Cross-Linguistic Grammar Engineering
Emily M. Bender
February 25, 2003
Some Big Grammars

• LinGO ERG (English), as of Nov, 2001:
  • 17,430 types, including:
    • 8,756 lexical entries
    • 65 syntactic constructions
  • 10+ person-years of development
  • Developed for spoken dialogue system (Verbmobil)
  • Adapted to email text, planned adaptation to edited text

Primary author: Dan Flickinger
Other contributors: Emily Bender, Ann Copestake, Rob Malouf, Ivan Sag, Jeff Smith, and others
Some Big Grammars

- JACY (Japanese) as of Jan, 2002:
  - 6,849 types, including:
    - 5,103 lexical entries
    - 50 syntactic constructions
  - 5 person-years of development
  - Developed for spoken dialogue system (Verbmobil)
  - Adapted to email text

Primary author: Melanie Siegel
Other contributors: Emily Bender, Atsuko Shimada
Mining Grammars for Information

- Years of accumulated knowledge
- ... trapped under years of rubble analyses
- Poorly documented
- Authors not always available for consultation
The Grammar Matrix

- Distill the knowledge from existing grammars
- Document it
- Produce a starter-kit for developing new grammars
  - Quick start-up
  - Rapid growth towards wide coverage
The Matrix: Motivations

• Linguistic:
  • Bottom-up approach to language universals

• Computational:
  • Matrix-derived grammars could provide an interesting proving ground for various stochastic experiments

• Practical:
  • Standardization: grammars for different languages ready to talk to the same back-end machinery
  • Reduce hurdle to commercial deployment of wide-coverage grammars: the immense cost of developing the resource
The Matrix: Contents (Preliminary)

- Basic feature geometry & technical devices (diff-lists)
- General classes of rules
- Basic constructions
- Semantics:
  - relation types
  - compositional semantics
  - illocutionary force
  - constructional semantics (C-CONT or C-SEM)
- Configuration and parameter files for LKB

Notice anything missing?
Step Two: The Lexicon

- Put off in the first instance due to difficulty:
  - The bulk of the work of an HPSG is the lexicon
  - Many dimensions of linguistic information (part of speech, valence, semantics, linking)
  - Wide cross-linguistic variation
Current Research: Complementation Patterns

- 5 interacting phenomena:
  - Binding theory
  - Raising/control constructions
  - Optionality/obligatoriness of arguments
  - Intra-clausal scrambling
  - Mobility with respect to modifiers
Binding Theory

• *Sara told herself/*himself about Jason.*

• Motivates a level of representation (ARG-ST) where all arguments are on one list, separate from valence features (subject to cancellation).

• Leave mapping from ARG-ST to valence features (‘argument realization’) as language-particular.

• ARG-ST is also the locus of linking.
Raising/Control

- Alex tried/continued to annoy Dana.
- We take this phenomenon to be (near-)universal.
- As such, we should provide types for lexical entries instantiating these constructions.
- The relevant constraints are constraints on ARG-ST.
Optionality of Arguments

- English has relatively little optionality in the sense that any missing argument has to be licensed
  - constructionally: imperatives, ‘recipe-style’ object-drop, VP ellipsis
  - lexically: eat v. devour

- Japanese has pervasive optionality
  - Rampant pro-drop; indeed no overt inanimate personal pronouns
  - And yet, certain arguments are obligatory (complement of nominalizing heads, number argument to numeral classifiers)
Optionality of Arguments

- We need:
  - A way of indicating which arguments are optional
  - A mechanism for discharging optional arguments
  - A way to set the default value of optionality

- The matrix will provide:
  - A feature [OPT bool] of synsems
  - (At least one) unary rule discharging [OPT +] elements
  - The ability to set default values of OPT at varying points in the type hierarchy
Intra-clausal Scrambling

- Suzuki-ga Hanako-ni hon-wo ageta.
  Suzuki-NOM Hanako-DAT book-ACC gave
  ‘Suzuki gave Hanako the book.’

- Hanako-ni Suzuki-ga hon-wo ageta.

- Hon-wo Hanako-ni Suzuki-ga ageta.

- Hanako-ni hon-wo Suzuki-ga ageta.
Intra-clausal Scrambling

- Multiple Head-Complement Rules that can pick any argument out of the COMPS list.
- In particular languages, Head-Subject Rule can be allowed to interleave with Head-Complement Rule.
- In particular languages, non-initial HCR can be made sensitive to various properties of the complements.
- *Kim talked to Sandy about Alex.*
- *Kim talked about Alex to Sandy.*
Mobility With Respect to Modifiers

• The book was given (on Tuesday) to Kim (on Tuesday).

• The book was purchased (on Monday) by Sandy (on Monday).

• Provide a feature [MOB bool] on synsems

• Head-Modifier Rule is potentially sensitive to MOB value of remaining undischarged complements.

• We may find a type like moblist (i.e., list([MOB +])) useful.