Talk given at the 13th Nordic Conference on Computational Linguistics, NoDaLiDa '01, May 22, 2001, in Uppsala.

Token-based minimal recursion semantics

Mats Dahllöf, Uppsala University, Dept of Linguistics

I will describe a new computational approach to natural language semantics. I call it *Token-based minimal recursion semantics* (TBMRS). TBMRS can be seen as a version of *Minimal recursion semantics* (MRS, Copestake *et al* 1999), which is way of structuring semantic representations. Both the original and the present versions of MRS have been developed as integrated into a Head-Driven Phrase Structure Grammar (HPSG). However, a semantics of this kind could be supported by various formal grammars.

The original motivation behind TBMRS has been to implement the philosopher Donald Davidson's (1984) so-called *paratactic* treatment of the semantics for intensional constructions. As far as I know, this is the first attempt of turning Davidson's proposal into a computational model. Here, I will only mention briefly what the "paratactic" semantics is all about.

I've developed a TBMRS syntax and semantics for some English constructions, mainly intensional ones (Dahllöf 2001a). This grammar has been implemented with the help of my own system, *Prolog-embedding typed feature structure grammar* (Dahllöf 2001b, 2001c). A demo is available over the WWW (at `http://stp.ling.uu.se/˜matsd/tbmrs1/`).

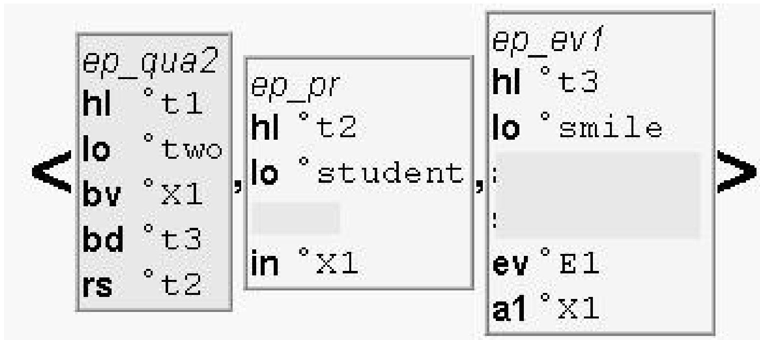
TBMRS is a formal and methodological framework. It's not a specific theory. It can be combined with various accounts of specific details.

Minimal recursion semantics, or MRS, is mainly a way of organizing semantic representations. The MRS method can be applied to almost any kind of underlying formal semantic framework.

The basic idea behind MRS is to use representations with a flat syntax. Syntactic embedding is eliminated. The elementary units of the representations are called *elementary predications (*EPs*)*. The EPs roughly correspond to the basic semantic units of ordinary logic. The MRS representations are sets or lists of EPs. Each EP is associated with a so-called *handle*, an identifying label. Some EPs, for instance, quantifiers, carry features which take handle values. This allows feature-handle connections to indicate operator scope. So, syntactic embedding is not needed for that purpose, as in ordinary predicate calculus.

The "ordinary" predicate EPs carry a handle value, as all EPs do. Another feature identifies the predicate, i.e. the lexeme. Additional features correspond to the logical arguments of the predicate. Their number and their kinds depend on which predicate we're dealing with. Apart from handle and lexeme features, generalized quantifiers involve a *bound variable*, a *restriction* and a *body*. The restriction and the body are scopal features, taking handle values. This is the way in which quantifier scope is captured, without the use of syntactic embedding.

This is illustrated by the simple example sentence “two students smiled”:



I’ve covered features which I don’t want to discuss here. For instance, I’m working on the treatment of tense and other aspects of verb inflection.

There are three EPs, with the handles `t1`, `t2`, and `t3`. The quantifier has `t2` for its restriction and `t3` for its body. The variable `X1` connects the quantifier binding and the argument places. in is the instance argument and a1 the first argument on the verb, i.e. the one corresponding to the subject.

An important motivation behind MRS has been to find a simple and efficient method for underspecification of quantifier scope. This is achieved by means of handle variables. Scope features are associated with handle variable values, in the underspecified cases. These variables may be associated with constraints concerning scope. Such constraints derive from the grammar. There are also more general constraints. (Object variables must be bound by a quantifier, for instance.) A computational advantage of this treatment of underspecification is that it allows scope resolution to be a monotonic process. Handle variables are simply unified with specific handle values.

The TBMRS approach is of a very Davidsonian character: First, it gives us an extensional and truth-conditional semantics. Secondly, it treats events Davidson-style (Davidson 1980). A more remarkable thing about TBMRS is that it provides a formalization Davidson’s “paratactic” analysis of intensional constructions. It can be illustrated by a sentence like “Galileo said that the earth moves”. As is well known, the subordinate clause cannot represent just a truth-value. There are many theories that claim that it represents a proposition, which is some kind of abstract structure.

Propositional theories are problematic for several reasons: They require sophisticated and speculative accounts of propositions and their components. Furthermore, if we assume that clauses represent propositions in certain contexts, the semantics becomes more complicated.

Davidson's (1984) "paratactic" treatment can be outlined with the help of an example like this sentence:

$$\underbrace{\text{Galileo said that}}_{S_1}\ \underbrace{\text{the earth moves}}_{S_2}.$$

Davidson's analysis is that the "said" relation relates Galileo to the subordinate clause *token* which is part of this sentence token. So, the token itself is the content, so to speak Propositions don't have to enter the picture. This gives us a logical form like:

$$\text{said}(\text{Galileo}, S_2)$$

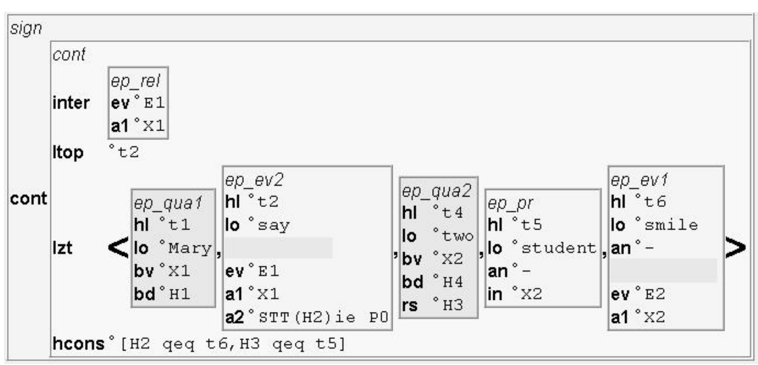
In order to formalize Davidson's proposal, we need a semantics in which tokens are available. This is one of the motivations behind TBMRS.

I've explored the idea of identifying the EP handles with the corresponding tokens. I assume, in as many cases as possible, that a word token corresponds to one EP. This is an assumption that is quite natural to make.

There can be exceptions: Some words don't carry any content, for instance, infinitival marks. There could also be "implicit" EPs. An example would be the implicit existential quantifiers which are often associated with Davidsonian event variables.

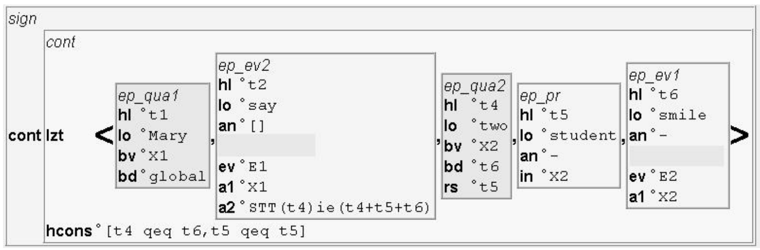
There are three kinds of semantic information in a TBMRS representation: (1) lexeme identification, (2) coindexing (by means of variables), and (3) scopal relations. These aspects are all represented by predicates which are ascribed to tokens. Tokens are classified and related to each other. So, this is clearly a very surface-oriented framework: Semantic information is tied to word (or morpheme) tokens, and the semantics makes minimal use of abstract entities.

Let's turn to an indirect speech example, "Mary said that two students smiled". It's given this underspecified representation:



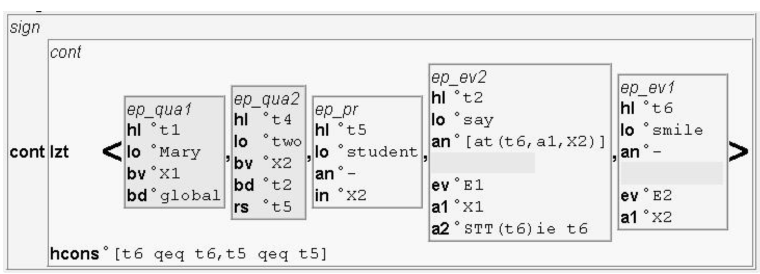
There are a number of handle variables here, the ones with `H`. `hcons` stands for "handle constraints". The relation `qeq`, *equality modulo quantifiers*, holds in the case of equality and when two handles are scopally connected by means of one or several quantifiers, to put it informally. This relation makes it easy to allow quantifiers to float. The a2 argument of the verb "say" is the intensional one, and carries a token value, as required by the "paratactic" analysis. The function STT (sub-tree token) assembles the possibly complex token formed by the token given as argument and the tokens which are scopally subordinate.

The previous underspecified representation may be specified in this way:



This corresponds to a reading which is *de dicto* with respect to the NP "two students". The intensional a2 argument, `STT(t4)=t4+t5+t6`, is the whole subordinate clause.

There is also a *de re* reading of the previous example sentence:



The intensional a2 argument now only includes the verb token of the subordinate clause. The NP "two students" outscopes the indirect speech verb. The an (anchor) argument serves the purpose of making the indirect speech relation sensitive to the "two" quantificational binding. Its value is a set (list) defining a function which associates the relevant token-argument slots with external entities. This makes it possible for a person to have said a token under a certain anchoring, but not under another one.

I want to conclude by saying why I think that the TBMRS approach is interesting:

- TBMRS provides a formalization of Davidson's "paratactic" analysis of intensionality. It also shows a way of dealing with quantification into intensional contexts.

- TBMRS is a conceptually and ontologically very parsimonious approach. It's a minimal version of truth-conditional semantics, yet a powerful one.
- The token-based nature of TBMRS allows various kinds of linguistic information to be integrated within the same framework.
- TBMRS supports a computationally attractive treatment of underspecification, in the same way as other versions of MRS.
- TBMRS supports a simple and straightforward compositional semantics. (The EPs are collected by means of set union thus percolating from syntactic daughters to mother.)
- The TBMRS principles work very well together with HPSG-style grammars.
- TBMRS provides a way of systematically tracing semantic information to the tokens that carry the information. This is a useful facility in language engineering applications.

REFERENCES

Copestake, A., D. Flickinger, I. A. Sag, och C. Pollard (1999), 'Minimal Recursion Semantics: An Introduction'.
WWW: `http://www-csli.stanford.edu/~aac/papers.html`

Dahllöf, M. (2001a), 'The paratactic treatment of intensional constructions: making it work', submitted to *Journal of Semantics*. [Revised as: Dahllöf, M. (forthcoming), 'Token dependency semantics, and the paratactic analysis of intensional constructions', *Journal of Semantics*.]

Dahllöf, M. (2001b), 'Prolog-embedding typed feature structure grammar and grammar tool. Version 2.1',
WWW: `http://stp.ling.uu.se/~matsd/petfsg2/`.

Dahllöf, M. (2001c), 'An implementation of token-based minimal recursion semantics using the PETFSG grammar tool',
WWW: `http://stp.ling.uu.se/~matsd/tbmrs1/`.

Davidson, D. (1984), *Inquiries into Truth and Interpretation*, Clarendon Press, Oxford.

Davidson, D. (1980), *Essays on Actions and Events*, Clarendon Press, Oxford.