Verb-headed Multi-Word Expressions (MWEs) in the Norwegian HPSG grammar Norsource

Lars Hellan, NTNU

PARSEME 2nd general meeting, Athens, March 10–11, 2014. WG2

The computational grammar *Norsource* at NTNU, Trondheim, is an HPSG grammar for Norwegian, using the LKB platform (Copestake 2002) with a feature structure based on the *HPSG Grammar Matrix* (Bender et al. 2002, 2010), a semantics using MRS (cf. Copestake et al.2005), and with an overall technical infrastructure as supported by the *DELPH-IN* network. Its lexicon consists of more than 80,000 words, thereof about 12,500 verb entries.

As is common in the 'head driven' design, when a verb can occur in x many construction types differing in features reflecting argument structure, there will be x many entries of this verb each one 'programmed' for one specific construction frame. In this lexicon, entries of verb-headed MWEs will consist of frames where specific lexical items are specified in the slots designed for dependent items. The specification of MWEs in Norsource falls into three categories:

1. Frames which, when a specific word is used in a given slot, project a specific grammatical feature such as aspect to the construction as a whole.
2. Frames which, when a specific word is used in a given slot, induce a meaning for the construction as a whole which is not perceived as 'compositional' relative to frequent uses of the words involved and the way they are combined.
3. Frame structures which, relative to the given verb, obtain only when a specific word is used in a given slot, and there is otherwise no special information projected from the specification..

Category 1 can be illustrated by a screenshot for the sentence *Regnet holder opp* ('the rain holds up' = 'the rain ceases') with the syntactic tree and semantic representation produced by the appropriate parse, and the lexical entry specification for the relevant use of the verb *holde*. Here the lexical type for the verb, *v-intrPrtcl-COMPLETEDACTIVITY*, induces a structure with subject and an adverbial particle as complement, and with the Aktionsart type 'completed activity' (see the part 'SIT-TYPE: COMPLETED +' in the top line of the MRS), and the entry itself specifies *opp* as the item filling the particle slot (as value of the attribute 'KEY-SPEC', an entrance point for such information), and thereby being what induces the Aktionsart value in this case.

Norsource has nearly 1800 entries of category 3, where the only reason for using an entry specifying the adverb or preposition is to avoid excessive parse forests: the verb in question may have other entries defining other environments where also adverbs or prepositions can serve as heads of admitted constituents, and the entry in question could be wrongly activated for these if the specific adverb or preposition is not indicated.

Category 2, finally, is amply represented in the language; however, the grammar Norsource has only three cases encoded, just to secure a formal procedure for their treatment. Why so few? We may use (1) as example:

(1) Du tar feil
You take wrong (= "You are wrong" (English)

In the analysis of such cases, one will want to represent the circumstance that the meaning of the verb is not compositionally related to other 'standard' uses of the same word. A standard mechanism one could use is to mark words with 'sense indices', so that in "du tar feil", the verb "ta" would carry a different sense index than it does in "jeg tar mat" ('I take food'). The standard way of assigning such marking in the MRS style is by defining PRED-values distinguished by integers, such as in:
_ta_v_1_rel, _ta_v_2_rel, _ta_v_3_rel, ...

The "ta" in "du tar feil" could then for instance be number *16* in such an inventory, and none of the semantic expectations going along with the other "ta"-variants 'ta #1, 2, 3 ... 15', and 'ta #17, 18, ...', would carry over to this case, thus, e.g., excluding inferences for (1) which imply taking possession or control over something.
The lexical entry for this distinguished variant of *ta* would include the following parts:

(2) ... [COMPS < [... HEAD noun [KEY feil]]>,
... [arg12-relation [PRED _ta_v_16_rel]]

'KEY feil' identifies the noun *feil*, by virtue of this noun having an entry with 'feil' as value for KEY, and this entry is the sole entry carrying this information for this attribute. The information in (2) will thus make sure that the only case where the meaning '_ta_v_16_rel' obtains will be when the verb ta has an object headed by the noun *feil*.

However, to avoid parse-forest explosions for a sentence like *Han tok feil*, one has to make sure that one avoids getting a number of parses where all the other variants of *ta* in effect combine with *feil*[2]. The problem is that with the mechanism mentioned, one can specify what an item should combine with, but not what it *should* ***not*** combine with: there is no way of instructing all of '_ta_v_1_rel', ... '_ta_v_15_rel' to *not* combine with *feil*.

The issue probably has to be approached at a pre-processing level: One lists a large number of V+N sequences of the type of (1), with a designated lexical entry aligned with the main verb of each member of the list. If the parse input string matches an item in the list, only that variant of the main verb is allowed in a parse which matches the entry specification which has been previously paired with this item in the list. Thus, a member of this list could be "ta feil", aligned with the entry in (2), and once an input string compatible with "ta feil" is encountered (like Ola tok feil), the only variant of *ta* selected for analysis is the one stated in (2).

Back to the semantic part of (2), it is obvious that a plain numbering of verb senses inside of a monolingual grammar provides little basis for obtaining a multi-lingually interesting representation of meaning – the numbering even in its own enumeration is arbitrary, and since verbs are not shared between languages, there are not even sequences of numberings to compare. One rather will need an ontology of predicates, or situation types, abstractly defined, or through cross-linguistic networks of senses on the format '_ta_v_16_rel'. The lexical representation of the ta in question should then be on the form (3) rather than (2), where the SIT value *xyz* represents a relevant point to which _ta_v_16_rel corresponds in the situation type hierarchy.

(3) Ta ... [COMPS < [... HEAD noun [KEY feil]]>,
... [arg12-relation [PRED _ta_v_16_rel]],
... [SIT xyz]

This suggests that the main analytic resources for dealing with category 2 MWEs will be slightly off the deep grammar as such: as preprocessing for the syntactic part, and in an independent semantic situation network for the semantic part, both remaining to be developed. The syntactic string list will be particular to Norwegian, however the architecture envisaged has not been developed yet relative to any DELPH-IN grammar. A candidate for serving as a situation type network probably does not yet exist.

Relative to category 2, we have answered the question why, so far, there is so little done on this type of verbal MWEs in Norsource. A remaining question is: how can the grammar survive at all without a comprehensive treatment of these MWEs? The answer is that for these MWEs, a monolingual grammar does not have to recognize them, as long as syntactic parses are what count, together with a semantics not involving a mapping to a system independent of the language analyzed. In the case of category 1, such a mapping is at hand inside of the grammar in the form of a closed aspectual system, hence this aspect of MWEs has been implementable.

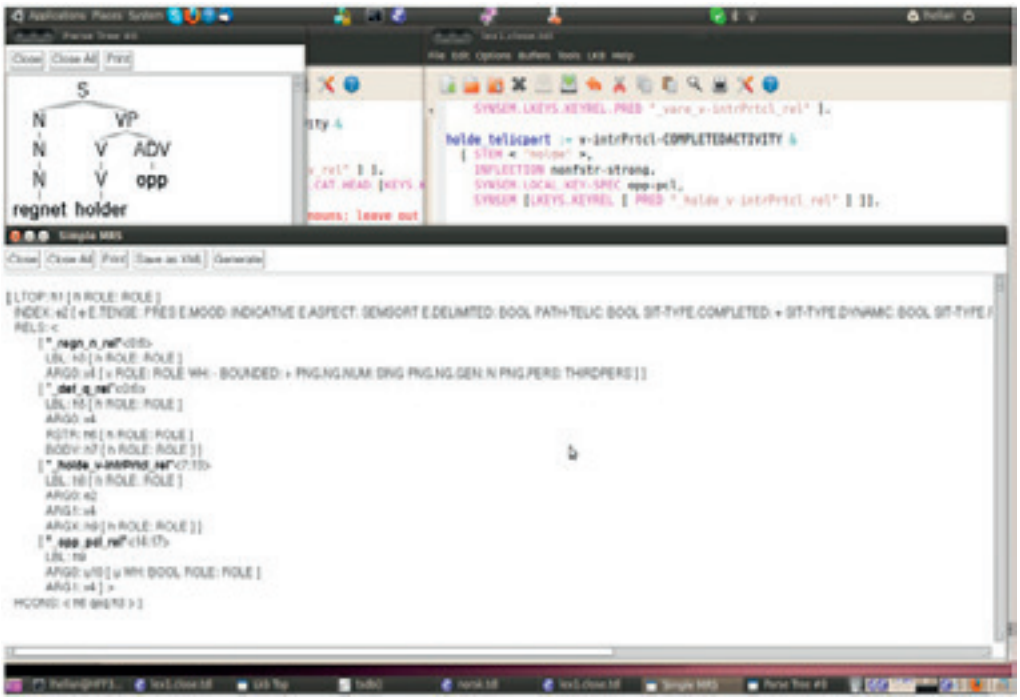

Fig 1. Category 1

[1] They also contrast with what one may call 'systematic abstract extensions' as in expressions like "think through the problem", where the preposition through is used with exactly the same spatial coordinates as in 'concrete' uses, only relative to an abstract space.

[2] There could of course be an instance where ta in the sense 'grasp' would be the intended item, in a described situation where someone grasps the wrong object. (For instance, in one of the Indiana Jones movies, the villain is realizing he drank from the wrong chalice; if uttering 'I took the wrong one'; a by-standing Norwegian could say, duly gleefully: "Du har rett. Du tok feil (kalk).")

[3] A drawback of this approach is that 'non-MWE' interpretations like exemplified in footnote 2 will not be recognized.

[4] A partial contribution done in the environment of the present grammar is however a multilingual valence repository, see http://typecraft.org/tc2wiki/Multilingual_Verb_Valence_Lexicon.

Links and references:

http://moin.delph-in.net/
http://typecraft.org/tc2wiki/Norwegian_HPSG_grammar_NorSource
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 Proceedings of the Workshop on Grammar Engineering and Evaluation, Coling 2002, Taipei.
Bender, E. M., Drellishak, S., Fokkens, A., Poulson, L. and Saleem, S. 2010. Grammar Customization. In Research on Language & Computation, Volume 8, Number 1, 23–72.
Copestake, A. 2002 Implementing Typed Feature Structure Grammars. CSLI Publications.
Copestake, A., D. Flickinger, I. Sag and C. Pollard. 2005. Minimal Recursion Semantics: an Introduction. Journal of Research on Language and Computation. 281–332.

NTNU – Trondheim
Norwegian University of Science and Technology