Verb-headed Multi-Word Expressions in the Norwegian HPSG grammar *Norsource*

Lars Hellan, NTNU

PARSEME workshop March 10-11, 2014   Athens   WG2

The computational grammar *Norsource* maintained 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) and with an overall technical infrastructure as supported by the *DELPH-IN* network (http://moin.delph-in.net/). The following is a link to a wiki page about the grammar: http://typecraft.org/tc2wiki/Norwegian_HPSG_grammar_NorSource , which provides links to web demos for parsing done in the grammar, and to two applications carried by the grammar, one for grammar checking, http://typecraft.org/tc2wiki/Classroom:Norwegian_Grammar_Checking, and one one for a tri-lingual aligned valence database. http://typecraft.org/tc2wiki/Multilingual_Verb_Valence_Lexicon .

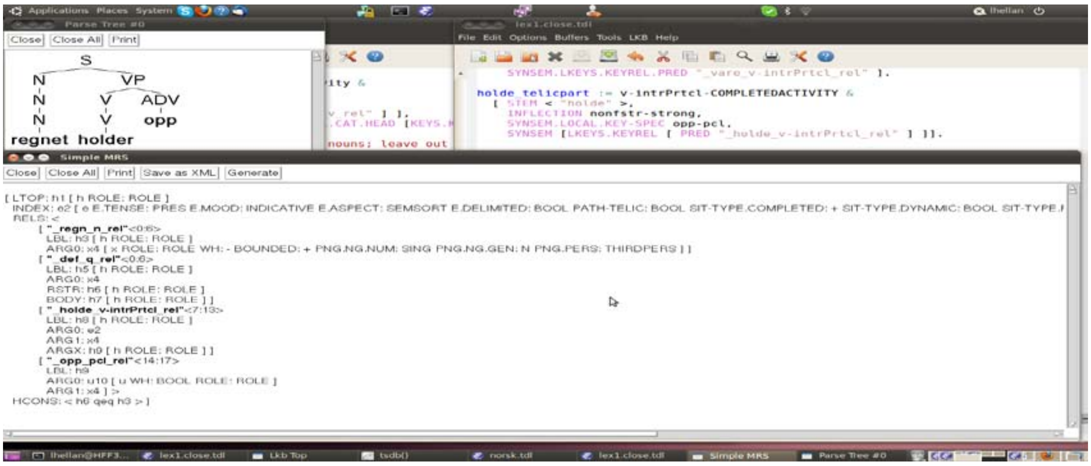
Its lexicon consists of more than 80,000 words, thereof about 50,000 nouns, and verbs and adjectives with about 12,500 entries each. As is common in the 'head driven' design, when a verb can occur in x many construction types differing in features reflecting argument structure, there will be x many entries of this verb each one 'programmed' for one specific construction frame (modulo notations for 'optionality' whereby a given entry can be a schema of more than one frame). To the extent that the grammar motivates rules/schemata of argument structure *derivation*, one can distinguish between basic entries and derived entries, where a set of derived entries may be derivationally related to a single basic entry. The 12,500 verb entries in the Norsource lexicon are basic entries in this sense, but no optionality marking is employed.

In this lexicon, entries of verb-headed MWEs ('Multi-Word Expressions') will consist of frames where specific lexical items are specified in the slots designed for dependent items. Thus, if there is an MWE where the verb occurs with a specific word as object, then that word is entered in the general slot designed for objects. There are no frames whose principal geometry is instantiated only by MWEs – frames for MWEs are always instantiations of general configurations.

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.

Cases of type 1 are interesting to the grammarian and semanticist: when aspect is involved, for instance, these will constitute cases of periphrastic aspect, of which one wants to have a general inventory for each language, and which pose certain challenges when it comes to representing the 'computing out' of the aspectual value from the items structurally involved. A screenshot from Norsource for the sentence *Regnet holder opp* ('the rain holds up' = 'the rain ceases') illustrates the syntactic tree and semantic representation (using MRS, cf. Copestake et al.) produced by the appropriate parse, and the lexical entry specification for the relevant use of the verb *holde*.


Applications Places System
lhellan
Parse Tree #0
Close Close All Print
S
N
VP
N
V
ADV
N
V
opp
regnet holder
lex1.close.tdl
File Edit Options Buffers Tools LKB Help
Simple MRS
Close Close All Print Save as XML Generate
lhellan@HFF3... lex1.close.tdl Lkb Top tsdb() norsk.tdl lex1.close.tdl Simple MRS Parse Tree #0

```
ity &

v_rel" ] ].
.CAT.HEAD [KEYS.K

nouns; leave out
```

```
    SYNSEM.LKEYS.KEYREL.PRED "_vare_v-intrPrtcl_rel" ].

holde_telicpart := v-intrPrtcl-COMPLETEDACTIVITY &
  [ STEM < "holde" >,
    INFLECTION nonfstr-strong,
    SYNSEM.LOCAL.KEY-SPEC opp-pcl,
    SYNSEM [LKEYS.KEYREL [ PRED "_holde_v-intrPrtcl_rel" ] ]].
```

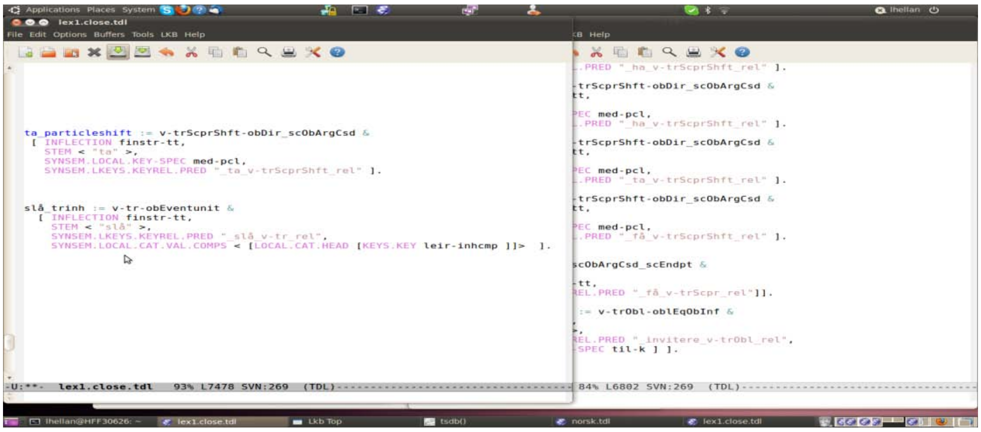
```
[ LTOP: h1 [ h ROLE: ROLE ]
  INDEX: e2 [ e E.TENSE: PRES E.MOOD: INDICATIVE E.ASPECT: SEMSORT E.DELIMITED: BOOL PATH-TELIC: BOOL SIT-TYPE.COMPLETED: + SIT-TYPE.DYNAMIC: BOOL SIT-TYPE.I
  RELS: <
        [ "_regn_n_rel"<0:6>
          LBL: h3 [ h ROLE: ROLE ]
          ARG0: x4 [ x ROLE: ROLE WH: - BOUNDED: + PNG.NG.NUM: SING PNG.NG.GEN: N PNG.PERS: THIRDPERS ] ]
        [ "_def_q_rel"<0:6>
          LBL: h5 [ h ROLE: ROLE ]
          ARG0: x4
          RSTR: h6 [ h ROLE: ROLE ]
          BODY: h7 [ h ROLE: ROLE ] ]
        [ "_holde_v-intrPrtcl_rel"<7:13>
          LBL: h8 [ h ROLE: ROLE ]
          ARG0: e2
          ARG1: x4
          ARGX: h9 [ h ROLE: ROLE ] ]
        [ "_opp_pcl_rel"<14:17>
          LBL: h9
          ARG0: u10 [ u WH: BOOL ROLE: ROLE ]
          ARG1: x4 ] >
  HCONS: < h6 qeq h3 > ]
```

Fig 1. Case 1

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’, 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. The grammar has about five cases of this type so far, although the number can easily reach 20 or 30 when one covers the phenomenon systematically – so far, the goal has only been to establish the principal coding procedure for such cases, and develop a semantic type system capable of holding the assumed full range of distinctions (partly reflected in the top line of the MRS specification).

In numerical contrast, Norsource has nearly 1800 entries of case 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 ‘MWE’ entry could be wrongly activated for these if the adverb or preposition is not specified. A sequence like *ta med*, rendered in its entry code format in fig. 2 below, exemplifies such a case.

Case 2, finally, is here illustrated by the expression *slå leir* (lit. ‘beat camp’ = ‘make camp’) in fig. 2, where a semantic assignment would signal that the meaning is not what *slå* and *leir* would yield in their more productive uses. The language is full of such combinations; however, the grammar has only this case and two others encoded, again just to secure a formal procedure for their treatment. Why so few? Because in the type of HPSG grammars in question, the semantic ‘PRED’ value assigned to a given word is essentially that word itself, without any mapping to, e.g., English predicates (as done, e.g., in some LFG grammars); the need for marking the special meaning of *slå leir* arises only when in the context of an MT system, or in an ontology mapping, or any other application where meaning has to be mapped to a meaning representation system independent of the language in question.



```
ta_particleshift := v-trScprShft-obDir_scObArgCsd &
 [ INFLECTION finstr-tt,
   STEM < "ta" >,
   SYNSEM.LOCAL.KEY-SPEC med-pcl,
   SYNSEM.LKEYS.KEYREL.PRED "_ta_v-trScprShft_rel" ].


slå_trinh := v-tr-obEventunit &
  [ INFLECTION finstr-tt,
    STEM < "slå" >,
    SYNSEM.LKEYS.KEYREL.PRED "_slå_v-tr_rel",
    SYNSEM.LOCAL.CAT.VAL.COMPS < [LOCAL.CAT.HEAD [KEYS.KEY leir-inhcmp ]]>  ].
```

```
-U:**-  lex1.close.tdl    93% L7478 SVN:269   (TDL)-----------------------------
```

```
L.PRED "_ha_v-trScprShft_rel" ].

-trScprShft-obDir_scObArgCsd &
tt,

PEC med-pcl,
L.PRED "_ha_v-trScprShft_rel" ].

-trScprShft-obDir_scObArgCsd &
tt,

PEC med-pcl,
L.PRED "_ta_v-trScprShft_rel" ].

-trScprShft-obDir_scObArgCsd &
tt,

PEC med-pcl,
L.PRED "_få_v-trScprShft_rel" ].


scObArgCsd_scEndpt &

-tt,
REL.PRED "_få_v-trScpr_rel"]].

 := v-trObl-oblEqObInf &
,
>,
REL.PRED "_invitere_v-trObl_rel",
SPEC til-k ] ].
```

```
84% L6802 SVN:269   (TDL)-----------------------------
```

Fig 2. Case 2 and 3

Conclusion:

Case 2 is obviously quite central in the study of MWEs, however for ‘deep’ grammars with ‘language immanent’ semantics like the DELPH-IN grammars, capturing their ‘non-compositional’ semantics inside of the grammar itself does not have a place. One could introduce another layer of semantic specification for use of English equivalents, or construct a language-independent space of semantic units to which meanings could in principle be mapped, and then specify such a mapping for every MWE of type 2 in the particular grammar. Both would be programs of action, the latter even for quite a bit of independent research. In the meanwhile, these grammars will address MWEs ‘on an ad hoc basis’, like in MT and similar activities through direct mapping of MWEs of one language to compositional expressions or MWEs of the other language.

http://moin.delph-in.net/   Copestake, A. 2002 Implementing Typed Feature Structure Grammars. CSLI Publications.