Semantic Representations of Syntactically Marked Discourse Status in Crosslinguistic Perspective

Emily M. Bender
David Goss-Grubbs
University of Washington (USA)
email: ebender@u.washington.edu

Abstract

This paper presents suggested semantic representations for different types of referring expressions in the format of Minimal Recursion Semantics and sketches syntactic analyses which can create them compositionally. We explore cross-linguistic harmonization of these representations, to promote interoperability and reusability of linguistic analyses. We follow Borthen and Haugereid (2005) in positing COG-ST ('cognitive status') as a feature on the syntax-semantics interface to handle phenomena associated with definiteness. Our proposal helps to unify the treatments of definiteness markers, demonstratives, overt pronouns and null anaphora across languages. In languages with articles, they contribute an existential quantifier and the appropriate value for COG-ST. In other languages, the COG-ST value is determined by an affix. The contribution of demonstrative determiners is decomposed into a COG-ST value, a quantifier, and proximity information, each of which can be contributed by a different kind of grammatical construction in a given language. Along with COG-ST, we posit a feature that distinguishes between pronouns (and null anaphora) that are sensitive to the identity of the referent of their antecedent and those that are sensitive to its type.

1 Introduction

In this paper, we discuss the compositional construction of semantic representations reflecting discourse status across a range of phenomena. Borthen and Haugereid (2005) propose COG-ST ('cognitive-status')[1] as a feature on the syntax-semantics interface to handle phenomena associated with definiteness. We explore how their approach leads to cross-linguistically unified treatments of demonstratives, overt pronouns and null anaphora as well. We find that cross-linguistic studies motivate different representations than we might have arrived at from just one language.

Our work grows out of the Grammar Matrix, a multilingual grammar engineering project (Bender et al., 2002; Bender and Flickinger, 2005) which strives to harmonize semantic representations across diverse languages. The Grammar Matrix is couched within the Head-driven Phrase Structure Grammar (HPSG) framework (Pollard and Sag, 1994). We use Minimal Recursion Semantics (Copestake et al., 2001, 2005) as our semantic representation system.

2 Background

2.1 Minimal Recursion Semantics

Grammar Matrix-derived grammars associate surface strings with MRS representations (or MRSs), in a bidirectional mapping that allows both parsing and generation. An MRS consists of a multiset of elementary predications (eps), each of which is a single relation with its associated arguments, labeled by a handle; a set of handle constraints relating the labels of eps to argument positions within other eps; and a top handle indicating which of the labels has outermost scope (Copestake et al., 2001, 2005). The MRSs produced by these grammars are underspecified for scope, allowing multiple different fully-scoped variants, according to the handle constraints.

Each ep has a predicate (PRED) value and one or more argument positions, usually labeled ARG0 through ARG*n*. By convention, we refer to elementary predications by their PRED values. For scope-taking eps (including quantifiers as well as clause-embedding predicates such as **_believe_v_rel** and scopal modifiers such as negation), at least one argument position is handle-valued, and related (in a well-formed structure) to the label of another ep. For non-scopal predications, the values of the argument positions are variables (also called indices) which may themselves be associated with 'variable properties', such as person, number and gender on individual variables, or tense, aspect, sentential force and mood on event variables.

One benefit of MRS is that it is designed to be compatible with feature-structure grammars. We build up MRSs through an HPSG implementation of the MRS algebra in Copestake et al. (2001), in which each constituent bears features recording the eps and handle constraints contributed within the constituent, as well as a set of properties exposed through the feature HOOK to facilitate further composition. These properties include pointers to the local top handle (LTOP), the constituent's primary index (INDEX), and the external argument, if any (XARG).

Eps are canonically contributed by lexical entries, with one ep per lexical entry. Lexical entries can, however, contribute more than one ep or no eps at all. In addition, syntactic constructions can also contribute eps of their own.

[1] Original feature name: COGN-ST.

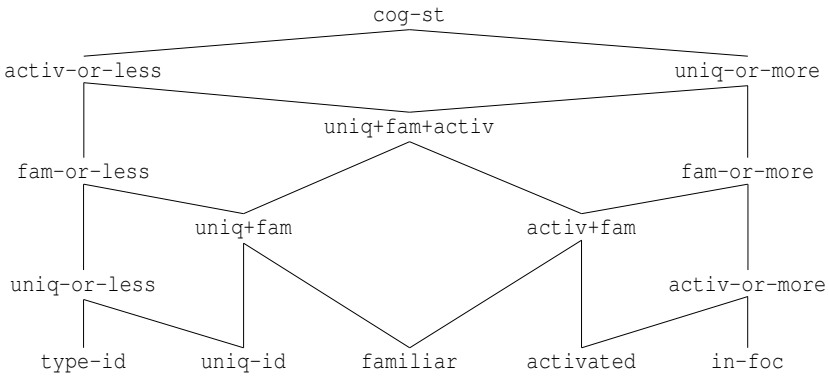

Figure 1: Cognitive status hierarchy

2.2 Harmonization of Representations

The semantic representations used in the Grammar Matrix were originally derived from those used in the English Resource Grammar (Flickinger, 2000), a wide-coverage grammar of English. In this paper, we propose to refine the semantic representations for phenomena connected to discourse status in light of the constraints on the syntax-semantics interface we find in a range of languages. This is not to say that we are promoting working towards an interlingua: indeed, even if it were possible to define a suitable interlingual set of representations, we believe it wouldn't be possible to map from surface strings to such representations in one compositional step.

Nonetheless, it is useful to harmonize representations across languages while still allowing for necessary differences, for at least two reasons. First, when semantic representations are as similar as they practically can be, this simplifies both the transfer component in transfer-based machine translation systems (e.g., Oepen et al., 2007) and the design of downstream components that make use of semantic representations in multilingual NLP systems in general. Second, harmonized semantic representations facilitate the creation of libraries in a resource like the Grammar Matrix, which in turn promotes both the reuse of analyses within implemented grammars and the exploration of computational linguistic typology.

2.3 Discourse/Cognitive Status

This paper builds on a tradition of work investigating the way the discourse status of referents influences the form of the referring expressions used to refer to them, or alternatively, the way that speakers use contrasts in form to signal to their interlocutors the discourse (or cognitive) status of their intended referents (Chafe, 1976, 1994; Prince, 1981; Gundel et al., 1993; Borthen and Haugereid, 2005; Arnold, 2008).

Borthen and Haugereid (2005) (henceforth B&H) present arguments from a range of languages that the discourse status associated with referring expressions can be constrained by multiple intersecting syntactic factors. They use this to motivate embedding the discourse status information within the semantic features of a sign, rather

than on the contextual features. They adapt the implicational scale proposed by Gundel et al. (1993) and Prince (1981), representing discourse referents as having a range of values from 'type identifiable' through 'in focus'. In Gundel et al. and Prince's work, this is an implicational scale, where a discourse status of 'in focus', for example, also entails a discourse status of 'activated'. B&H argue that it needs to be represented within the syntax by a type hierarchy that makes each discourse status type incompatible with the others, while also creating supertypes that represent ranges of discourse status values. Their intuition is that the syntactic constraints restrict the distribution of certain forms based on the highest discourse status they are compatible with, rather than on the actual discourse status of the referent they are used to evoke in a given context. The cognitive status hierarchy, as we adopt it from Borthen and Haugereid (2005) is shown in Fig 1.

3 Markers of Definiteness

The first phenomenon we consider is markers of definiteness. In English, these are syntactically identified with determiners, and thus the English Resource Grammar represents the semantic contrast between *the* and *a* with the PRED value of the ep contributed by the determiner: **_the_q_rel** vs. **_a_q_rel** (where 'q' stands for 'quantifier'). Crosslinguistically, however, definiteness is not always marked in lexical determiners which might plausibly contribute quantifier relations. For example, in Norwegian, definiteness is signaled in part by an affix on the noun:

(1) Jeg så bilen.
I saw car.DEF
'I saw the car.' [nob]

This does not lend itself to the analysis of definiteness in English provided by the ERG: First, the definite suffix can co-occur with something else in the determiner role, as in (2).[2] Second, even if the affix did contribute a **_def_q_rel**, this would lead to ill-formed MRSs as soon as there were any intersective modifiers: Eps introduced by intersective modifiers (such as *nye* in (2)) should be labeled with the same handle as the ep introduced by the noun. But according to the MRS model of semantic compositionality, the label of the noun's relation is not available for further composition once the quantifier has attached.

(2) Jeg så den nye bilen
I saw the new.DEF car.DEF
'I saw the new car.' [nob]

Third, adjectives can also take definite forms. We would like to enforce the compatibility of this information, rather than having each instance of the definite suffix contribute an additional ep. Per B&H, this supports treating definiteness in terms of a feature rather than through eps.

[2]Note that the determiner is required when there is an adjective in a definite NP, and pragmatically very restricted when there is not.

Following B&H, we note that the apparently binary distinction between definites and indefinites is better assimilated to the cognitive status hierarchy. There are morphosyntactic phenomena in various languages which divide the cognitive status hierarchy into two separate ranges, though the division point may vary across languages and within languages across phenomena. Using a single feature for cognitive status that takes its values from the type hierarchy in Fig 1 allows these various distinctions to be modeled elegantly.

B&H propose wrapping semantic indices in a new structure *ref-prop*, which contains COG-ST as well as other features related to the referential properties of a noun phrase. In this paper, we focus on COG-ST and leave the other dimensions to future work. However, we differ from B&H in proposing that COG-ST, at least, should be a feature of semantic indices, rather than inside a parallel structure (i.e., their *ref-prop*). This has the benefit of causing the COG-ST information from particular words or affixes to be included in the compositionally created semantic representations of phrases and sentences without any further effort: wherever the index so marked appears, it will carry its COG-ST value with it. It also makes the (correct, we believe) prediction that whenever an index appears in multiple places in the semantic representation, it should bear the same cognitive status information in all of them. For example, the MRS for (3) is as in (4), where the variable 'x5' represents the cat, and appears as a value in four separate elementary predications: ARG0 of **_cat_n_rel**, ARG0 of **_exist_q_rel**, ARG1 of **_want_v_rel**, and ARG1 of **_go+out_v_rel**. We claim that in all of these guises, the cognitive status of the referent is the same; there is only one mental representation of the referent involved.

(3) The cat wanted to go out.

(4)
$$\begin{bmatrix} \text{LTOP} & \text{h0} \\ \text{INDEX} & \text{e1} \\ \text{RELS} & \left\langle \begin{bmatrix} \textbf{\_exist\_q\_rel} \\ \text{LBL} & \text{h3} \\ \text{ARG0} & \text{x5}[\textit{uniq-id}] \\ \text{RSTR} & \text{h6} \\ \text{BODY} & \text{h4} \end{bmatrix}, \begin{bmatrix} \textbf{\_cat\_n\_rel} \\ \text{LBL} & \text{h7} \\ \text{ARG0} & \text{x5} \end{bmatrix}, \begin{bmatrix} \textbf{\_want\_v\_rel} \\ \text{LBL} & \text{h8} \\ \text{ARG0} & \text{e2} \\ \text{ARG1} & \text{x5} \\ \text{ARG2} & \text{h9} \end{bmatrix}, \begin{bmatrix} \textbf{\_go+out\_v\_rel} \\ \text{LBL} & \text{h10} \\ \text{ARG0} & \text{e11} \\ \text{ARG1} & \text{x5} \end{bmatrix} \right\rangle \\ \text{HCONS} & \left\langle \text{h6} =_q \text{h7}, \text{h9} =_q \text{h10} \right\rangle \end{bmatrix}$$

B&H consider this possibility and dismiss it on the grounds that coreferential noun phrases don't necessarily share the same cognitive status. However, placing the COG-ST value on the index does not necessarily entail that the expressions *The cat*, *herself*, and *her* impute the same cognitive status to their discourse referent in (5). As far as the syntactic processing is concerned, these expressions introduce distinct indices. It is up to a separate reference resolution component to identify them, and that component could merge their COG-ST values or not, as appropriate.

(5) The cat opened the door herself with her paw.

Thus rather than having English *the* and similar elements introduce a specialized quantifier relation, we instead do a small amount of semantic decomposition: *the* introduces just an existential quantifier (**_exist_q_rel**), but constrains the variable it

```
a. def-noun-lex-rule := inflecting-lexeme-to-word-rule &
   %prefix (ha- *)
     [ SYNSEM.LOCAL.CONT.HOOK.INDEX.COG-ST uniq-id,
       DTR noun-lex ].
b. def-adj-lex-rule := inflecting-lexeme-to-word-rule &
   %prefix (ha- *)
     [ SYNSEM.LOCAL.CAT.HEAD.MOD.LOCAL.CONT.HOOK.INDEX.COG-ST
       uniq-or-more, DTR adj-lex ].
c. indef-noun-lex-rule := constant-lexeme-to-word-rule &
     [ SYNSEM.LOCAL.CAT.HEAD.MOD.LOCAL.CONT.HOOK.INDEX.COG-ST
       type-id, DTR adj-lex ].
```

Figure 2: Sample lexical rules for definiteness affixes

binds to be [COG-ST *uniq-id*]. This signals to the hearer that s/he should be able to assign a unique representation to the referent (but not that the referent itself is unique in the world or in the previous discourse, cf. Gundel et al., 2001).

In other languages affixes can also constrain COG-ST to *uniq-or-more* or *uniq-id*. We illustrate here with the Hebrew definite prefix *ha-*, shown in (6) (from Wintner, 2000:322).

(6) koll šešš ha-smalot ha-yapot ha-'elle šelli mi-'rhb
all six DEF-dresses DEF-nice DEF-these mine from-US
'all these six nice dresses of mine from the US' [heb]

ha- is added by a lexical rule (sketched in Fig 2a) which adds information about the COG-ST to the noun's own INDEX value.[3] When *ha-* attaches to an adjective in Hebrew, it instead adds the information that the noun the adjective is modifying must have the COG-ST value *uniq-or-more*, as sketched in Fig 2b. This rule is paired with a non-inflecting lexical rule Fig 2c which produces adjectives which can only modify nouns that are [COG-ST *type-id*], i.e., indefinite. This will enforce definiteness agreement across the noun phrase.[4]

This section has briefly outlined an adaptation of B&H's proposal for definiteness marking. The main difference to their proposal is in the location of COG-ST in the feature geometry. In the following two sections, we extend the approach to demonstratives and a variety of null anaphora.

4 Demonstratives

Demonstratives can stand alone as noun phrases (demonstrative pronouns) or function as nominal dependents. Starting again with English, we find that demonstratives in their nominal-dependent guise, like the markers of definiteness, fill the spec-

[3]The lexical rules in Fig 2 are non-branching productions that apply at the bottom of the parse tree, before any syntactic rules can apply. The SYNSEM value represents the mother and the DTR value the daughter. The types they inherit from (e.g., *inflecting-lexeme-to-word-rule*) enforce identity of most of the information between mother and daughter. The rules add information about COG-ST, which must be compatible with what's provided by the lexical entries for the rules to apply.

[4]For the rule for unmarked nouns, see §4 below.

ifier slot of the noun phrase and function as determiners. Accordingly, the ERG represents their semantic contribution through the PRED value of the quantifier relation: **_this_q_dem_rel** and **_that_q_dem_rel**. Crosslinguistically, however, demonstratives functioning as nominal dependents can also appear as adjectives or affixes (Dryer, 2008). In such languages, within the general constraints of composition of MRS, it is not elegant or natural-seeming to have an adjective contribute a quantifier relation or constrain the PRED value of a relation contributed by a separate determiner or non-branching NP construction.

Instead, it seems more appropriate to decompose the semantic representation of determiners into a quantifier relation (**_exist_q_rel**) and a separate one-place modifier relation (e.g., **_distal+dem_a_rel**, for 'that'). In languages with demonstrative adjectives, the demonstrative form contributes only the modifier relation. In languages with demonstrative determiners, the demonstrative forms contribute both.

Demonstratives also constrain the COG-ST value of the nouns the modify, typically to *activ-or-fam*. In some languages, (e.g., Irish Gaelic), the demonstratives require additional marking of discourse status. Typically this takes the form of a definite article (see (7) from McCloskey (2004)), but demonstratives can also attach to pronouns and proper nouns (McCloskey, 2004).

(7) an fear mór téagartha groí seo
the man big stocky cheerful DEM
'this big stocky cheerful man' [gle]

(8) *fear mór téagartha groí seo

Such languages are straightforwardly countenanced within this system: the definite article and article-less NPs have incompatible COG-ST values, and only the former is compatible with the COG-ST constraints contributed by the demonstrative adjective.[5]

The situation in Hebrew is slightly more complex: Demonstratives can occur with or without the *ha-* prefix, so long as they agree with the noun they modify. Conversely, nouns without the *ha-* prefix are interpreted as indefinite, unless they are modified by a demonstrative adjective. It is unclear at this point whether there is a difference in interpretation between (9) and (10) (from Wintner, 2000:334), but it seems likely that *type-id* is not the correct cognitive status for (9); that is, it is most likely not an indefinite.

(9) sepr ze nimkar heiteb
book this is.sold well
'This book sells well.' [heb]

(10) ha-sepr ha-ze nimkar heiteb
DEF-book DEF-this is.sold well
'This book sells well.' [heb]

[5] McCloskey points out that the demonstratives can attach to coordinated NPs, each with their own article. This raises difficulties for treating the demonstratives as adjectives, as it would require the demonstrative adjectives to attach outside the determiner (cf. Bender et al., 2005). We leave this issue to future work.

Here, we postpone the assignment of a COG-ST value to an unmarked noun until the NP level, filling in *type-id* in case no demonstrative has attached. This requires an additional syntactic feature to control the application of the NPs rules, but this seems motivated: As Wintner notes, *ha-* is functioning as an agreement marker; its distribution has become grammaticized and drifted somewhat from what purely semantic constraints would predict.

To provide complete representations for demonstratives, we also need to address the additional information they carry in many languages, such as the relative proximity of the referent to the speaker and/or the hearer, its visibility or elevation (Diessel, 1999). These distinctions appear to be at least partially independent of the COG-ST dimension. In addition, in the absence of any evidence for syntactically-mediated agreement between elements of a sentence along this dimension, for now we represent this part of the meaning of demonstratives as an elementary predication rather than as a feature.

Some languages (e.g., Lithuanian) have a demonstrative element which does not express any distance contrast, in addition to ones that do (Diessel, 2008). In this case, it might make sense to reduce the contribution of the former sort of element to the constraints it places on the noun's COG-ST value. However, in the interests of uniformity within the system, we continue to assign it an elementary predication.

Other languages (e.g., French and German) don't mark any distance contrast on the primary demonstrative element. In all such languages, there are optional, deictic adverbials which can be added to mark the contrast (Diessel, 2008).

(11) Das Bild hier gefällt mir besser als das da.
DEM picture here like me better than DEM there.
'I like this picture better than that one (over there).' [deu]

In light of such data, we could decompose demonstratives with distance contrasts in all languages into separate demonstrative and deictic/distance relations. Alternatively, we could do that decomposition only in languages like German and French. To the extent that the deictic elements (e.g., German *hier* and *da*) have other uses as ordinary adverbs which can be syntactically assimilated to the same lexical entry, we would want to at least make sure that the ep they contribute is the same in both cases.

5 Overt pronouns and zero anaphora

Pronouns in the ERG are currently represented by an index which is bound by the quantifier **_pronoun_q_rel** and modified by **_pronoun_n_rel**. The quantifier ep marks the pronoun as definite, and the modifier ep serves as the restriction for the quantifier as well as identifying the index as a pronoun.

Following the treatment of other nominals presented here, however, we do away with the quantifier ep in favor of the COG-ST feature. Similarly, we replace the modifier ep with a feature PRON-TYPE, which indicates whether an index is to be interpreted as pronominal, and if so, the type of the pronoun (as discussed below). Not only is this representation simpler, there is no prediction that pronouns participate in quantifier scope relations, as there is when using **_pronoun_q_rel**.

Overt pronouns, clitics and zero pronominals are generally assumed to take a COG-ST value of *in-focus* (Gundel et al., 1993; Borthen and Haugereid, 2005). In general, we agree. We assume that most overt pronouns and many forms of zero anaphora do take that value. However, there are forms which require us to make exceptions to this.

First let us consider the English indefinite pronoun *one*, as in (12). Clearly in this case the referent of *one* is not in focus. Rather, such a pronoun should bear the COG-ST value *type-id*.

(12) Kim bought a computer and Sandy borrowed one.

B&H make a distinction between what they call *token pronouns* and *type pronouns*, where the former are the standard pronouns, which corefer with their antecedents, and the latter are like English *one*, which refer to a new token whose type is taken from its antecedent. We propose that the PRON-TYPE feature take a value of type *pron-type*, with subtypes *not-pron* for non-pronouns and *type-or-token* for pronouns. The latter will have two further subtypes, *token-pron* and *type-pron*. English *one* will be lexically specified as [PRON-TYPE *type-pron*].

Certain cases of zero anaphora similarly get their type information from their antecedents. A couple of instances of the Italian null subject construction appear in (13) and (14).

(13) John ha fatto la torta. La-ha mangiata
John has make.PPRT the cake. it-has eat.PPRT
'John baked the cake. (He) ate it.' [ita]

(14) Se uno bambino vuole un biscotto, gli-arriva
if a child wants a cookie to.him-arrives
'If a child wants a cookie, he gets one.' [ita]

In (13), the referent of the null subject is indeed an entity which is in focus, namely John. On the other hand, in (14) the referent of the null subject is a new token of a type which is in focus, namely the type 'cookie'.

To handle this situation, we propose that Italian null subjects are associated with COG-ST *in-focus*, and with PRON-TYPE *type-or-token*. The grammar for Italian contains a 'subject drop' construction which discharges the subject requirement of the verb without realizing any overt dependent. Because the verb will have linked the appropriate argument position of its own ep to the HOOK.INDEX value inside the feature recording its subject requirement, the subject drop construction can constrain the properties of this index. In particular, it will specify that its PRON-TYPE is *type-or-token* (i.e., it is a pronominal), and that its COG-ST is *in-focus*. The subject-drop construction is sketched in Fig 3. When further processing determines the nature of the antecedent, the PRON-TYPE value will get further specified. If it is a non-specific indefinite, e.g. it is an indefinite in an intensional context, the pronominal will be specified *type-pron*, otherwise it will be specified *token-pron*.

The next type of zero pronominal we consider are Japanese dropped arguments, which present a counterexample to Gundel et al. (1993)'s claim that all zero pronominals are COG-ST *in-focus*. To be sure, Japanese zero anaphora can be understood

```
head-opt-subj-phrase := head-valence-phrase & head-only &
[ SYNSEM.LOCAL.CAT.VAL.SUBJ < >,
  HEAD-DTR.SYNSEM.LOCAL.CAT [ HEAD verb & [ FORM fin ],
                              VAL.SUBJ < [ LOCAL.CONT.HOOK.INDEX
                                             [ COG-ST in-focus,
                                               PRON-TYPE type-or-token ]] > ]].
```

Figure 3: Subject drop construction for Italian

similarly to overt token pronouns, as in (15). However, there are also examples where it can be understood like an overt type pronoun, like English *one*, as in (16). Note that (16) is different from (14) in that the antecedent of the null anaphor is not in an intensional context.

(15) Mi-ta.
see.PAST
'(He/she) saw (it).' [jpn]

(16) Zyon-wa konpyuutaa-o kat-ta. Mearii-wa kari-ta.
John.TOP computer.ACC buy.PAST Mary.TOP borrow.PAST
'John bought a computer. Mary borrowed one.' [jpn]

We propose that Japanese dropped arguments are underspecified with respect to cognitive status and pronoun type. They are associated with indices specified as COG-ST *cog-st* and PRON-TYPE *type-or-token*.

Finally, we turn to lexically licensed null instantiation in English, beginning with definite null instantiation. Fillmore et al. (2003) define definite null instantiation as a phenomenon whereby some conceptually necessary participant in a situation is left unexpressed, but its identity is derivable from context. In lexically licensed null instantiation, the possibility of argument drop and the interpretation of the dropped argument are dependent on the selecting head. In English, lexically licensed DNI is typically a kind of token pronominal, as in (17). But some items can also license type-pronominal DNI, as in (18). In (17), the thing that was won is the previously mentioned game. In (18), there is no particular job that is being sought, although we do know from the context that it is a job.

(17) Kim played a game with Sandy, and Sandy won.

(18) I can't find a job, but I'm still looking.

We model lexical licensing of null instantiation through a feature called OPT which allows selecting heads to record whether or not their arguments are 'optional'. Since the interpretation of dropped arguments is also constrained by the lexical heads, we propose two additional features OPT-CS and OPT-PT which encode the cognitive status and pronoun type to assign to that argument in case it is dropped. The complement-drop construction and the lexical constraints on *look* are sketched in Fig 4a-b.

In this figure, strings prefixed with # indicate reentrancy in the feature structure. The feature KEYREL in lexical entries is a pointer to the main ep they contribute. The

```
a.  head-opt-comp-phrase := head-valence-phrase & head-only &
    [ SYNSEM.LOCAL.CAT.VAL.COMPS #comps, HEAD-DTR.SYNSEM.LOCAL.CAT
      [ VAL.COMPS [ FIRST [ OPT +,
                            OPT-CS #cog-st,
                            OPT-PT #pron-type,
                            LOCAL.CONT.HOOK.INDEX [ COG-ST #cog-st,
                                                    PRON-TYPE #pron-type ]],
                    REST #comps ]]].
b.  look := pp-transitive-verb-lex &
    [ STEM < "look" >,
      SYNSEM [ LOCAL.CAT.VAL.COMPS < [ OPT-CS in-focus,
                                       OPT-PT type-or-token ] >,
               LKEYS.KEYREL.PRED "_look_v_rel" ]].
c.  read := transitive-verb-lex &
    [ STEM < "read" >,
      SYNSEM [ LOCAL.CAT.VAL.COMPS < [ OPT-CS type-id,
                                       OPT-PT non-pron ] >,
               LKEYS.KEYREL.PRED "_read_v_rel" ]].
d.  devour := transitive-verb-lex &
    [ STEM < "devour" >,
      SYNSEM [ LOCAL.CAT.VAL.COMPS < [ OPT - ] >,
               LKEYS.KEYREL.PRED "_devour_v_rel" ]].
```

Figure 4: Lexically licensed complement drop for English

type *transitive-verb-lex* inherits from its supertypes the linking constraints which identify the HOOK.INDEX values of the syntactic arguments with the appropriate ARG*n* values in the ep contributed by the verb.[6]

Indefinite null instantiation is similar, except that the identity of the missing element is either unknown or immaterial. An example of this is (19). INI differs from other null nominals in that it is not a kind of anaphor. There is nothing in the context that helps to identify its referent.

(19) Kim is reading.

We propose that indices in INI constructions are specified as COG-ST *type-id* and PRON-TYPE *non-pron*. In English, these constructions are also lexically licensed, and can be handled with the same features described for DNI. The lexical constraints on *read* are illustrated in Fig 4c. For completeness, we also include in Fig 4d an example of a lexical item which does not license missing complements.

6 Summary and Future Work

In this paper we have explored the construction of semantic representations for a variety of forms of referring expressions. Building on Borthen and Haugereid (2005)'s proposal to treat cognitive status as a semantic feature within HPSG, we have developed representations for definite, demonstrative and null NPs, and sketched means of arriving at them compositionally.

[6]The constraints shown on the COMPS value of lexical entries would actually be implemented as constraints on types that the lexical entries inherit from, allowing the grammar to capture generalizations across lexical entries. They are shown as constraints on the lexical entries here for ease of exposition only.

In future work, we plan to expand the range of these analyses to cover phenomena such as Irish demonstratives taking scope over coordinated noun phrases and cross-linguistic variation in the marking of generics as definite or indefinite.

On the basis of these analyses, we plan to develop libraries for the Grammar Matrix customization system covering the topics discussed here. The Grammar Matrix customization system (Bender and Flickinger, 2005; Drellishak and Bender, 2005) presents the linguist-user with a typological questionnaire which elicits information about the language to be described. On the basis of the user's responses to the questionnaire, the customization system compiles a working starter grammar out of the Matrix core grammar and analyses stored in libraries. The new libraries will cover argument optionality (both general pro-drop and lexically-licensed), as well as demonstratives of different syntactic types (pronouns, determiners, adjectives and affixes), the marking of definiteness, and definiteness agreement.

Acknowledgments

We would like to thank Toshiyuki Ogihara, Laurie Poulson, Jeanette Gundel, Jennifer Arnold, Francesca Gola, and the reviewers for STEP 2008 for helpful comments and discussion. Any remaining errors are our own. This material is based upon work supported by the National Science Foundation under Grant No. BCS-0644097.

References

Arnold, J. E. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes 23*(4), 495–527.

Bender, E. M., M. Egg, and M. Tepper (2005). Semantic construction for nominal expressions in cross-linguistic perspective. In *IWCS-6*.

Bender, E. M. and D. Flickinger (2005). Rapid prototyping of scalable grammars: Towards modularity in extensions to a language-independent core. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing IJCNLP-05 (Posters/Demos)*, Jeju Island, Korea.

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. In J. Carroll, N. Oostdijk, and R. Sutcliffe (Eds.), *Proceedings of the Workshop on Grammar Engineering and Evaluation at the COLING19*, Taipei, Taiwan, pp. 8–14.

Borthen, K. and P. Haugereid (2005). Representing referential properties of nominals. *Research on Language and Computation 3*(2), 221–246.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. Li (Ed.), *Subject and Topic*, pp. 25–56. New York: Academic Press.

Chafe, W. (1994). *Discourse, Consciousness, and Time*. Chicago: Chicago University Press.

Copestake, A., D. Flickinger, C. Pollard, and I. A. Sag (2005). Minimal recursion semantics: An introduction. *Research on Language and Computation 3*(4), 281–332.

Copestake, A., A. Lascarides, and D. Flickinger (2001). An algebra for semantic construction in constraint-based grammars. In *Proc. ACL*.

Diessel, H. (1999). *Demonstratives: Form, Function, and Grammaticalization*. Amsterdam: John Benjamins.

Diessel, H. (2008). Distance contrasts in demonstratives. In M. Haspelmath, M. Dryer, D. Gil, and B. Comrie (Eds.), *The World Atlas of Linguistic Structures Online*, Chapter 41. Munich: Max Planck Digital Library.

Drellishak, S. and E. M. Bender (2005). A coordination module for a crosslinguistic grammar resource. In S. Müller (Ed.), *Proc. HPSG*, Stanford, pp. 108–128. CSLI Publications.

Dryer, M. S. (2008). Order of demonstrative and noun. In M. Haspelmath, M. Dryer, D. Gil, and B. Comrie (Eds.), *The World Atlas of Linguistic Structures Online*, Chapter 88. Munich: Max Planck Digital Library.

Fillmore, C., C. Johnson, and M. Petruck (2003). Background to FrameNet. *International Journal of Lexicography 16*, 235–250.

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

Gundel, J., N. Hedberg, and R. Zacharski (1993). Cognitive status and the from of referring expressions in discourse. *Language 69*, 274–307.

Gundel, J., N. Hedberg, and R. Zacharski (2001). Definite descriptions and cognitive status in English: Why accommodation is unnecessary. *English Language and Linguistics 5*, 273–295.

McCloskey, J. (2004). Irish nominal syntax I: Demonstratives. UC Santa Cruz.

Oepen, S., E. Velldal, J. T. Lønning, P. Meurer, V. Rosén, and D. Flickinger (2007). Towards hybrid quality-oriented machine translation. On linguistics and probabilities in MT. In *TMI 2007*, Skövde, Sweden.

Pollard, C. and I. A. Sag (1994). *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. Chicago: University of Chicago Press.

Prince, E. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical Pragmatics*, pp. 223–255. New York: Academic Press.

Wintner, S. (2000). Definiteness in the Hebrew noun phrase. *Journal of Linguistics 36*, 319–363.