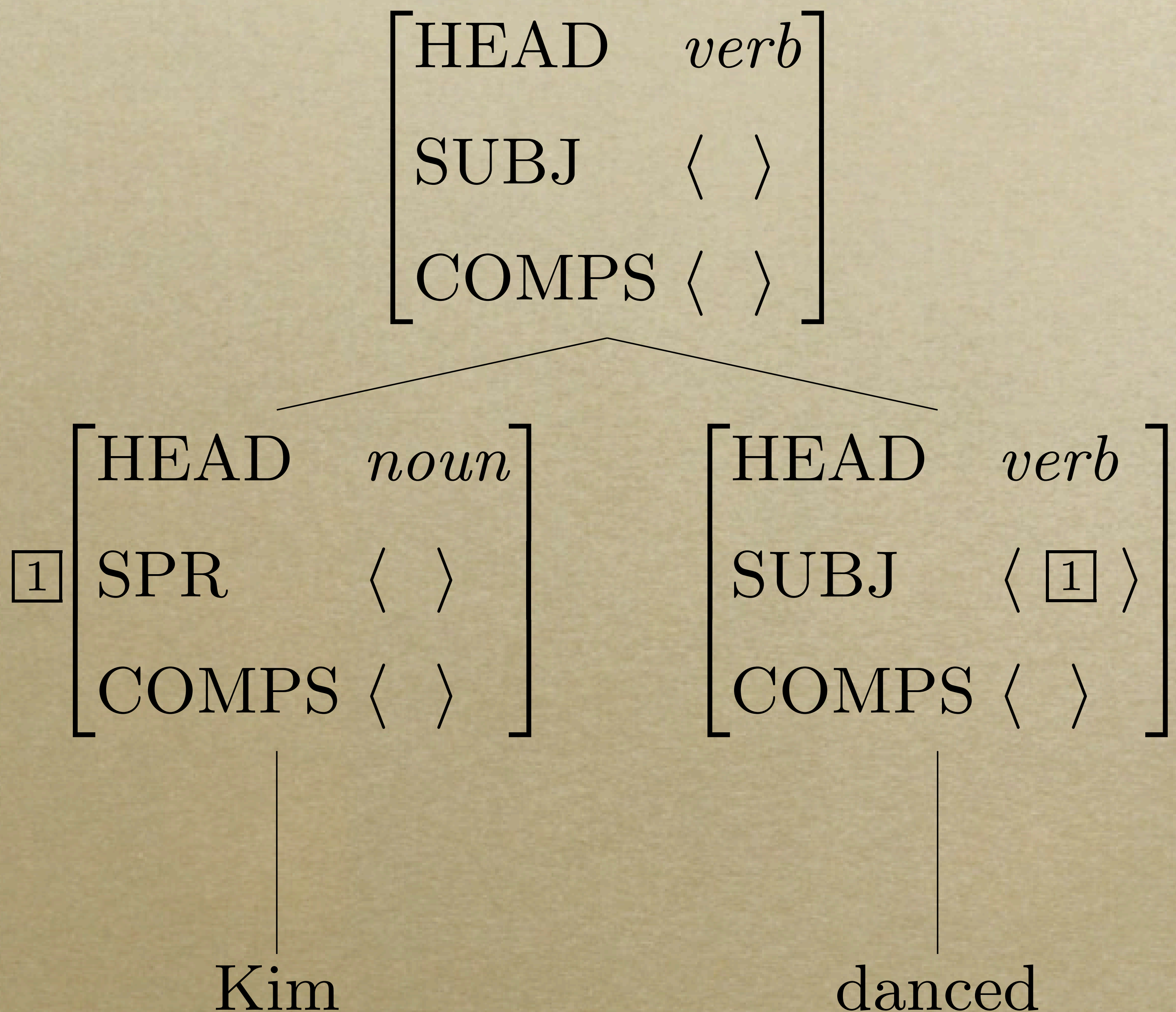
Example rule type

head-final:

$$\begin{bmatrix} \textit{binary-headed-phrase}\ \& & \\ \text{HEAD-DTR} & \boxed{1} \\ \text{NON-HEAD-DTR} & \boxed{2} \\ \text{ARGS} & \langle\ \boxed{2}, \boxed{1}\ \rangle \end{bmatrix}$$

subj-head: *head-subj-phrase* & *head-final*

Example parse



The Grammar Matrix: Overview

- *Motivation*
- *HPSG*
- *Semantic representations*
- *Cross-linguistic core*
- *Modules*

Semantic Representations

- *Not going for an interlingua*
- *Not representing connection to world knowledge*
- *Not representing lexical semantics (the meaning of life is life')*
- *Making explicit the relationships among parts of a sentence*

Semantic Representations

- *Kim gave a book to Sandy*
- *give(e,x,y,z), name(x,'Kim'), book(y), name(z,'Sandy'), past(e)*

Semantic Representations

- *Sandy was given a book by Kim*
- *Kim continues to give books to Sandy*
- *This is the book that Kim gave Sandy*
- *Which book did Kim give Sandy?*
- *Which book do people often seem to forget that Pat knew Kim gave to Sandy?*
- *This book was difficult for Kim to give to Sandy.*

Semantic representations

- *Minimal Recursion Semantics* *(Copestake et al, forthcoming)*
  - *Expressive adequacy*
  - *Computational tractability*
  - *Grammatical compatibility*
  - *Underspecifiability*

Semantic representations

- *MRS specifies well-formedness*
- *Matrix specifies representations*
    - *Nominal v. verbal predicates*
    - *Quantifiers*
    - *Illocutionary force*
    - *Coordination*

Semantic representations

- *Languages may still differ:*
  - *Lexical predicates*
    - *Japanese: kore, sore, are*
  - *Grammaticized tense/aspect, discourse status*
  - *Ways of saying*
    - *make a wish, center divider*

Design criteria

- *Strip all syntactic information*
- *Stay lexically close to the surface (for hybrid deep/shallow systems)*
- *Encode all distinctions marked in the surface from*
- *Leave underspecified all else that can be computed*

The Grammar Matrix: Overview

- *Motivation*
- *HPSG*
- *Semantic representations*
- *Cross-linguistic core*
- *Modules*

Matrix: Cross-linguistic core

- *Types defining feature geometry*
- *Types encoding compositional semantics*
- *General classes of phrase structure rules*
- *General classes of lexical items*
- *Configuration and parameter files for LKB (Copestake 2002) and PET (Callmeier 2000)*

Matrix: Hypothesized universals

- *Words and phrases combine to make larger phrases.*
- *The semantics of a phrase is determined by the meaning of its parts and how they're put together.*

Matrix: Hypothesized universals

- *Some rules for phrases add semantics, some don't.*
- *No rule can remove semantic information.*
- *Most phrases have an identifiable head daughter.*

Matrix: Hypothesized universals

- *Heads determine the type of arguments they require, and how they combine semantically with those arguments.*
- *Modifiers determine the type of heads they modify, and how they combine semnatically with the head.*

The Grammar Matrix: Overview

- *Motivation*
- *HPSG*
- *Semantic representations*
- *Cross-linguistic core*
- *Modules*

Modules: Motivation

- *Many patterns are not universal, yet recurring*
  - *Systems represented in every language:*
    - *word order, negation, questions*
  - *Systems/patterns represented in some languages:*
    - *noun incorporation, numeral classifiers, verb particle construction*

Modules: Motivation

- *Promote reuse of code*
- *Promote consistency of analyses*
- *Sometimes the same technical solution is useful in different constructions across different languages.*

Modules: Motivation

- *Both Basque and Latin have free word order*
- *Except:*
    - *Basque embedded clauses are verb-final*
    - *Latin yes-no questions are verb-initial*

Modules: Open issues

- *How independent can modules be?*
- *How do we design a UI allowing the linguist to find the relevant modules?*

Modules: Proof of concept

- *Implemented modules for word order, negation, yes-no questions*
- *Tested against a convenience sample of 7 languages*
- *Developed abstract test suites for each language*

Modules: Proof of concept

Language	Word order	Negation	Yes-no Q
Hindi	SOV	pre-V adv	sentence-initial particle
Japanese	V-final	verbal suffix	sentence-final particle
Mandarin	SVO	pre-V adv	sentence-final particle, A-not-A
Polish	free	pre-V adv	sentence-initial particle
Slave	SOV	post-V adv	sentence-initial particle
English	SVO	post-aux adv	aux inversion
Spanish	SVO	pre-V adv	main verb inversion

Modules: Proof of concept

Language	Pos.	Coverage	Neg.	Over-generation
Hindi	5	100%	10	10%
Japanese	6	100%	8	0%
Mandarin	4	75%	9	0%
Polish	14	100%	8	0%
Slave	3	100%	6	0%
English	5	80%	11	45%
Spanish	5	80%	8	25%

Further planned modules

- *Coordination*
- *Content questions*
- *Relative clauses*
- *Case, agreement*
- *Tense, aspect, mood*
- *Marking of discourse status*

Outlook:
Assisting endangered language documentation

Current state of the art

- *Existing crosslinguistic core and modules sufficient for rapid prototyping*
- *Results suitable as basis for sustained development*

This year’s Ling 567

- *Basic word order*
- *Case, agreement*
- *Adjectival and adverbial modifiers*
- *Matrix/embedded statements/questions*
- *Coordination*
- *Sentential negation*

The Montage vision

- *A field linguist working on an endangered language*
- *Builds a precision grammar by selecting modules as she learns the facts of the language*
- *Uses the precision grammar to test hypotheses against collected texts, find relevant examples*

The Montage vision

- *Works with a grammar engineer to further fine-tune the precision grammar*
- *Produces language resources (annotated corpora, prose grammar, precision grammar) which are ontologically indexed for smart searching*

The Montage vision

- *Word-wide database of linguistic data and analyses*
- *Machine-readable language resources for minority languages*

Work to be done

- *More modules*
- *Module UI*
- *Data-exchange infrastructure*
- *Ontological indexing of complex objects*
- *Robust processing with partial grammars*

Acknowledgments

- *Matrix: Dan Flickinger, Stephan Oepen, Scott Drellishak*
- *Montage: Jeff Good, Laurie Poulson, Anya Dormer, David Goss-Grubs*
- *Linguistics 471 (2004) and 567 (2005)*

References

- *http://www.delph-in.net/matrix/*
- *A version of these slides with full bibliography will be available online: http://faculty.washington.edu/ebender/*

References

Baldridge, J. and M. Osborne. 2003. Active learning for HPSG parse selection. In Proceedings of the 7th Conference on Natural Language Learning.

Baldwin, T., J. Beavers, E.M. Bender, D. Flickinger, A. Kim and S. Oepen. In press, 2005. Beauty and the Beast: What running a broad-coverage precision grammar over the BNC taught us about the grammar---and the corpus. Kesper, Stephan and Marga Reis (eds). *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*. Mouton de Gruyter.

Baldwin, T. To appear. The Deep Lexical Acquisition of English Verb-particle Constructions, *Computer Speech and Language, Special Issue on Multiword Expressions*.

Baldwin, T. and F. Bond (2003) Learning the Countability of English Nouns from Corpus Data, In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, pp. 463–70.

Bateman, John A., Ivana Kruijff-Korbayová, and Geert-Jan Kruijff. In press. Multilingual resource sharing across both related and unrelated languages: an implemented, open-source framework for practical natural language generation. *Journal of Research on Language and Computation*.

Bender, E.M., D. Flickinger and S. Oepen. 2002. The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. Carroll, J., N. Oostdijk, and R. Sutcliffe, eds. *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*. Taipei, Taiwan. pp. 8-14.

Bender, E.M., D. Flickinger, J. Good and I.A. Sag. 2004. Montage: Leveraging Advances in Grammar Engineering, Linguistic Ontologies, and Mark-up for the Documentation of Underdescribed Languages. *Proceedings of the Workshop on First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, LREC 2004*, Lisbon, Portugal.

Callmeier, U. 2000. PET - A platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering*, 6 (1) (Special Issue on Efficient Processing with HPSG):99-108.

Carroll, J., A. Copestake, D. Flickinger and V. Poznanski, 1999. An Efficient Chart Generator for (Semi-)Lexicalist Grammars. In: Proceedings of the 7th European Workshop on Natural Language Generation (EWNLG'99), Toulouse.

Copestake, A. 2002. *Implementing Typed Feature Structure Grammars*. Stanford: CSLI Publications.

References

Copestake, A., D. Flickinger, C. Pollard, and I.A. Sag. ms, 2003. Minimal Recursion Semantics: an Introduction.
Flickinger, D. and E.M. Bender. 2003. Compositional Semantics in a Multilingual Grammar Resource. In E. M. Bender, D. Flickinger, F. Fouvry, and M. Siegel (Eds.), *Proceedings of the Workshop on Ideas and Stratgies for Multilingual Grammar Development*, ESSLLI 2003, Vienna. pp.33-42.
Hellan, L., and P. Haugereid. 2003. Norsource: An exercise in matrix grammar-building design. In E. M. Bender, D. Flickinger, F. Fouvry, and M. Siegel (Eds.), *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development*, ESSLLI 2003, 41–48, Vienna, Austria.
Joshi, A., K. Vijay-Shanker, and D. Weir. 1991. The Convergence of Mildly Context-Sensitive Grammar Formalism. In P. Sells, S. Shieber, and T. Wasow (eds). *Processing of Linguistic Structure*. MIT Press. pp. 31-81.
King, T.H., M. Forst, J. Kuhn and M. Butt. In press. The Feature Space in Parallel Grammar Writing. *Journal of Research on Language and Computation.*
Kordoni, V., and J. Neu. 2003. Deep gramamr development for Modern Greek. In E. M. Bender, D. Flickinger, F. Fouvry, and M. Siegel (Eds.), *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development*, ESSLLI 2003, 65–72, Vienna, Austria.
Oepen, S., D. Flickinger, J. Tsujii, an H. Uszkoreit, editors. *Collaborative Language Engineering. A Case Study in Efficient Grammar-Based Processing*. CSLI Publications, Stanford, CA, 2002.
Oepen, S., D. Flickinger, K. Toutanova, and C.D. Manning. forthcoming, 2005. LinGO Redwoods: A Rich and Dynamic Treebank for HPSG. To appear in *Research on Language and Computation*.
Osborne, M. and J. Baldridge. 2004. Ensemble-based Active Learning for Parse Selection. In Proceedings of HLT-NAACL 2004.
Pollard, C. and I.A. Sag. 1994. *Head-Driven Phrase Structure Grammar.* Chicago: University of Chicago Press.
Riezler, S., T.H. King, R.M. Kaplan, R. Crouch, J.T. Maxwell, and M. Johnson. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadephia, PA.
Siegel, M. and E.M. Bender. 2004. Head-Initial Constructions in Japanese. Müller, S. (ed). *Proceedings of the 11th International Conference on Head-Driven Phrase Structure Grammar*. Stanford: CSLI. pp.244-260.
Toutanova, K., C.D. Manning, D. Flickinger, and S. Oepen. forthcoming, 2005. Stochastic HPSG Parse Disambiguation using the Redwoods Corpus. To appear in *Research on Language and Computation*.