Dan Flickinger & Emily M. Bender Semantic Selection in a Cross-linguistic Framework Workshop on Semantics in Grammar Engineering HPSG 2004, Leuven, Belgium August 3, 2004





### Overview

- The LinGO Grammar Matrix
- Grammaticized selection v. defeasible semantic constraints
- Mechanisms & Examples
- Conclusion

#### The LinGO Grammar Matrix: Overview

- A cross-linguistically useful core grammar/starter-kit
- Jump start the development of precision grammars
- Facilitate steady expansion to broad coverage
- Facilitate exchange of analyses across grammars
- Support mapping to standardized semantic representations (in MRS; Copestake et al. 2003)

#### The LinGO Grammar Matrix: Languages (1/2)

- Developed initially on the basis of ERG (Flickinger 2000) and JACY (Siegel and Bender 2002)
- Early deployment in Norwegian (Hellan and Haugereid 2003), Modern Greek (Kordoni and Neu 2004) and Italian (CELI)
- In grammar engineering courses, used for small grammars of Akan, Arabic, Armenian, Basque, Cantonese, French, Farsi, Haitian Creole, Hindi, Hungarian, Japanese, Latin, Mongolian, Navajo, Polish, Portuguese, Russian, Spanish, Swedish, Tigrinya, and Turkish.

#### The LinGO Grammar Matrix: Languages (2/2)

- Most recently, work has begun on a Matrix-based grammar for Slave (Athabaskan)
- Intended development into component of Montage toolkit for endangered language documentation

#### The LinGO Grammar Matrix: Components (1/2)

- Types defining basic feature geometry
- Types defining somewhat underspecified phrase structure rules, very underspecified lexical rules
- Extensive support for the syntax-semantics interface
- Somewhat underspecified set of lexical types (argument structure, linking)

The LinGO Grammar Matrix: Components (2/2)

- Collateral files for interaction with the LKB (Copestake 2002)
- Future work:
  - More lexical types
  - "Modules" for non-universal recurring patterns
  - Solution to dilemma of head-types, and corresponding additions to lexical type hierarchy

# Grammaticized selection v. defeasible semantic constraints

- Most constraints in the category "semantic selection" are defeasible in the right context (e.g., fairy tales).
- A few seem to genuinely related to grammaticality/ungrammaticality.
- Unification only seems appropriate to the latter.
- The former might be tractable with persistent defaults or (more likely) stochastic methods (e.g., Resnik 1996).
- Strategy: Minimize spurious ambiguity while still allowing all grammatical sentences, even unlikely ones

#### Grammaticized selection

- Examples
  - Verbs selecting prepositions (English)
  - Prepositions selecting temporal NPs (English)
  - Relativizers (English)
  - Cranberry words (German)
  - Athabaskan "gender" (Slave)
- Eagerly awaiting the results of an Autotyp project on this topic.

## Verbs selecting prepositions

• Empty or 'marker' prepositions:

We dispensed with/\*using/\*by means of needless formalities.

• Contentful yet still selected prepositions:

They presented Jack with/\*using/by means of a check for \$2 million.

cf. They appeased Jack with/using/by means of a check for \$2 million.

(Tseng 2000)

#### Prepositions selecting temporal NPs

- Examples
  - We met in/\*on/\*at October.
  - We met \*in/on/\*at Tuesday.
  - We met \*in/\*on/at five o'clock.
- Each of the prepositions (in its temporal use) selects for a particular class of nouns.

### Relativizers

- Examples
  - I know a book that/which/??who likes to eat soup.
  - The day that/\*which/\*who I arrived was sunny.
- Contrast the corresponding question words, which seem easier to contextualize.
  - Who wants to eat soup? [To a shelf of books.]
  - Who will take Myday? [In the context of *The Day that Monday Ran Away*]
- If *day* takes a *who/which* relative clause, it can't be an extracted adjunct.
  - The day who took Myday was very sleepy.

## Cranberry words

• German (Soehn and Sailer 2003:151):

Das Mädchen schmetterte ein Lied aus/\*von

the girl belted.out a song from/\*with

Herzenslust und Leibeskräften

heart.delight and body.power 'The girl belted out a song to the top of her bent and with all her heart.'

- Herzenslust, Leibeskräften occur only in PP[aus].
- Possibility of coordination (with restricted and non-restricted NPs) rules out a listing analysis.
- Syntax is regular, semantics is compositional.

## Athabaskan "gender" (1/2)

- Some Slave verbs (as well as postpositions and noun modifiers) show agreement with noun class, where the classes are "area", "wooden", and everything else (unmarked).
  - kộé gogháyįda
  - house 2sg.see.area
  - 'You sg. look at the house.'
  - nǫhbáli gháyįda
  - tent 2sg.see
  - 'You sg. look at the tent.'

(Rice 1989:1024)

### Athabaskan "gender" (2/2)

- In the literature, this is called gender, but it appears more semantically motivated than the arbitrary gender systems found in European languages.
- Furthermore, there is lexical idiosyncrasy concerning which verbs take the gender markers (Rice 1989:1023).
- Finally, the class of nouns requiring gender agreement appears somewhat arbitrary.
- Thus these data might be more amenable to an analysis in terms of grammaticized semantic selection rather than gender agreement.

#### Grammaticized selection: summary

- Violation results in strict ungrammaticality
- Possibly grammaticized from earlier patterns of semantic selection
- Retain some degree of productivity/compositionality
- In some cases particular words are selected for (canonically closed-class words)
- In other cases, more or less semantic classes of words are selected

#### Defeasible semantic constraints

- Under normal circumstances, violations cause anomaly.
- The same strings are perfectly grammatical (with the ordinary meanings of each word) in unusual contexts.
  - Animacy, concreteness, etc. (English)
  - Numeral classifiers associated with particular classes of nouns (Japanese)
  - Certain determinerless PPs, where the P selects an N' (English)

Animacy, concreteness, etc.

- Examples:
  - # That idea is green.
  - # Who wants to eat soup? (To a shelf of books)
- Contextualizable in proper fairy-tale contexts, even with ordinary meanings of all of the words.

Japanese numeral classifiers (1/2)

- Example:
  - san biki no gakusei ga hanashiatta
     three NumCl GEN students NOM discussed
     'The three students (who were small animals) discussed (it).'
- Decidedly odd in normal context.
- Somewhat acceptable in a story where the school is populated with animals (although the numeral classifier *-nin* 'person' would be more common there).

Japanese numeral classifiers (2/2)

- Acceptable, though rude, to express contempt of the students.
- Attested example: 1964 movie *Sanbiki no samurai*, in which the samurai live like animals.
- Nonetheless, numeral classifier selection is an important part of fluent prose generation (Bond and Paik 2000).

### Determinerless PPs (1/2)

- Determinerless PPs (Baldwin et al. To appear)
  - They arrived by/\*with/\*in train/hydrofoil/pogo stick...
  - They arrived by carpet/pencil/mouse pad/room...
- NB: These nouns don't occur without determiners in other contexts

#### Determinerless PPs (2/2)

- Baldwin et al. (To appear) propose an entry for *by* which selects an N' complement, and provides the means/instrument semantics.
- Some nouns are more easily construed as means/instruments than others, but it appears that in suitable contexts, any noun can be so construed.

#### Defeasible semantic constraints: Summary

- Languages abound with very specific semantic selectional constraints.
- These constraints tend to be *graded* and *defeasible*.
- ... which suggests that unification is not the right tool.
- Possible alternatives include unification with persistent defaults (cf. Terkourafi and Villavicencio 2003), stochastic parse/realization ranking (Oepen et al. 2002), and corpus-based learning of selectional associations (Resnik 1996).
- Links between semantic predicates and an ontology might facilitate some of these alternatives.

## Mechanisms I: HEAD.KEYS.KEY

- We have narrowed the domain to cases where particular words or classes of words are non-defeasibly selected.
- In many cases, the required word will be the lexical head of the selected constituent.
- Treat these situations as sub-classifications of HEAD values, via a HEAD feature, KEYS.KEY.
- Akin to PFORM feature from GPSG, but generalized.
- For closed-class items, KEY and LKEYS.KEYREL might be in one-to-one relationship.
- Allow, e.g., on and upon to appear in the same contexts.

#### HEAD.KEYS.KEY: Example

```
on_day := p_temp_le &
  [ STEM < "on" >,
   SYNSEM [ LKEYS [ --COMPKEY day_rel,
        KEYREL.PRED _on_p_temp_rel ]]].
```

```
n_day_of_week_le := norm_n_temp_ppcomp_lexent &
  [ ...KEY dofw_rel ].
```

#### Mechanisms II: INDEX.SORT

- In other cases, the required item is not the syntactic head of the selected constituent.
- Therefore, we also provide for semantic typing on the *index*, which is not necessarily contributed by the same lexical element as the HEAD value.

#### INDEX.SORT: Example

- It usually lasts from two o'clock to three o'clock.
- I walked from the house to the car.
- •\*I walked from the house to three o'clock.

p\_ditrans\_from\_to\_le := p\_ditrans\_lexent &
 [ ...COMPS < [ ...INDEX.SORT #sort ],
 [ ...KEYS.KEY selected\_rel,
 ...INDEX.SORT #sort ] > ].

#### Mechanisms: Summary

- Allow identification of classes of words or of particular (most likely closed-class) words via INDEX.SORT and HEAD.KEYS.KEY.
- By hypothesis, any grammaticized selection will be in situations where either the INDEX or the HEAD value of the selected item are visible.
- Other phenomena under the rubric of 'semantic selection' are best handled with other means, e.g., stochastic parse/realization ranking.

### Conclusions (1/2)

- Strong notion of grammaticality prevents ruling out *The teapot giggled*.
- Still can rule out \**We'll meet on October*, since it's not contextualizable (with same senses of words).
- Linguistic advantages: Monotonicity, accuracy of model
- Engineering advantages: Reusability
- Increased ambiguity is manageable through stochastic methods.

## Conclusions (2/2)

- Provide two mechanisms for identifying words or classes of words, along syntactic and semantic head paths.
- While the phenomena captured by these mechanisms probably reflect historically grammaticized semantic selection, we don't see them as synchronically semantic.
- We predict that these two mechanisms will be cross-linguistically sufficient for the kind of selection discussed here (idioms are another matter, cf. Riehemann 2001).
- We expect the particular classes/individual words subject to this kind of selection to be language particular.

## References

Baldwin, Timothy, John Beavers, Leonoor van der Beek,
Francis Bond, Dan Flickinger, and Ivan A. Sag. To
appear. In search of a systematic treatment of
determinerless PPs. In P. Saint-Dizier (Ed.), *Computational Linguistics Dimensions of Syntax and Semantics of Prepositions*. Kluwer Academic.

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 *Proceedings of the Workshop on Grammar Engineering and Evaluation*

at the 19th Internation Conference on Computational Linguistics, 8–14, Taipei, Taiwan.

Bond, Francis, and Kyoung-Hee Paik. 2000. Reusing an ontology to generate numeral classifiers. In *Coling 2000*, Saarbrücken, Germany.

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

Copestake, Ann, Daniel P. Flickinger, Ivan A. Sag, and Carl Pollard. 2003. Minimal Recursion Semantics. An introduction. Unpublished ms.

Flickinger, Dan. 2000. On building a more efficient grammar by explointing types. *Natural Language Engineering* 

6(1).

- Hellan, Lars, and Petter Haugereid. 2003. Norsource: An exercise in the matrix grammar building design. In
  E. M. Bender, D. Flickinger, F. Fouvry, and M. Siegel (Eds.), *Proceedings of the ESSLLI 2003 Workshop Ideas and Strategies for Multilingual Grammar Development*.
- Kordoni, Valia, and Julia Neu. 2004. Deep analysis of modern greek. In *Proceedings of the First International Joint Conference on Natural Language Processing* (*IJCNLP-04*), Sanya City, Hainan Island, China.
- Oepen, Stephan, Kristina Toutanova, Stuart Shieber, Chris Manning, Dan Flickinger, and Thorsten Brants. 2002. The LinGO Redwoods treebank. Motivation and

preliminary applications. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Resnik, Philip. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61:127–159.

Rice, Keren. 1989. *A Grammar of Slave*. New York: Mouton de Gruyter.

Riehemann, Susanne. 2001. A Constructional Approach to Idioms and Word Formation. PhD thesis, Stanford University.

Siegel, Melanie, and Emily M. Bender. 2002. Efficient deep

processing of japanese. In *Proceedings of the 3rd Workshop on Asian Language Resources and Standardization at the 19th Internation Conference on Computational Linguistics*, Taipei, Taiwan.

Soehn, Jan-Philipp, and Manfred Sailer. 2003. At first blush on tenterhooks: About selectional restrictions imposed by nonheads. In G. Jäger, P. Monachesi, G. Penn, and S. Wintner (Eds.), *Proceedings of Formal Grammar* 2003, 149–161.

Terkourafi, Marina, and Aline Villavicencio. 2003. Toward a formalisation of speech act functions of questions in conversation. In R. Bernardi and M. Moortgat (Eds.), *Questions and Answers: Theoretical and Applied*  Perspectives. Utrecht Institute of Linguistics OTS.

Tseng, Jesse L. 2000. *The Representation and Selection of Prepositions*. PhD thesis, University of Edinburgh.