SEM-I Rational MT: Enriching Deep Grammars with a Semantic Interface for Scalable Machine Translation

Dan Flickinger[♣♠], Jan Tore Lønning[♣], Helge Dyvik[◇], Stephan Oepen[♣♠], and Francis Bond[♡♣]

[♣]Universitetet i Oslo, Boks 1102 Blindern; 0317 Oslo (Norway)
[◇]Universitetet i Bergen, Sydnesplassen 7, 5007 Bergen (Norway)
[♡]NTT Communication Science Laboratories, 2-4 Hikaridai, Kyoto 619-0237 (Japan)
[♠]Center for the Study of Language and Information, Stanford, CA 94305 (USA)

{ danf | jtl | helge | oe | bond }@emmtee.net

Abstract

In the LOGON machine translation system where semantic transfer using Minimal Recursion Semantics is being developed in conjunction with two existing broad-coverage grammars of Norwegian and English, we motivate the use of a grammar-specific semantic interface (SEM-I) to facilitate the construction and maintenance of a scalable translation engine. The SEM-I is a theoretically grounded component of each grammar, capturing several classes of lexical regularities while also serving the crucial engineering function of supplying a reliable and complete specification of the elementary predications the grammar can realize. We make extensive use of underspecification and type hierarchies to maximize generality and precision.

1 Introduction

In this paper we introduce two interesting features of the Norwegian-to-English machine translation system LOGON. (1) It is the first system to use the full power of Minimal Recursion Semantics in translation (originally introduced by Copestake, Flickinger, Malouf, Riehemann, & Sag, 1995). (2) The transfer modules use the SEM-I, an interface specification designed to allow the use of deep grammars in various applications without knowledge of the grammar internals (Copestake & Flickinger, 2003).

The motivation for the SEM-I is threefold. First, it allows the semantic representation to be underspecified. In the LOGON system, if the analysis system does not have enough information to commit to an interpretation then the ambiguity is retained. Doing this by naively expanding all interpretations is so inefficient as to be unworkable. Second, it exposes only the information that is relevant to semantic interfaces, and hides grammar internal specifics. For example, the fact that the English degree specifier *enough* idiosyncratically follows its head (*That mountain is high enough*) need not be a concern of someone working with the semantics. Thirdly, having a well-defined SEM-I allows the various components of the system to maintain consistency in the semantic representations that they accept and produce even as the coverage of these modules is extended over time.

The rest of the paper is organized as follows. In section 2 we give some background on semantic transfer. In section 3 we provide an overview of the LOGON project and show how semantics is integrated. In section 4 we introduce the concept of a SEM-I (semantic interface) and show how it plays a key role in the system. We illustrate in section 5 how our approach works and in particular the interaction between SEM-Is and underspecification.

2 Background: Semantic Transfer and Underspecification

LOGON (Oepen et al., 2004) is an experimental machine translation system from Norwegian to English. The core strategy is based on semantic transfer where the semantic representations are expressed in Minimal Recursion Semantics (MRS; Copestake, Flickinger, Pollard, & Sag, in press). Semantic transfer shares with interlingual approaches the assumption that translation is at its core a semantic activity, but departs from interlingua in emphasizing that different languages carve up reality differently. One has to consider these differences one language pair at a time.

Many syntactic properties which are important for getting the correct syntactic and semantic analysis are irrelevant for translation. An approach based on semantic transfer can more easily abstract away from such properties than

an approach based on syntactic transfer. We will give examples below where the transfer step is simplified while the correct output is guaranteed by the generation grammar.

Semantic representations based on logic have known problems when used in machine translation. First, there are many logically equivalent formulas. If one of them is the canonical representation of the meaning of the sentence constructed during analysis, how can we guarantee a successful generation from a syntactically different but semantically equivalent formula which is the output from the transfer module? Second, one syntactic analysis of a sentence may correspond to many non-equivalent formulas. One could calculate all these different formulas and use all of them as input to transfer but that is both highly inefficient and also unnecessary. In many cases the different formulas may result in the same translation. And if they do not result in the same translation, it does not help much to try to specify scope if we do not have cues to choose the right one. MRS was proposed as a representation format that overcomes these problems and hence is useful for semantic transfer (Copestake et al., 1995). An MRS structure is a flat, unordered structure which is underspecified for scope. We believe we are the first to actually build an MT system where transfer is done on the MRS level.

Example (2) is an MRS structure for (1). The building blocks are elementary predications (EPs), like $\mathsf{ski\_n}(x_1)$ and $\mathsf{try\_v\_out}(x_1, x_2)$, corresponding to atomic formulas.

(1) Try out these skis of mine.

(2) $\langle h_1, \{h_1\colon \mathsf{imp\_message}(h_5),$
$h_6\colon \mathsf{pronoun\_q}(x_1, h_7, h_8),$
$h_{10}\colon \mathsf{pron}(x_1\{\text{PERS } \textit{2}\}),$
$h_{11}\colon \mathsf{try\_v\_out}(e_1, x_1, x_2),$
$h_{14}\colon \mathsf{these\_q\_dem}(x_2, h_{16}, h_{15}),$
$h_{17}\colon \mathsf{ski\_n}(x_2),$
$h_{17}\colon \mathsf{poss}(e_2, x_2, x_3),$
$h_{20}\colon \mathsf{pronoun\_q}(x_3, h_{21}, h_{22}),$
$h_{23}\colon \mathsf{pron}(x_3\{\text{PERS } \textit{1}, \text{NUM } \textit{sg}\}),\ \},$
$\{h_5 =_q h_{11}, h_7 =_q h_{10}, h_{16} =_q h_{17},$
$h_{21} =_q h_{23},\ \}\rangle$

The MRS is considered a bag of EPs, hence the ordering between the EPs is without importance. Even though MRS representations are not represented in an interlingua, there is some semantic decomposition. The imperative force is shown in the message type ($\mathsf{imp\_message}(h_5)$) and the elided second person pronoun is inserted. Both the elided *you* and the overt *mine* are represented as the same kind of relation $\mathsf{pronoun\_q}(x, h_A, h_B)$ with different constraints on the PERS(on) and NUM(ber) properties. In addition, the verb-particle construction *try out* is represented as a single predicate $\mathsf{try\_v\_out}(e_1, x_1, x_2)$.

Finally, all EPs are labeled with *handles*, e.g. h_{17} is the label on the predication $\mathsf{ski\_n}(x_2)$. Quantifiers introduce special relations in an MRS corresponding to generalized quantifiers. The first argument of a quantifier relation is the bound variable, the second is the restriction, and the third is the body. Scope underspecification is represented in MRS by having handles in argument positions of EPs that are not the labels of other EPs.

Several other approaches to scope underspecification were proposed at about the same time as MRS, including Hole Semantics (Bos, 1995) and UDRS (Reyle, 1993). The latter was used at the transfer level in VerbMobil (Wahlster, 2000). In the LOGON project we take underspecification one step further. In addition to underspecification of structural logical properties, it is desirable for MT systems to be able to underspecify lexical semantic properties, e.g. the different readings of a polysemous word or the count vs. mass distinction (cf. Bunt, 2003). In the LOGON system, the implementation of the MRSs and transfer rules is done in a typed formalism with inheritance. This enables underspecification over classes of predicates, and thereby allows MT components to defer the resolution of ambiguity, as we will illustrate below.

MT research has been dominated by statistical methods on relatively shallow input for the past decade or so, but we doubt the value of pure statistical approaches in the long run. A semantic analysis is necessary to preserve the semantic content of the input and a grammatically based generator is needed to secure the wellformedness of the output. Computational precision grammars have a vastly larger coverage than a few years ago and—thanks to improved algorithms and faster hardware—they can now be put to practical use.

In the LOGON project we reuse and refine two computational grammars which have both been developed over several years within various projects. For Norwegian, we use the NorGram based on LFG, developed on the Xerox Linguistic Environment (XLE) platform, whose parser

is integrated in the LOGON system, and for English, we use the ERG, based on HPSG. The ERG has, for example, been used for several different purposes, including MT, email autoresponse, deep information extraction, and ontology building for the semantic web. A major reason for the growing success of the deep approach is that computational grammars are steadily developed, extended, and refined over several years in this way.

3 LOGON—MRS-Based MT

The Norwegian LOGON initiative (Oepen et al., 2004) aims to deliver a high-quality, domain-adapted MT system for written texts. The project involves research groups at the Universities of Oslo, Bergen, and Trondheim and targets the domain of tourism-related information—specifically the translation of Norwegian instructional documents on back-country activities into English.[1]

Emphasizing translation quality more than breadth of coverage, the consortium has adapted a relatively conventional approach, viz. semantic transfer of logical-form meaning representations obtained from a broad-coverage parser of Norwegian and subsequent grammar-based generation from target language semantics. On top of this symbolic backbone, LOGON incorporates stochastic components at all processing levels, primarily to rank and select among competing hypotheses and, to a lesser degree, increase end-to-end robustness. Given significant progress in computational linguistics in the past two decades, a central goal in LOGON is to evaluate state-of-the-art grammatical frameworks and processing schemes as to the contribution they can make to a high-quality end-to-end MT system.

3.1 Integrating Diverse Approaches

LOGON builds on independently-developed grammars couched in diverse grammatical frameworks (LFG for Norwegian, HPSG for English) for the analysis and generation components, respectively. The choice of semantic transfer and specifically Minimal Recursion Semantics provides a fertile testbed for the ability of the approach to abstract from theory- and grammar-internal representations. The overall LOGON set-up encourages a relatively pure, linguistically 'deep' approach to transfer and puts central emphasis on aspects of semantic representation and their feasibility for MT. Plurality of approaches to grammatical description and re-usability of individual resources are among the strong points of the LOGON collaboration. Accordingly, the potentially conflicting desiderata of linguistic adequacy vs. task-specific representations, and of grammar-internal vs. system-wide requirements, inevitably govern the design decisions concerning the grammar interfaces and the transfer component.

Figure 1 provides a schematic overview of the LOGON translation pipeline. Source language analysis, transfer, and target language realization all use MRS as their interface representation. Despite our use of underspecification where applicable, each component will typically output multiple hypotheses—corresponding to distinct parses, for example, in analysis. Cascading ambiguity through the translation pipeline yields a fan-out tree, where specialized stochastic processes for each step allow ranking and pruning of intermediate results (e.g. Velldal, Oepen, & Flickinger, 2004).

In contrast, many NLP systems are forced to prune relatively early. It is common for rule-based MT systems (e.g., **ALT-J/E**—Ikehara, Shirai, & Bond, 1996) to prune early—after morphological analysis and then after syntactic/semantic analysis (before transfer). Similarly, most statistical systems end up using some kind of beam search, either during analysis or decoding. This can lead to a solution being found that is worse than the global optimum.

Transfer in LOGON is realized as a resource-sensitive rewrite process, where rules replace MRS fragments (SL to TL) in a step-wise manner. The general set-up is similar to transfer in Verb*mobil* (Wahlster, 2000), but operating on semantic representations only and adding two new elements: (i) the use of typing for hierarchical organization of transfer rules and (ii) a chart-like treatment of transfer-level ambiguity.

As an example of how transfer with MRS simplifies translation, consider example (3), the Norwegian equivalent of (1). There are several differences syntactically: Norwegian uses a single verb *prøve* 'to try out', and the noun phrase structure is quite different. However, the MRSs (shown here omitting some scope in-

[1] See the public project web pages at 'http://www.emmtee.net' for additional information. To investigate the portability of the general approach and component re-usability, a bi-directional Japanese – English instantiation of the system serves as a secondary test-bed, but for the time being it is far less developed than the Norwegian – English main branch.

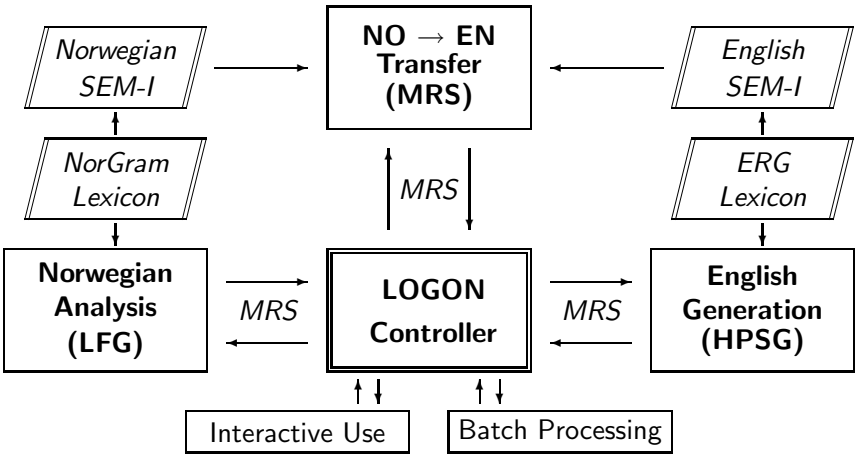

Figure 1: Schematic LOGON system architecture: the three core processing components are managed by a central controller that passes intermediate results (MRSs) through the translation pipeline. Both the analysis and generation grammars 'publish' their interface to transfer—i.e. the inventory and synopsis of semantic predicates—in the form of a Semantic Interface specification ('SEM-I'), such that transfer can operate without knowledge about grammar internals.

formation) are almost identical, and the transfer component has only to rewrite three predicates: prøve_v → try_v_out, denne_q_dem → these_q_dem and ski_n → ski_n.

(3)

Prøv	*disse*	*skiene*	*mine!*
verb	demve	noun	poss
try-out	these	skis	mine

Try out these skis of mine!

(4) $\langle h_1\{h_1$: imp_message(h_5),
h_6: pronoun_q(x_1, h_7, h_8),
h_{10}: pron$(x_1\{2nd\})$,
h_{11}: prøve_v(e_1, x_1, x_2),
h_{14}: denne_q_dem(x_2, h_{16}, h_{15}),
h_{17}: ski_n(x_2),
h_{17}: poss(e_2, x_2, x_3),
h_{20}: pronoun_q(x_3, h_{21}, h_{22}),
h_{23}: pron$(x_3\{1st, sg\})$, $\}$, $\ldots\}$ $\rangle$

3.2 Augmenting LFG with MRS

The Norwegian MRS representations which serve as input to the transfer component are derived by means of augmenting the Norwegian LFG-based resource grammar NorGram with an MRS component. To our knowledge this is the first attempt to couple LFG syntax with MRS semantics, but the attempt bears some resemblance to earlier work on mapping LFG f-structures to other kinds of underspecified semantic representations, notably Quasi Logical Forms (QLF) and Underspecified Discourse Representation Structures (UDRS). Genabith & Crouch (1997), for example, show that subsets of the LFG and UDRS formalisms can be brought into one-to-one correspondence, i.a. based on the fact that LFG f-structures, like UDRSs, contain predicate-argument structure information, abstract away from linear order, and leave quantifier scope underspecified.

While these results provide a starting point, they cannot be applied directly to the task of deriving MRS representations within the LOGON project. The main reason is that even if it may be possible to define a similar simple mapping from well-formed f-structures to corresponding well-formed MRS-structures, this would only demonstrate the relationship between the two formalisms in general, while our task goes beyond that. We need to interrelate specific f-structures and MRS representations which are not only well-formed, but which also satisfy further, mutually independent constraints. In the first place, already the fact that f-structures are syntactic representations and MRSs semantic representations designed to capture translational relations frequently motivates different packagings of information on the two levels. Furthermore, the NorGram f-structures meet the requirements for f-structures developed within the ParGram project (Butt, Dyvik, King, Masuichi, & Rohrer, 2002; Dyvik, 2003), while the NorGram MRS representations are constructed according to the same general principles as the MRS representations of the target ERG grammar. As a result the f-structure and MRS analyses of the same sentence are not always in a simple structural correspondence with each other. One example is nominal phrases with several specifiers in the f-structure, and phrases with no specifiers (Norwegian bare sin-

gulars), contrasted with the MRS requirement that a variable must always be bound by one single quantifier; other examples involve different dominance relations among predicates and decomposition of predicates.

The implementation of the MRS module exploits the projection architecture of LFG by projecting the MRS representation off the f-structure by co-description, and subjecting the resulting structure to a limited amount of post-processing to convert it to the LOGON interface format. It demonstrates the feasibility of deriving structures meeting external specifications from LFG resource grammars.

3.3 Facilitating Transfer by Normalization

The fact that MRS structures are semantic representations means that much syntactic variation in the source and target languages has already been dealt with during the derivation of the MRS representations. This contributes towards a simpler transfer component. For example, both languages show a variety of possessive constructions; Norwegian has the following main types (where all the examples mean 'the crocodile's tail'): *krokodillens hale* (lit. the-crocodile's tail); *halen til krokodillen* (lit. the-tail to the-crocodile); *krokodillen sin hale* (lit. the-crocodile his/her/its$_{[\text{REFL}]}$ tail). All these types are mapped to the same SEM-I predicate poss(e, x_i, x_j) in the Norwegian MRS representation, a predicate which also belongs to the English SEM-I. Hence there is a null-transfer of the relevant EP in this case, yielding English equivalents like *the crocodile's tail* and *the tail of the crocodile*.

Similarly the MRS analysis may exempt the transfer component from recovering bound antecedent-anaphor relations in some cases. Norwegian has a reflexive possessive which is unmarked for natural gender: *Han tok **sin** hatt* 'He$_i$ took **his**$_i$ hat', *Hun tok **sin** hatt* 'She$_i$ took **her**$_i$ hat'. GEND(er) is one of the properties associated with referential variables in both Norwegian and English MRSs. The Norwegian analysis recovers the antecedent in cases of bound anaphora and associates its GEND value with the referential variable of the anaphoric possessor; for the first example this yields a constraint like $x\{\text{GEND } m\}$ on the referential index associated with the possessor. Again null transfer of the corresponding EPs ensures the correct choice among *his* and *her* in the translation.

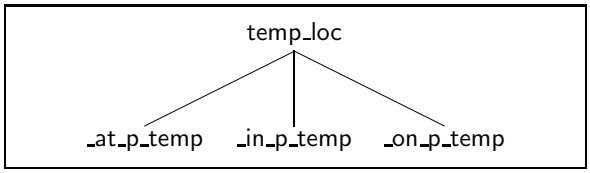

Figure 2: Excerpt from predicate hierarchy provided by the English SEM-I. Temporal, directional, and other usages of prepositions give rise to distinct, but potentially related, semantic predicates, and abstractions like the temp_loc predicate facilitate underspecification over grammar-internal variation.

4 The *SEM*antic *I*nterface (SEM-I)

In LOGON, the source and target language grammars evolve partly independent from the transfer component and MT system as a whole. This is to some degree caused by the distribution across sites, but more importantly comes as a consequence of the general aim for modularity and re-usability. Parallel to LOGON, our resource grammars continue to be applied in additional domains and for tasks other than machine translation. Accordingly, it is vital to users of such grammars that the external interface be documented in sufficient detail and that there exist procedures to validate analyses obtained from a grammar or given as input to the generator. A related practical requirement lies in the continuous evolution of resource grammars: while extensions and sometimes revisions of earlier analyses are desirable, external consumers of a grammar require automated tools to detect such changes over time and adjust their use of the grammar accordingly.

In the LOGON machine translation system, the transfer component can be conceptualized as an 'external' consumer to both the Norwegian and English grammars. In this view, the SEM-Is for the two grammars serve analogous to an application programmer interface (API) to the analysis and generation components: transfer can operate without knowledge about grammar internals, as long as all the relevant aspects of the semantic interface to either grammar are published in the SEM-I. The SEM-I is maybe best conceptualized as a large table, enumerating the complete list of semantic predicates and their terms of use. For each predicate, these minimally include the set of valid semantic roles (including an indication of optionality of arguments) and their value constraints, if any. Furthermore, the SEM-I provides generalizations over classes of predicates—e.g. hierarchical relations like those depicted in Figure 2—that play an important role in the organization of MRS

Predicate	Synopsis
_try_v_1	ARG0: e ARG1: x [ARG2: u]
_try_v_for	ARG0: e ARG1: x ARG2: x
_try_v_out	ARG0: e ARG1: x ARG2: x

Table 1: Sample SEM-I entries for various *try* predicates. While the two variants selecting for a semantically vacuous preposition supply separate senses, all other syntactic variants of *try* (e.g. the simple transitive and subject-equi ones) introduce the same predicate in the semantics. Value constraints and optionality of semantic arguments are indicated as part of the predicate synopsis. Optional arguments are shown in square brackets, and sortal constraints as different types of MRS variables.

Predicate	Synopsis
se_v_1	ARG0: e ARG1: x ARG2: x
se_v_1	ARG0: e ARG1: x ARG2: h

Table 2: Sample SEM-I entries for the Norwegian verb *se* ("see"). Both entries introduce the same semantic predicate, with the first corresponding to the simple transitive verb with an NP complement, and the second entry corresponding to the sentential-complement variant, where the second semantic argument is the handle of the complement, rather than a referential index.

transfer rules. Table 1 shows a simplified example for the English verb *try*, and Table 2 shows the entry for the Norwegian verb *se* ("see").

Role labels in MRS are typically drawn from a small inventory of bleached names ranging from ARG_0 to ARG_n, where the interpretation of thematic roles is assumed to be grammar-external, but largely homogeneous for clusters of predicates in the SEM-I. A minimal hierarchy of variable types—comprising events (e), referential indices (x), scopal variables dubbed handles (h), and generalizations over these (i or u)—serves to encode elementary ontological distinctions on argument positions. Further constraints on semantic arguments can be spelled out in terms of variable properties, where a specification like ARG0: $x\{\text{NUM } pl\}$ could be associated with the SEM-I entry for a pluralia tantum like woods_n.

5 The SEM-I in Action

Once the SEM-I's for both source and target language grammars are in place, they enable improved levels of robustness and precision in the set of transfer rules within an MT system, as well as additional efficiency benefits in parsing and generation. In this section we illustrate several uses of the SEM-I within the LOGON demonstrator framework.

5.1 Word Sense Underspecification

Given that surface words often have multiple senses which are not syntactically distinguished in one language but are in another, the SEM-I enables a grammar to underspecify the semantics of many ambiguous lexical items. For a noun like the familiar English *bank*, which has one sense as a financial institution and another as the side of a river, the English grammar's lexicon used for parsing and generation will only have a single lexical entry whose semantic predicate is underspecified for these two senses. Thus the grammar will only assign a single syntactic analysis for the sentence *The park is near the bank.* with an MRS representation that includes an elementary predication whose predicate is simply bank_n. Since these two senses of the noun *bank* are never syntactically distinguished, the English SEM-I also contains only a single entry for the corresponding semantic predicate bank_n, where the suffix on the predicate name supplies a coarse-grained distinction between noun senses and verb senses. An MT transfer rule can then target the single noun *bank* via this SEM-I entry, even if the source language semantics supplied a more specific sense. The SEM-I as used here does not represent ontological information unless there is a grammaticization of that sense distinction.

This grammaticization can be seen in the SEM-I entries for the English noun *paper*, where two of its senses correspond to distinct syntactic properties. The 'academic work' sense of *paper* as in *She wrote a good paper* is associated with a count noun, while the 'writing material' sense of *The printer needs more paper* is supplied by a mass noun. By making use of the SEM-I to represent these correlations, we can avoid lexical ambiguity which is costly for parsing and generation, while still providing the basis for transfer rules that produce high-quality output. The SEM-I has three entries corresponding to the one lexical entry for the noun *paper*: one for each of the two senses, and one for the underspecified sense that the lexical entry supplies directly.

paper_n	ARG0: x
paper_n_inform	ARG0: $x\{\text{IND} +\}$
paper_n_subst	ARG0: $x\{\text{IND} -\}$

The SEM-I entries for each of the two specific senses also include a constraint on the boolean property IND (for 'individuated') of the refer-

ential index introduced.[2]

In Norwegian, these two senses are supplied by distinct words, so the transfer rules can identify the correct target sense in English, and the SEM-I entries above ensure that the appropriate IND value is visible to the grammar when generating the English output. Since this IND property is associated by the English grammar with the obligatory presence of a determiner for singular count {IND +} nouns, we will generate for example *She wrote a good paper* but not **She wrote good paper*. The underspecified first entry above will be useful in translating between English and, say, German, where in both languages one word can be used for both senses, and where for some sentences no disambiguation is required, as in *That paper is good*. This approach differs from many semantic-transfer based systems, particularly knowledge-based MT. In these systems lexical semantic information (including word senses, semantic classes and selectional preferences) is used to disambiguate input (Mahesh et al., 1997; Ikehara et al., 1996). The interface between transfer and generation is then a single fully specified semantic representation (although often with scope issues ignored), rather than an underspecified one. The SEM-I is compatible with such an approach — it would then link the grammars to the ontology of lexical semantic information and provide the input to the sense disambiguation module.

5.2 Productive Derivation

Languages typically include productive derivational processes which result in multiple words whose syntactic and semantic properties are related but distinct. One such common process is 'Grinding', which relates a noun like *crocodile* (denoting the individual) to a corresponding noun which denotes some portion of the matter which comprises that individual (Pelletier, 1979). Here the two senses correspond to two words with distinct syntactic properties in English, where the 'individual' sense is again associated with a singular count noun which requires a determiner, while the derived 'substance' sense is expressed by a mass noun whose determiner is optional. One expensive approach to treating this alternation would be to add a rule to the grammar introducing a mass noun lexical entry for every count noun entry. Such an approach would lead to potentially exponential increases in processing cost, even though the syntactic context for each noun will often ultimately disambiguate the two senses: *I saw a crocodile* vs. *She doesn't eat crocodile*.

The SEM-I presents an opportunity to avoid the increased processing cost threatened by treating the Grinding rule as a lexeme-to-lexeme derivational rule. Instead, we treat Grinding as a productive alternation that applies to entries in the SEM-I, rather than to lexical entries. Thus the lexicon contains just one lexical entry for *crocodile* which is underspecified both for its predicate and for the count/mass distinction, which corresponds to obligatoriness of the noun's determiner. Every sentence with the word *crocodile* will have an MRS including an EP with the underspecified predicate crocodile_n, but for some sentences the syntactic constraints of the grammar will result in an MRS with a more specific value for the boolean property IND encoding the mass/count distinction on a given referential index. For example, a sentence like *The crocodile was expensive* has a single analysis whose MRS leaves the IND value for the 'crocodile' EP underspecified; in contrast, the MRS for *I saw a crocodile* includes the constraint {IND +}, while the MRS for *She doesn't eat crocodile* says {IND −}. This approach avoids lexical ambiguity for the Grinding alternation, yet preserves enough information in the resulting MRSs so the SEM-I can provide disambiguated word senses when possible.

The SEM-I includes the following two entries for the semantic predicate crocodile_n, where the second is automatically derived from the first via the Grinding rule, either in a precompilation of the SEM-I extracted from the grammar, or on the fly when processing: Both entries are also assigned a value for the boolean property DRV ('derived') to distinguish basic senses from coerced ones, a distinction which can be exploited by stochastic methods for selecting preferred parses or realizations.

crocodile_n	ARG0: x{IND +, DRV −}
crocodile_n	ARG0: x{IND −, DRV +}

Of course, this sketch of our approach only addresses the tip of the iceberg on the Grinding alternation, and important related alternations for packaging (*He ordered two coffees*) and kinds (*This printer needs a better paper*), to say nothing of exceptions, lexicalizations, or sub-

[2]Each entry in the SEM-I also contains fields for optionality, example sentences, documentation, etc, suppressed here for brevity, but discussed in (Copestake & Flickinger, 2003).

regularities. But locating such alternations in the SEM-I enables us to express both idiosyncrasies and productive correlations between syntactic form and word sense while minimizing increased search costs in parsing and generation.

5.3 Closed-class predicates

The grammar for a given language will include a set of semantic predicates introduced by closed-class words or morphemes like determiners, modals, prepositions, pronouns, etc. As seen above, these predicates are organized in a hierarchy like the one illustrated for temporal prepositions, allowing transfer rules to be stated at a level of abstraction best suited for the relevant correspondences in a given language pair. As another example consider the Norwegian modal verb *kunne* which can be translated into either English *can* or *be able to*. The two semantic predicates introduced by the lexical entries for the verb *can* and the adjective *able* will both be subtypes of a more general predicate which is the target in the appropriate transfer rule. The SEM-I enables both lexical entries to be made available to the generator given this single abstract predicate as input, and then grammar-internal constraints will guarantee that *can* is only used in present-tense verb phrases, while *able* realizes non-tensed or future tense phrases, as in *Will we be able to climb that mountain?*.

6 Conclusion

To develop a large-scale MT system using linguistically rich grammars with semantic transfer, it is essential to provide a complete semantic interface specification for each of the grammars. We treat the SEM-I as a first-class component of each grammar, enabling robust scalable transfer rules over MRS representations.

Acknowledgments

The LOGON initiative is funded by the Norwegian Research Council as part of the KUNSTI (*Knowledge Development for Norwegian Language Technology*) programme. This work is in part supported by the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford.

References

Bos, J. (1995). Predicate logic unplugged. In *Proceedings of the Tenth Amsterdam Colloquium* (pp. 133 – 142). Amsterdam, The Netherlands.

Bunt, H. (2003). Underspecification in semantic representations: Which technique for what purpose? In H. Bunt, I. van der Sluis, & R. Morante (Eds.), *Proceedings of the Fifth International Workshop on Computational Semantics IWCS-5* (pp. 37 – 54). Tilburg, The Netherlands.

Butt, M., Dyvik, H., King, T. H., Masuichi, H., & Rohrer, C. (2002). The parallel grammar project. In *Proc. of 19th Int. Conference on Computational Linguistics, Coling 2002.* Taipei, Taiwan.

Copestake, A., & Flickinger, D. (2003). *The semantic interface to the ERG and other LinGO grammars.* DeepThought draft tech report, University of Cambridge, UK.

Copestake, A., Flickinger, D., Malouf, R., Riehemann, S., & Sag, I. (1995). Translation using Minimal Recursion Semantics. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation* (pp. 15–32). Leuven, Belgium.

Copestake, A., Flickinger, D., Pollard, C., & Sag, I. A. (in press). Minimal Recursion Semantics. An introduction. *Research on Language and Computation.*

Dyvik, H. (2003). ParGram. Developing parallel grammars. *ELSNews*, *12*(2), 12 – 14.

Genabith, J. van, & Crouch, R. (1997). Interpreting f-structures as UDRSs. In *Proc. of the 35th Assoc. for Computational Linguistics and the 7th European ACL* (pp. 402 – 409). Madrid, Spain.

Ikehara, S., Shirai, S., & Bond, F. (1996). Approaches to disambiguation in **ALT-J/E**. In *Int. seminar on multimodal interactive disambiguation: Middim-96* (pp. 107–117). Grenoble.

Mahesh, K., Nirenburg, S., Beale, S., Viegas, E., Raskin, V., & Onyshkevych, B. (1997). Word sense disambiguation: Why statistics when you have these numbers? In *Proc. of 7th Int. Conf. on Theoretical and Methodological Issues in Machine Translation: TMI-97* (pp. 151–159). Santa Fe.

Oepen, S., Dyvik, H., Lønning, J. T., Velldal, E., Beermann, D., Carroll, J., Flickinger, D., Hellan, L., Johannessen, J. B., Meurer, P., Nordgård, T., & Rosén, V. (2004). Som å kapp-ete med trollet? Towards MRS-based Norwegian – English Machine Translation. In *Proc. of 10th Int. Conf. on Theoretical and Methodological Issues in Machine Translation* (pp. 11 – 20). Baltimore, MD.

Pelletier, F. J. (Ed.). (1979). *Mass terms: Some philosophical problems.* Dordrecht: Reidel.

Reyle, U. (1993). Dealing with ambiguity by underspecification. Construction, representation, and deduction. *Journal of Semantics*, *10*, 123 – 179.

Velldal, E., Oepen, S., & Flickinger, D. (2004). Paraphrasing treebanks for stochastic realization ranking. In *Proc. of 3rd Workshop on Treebanks and Linguistic Theories.* Tübingen, Germany.

Wahlster, W. (Ed.). (2000). *Verbmobil. Foundations of speech-to-speech translation.* Berlin: Springer.