

Semantic selection in a cross-linguistic framework

Dan Flickinger

CSLI
Stanford University
Stanford CA 94305-2150

`danf@csli.stanford.edu`

Emily M. Bender

University of Washington
Dept of Linguistics
Box 354340
Seattle WA 98195-4340

`ebender@u.washington.edu`

The Matrix grammar starter-kit (Bender et al. 2002) is described as a language-independent core grammar designed to facilitate the rapid initial development of grammars for natural languages, with foundations solid enough to support steady expansion to broad coverage of the linguistic phenomena in these languages. Such grammars are particularly valuable because they can assign semantic representations to linguistic input, providing the foundation for applications which require natural language understanding. As grammatical coverage expands, one class of linguistic phenomena that emerges involves constraints by selectors on arguments based on their semantic properties. In this paper we explore the descriptive devices currently made available in the Matrix for semantic selection, and study their usefulness within two wide-coverage HPSG grammars, the English Resource Grammar (ERG: Flickinger 2000) and the JACY Japanese grammar (JACY: Siegel and Bender 2002).

1 Composition using Minimal Recursion Semantics

To provide the context for the role and mechanisms of semantic selection within the Matrix, an overview of semantic composition in this framework will be useful, following the summary given by Flickinger and Bender 2003. The Matrix is constructed within the formal system of typed feature structures defined in (Carpenter 1992), using the single operation of unification to build phrases from the words and phrases they contain. Minimal Recursion Semantics was designed to enable semantic composition using only this same unification of typed feature structures, producing for each phrase or sentence a description of the meaning representation sufficient to support logical inference. The type definitions for signs in the Matrix include a semantic component which is an implementation of MRS, and more specifically of the elaboration of a semantic algebra for MRS presented in Copestake et al. 2001. In addition, MRS was designed to answer the competing demands of expressive adequacy and computational tractability, as well as to allow underspecification where it facilitates computational applications, such as machine translation.

The flat semantic representations assigned to each word or phrase in MRS comprise three components:

1. RELS - a bag of atomic predications, each with a label (for scope relations) and one or more roles;
2. HCONS - a set of handle constraints which reflect syntactic limitations on possible scope relations among the atomic predications;
3. HOOK - a group of distinguished externally visible attributes of the atomic predications in RELS, used in combining the semantics of this sign with the semantics of other signs.

Thus objects of type *mrs* (i.e., the value of the features `CONT(ENT)` and `C(ONSTRUCTIONAL)-CONT` of *signs* in the Matrix) are constrained as in (1).

$$(1) \quad mrs: \begin{bmatrix} \text{HOOK} & \textit{hook} \\ \text{RELS} & \textit{diff-list} \\ \text{HCONS} & \textit{diff-list} \end{bmatrix}$$

All relations, representing elementary predications in RELS, bear values for the two features introduced on the type *relation*, as shown in (2):

$$(2) \quad relation: \begin{bmatrix} \text{LBL} & \textit{handle} \\ \text{PRED} & \textit{string} \end{bmatrix}$$

The value of LBL is a *handle*, which is used to express scope relations. The value of PRED can be a string or a sort, used to distinguish particular relations. Earlier versions of the ERG included a separate type for each distinct relation, leading to a very flat (and very large) type hierarchy. We have since found it preferable to distinguish relations for open-class lexemes via a string value for the PRED feature, and only assign a type to this feature for closed-class lexemes, drawing from a small hierarchy of predicate types. For compatibility with Robust MRS (RMRS) ((Copestake 2003)) and software designed to integrate deep and shallow processing, the PRED string values conform to the template `_orth_pos_sense_rel`, enabling automatic construction of lexemes for unknown words, given a part-of-speech tag.

To sustain the composition of the semantics of a phrase from its parts, *mrs* objects introduce the attribute HOOK, which identifies the externally visible semantic properties of a sign. Following Copestake et al. 2001, the value of HOOK is a type which introduces three such properties:

$$(3) \quad hook: \begin{bmatrix} \text{LTOP} & \textit{handle} \\ \text{INDEX} & \textit{individual} \\ \text{XARG} & \textit{individual} \end{bmatrix}$$

The value of LTOP is the local top handle, the handle of the relation(s) with the widest scope within the constituent, modulo quantifiers. The value of INDEX is the distinguished non-handle variable supplied by the sign, identified with the INDEX of the semantic head

daughter, and usually the ARG0 of the main relation introduced by the syntactic head of the constituent. The value of XARG (mnemonic for ‘external argument’) is the index of the single argument in a phrase which can be controlled.

With every word or phrase providing a semantics which consists of HOOK, RELS, and HCONS, the principles of semantic composition in phrase structure rules can be stated (and implemented) quite elegantly, following the definitions in Copestake et al. 2001:

1. The value for RELS on the mother of a phrase is the result of appending the RELS values of all of its daughters.
2. The value for HCONS on the mother of a phrase is the result of appending the HCONS values of all of its daughters.
3. The value for HOOK on the mother of phrase is identified with the HOOK value of its semantic head daughter, where each phrase type uniquely determines which of the daughters is the semantic head.

These principles are encoded as constraints on the relevant phrasal types in the Matrix, ensuring that all phrasal signs preserve and usually extend the semantic content of their parts.

2 Semantic selection

Implicit in the overview given above is the claim that no other properties of the semantics of a sign outside of those in HOOK will be necessary for semantic composition. However, the Matrix does provide one additional device for constraining the selection of a complement or modified phrase based on (limited) semantic properties of that argument. Recall that the value of the attribute PRED in each elementary predication will be a simple string for most open-class lexemes, but for closed-class items this value is usually a type drawn from a relatively small grammar-specific hierarchy. The Matrix enables a grammar writer to make these lexeme-specific semantic types visible for selection by allowing the identification of the value of PRED in the main relation of a given lexical item with the value of a distinguished (syntactic) head feature called KEY. The effect of this reentrancy is to provide for very fine-grained subtyping of part-of-speech types, which we illustrate first with the treatment of preposition selection in the ERG.

Consider the following example sentence in which the verb *rely* lexically subcategorizes for a complement prepositional phrase headed by the specific preposition *on*:

- (4) The professor relied on the students.

To ensure that no other preposition heads the complement PP, the grammar must appropriately constrain the lexical entry for *rely*; in the ERG, consistent with the Matrix, this is expressed by requiring the KEY value of the complement PP to be the specific predicate type introduced by the lexical entry for the preposition *on*. The lexical types for both *rely* and *on* thus include the following constraints:

$$(5) \begin{array}{l} \left[\begin{array}{l} \textit{rely} \\ \text{COMPS} \quad \langle [\text{CAT.HEAD.KEY} \quad \textit{on_p_rel}] \rangle \end{array} \right] \\ \\ \left[\begin{array}{l} \textit{on} \\ \text{CAT.HEAD.KEY} \quad \boxed{1} \textit{on_p_rel} \\ \text{CONT.RELS} \quad \langle [\text{PRED} \quad \boxed{1}] \rangle \end{array} \right] \end{array}$$

Since the `KEY` attribute is a head feature, it will propagate from the lexical preposition head to the prepositional phrase *on the students*, so the type constraint will be visible when the verb combines with the PP. This use of the `KEY` feature for preposition selection is superficially similar to the use of the `PFORM` feature originally used in GPSG analyses, but the same mechanism extends to more interesting cases of semantic selection.

As a second illustration, this use of the `KEY` head feature enables temporal prepositions in English to constrain the semantic types of the noun phrases they can take as complements, as shown in the following examples:

- (6) We met *in/*on/*at* October.
 We met **in/on/*at* Tuesday.
 We met **in/*on/at* five o'clock.

The ERG assigns the appropriate types to the `KEY` attribute in the lexical entries for temporal nouns like *October* and *Tuesday*, and constrains the lexical entries for the prepositions *in*, *on*, *at* so they require the relevant values for their NP complement's `KEY`. Again, the head feature principle will ensure that this `KEY` value will propagate if the NP containing the noun is more complex, as in the example

- (7) We met **in/on/*at* the first Tuesday of the month.

Here the noun *Tuesday* combines with a prenominal adjective and a modifying prepositional phrase, yet remains the syntactic head of the phrase so the `KEY` type introduced by the lexical entry for *Tuesday* will be reentrant with the value of `KEY` on the NP *the first Tuesday of the month*. Using the same value for `KEY` in semantically related lexical entries for days of the month like *nineteenth*, the grammar correctly and efficiently predicts the grammaticality judgments for the following example:

- (8) We met **in/on/*at* the nineteenth of December.

This example with preposition selection for temporal nouns underscores the benefits of this analysis for characterizing the idiosyncratic yet semantically coherent classes which are found in constrained collocations across languages.

3 Conclusion

As we have illustrated with the preceding examples from the ERG, the syntax-semantics interface of a precision broad-coverage grammar must encode not only general principles of semantic composition but also the more fine-grained details of semantic selection. While some selectional constraints are arguably more pragmatic (and therefore defeasible in appropriate fairy-tale contexts), we find that there are some stricter constraints such as the ones illustrated here. Without a means to encode them, we are faced with increased ambiguity, creating both a more difficult parse-selection problem when parsing and the potential for mal-formed output when using the same grammar to generate.

These issues are not, of course, specific to English. We find similar effects in Japanese, e.g., in the domain of numeral classifiers Bond and Paik 2000, which will be discussed in detail in the full paper. Here especially an issue arises with respect to generation because the relevant elements (the numeral classifiers) are not distinguished in their semantic contributions but only in their semantic representation.

We will go on to show how the Matrix can be extended to codify the general means of semantic selection with appropriate features and types of feature values (e.g., strings or sorts as the value of KEY). As grammars based on the Matrix continue to grow, we will be able to verify the validity of the approach cross-linguistically, and search for cross-linguistically relevant groupings of semantic types.

References

- 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 International 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.
- Carpenter, Bob. 1992. *The Logic of Typed Feature Structures*. Cambridge, UK: Cambridge University Press.
- Copestake, Ann. 2003. Report on the design of RMRS: Deep Thought project report D1.1. Unpublished ms.
- Copestake, Ann, Alex Lascarides, and Dan Flickinger. 2001. An algebra for semantic construction in constraint-based grammars. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics*, Toulouse, France.
- Flickinger, Dan. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering* 6(1).

- Flickinger, Dan, and Emily Bender. 2003. Compositional semantics in a multilingual grammar resource. In *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development, ESSLI 2003*, Vienna, Austria.
- 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 International Conference on Computational Linguistics*, Taipei, Taiwan.