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

Lars Hellan, NTNU

The computational grammar Norsource at NTNU, Trondheim, is an HPSG grammar for Norwegian, using the LKB platform (Copestake 2003) 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 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 from the sentence Regnet holder opp ‘(the rain holds up = ‘the rain causes’) with the syntactic tree and semantic representation produced by the appropriate parse, and the lexical entry specification for the relevant use of the verb holder. Here the lexical type for the verb, v-intrPrtclCOMPACTIVITY, induces a structure with a subject and an adverbial particle as complement, and with the Aktionart type ‘completed activity’ being the part ‘SIT-HAS-COMPLETED’ in the top line of the MRS, and the entry itself specifies: the item filling the particle slot (as value of the attribute ‘KEY-SPEC’, an entry point for such information), and thereby being what induces the Aktionart 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 for-ests: 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 simply represented in the language; however, the grammar Norsource has only three cases encoded, just to formalize a procedure for their treatment. Why so few? We may use (1) as example:

(1) Du tar feil
You take wrong [v = 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 indicators’, so that in ‘Ta tar feil’, the word ‘ta’ would carry a different sense index than it does in ‘Jeg tar’ (I take hold). The standard way of assigning such marking in the MRS system is by defining PRE-value distinguished by integers, such as in ta_v_1_rel, ta_v_2_rel, ta_v_3_rel

... the ‘ta’ in ‘Ta tar feil’ could then for instance be number 16 such an inventory, and none of the semantic expectations going along with the other ‘ta’-variants ‘ta’_#1, 2, 3, ... 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:

... [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 _ta_v_16_rel refers will be when the verb has an object headed by the noun ‘feil’.

However, to avoid parse forest explosions for a sentence like ‘Man tar 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’. The problem is that with the mechanism men-

Fig 1. Category 1

2. Frames which, when a specific word is used in a given slot, induce a meaning combined.

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

... [arg12-relation [PRED _ta_v_16_rel]...]

They also contrast with what one may call ‘systematic abstract extensions’ as in expressions like ‘think through the problem’, where the proposition is used with exactly the same spatial coordinates as ‘concrete’ uses, only relative to an abstract space.

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 drunk 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 (kaik).’)

The problem is that with the mechanism men-

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 in-

volving 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 encompassed system, hence this aspect of MWEs has been implementable.

A drawback of this approach is that ‘non-MWE’ interpretations like exemplified in footnote 2 will not be recognized.

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.

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 numbersings 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’_#1, ‘ta’_#16_rel. The lexical representation of the ‘ta’ in question should then be in the form (3) rather than (2), where the SIT value corresponds to a relevant point to which ‘ta’_#16_rel corresponds in the situation type hierarchy.

(3) Ta  ... [COMPS < [... HEAD noun [KEY feil]...>]

[arg12-relation [PRED _ta_v_16_rel]]...

... [SIT ...]

This suggests that the main analytic resources for dealing with category 2 MWEs will be slightly off the deep grammar as such: as prepossessing 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,