Compositional semantics in a multilingual grammar resource

Dan Flickinger and Emily M. Bender
Center for the Study of Language and Information
Stanford University, Stanford, CA 94305 USA
danf@csli.stanford.edu, bender@csli.stanford.edu

1 Introduction

The Matrix grammar starter-kit (Bender et al., 2002) is 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 such, a central component of the Matrix is the collection of resources it contains for simplifying the implementation of semantic composition within each language and supporting the development of a standardized description language for meaning representations, which can provide an effective interface for practical applications. The resources in the Matrix further enable the meaning representations to keep pace as the syntactic analyses of a grammar grow in complexity. In this paper we present the methodology and mechanisms employed for semantic composition in the Matrix grammar starter-kit, using Minimal Recursion Semantics (MRS) with grammars written in the Head-driven Phrase Structure Grammar (HPSG) framework.

2 Background

The goal of the Matrix grammar starter-kit is to provide the necessary definitions of core linguistic types at a level of generality which enables quick specialization to encode the additional basic grammatical constraints for a particular language. With this language-specific tuning, it should be possible to construct a grammar within an afternoon which can be used to parse non-trivial sentences of a given language, then use that same implementation as the basis for the development over time of a semantically precise, broad-coverage grammar. The existing Matrix release also includes software links and parameter settings for one particular grammar development system, the LKB (Copestake, 2002), which includes an efficient parser and generator, but grammars built on the Matrix can be read and used by a number of other parsers (cf. Oepen et al. (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 (MRS: Copestake et al. (1999)) 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). Very briefly, this framework assigns a flat semantic represen-

tation to each word or phrase, consisting of

1. RELS - a bag of atomic predications introduced by lexical entries or by syntactic constructions, each with a "handle" (used to express 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.

3 Implementation

3.1 Architecture of Matrix CONT

The overall architecture of a sign in the Matrix follows closely the definition in Pollard and Sag (1994), with a SYNSEM attribute consisting of (at least) CAT and CONT attributes, for syntactic and semantic constraints, respectively. The value of CONT is a feature structure of type *mrs*, with three attributes HOOK, RELS, and HCONS, encoding the three parts of an MRS representation outlined above. The values of RELS and HCONS are implemented for convenience as lists, since they represent information that is accumulated as one moves up the tree (see below), while the value of HOOK is a feature structure that introduces attributes for each of the four externally visible semantic elements that a sign may present for further composition. These four attributes are as follows:

1. LTOP - the handle of the relation in RELS with highest scope.

2. INDEX - instance or event variable introduced by the lexical semantic head, analogous to lambda variable.

3. E-INDEX - an additional event variable required for some signs like gerunds or predicative phrases, which need to expose both a nominal instance and an event, for control and modification.

4. XARG - the semantic index of the sign's externally visible argument, if any (typically the subject of a verb phrase or other controlled complement).

All of the above is implemented in the type *mrs*, defined below, where *individual* is the supertype to both *event* and *index*:

(1)
$$mrs: \begin{bmatrix} \text{HOOK} & \begin{bmatrix} hook & \\ \text{LTOP} & handle \\ \text{INDEX} & individual \\ \text{E-INDEX} & individual \\ \text{XARG} & individual \end{bmatrix} \\ \text{RELS} & \text{diff-list} \\ \text{HCONS} & \text{diff-list} \end{bmatrix}$$

3.2 Semantics principles for phrases

With every word or phrase providing a semantics which consists of these three parts (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.

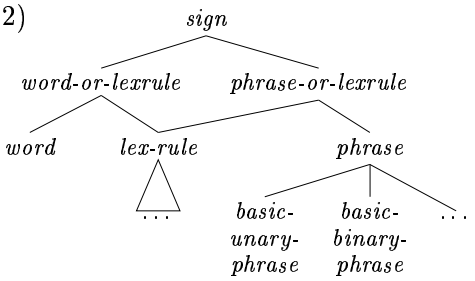
2. The value for HCONS on the mother of a phrase is the result of appending the HCONS values of all of its daughters.

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.

In the Matrix (version 0.4, March 14, 2003), principles 1 and 2 are implemented as constraints on a few high-level types (*lex-rule*, *basic-unary-phrase* and *basic-binary-phrase*) within the *sign* subhierarchy (sketched in (2)), such that they are inherited by all phrases and lexical rules. In

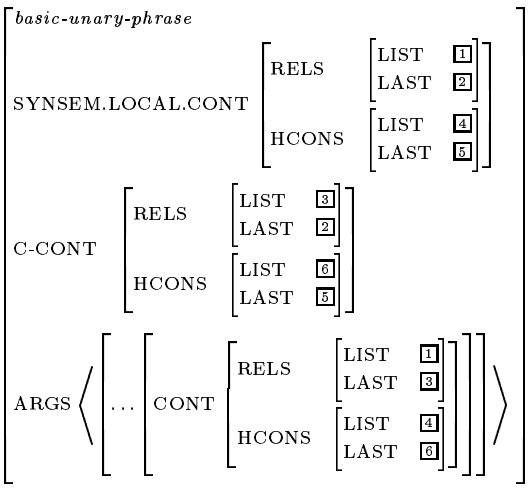
addition, the type *phrase-or-lexrule* identifies the mother's HOOK with the HOOK of the semantics provided by the rule itself (the value of a feature called C-CONT; see §3.4 below). More specialized phrase types identify the HOOK of the C-CONT with the HOOK of the head or non-head daughter.

(2)



Principles 1 and 2 require the accumulation of RELS and HCONS values from daughter to mother in a phrase. The values of these features are implemented as difference lists (typed feature structures which bear values for two attributes LIST and LAST), allowing us to state the accumulation of values using the same single operation of unification of typed feature structures. (3) shows the constraints on the type *basic-unary-phrase*, including the 'diff-list appends' that implement principles 1 and 2.

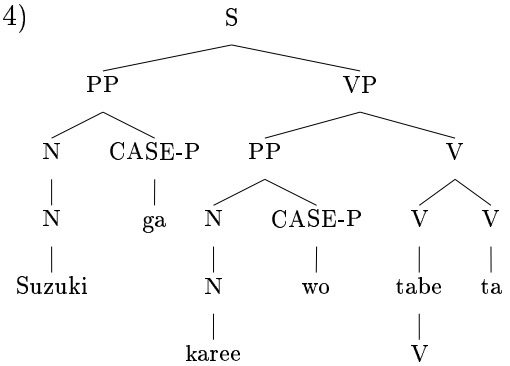
(3)



3.3 Composing heads and arguments

All remaining particulars about the semantics of a phrase are determined by the reentrancies for semantic attributes specified by that phrase, or by the signs which unify in as daughters of the phrase. For example, in the Japanese sentence in (4),[1] the verb *tabe-* ('eat') identifies the HOOK|INDEX of its subject with the ARG1 of the 'taberu' relation, and the HOOK|INDEX of its object with the ARG2.

(4)



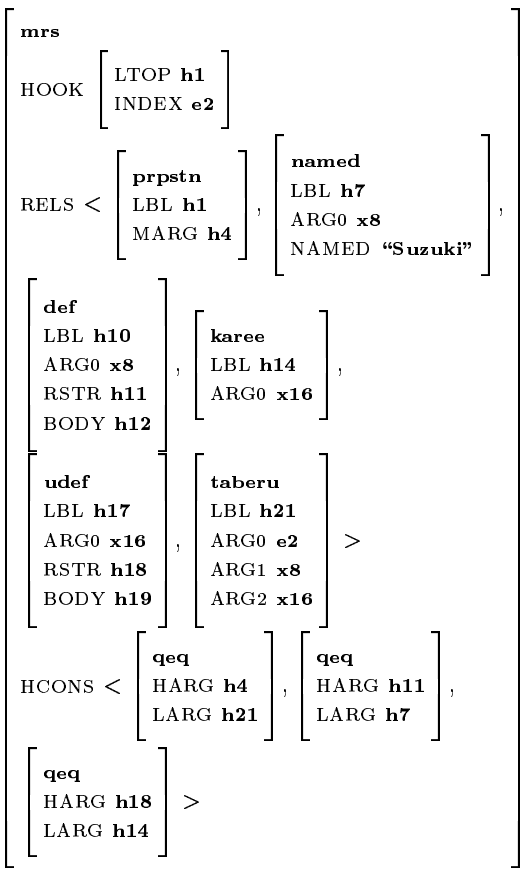
Suzuki-NOM curry-ACC eat-PAST

The rules which build the S and the VP (Japanese-specific specializations of the head-subject and head-complement rules) identify the *synsem* of the PP with the relevant valence requirement (e.g., COMPS) of the verbal head daughter. Since the *synsem* includes the CONT value, a cascade of identities of the familiar kind relates the indices of *Suzuki* and *karee* with the correct roles in the 'taberu' relation. That is, the verb's lexical entry imposes both syntactic and semantic constraints on the synsems of its subject and complement, and the syntactic rules propagate those constraints to the phrases corresponding to the subject and complement, gathering up the semantic relations introduced by the lexical entries and establishing via simple unification the reentrancies that link the semantics of these phrases as intended. Once the verbal

[1] The tree given is a simplification of the tree assigned to this sentence by the JACY grammar (Siegel, 2000; Siegel and Bender, 2002).

head has combined with its complement PP, the semantic variable introduced by that PP's noun as its (externally visible) index is now unified with the appropriate semantic argument position in the relation introduced by the verb. The subject-head rule has an analogous effect on the semantic index of the subject PP, ensuring that it is identified with the other argument position in the verb's relation. The resulting MRS is shown in (5):

(5)



In this representation, the *prpstn* relation (indicating the illocutionary force of the utterance, see Ginzburg and Sag (2000)) bears the top handle of the sentence as its LBL value, and takes as its argument the label of the *taberu* relation, mediated by a *qeq* constraint[2] in the HCONS list of handle constraints to allow for possible intervening quantifiers. Two quantifiers are introduced in this example: a definite quantifier for the proper name and an underspecified one for the bare noun. Each of the two noun relations introduces a variable as its ARG0 value, with each of these variables bound by the appropriate quantifier relation, which also identifies its restrictor value as the label of that noun's relation (again mediated by *qeq* constraints). Since quantifier scope is left underspecified in the grammar, the BODY attributes of the two quantifiers are left with unbound values.[3] The ARG0 variable introduced by *Suzuki* is identified with the ARG1 of the *taberu* relation, and that of *karee* with the ARG2 role. These radically underspecified role names in the *taberu* relation will be interpreted by the grammar-specific semantic interface to indicate that Suzuki does the eating and the curry gets eaten.

3.4 Semantic contributions of constructions

Since some phrase types may introduce semantic content which is not drawn from any of the daughters of the phrase, the MRS framework provides an attribute for phrasal signs called C-CONT (for construction content), which behaves with respect to the semantics principles just as though it were another daughter of the phrase (see, for example, (3) above). C-CONT is implemented in the Matrix as a top-level attribute of phrases and lexical rules, introduced on the type *phrase-or-lexrule*. Like CONT, its value is of type *mrs*.

If a phrase does not introduce any additional semantic content of its own, the values for the attributes RELS and HCONS in C-CONT will be empty lists, so unary and

[2] *Qeq* constraints state that the HARG and LARG are of equal scope, unless a quantifier scopes in between, in which case the HARG outscopes the quantifier which outscopes the LARG.

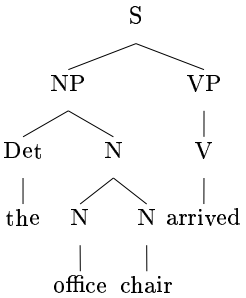
[3] The underspecified MRS in (5) is compatible with two fully resolved MRSs, representing the two possible scopings of the two quantifiers.

binary phrases can safely always append these values to those supplied by the syntactic daughters. Likewise, the HOOK value of a phrase is always identified with its C-CONT's value for HOOK, where for most phrases this HOOK in C-CONT will simply be identified with that of one of the daughters of the phrase, namely the semantic head daughter.

One example of a phrase type in the Matrix that does introduce its own semantic content is the type for (non-relative) clauses (*non-rel-clause*), which introduces a relation encoding the illocutionary force of the clause. Such relations (e.g., *prpstn-rel*, mentioned above) are of type *message*, following the analysis in Ginzburg and Sag (2000).

We illustrate construction-introduced semantic content further with the treatment of noun-noun compounds in the English Resource Grammar (Flickinger, 2000), which uses the same principles of semantic construction as the Matrix toolkit. In the analysis of the sentence *the office chair arrived*, the phrase *office chair* is built using a syntactic rule specifically for noun-noun compounds, and this rule introduces a generic two-place relation *n-n-cmpnd* which relates the variables introduced by the two nouns. The syntactic structure is sketched in (6), where the head-specifier rule is used to combine the determiner and the compound noun, while the head-subject rule combines the full NP with the verb phrase *arrived*. The corresponding MRS semantics is shown in (7):

(6) *The office chair arrived*



(7)

$$\left[\begin{array}{l}\textbf{mrs}\\ \text{HOOK}\ \left[\begin{array}{l}\text{LTOP } \textbf{h1}\\ \text{INDEX } \textbf{e2}\end{array}\right]\\ \text{RELS} < \left[\begin{array}{l}\textbf{prpstn}\\ \text{LBL } \textbf{h1}\\ \text{MARG } \textbf{h4}\end{array}\right], \left[\begin{array}{l}\textbf{def}\\ \text{LBL } \textbf{h10}\\ \text{ARG0 } \textbf{x8}\\ \text{RSTR } \textbf{h11}\\ \text{BODY } \textbf{h12}\end{array}\right],\\ \left[\begin{array}{l}\textbf{chair}\\ \text{LBL } \textbf{h7}\\ \text{ARG0 } \textbf{x8}\end{array}\right], \left[\begin{array}{l}\textbf{office}\\ \text{LBL } \textbf{h14}\\ \text{ARG0 } \textbf{x16}\end{array}\right], \left[\begin{array}{l}\textbf{udef}\\ \text{LBL } \textbf{h17}\\ \text{ARG0 } \textbf{x16}\\ \text{RSTR } \textbf{h18}\\ \text{BODY } \textbf{h19}\end{array}\right],\\ \left[\begin{array}{l}\textbf{n-n-cmpnd}\\ \text{LBL } \textbf{h7}\\ \text{ARG1 } \textbf{x16}\\ \text{ARG2 } \textbf{x8}\end{array}\right], \left[\begin{array}{l}\textbf{arrive}\\ \text{LBL } \textbf{h21}\\ \text{ARG0 } \textbf{e2}\\ \text{ARG1 } \textbf{x8}\end{array}\right] >\\ \text{HCONS} < \left[\begin{array}{l}\textbf{qeq}\\ \text{HARG } \textbf{h4}\\ \text{LARG } \textbf{h21}\end{array}\right],\\ \left[\begin{array}{l}\textbf{qeq}\\ \text{HARG } \textbf{h11}\\ \text{LARG } \textbf{h7}\end{array}\right], \left[\begin{array}{l}\textbf{qeq}\\ \text{HARG } \textbf{h18}\\ \text{LARG } \textbf{h14}\end{array}\right] >\end{array}\right]$$

Note that this MRS representation for the English example is in many respects similar to that of the earlier Japanese example, again introducing two noun relations and supplying a quantifier relation to bind each of these two variables. Here, however, the two noun variables are identified with the ARG1 and ARG2 attributes of the *n-n-cmpnd* relation, and the variable for the head noun *chair* is also the value of the single argument of the *arrive* relation. The *n-n-cmpnd* relation is introduced by the grammar in the RELS attribute of the C-CONT of the grammar rule for noun-noun compounds, which also identifies the assignments of the two nominal instance variables (supplied by its two daughters) to the ARG0 and ARG1 attributes of that *n-n-cmpnd* relation. The relevant constraint on the grammar rule is sketched in (8):

(8)
$$\begin{bmatrix} \text{HEAD-DTR...HOOK} & [\text{INDEX } \boxed{1}] \\ \text{NON-HEAD-DTR...HOOK} & [\text{INDEX } \boxed{2}] \\ \text{C-CONT} & \left[\text{RELS.LIST} \left\langle \begin{bmatrix} \textit{n-n-cmpnd} \\ \text{ARG1} \quad \boxed{1} \\ \text{ARG2} \quad \boxed{2} \end{bmatrix} \right\rangle \right] \end{bmatrix}$$

As discussed above, general principles of semantic composition that are encoded in the Matrix types ensure that rule-specific relations are gathered up along with the relations supplied by the daughters of the rule, and that the appropriate external semantic hooks (the LTOP and INDEX values) are identified on the phrase itself, ready for further composition.

3.5 Lexical rules

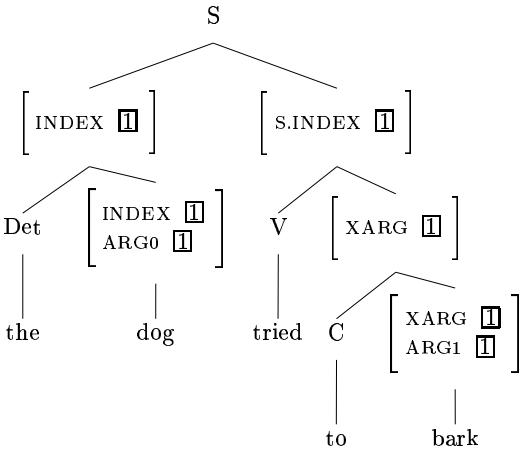
Lexical rules are treated in the Matrix as a particular type of unary rule, in most respects like syntactic unary rules, though lexical rules are prevented from interleaving with syntactic rules. Thus semantic composition for lexical rules is implemented using the same principles as outlined above for syntactic phrases. A lexical rule may or may not introduce semantic content of its own; if it does, that content is found in the C-CONT attribute of the rule and is combined with the semantic content of the (single) daughter (the 'input' to the lexical rule) by those same principles.

Note that this approach imposes a strong constraint on the directionality of lexical rules in the Matrix, since the principles of composition guarantee monotonic accumulation of atomic predications, so no semantic content from a daughter in a phrase or lexical rule is ever lost. For example, a lexical rule relating the causative/inchoative alternation in English for verbs like *break* will have to treat the inchoative lexical entry as the 'input' to the lexical rule (that is, the daughter), and the semantically richer causative lexical entry as the 'output' (the SYNSEM value of the lexical rule).

3.6 External arguments

In the examples above, we made reference to two of the *hook* attributes, LTOP and INDEX, both of which play a crucial role in the semantic construction of every phrase. A third attribute, XARG, is relevant for control phenomena such as equi and raising, since it identifies the semantic index of a phrase's external argument (usually the subject of a verb phrase). Identifying this property of a phrase as part of the hook allows our general principles of semantic composition to make this attribute visible for control of subject-unsaturated complements (VPs, predicative PPs, etc.) and also for agreement even at the sentence level, as for example in tag questions in English (Bender and Flickinger, 1999). An example using this XARG attribute in composition is given in (9), with the lexical type for subject-equi verbs given in (10).

(9) The dog tried to bark

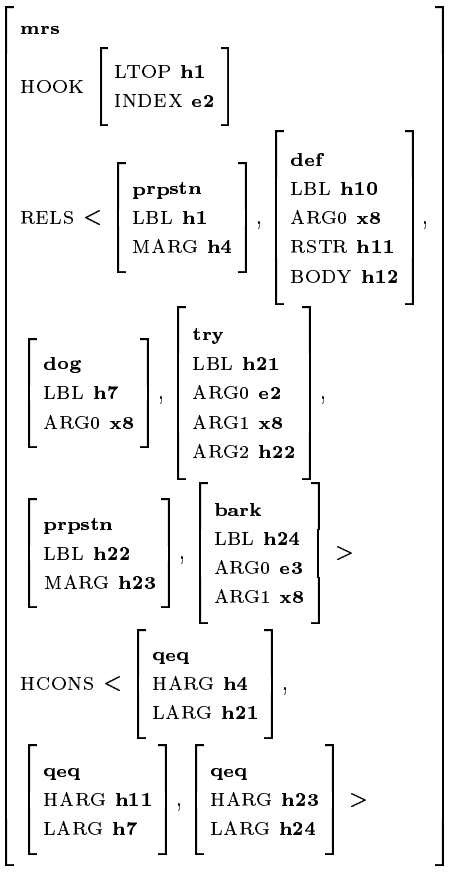


(10) Type for subject-equi verbs like *try*

$$\begin{bmatrix} \textit{subj-equi-verb} \\ \text{SUBJ} < \left[\text{HOOK.INDEX } \boxed{1} \right] > \\ \text{COMPS} < \left[\text{HOOK.XARG } \boxed{1} \right] > \\ \text{CONT.RELS} < \left[\text{ARG1 } \boxed{1} \right] > \end{bmatrix}$$

Here the lexical entry for the verb *try* identifies its VP complement's semantic external argument (XARG value) with its subject's semantic index (INDEX value), and further identifies that index with the appropriate role (the ARG1) in the lexical relation introduced by the verb. The MRS semantics constructed for this example is given in (11).

(11)



The construction of this representation is the result of the same general principles of semantic composition presented above. The head-complement rule unifies the verb *try*'s constraints on its complement with those of the VP phrase *to bark*, which results in the identification of the XARG value of that VP with the INDEX of the subject of *try*. The constraints on *try*'s subject are propagated up to the verb phrase *tried to bark* from the head-daughter *tried* by the head-complement rule, and the semantics of this verb phrase preserve the semantic properties of its daughters, including the desired re-entrancies with the subject index. Hence when the head-subject rule combines *the dog* with *tried to bark*, the syntactic and semantic constraints of the noun phrase are unified with those in the SUBJ attribute of the verb phrase, resulting in the identification of the ARG0 value introduced by it dog with the ARG1 values in both the *try* relation and the *bark* relation.

3.7 Future work: lexical semantics

The primary focus of the current Matrix machinery for the syntax-semantics interface is on the definitions of the semantic principles for composition in the construction of syntactic phrases. In its current early form, the Matrix provides only minimal support for defining the semantics of lexical entries, though work is underway to add a partial hierarchy of lexical types to express generalizations about some standard subcategorization phenomena, including the linking of syntactic arguments to semantic roles in atomic predications. This will build on the set of primary types of lexical relations (*arg1-rel*, *arg2-rel*, *arg12-rel*, ...) presently provided. These types introduce a perhaps surprisingly sparse set of role names that are nonetheless meant to be (nearly) exhaustive.

The task of linking syntactic arguments to these semantic roles is still left as an exercise for the developers of each individual grammar. Developers are encouraged to avoid the introduction of any additional role names, and to avoid the use of multiple relations for individual lexical entries for open-class lexeme types, so that grammars being developed using the current Matrix will be compatible with the next release aimed at carrying more of the burden of lexical type definitions.

4 Conclusion

In this paper, we have illustrated how the implementation of the syntax-semantics interface in the Matrix provides support for the rapid development of broad-coverage precision grammars. We have identified the principles of semantic composition adopted in the Matrix using the framework of Minimal Recursion Semantics, and illustrated their interaction with the semantic properties introduced by lexical entries and syntactic constructions. Initial experiments applying the Matrix to grammars of Norwegian, Italian and Greek, some of which are reported on in this workshop, have validated the benefits of the machinery already provided for semantic composition in speeding up the development of semantically rich grammars. The experience of these grammar writers also underscores the significant potential benefit of planned extensions to the Matrix including a lexical type hierarchy, which will help to sustain the standardization across implementations and languages which is one of the chief goals of the Matrix.

5 Acknowledgements

We have benefited greatly in this work from conversations with Ann Copestake, Alex Lascarides, Stephan Oepen, and Ivan Sag on general issues of semantic composition; from insights provided by Melanie Siegel and Berthold Crysmann who are developing broad-coverage grammars for Japanese and German, respectively; and from our interactions with the resourceful and cooperative developers of Matrix-based grammars for Norwegian and Italian.

References

Emily Bender and Dan Flickinger. 1999. Peripheral constructions and core phenomena. In Gert Webelhuth, Andreas Kathol, and Jean-Pierre Koenig, editors, *Lexical and Constructional Aspects of Linguistic Explanation*. CSLI Publications, Stanford.

Emily Bender, 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*, pages 8–14, Taipei, Taiwan.

Bob Carpenter. 1992. *The Logic of Typed Feature Structures*. Cambridge University Press, Cambridge, UK.

Ann Copestake, Daniel P. Flickinger, Ivan A. Sag, and Carl Pollard. 1999. Minimal Recursion Semantics. An introduction.

Ann Copestake, 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.

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

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6 (1) (Special Issue on Efficient Processing with HPSG):15 – 28.

Jonathan Ginzburg and Ivan A. Sag. 2000. *Interrogative Investigations: The form, meaning and use of English interrogatives*. CSLI Publications, Stanford, CA.

Stephan Oepen, Daniel Flickinger, J. Tsujii, and Hans Uszkoreit, editors. 2003. *Collaborative Language Engineering. A Case Study in Efficient Grammar-based Processing*. CSLI Publications, Stanford, CA.

Carl Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. Chicago University Press, Chicago.

Melanie Siegel 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.

Melanie Siegel. 2000. HPSG analysis of Japanese. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin.