“Under the word belt”

Lars Hellan

NTNU, Trondheim, Norway

Research Group in Digital Linguistics

COST Action IC1207 Meeting,

Warsaw, September 17, 2013

- An intriguing aspect of some MWEs
- Grimness of parsing them
- Two prospects of doing semantics on them
- A possible approach to multilingual MWE studies

Background:
http://typecraft.org/tc2wiki/Norwegian_HPSG_grammar_NorSource

Expressions like (1) presumably mean the same:

(1)
“You are wrong” (English)
“Du tar feil” (Norwegian) (literally: ‘you take wrong’)
“Tu te trompes” (French) (literally ‘you wrong yourself’)

They establish their content by means of word combinations that in some sense of ‘literal meaning’ compose the content in
quite different ways, and for none of these ways can one say that one is more or less ‘literal’ or ‘figurative’ than any of the others.
They are as ‘direct’, and in a Saussurean sense ‘arbitrary’,
as word level entities normally are,
and yet they are composed of more than one word.

They contrast with commonly quoted locutions like ‘kick the bucket’, which, although ‘kick’ and ‘bucket’ in no way compose to convey the content ‘die’, is a fully compositional expression expressing a situational image which, by convention of ‘preserved metaphor’, is used as label of a specific situation type.

There also happens to be a word “die” in this same language, naming this same meaning and counting as the official way of expressing it. Not so for the expressions in (1).

The expressions in (1) are like lexical units in being ‘minimal’ and ‘basic’ carriers of the meaning in question. But they are composed by syntactic units through recognized rules of composition, supposed to create the full meaning from the meanings of the parts; only not so here.

We discuss the case relative to MRS (Copestake et al. 2005).

Minimal Recursion Semantics (Copestake et al. 2005)

$$
\begin{bmatrix}
\text{LTOP} : \text{h1} \\
\text{INDEX e2}\left[\text{E}\left[\text{TENSE} : \text{PRES}\right]\right] \\
\\
\text{RELS} : \left\langle
\begin{bmatrix} \text{\_def\_q\_rel} \\ \text{LBL h5} \\ \text{ARG0 x4} \\ \text{RSTR h6} \\ \text{BODY h7} \end{bmatrix},
\begin{bmatrix} \text{\_boy\_n\_rel} \\ \text{LBL h3} \\ \text{ARG0 x4} \end{bmatrix},
\begin{bmatrix} \text{\_throw\_v\_rel} \\ \text{LBL h8} \\ \text{ARG0 e2} \\ \text{ARG1 x4} \\ \text{ARG2 x9} \end{bmatrix},
\begin{bmatrix} \text{\_def\_q\_rel} \\ \text{LBL h11} \\ \text{ARG0 x9} \\ \text{RSTR h12} \\ \text{BODY h13} \end{bmatrix},
\begin{bmatrix} \text{\_ball\_n\_rel} \\ \text{LBL h10} \\ \text{ARG0 x9} \end{bmatrix}
\right\rangle \\
\\
\text{HCONS} : \left\langle \text{h6 QEQ h3, h12 QEQ h10} \right\rangle
\end{bmatrix}
$$

In MRS, each word of the input string is represented as an ‘Elementary Predication’ (EP), normally using the word itself as part of the predicate name, and thus as identifier of the meaning.

One of course will want an MRS representation to be not just a mechanical replica of the syntactic words used. So for one thing, in the analysis of the cases in (1), one will want to somehow represent the circumstance that the full meaning of any of the examples in (1) is not compositionally reducible to other ‘standard’ uses of the same words.

A standard mechanism one could use is to mark words with ‘sense indices’ consistently 1-to-1-related to their meanings, 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
_ta_v_4_rel

The “ta” in “du tar feil” could then for instance be number *16* in such an inventory, and one would know that 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 which imply taking possession or control over something. This is one way of tackling the issue of ‘non-compositionality’.

(Rather than integers, for the sake of mnemonics, one could in this case write ‘_ta_v_feil_rel’ rather than ‘_ta_v_16_rel’, if one can assume that this sense of “ta” is tied uniquely to the appearance of this noun. A mix of the strategies would go.)

To avoid parse-forest explosions, each such “ta”-variant must be contextually restricted so that at most two or three variants (including the right one) turn up in a parse result. Our ‘_ta_v_16_rel’, for instance, would have a COMPS list consisting of an NP required to be headed by “feil” (perhaps also to be a ‘bare’ singular).
In many strings of the form ‘... ta NP ...’ this would be checked easily enough, but for any construction where a gap could be hypothesized ‘following’ “ta”, one’s gap-filling/binding rules will need to be carefully defined such that whatever information is used in the COMPS-list for the NP must be compatible with information present in whatever item is ‘filler’-related to the gap.

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 shared 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.

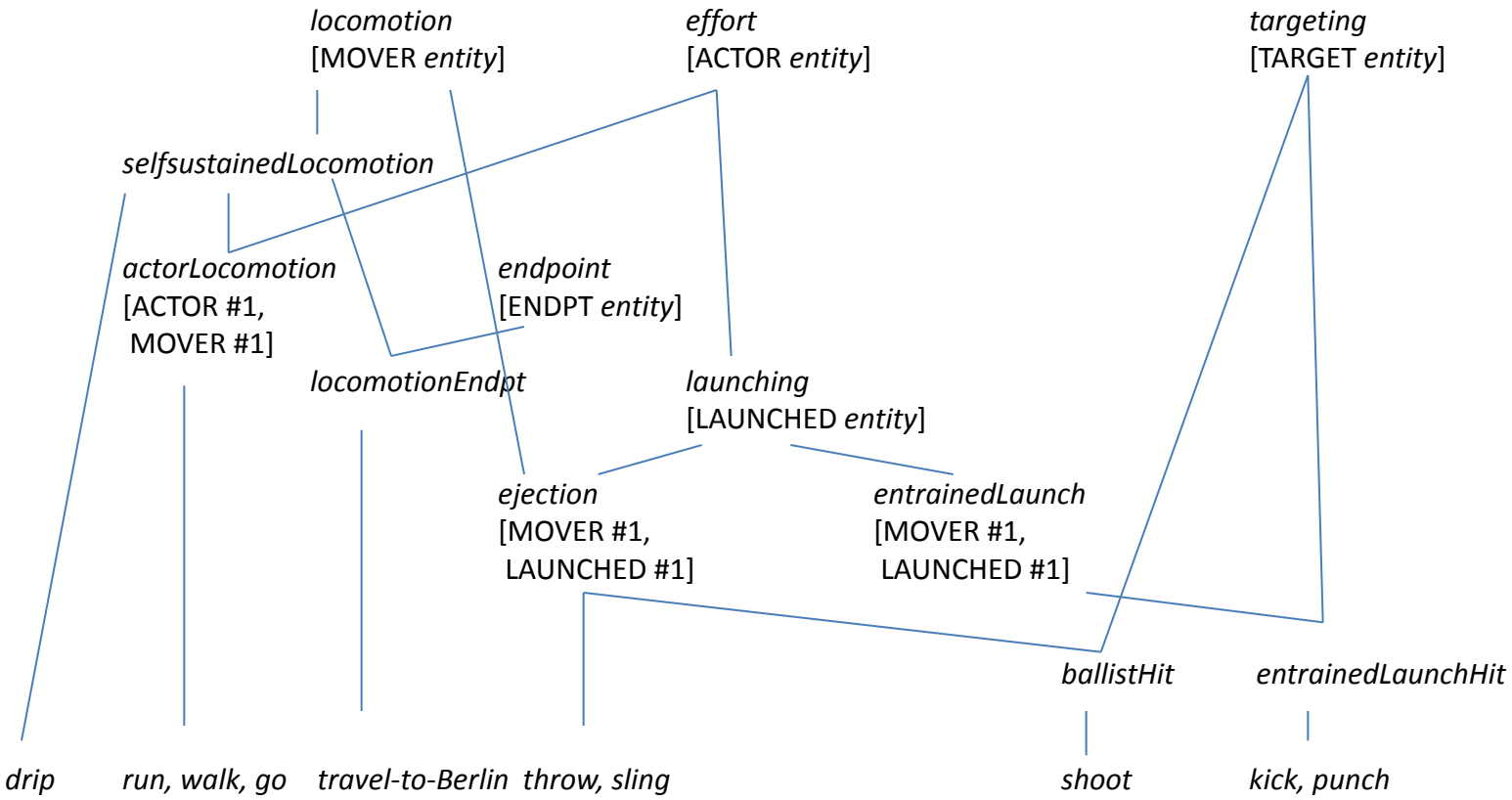
In the MRS-driven framework, multilingual applications are mainly limited to MT for language pairs, where pairings are defined predicate by predicate in so-called **transfer rules**. Schematically, for cases like (1), a rule would be essentially as in (2), for Norwegian to English (’47’ being an arbitrary guess just as ’16’):

(2) arg12-relation [PRED _ta_v_16_rel] =>
arg?-relation [PRED _be_v_47_rel

Using transfer rules may be called a *procedural* way of representing cross-linguistic sameness of meaning. One never gets that supposed *same* meaning represented as such, one is only lead between expressions that actually (or supposedly) mean the same.

To indeed establish a representation of the *common meaning* of cases like (1), one would have to construct a point in some semantic space representing this exact meaning.
This would have to be in an ontology of predicates, or situation types. Node names in such an ontology would probably be English, but without any pretence that these are the ‘actual’ meaning identifiers. Below is an example of how this might look:

Excerpt of a possible situation-type hierarchy



Creating such a typology would be a challenge. But supposing it can be done, how would it become relevant in an approach as sketched above?
Instead of an entry of “ta” for the example in (1) schematically as in (3), one could consider an entry as in (4),

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

(4) Ta ... [COMPS < [... HEAD noun [KEY feil]]>,
... [xyz-relation [PRED _uvw_rel]]

where *xyz* and *uvw* represent points in a hierarchy like above, and are shared between the specifications for the verbs in (1).

However, it is an essential strategy in MRS representations to preserve the links to grammatical information as far as possible into the semantic analysis, which will include PRED-values. Thus, rather than (4), we would need something in the style of (5), retaining (3) but bringing in situation-type as an additional layer of representation,

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

where the SIT value *xyz* still represents the relevant point in the situation type hierarchy.
One can then in practice combine the transfer-rule strategy with involvement of situation types, as these start taking form.

One potential general advantage of establishing a situation-type based inventory/ontology of construction types is to enable a more 'object-driven' approach to multilingual analysis, including MT.
A substructure of such a construction inventory can be a **Valence inventory**, which can be partially populated from lexically driven grammars like the LKB grammars.
An example:
http://regdili.idi.ntnu.no:8080/multilanguage_valence_demo/multivalence
Cases like those in (1) will be no exception:
from representations in the style of (5), even these very specific frames will have a place in an over-all multilingual valence inventory, with alignment via situation types to the way other languages express each meaning; a broader typology of expressions like those in (1) can then be entertained – and perhaps more readily than if based on a network of transfer-rule pairs.