Word Sense Disambiguation and Semantic Disambiguation for Construction Types in Deep Processing Grammars

Dorothee Beermann and Lars Hellan

NTNU
Trondheim, Norway
dorothee.beermann@hf.ntnu.no, lars.hellan@hf.ntnu.no

Abstract

The paper presents advances in the use of semantic features and interlingua relations for word sense disambiguation (WSD) as part of unification-based deep processing grammars. Formally we present an extension of Minimal Recursion Semantics, introducing sortal specifications as well as linterlingua semantic relations as a means of semantic decomposition.

1. Introduction

Not only 'word sense disambiguation' (WSD) but also the disambiguation of *Construction* senses are necessary to successfully accomplish most of the NL processing tasks. Moreover, aside from keeping senses apart, one ideally also wants to represent them by interesting criteria of *correctness*. Only for semantically atomic items will representations of the form 'Item-sense1' vs 'Item-sense2' be in any way adequate, whereas for constructs of internal complexity, a base-line requirement will be that elementary predications and coreference as well as variable identity be represented adequately. Grammar implementations within LFG and HPSG meet this goal (for HPSG, see below), within what is referred to as 'deep' language processing. To be presented here is a design that enriches flat predicate-logic semantic representations (Minimal Recursion Semantics; see below) by lexical semantic information and thus may serve as a cross-linguistically valid word and construction sense disambiguation tool. It will be illustrated in the following with examples from the Norwegian HPSG grammar 'NorSource' which is a grammar implementation that uses the development platform LKB (Copestake 2002) and that is situated within the DELPH-IN consortium of grammar development (http://www.delph-in.net/). For this paper we focus on a fine-grained semantic representation of spatial relations, and of comparative constructions. In section 2 we present desiderata on the representation of semantic distinctions. In section 3 we discuss the implementation of some of these desiderata.

2. WSD adequacy

In our view a WSD tool should be 'calculation-sharp'; for example, when it comes to expressions of measurement, the representations delivered should provide a break-down of exactly what is being measured and which values are being assigned. For instance, for *This building is 5 meters higher than the church*, a calculation-sharp analysis should provide a representation corresponding to the quasi-paraphrase:

(1) "For a degree *d1* and a degree *d2* such that *d1* is the height of the building and *d2* is the height of the church, *d1* exceeds *d2* to an extent *d3*, and *5 meters* measures out *d3*."

(1) stands in contrast to, for instance, a representation such as:

"5-meters-taller-than (the building, the church)"

where the naming of a function suggests a meaning rather than providing it.

Similarly, a semantic disambiguation tool should be able to recognize spatial uses of prepositions and distinguish the directional use from the locative one, so as to capture the ambiguity of, e.g., *jump in the car*. Within the directional reading we need to be able to account for different thematic constraints on subjects of movement verbs, such as 'mover' and 'line' concepts to account for the fact that *lead* is a synonym of *go* in *The way leads along the ridge* but not in *The guide leads along the ridge*. How many of the above and other, similar distinctions grounded in the lexical semantics of the items processed one is willing to make, will decide on the fine-grainedness of the WSD tool. Together with calculation sharpness, fine-grainedness will decide if one is able to represent, e.g., for a sentence such as *Walk along the ridge 5 kilometers towards the south* the understood 'event-overlap', that is a 'mover' who, in one and the same action, walks along a ridge *and* changes her position to a place whose degree of 'toward-south' exceeds the position held at the beginning of the action by (= 'measured-out' by) 5 kilometers. Given that one is able to make these necessary distinctions, one is then also able to represent the distinct meaning of *Walk along the ridge* ***and then*** *5 kilometers further south*, that is applying the above made descriptions to *time consecutive* actions.

In the following we are going to show how some of the above mentioned constraints can be stated within the formal frame of Minimal Recursion Semantics (MRS; cf. Copestake et al. (to appear)) and as part of an HPSG deep processing grammar. Since all specifications to be discussed in the following are clearly within the scope of a linguistic semantics (and not an aspect of physical or other analysis), they should be recognized by any linguistically

realistic language processing tool, and in particular by a deep processing grammar. But although linguistically desirable, it is rather obvious, given present limitations of deep processing grammars, that WSD cannot exclusively reside within a deep processing tool - corpus based statistically WSD is indispensable. In fact, interesting 'schnittpunkte' can be found. Siegel (1999), working on a corpus-based statistical WSD tool, points out that verbal sense disambiguation in English may rely on aspectual class discriminators. Related to the above made distinctions between 'movers' and 'line' subjects of motion verbs, the verb *reach* has a dynamic as well as a stative use, as exemplified by *The train is now reaching Trondheim* versus *The pilger road reaches Trondheim.* English allows the use of progressive tense with the dynamic reading but not with the stative reading, as observed, e.g., by Levin (1993). In corpus based WSD machine learning for English, progressive tense markers can thus be used as an indicator for the occurrence of the dynamic verb sense. However, the same indicator for a sense distinction will fail relative to languages where the progressive use is highly unusual with either of the two readings. A sense disambiguation tool like the one presented here, however, where items will be marked for the relevant sense distinctions as long as they can be parsed, the necessary distinctions will be part of the MRS output for any language, independent of syntactic encoding distinctions. It thus might in general be the case that also for WSD, a hybrid approach should be chosen, making the present method of sense disambiguation at the level of a precise logical-form meaning representation an interesting addition to statistical WSD tools. Moreover, it should be pointed out that WSD on the level of MRS, that is at a semantic level, will be of particular interest for all applications that use a robust semantic interchange format between NLP applications of different depth. The Deep Though project (http://www.progect-deepthought.net) represents an enterprise in this direction (for further reference to the use of MRS and its extension RMRS for hybrid language processing, see, e.g., Callmeier et.al (2004)). Within MT, applications based on semantic transfer such as the Norwegian 'LOGON' project (http://www.norskdok.uib.no/projects/?logon&lang=en) are of special interest. LOGON uses Minimal Recursion Semantics and base-generation by a deep-processing grammar of English as translation tools, which means that semantically based WSD information is directly accessible at the level of transfer from Source Language to Target Language.

3. The Tool

The tool we are describing outputs a semantic representation within the bounds of MRS as described by Copestake et.al (op.cit.). The two representations given below in figure 1 and figure 2 show partial semantic descriptions for the directional construction (Norwegian) *Han springer til skogen* ('He runs to the forest') and the comparative expression *The building is 5 meters higher than the church*, respectively. We first comment on figure 1:

Figure 1 *Minimal Recursion Semantic Representation of the Norwegian sentence: 'Han springer til skogen' (He runs to the forest)*

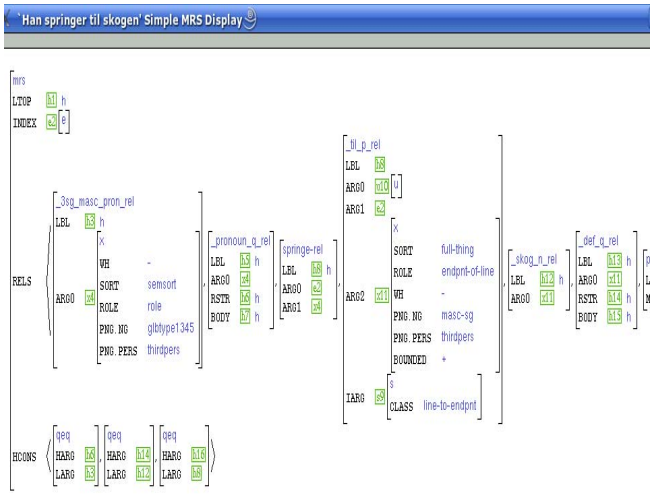


Figure 1 is a screenshot of an MRS structure from the Norwegian NorSource grammar. The attribute RELS provides a non-ordered list of the meaning-bearing elements presented as 'elementary predications' ('EP'), whose specifications are enclosed in square brackets. Inside each EP, the value of LBL identifies the EP itself , the attribute ARG0 provides a semantic index for the predication, being of type 'event' or 'ref-ind', according to whether what is specified by the EP is of a propositional or 'thing'- nature. ARG1 specifies a participant of the event and ARG2 a possible second participant, and so forth. Notice that each variable of the type *x* is bound, so that the pronominal subject of the clause is represented by a *3sg-masc_pron_rel* and a *pronoun_q_rel*, while the object 'skog' (forest) is bound by a *def_q_rel* in accordance with general constraints on MRS structures. An MRS remains underspecified for scope, and the according constraints are encoded under the attribute HCONS, an aspect that we will not further comment on here (for more see Copestake et. al., cited). The relevance of figure 1 in the present context arises from the specifications that the preposition *til* 'to' and its dependence receive. In essence what is illustrated is that all spatial relations are typed according to the conceptual type of the arguments they select. Let us start with the specifications that are embedded under the attribute IARG (for 'internal argument'). *Til* is specified as a preposition of type ***line-to-endpoint*** where ***line*** and ***endpoint*** stand for concepts imposed on the two arguments of the preposition. Prepositions are thus typed according to the conceptual restrictions they impose on their arguments. Specifications such as ***line-to-endpoint*** induce a further level of classification of the semantic variables of spatial relations and therefore must be thought of as a set of specifications induced on the identity variable of spatial relations. This means that specification such as the ***line-to-endpoint*** should be part of the ARG0 of the relations in question. However, since the ARG0 in an MRS serves in the general combinatorial apparatus that, e.g., allows the underspecification of scope, we have chosen instead to represent the information in question in a 'safe place' within the elementary predication, now called IARG for

'internal argument' (for more discussion of this and related topics see also Hellan and Beermann (2005)).

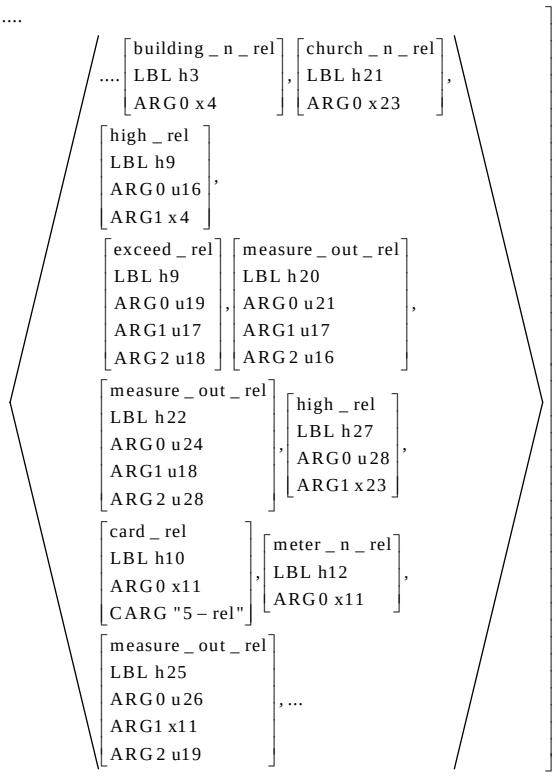
Returning to the preposition *til* notice that specifications such as *line* and *endpoint* live in an ontology where the top level distinction is that between *line* an object of any dimension (*xdim*). *Endpoint* together with *startpoint* and *viapoint* is a specification under *xdim* that plays a roll in the semantics of directions. The *line* concept is imposed on the first argument of the spatial relation. In figure 1 the *line* concept is satisfied by the motion verb. Notice that the prepositional phrase is identified as an event modifier in figure 1. In MRS terms that means that its label argument is identified with the label of the event which in the present case is instantiated by the verbal EP *springe-rel*. It is thus the motion event directly that satisfies the line requirement imposed by the preposition. In Jackendovian terms (cf.,eg., Jackendoff 1987), contrary to the above view, the so called 'path-argument' in a directional construction will be provided by the mover, that is, the subject of the motion event. Notice that nothing in the MRS formalism prevents a Jackendovian rendering of directions, which means that the ARG1 of the directional preposition would be directly identified with the ARG1 of the subject rather than with the event variable. Independent of these two distinct views of directional semantics, what remains essential for us here is that the verb directly or indirectly satisfies the *line* requirement of the spatial predicate. The *endpoint* concept on the other hand needs to be satisfied by the second argument, that is the object of the preposition, and, as again can be observed in figure 1, the *endpoint* concept is matched by the ROLE specification of the *x*-variable provided by *skog.* In summary, figure 1 illustrates how a standard MRS semantic output can be enriched by a system of spatio-temporal types, providing a system that distinguishes lines and spans from objects. As a consequence, directional expressions receive a different semantics from event-modifiers (the latter are of type *xdim-to-xdim*). Notice that that even holds for those MRS implementations where VP-modifying prepositional phrases are interpreted as event modifiers throughout, due to the additional specifications given in the IARG of spatial relations. Figure 1 thus illustrates *word sense disambiguation* beyond arbitrary indexation and naming conventions that use English as the interlingua, where, e.g., *run_v1_rel* and *to_p2_rel* would exhaust the information provided about predicate types. Such a system of typed predicate labels will allow the classification of senses, but has a clear limit for a further semantic break-down.

For applications that use MRS representations as input (e.g., in information extraction), the design offers potential advantages in two respects. For one thing, to the extent that semantic information is channelled through a finite set of defined features, search can be principally restricted relative to this set, rather than oriented towards the potentially more unrestricted array of predicates like *run_v1_rel* and *to_p2_rel*. A test demo along these lines is described in Beermann et al. (2004), where mountain hiking routes are rendered in pictographic form via RMRS grammar outputs and XML conversion.

Secondly, to the extent that the system underlying the annotations exemplified in figure 1 is cross-linguistically applicable, it opens the possibility for grammars of different languages to produce MRSes with identical annotations for sentences with essentially the same content in relevant respects. Once a given extraction algorithm has been defined for MRSes of this kind for one language, the same algorithm may then in principle be applicable for the corresponding MRSes for a different languages, enhancing the generic value of such an application module.

We next turn to the dimension of calculation-sharpness of semantic representations. In figure 2 below, this factor is more clearly brought into play, in a representation of comparison where the dimension of comparison is height, viz. the height of the building vs. the height of the church, explicating the paraphrase suggested in (1) for the sentence *This building is 5 meters higher than the church* (we here use English counterparts of the structures created by the Norwegian grammar):

Figure 2. (Part of) MRS representation for the sentence *This building is 5 meters higher than the church*:



The first two EPs introduce *x4* and *x23* as variables representing the building and the church, respectively. Next, *u16* is introduced as a variable representing *x4*'s height, and further down, *u28* as representing *x23*'s height. The EPs with predicate value *measure_out_rel* assign the measure values *u17* and *u18* to these respective heights, and the EP *exceed_rel* states that *u17* exceeds *u18*, i.e., that the building is higher than the church. The ARG0 of this EP - *u19* - in turn represents the extent to which this difference holds, and this value is associated with the measure *x11* (last EP), with the EPs preceding stating that this is meter in a number of five. Of these EPs, those for

measure_out_rel and *exceed_rel* clearly do not stand in a one-to-one relation to occurring words or morphemes: the last *measure_out_rel* reflects a syntactic specifier-head constellation between *five metres* and *higher*, and the other two plus the *exceed_rel* EP emerge from the comparative morpheme *-er*. The analysis mirrors proposals made in formal semantics, and indicates down to which detail insights from this field can also be implemented in a computational tool.

Potential advantages of such a degree of preciseness lie not only in the sheer adequacy of the representations, but also in the way such MRSes, suitably rewritten via XML conversion or other means, may be exchanged into calculable descriptions or instructions to natural or artificial agents. For such an endeavor to be feasible, the domain will have to be quite restricted, and with a highly predictable vocabulary. Still, once instructions or descriptions can be delivered using natural language text (written or spoken), interesting prospects clearly arise, and a prerequisite for making use of such a resource will be a calculation-sharp semantics, as here illustrated.

4. Final remarks

An open issue is to which extend lexical semantic information could and should be computed by a core grammar, or, to formulate the same question slightly different, which aspects of lexical semantic information should be handled by the core grammar because, for example, it may be needed for constraining the grammars own parse selection, and which of the cross-linguistically important lexical semantic distinctions could or must reside in add-on components, such as stand-off annotations and ontologies. The semantic distinctions discussed here are at present computed by a grammar. It therefore might be worthwhile mentioning that the parsing grammar 'NorSource' with its 494 phrasal types and 950 word and lexeme types can host the here discussed distinctions without compromising general efficiency. However, in order to fully link the prepositional typology to information encoded by verbs and nouns, and in the latter case real-world knowledge concerning their inherent properties, one will have to resort to additional independent sources of information.

An interesting feature of a grammar producing fine-grained and precise semantics of the kind mentioned is a high degree of what one may call *construction control*: detailed MRSes should occur exclusively with exactly those constructions whose meaning they reflect. One thus needs an architecture allowing a close form - meaning correspondence. In implementing such a design, it is noticeable that the grammar internal specifications yielding the semantics exposed, i.e., the (R)MRS, are not necessarily sufficient to serve as the correlates for construction control. Without being able to go into any detail at this point, it nevertheless should be noted that depended on where the implementation mechanism allows semantic information to propagate, a certain amount of ad hoc replicates of semantic specifications may have to be used, as it is the case in the grammar referred to. In our view the concepts of semantic headedness and semantic locality in particular demand further scrutiny in the computational semantics research to come.

5. References

Beermann, D., J.A. Gulla, L. Hellan and A. Prange (2004) Extraction of spatial Information using a Deep Processing Grammar. Paper presented at CLIN 2004, Leiden; to appear in proceedings.

Callmeier, U., Eisele.A., Schäfer, U. and M. Siegel. The Deep Thought Core Architecture Framework. In: *Proceedings of LREC 2004.*

Copestake, A. (2002) *Implementing Typed Feature Structure Grammars*. Stanford: CSLI Publications.

Copestake, A., D. Flickinger, I.A. Sag and C, Pollard. (To appear) Minimal Recursion Semantics: an Introduction. htttp://www-csli.stanford.edu/~aac/papers.html.

Hellan, L. and D. Beermann (2005). Classification of Prepositional Senses for Deep Grammar Applications. In Kordoni, V. and A. Villavicencio (eds) *Proceedings of the Second ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications.* University of Essex.

Jackendoff, R. (1987) The Status of Thematic Relations in Linguistic Theory. *Linguistic Inquiry*.18. 369 – 411.

Levin, B. (1993) *English Verb Classes and Alternations.* The University of Chicago Press.

Siegel, E.V. (1999) Corpus-Based Linguistic Indicators for Aspectual Cassification. In *Proceedings of ACL 1999*, pages 112-119, MD. University of Maryland.