Lexical Ambiguity in Machine Translation: Using Frame Semantics for Expressing Systemacies in Polysemy

Bolette Sandford Pedersen
Center for Sprogteknologi
Njalsgade 80
DK-2300 Copenhagen S

Abstract

One of the most pervasive problems in machine translation is that of lexical ambiguity. In this paper we suggest to apply a Frame Semantic approach to the construction of the Danish monolingual dictionary used in the analysis phase. We show how such an approach - implemented in an HPSG-like formalism - allows us to express some of the systemacies in polysemy and thereby avoid the much criticized *sense enumeration* techniques traditionally seen in machine translation, where there is no distinction between unrelated meanings (homonomy) and related meanings (polysemy). Our case study for this experiment is that of Danish motion verbs which are in broad terms semantically very ambiguous and rather difficult to classify at hand. We show on the basis of corpus examples how these verbs can be grouped into four different kinds of polysemy, namely *meaning extension, meaning reduction, co-composition*, and *meaning focussing* . Having organised meaning changes into these categories, we can now generalise over these by means of a set of lexical rules and thus incorporate in the dictionary some of the productive mechanisms in language which are relevant for translation.

1 Lexical disambiguation in machine translation

Along with more and more refined parsing strategies, more computational power, and more robust mechanisms to account for unforeseen grammatical and lexical constellations, one of the remaining problems in machine translation persists to be the disambiguation of words, and as a consequence of this: the problem of selecting the correct translations in the target language.

In many operating translation systems, the problem of lexical ambiguity is either more or less ignored or mainly treated by means of *domain typing*. This obviously leads to incorrect translations in several cases, which are then to be corrected by a human translator in the postediting phase. However, since commercial machine translation systems are often tuned towards - or especially designed for - specific subdomains, a large number of the words that occur in the relevant input texts can be classified as domain specific terms and can thereby be made unambiguous. To give an example of such a domain specific sense of a word, consider the Danish motion verb *rulle* in the following example taken from a Danish word processing manual:

> *brug piletasterne for at rulle gennem brevet*
> (apply the arrow keys to scroll through the letter)

The most general translation of *rulle* would definitely be *roll*, however, in a context of computer manuals, this translation is overridden by the domain-specific term *scroll*.

However, when it comes to the general vocabulary (to which the majority of motion verbs belongs) disambiguation by means of subdomains is generally insufficient. Therefore, *semantic typing* on nouns is applied in many advanced machine translation systems; a typing which enables the dictionary coder to express *selectional restrictions* on valency-bearing words. The two examples below illustrate how (part of) the ambiguity of a Danish motion verb like *gå* (walk) can be described on the basis of the semantic characteristics of the first argument:

> *når julefrokosten har været så våd, at man*(HUMAN)*ikke kan gå*
> (when the Christmas dinner has been so

wet, that one(HUMAN) cannot walk)

når der er gået 7 dage(TIME)*efter at den sidste pille blev taget, begyndes på en ny pakke*
(when 7 days(TIME) have passed by since the last pill was taken a new package is opened)

Although we can now presumably overcome some aspects of verb amiguity, lexical ambiguities enforced by other contextual factors in the text still remains to be accounted for. In the following example a directional modifier has been added to the verb *gå* changing the focus from the manner component (to put one foot in front of the other fairly slowly on the ground) over to the direction component which is not inherent in the verb.

han gik pludselig ned i bådens agterende
(he suddenly walked/went down to the stern of the boat)

Such a change in focus is also relevant for translation since a change from manner to direction triggers different lexical choices in many languages. In English, for example, we can now choose between *walk* and *go*; in a Romance language, like Spanish, we would choose a completely different verb; namely a pure direction verb such as *bajar* (go down) and leave the manner component either unexpressed or to be specified as a modifier (see Talmy 1985):

De repente bajó (caminando) a la popa del barco
(he suddenly went down (walking) to the stern of the boat)

Although it may be argued that several ambiguities cannot be solved unless a 'deep' meaning analysis is performed, including maybe even extralinguistic knowledge, we in this paper suggest to apply a Frame Semantic approach in order to capture some basic semantic aspects relevant for translation and still come up with an *operationally feasible* solution.

In the following we shall first look at some influental approaches to lexical ambiguity that have inspired this work; then we shall look at some basic characteristics of Danish motion verbs seen in the light of a taxonomic analysis based on Frame Semantics. Having defined the basic apparatus, we procede to a corpus-based division of polysemous motion verbs into four different kinds of meaning changes, namely *meaning extension, meaning reduction, co-composition*, and *meaning focussing*. Finally, we discuss some exceptions to productive meaning changes and establish a blocking mechanism which we claim cannot be generally predicted but must be assigned on an idiosyncratic basis.

2 Some influental approaches to lexical ambiguity

Traditionally, we refer to two kinds of lexical ambiguity: *homonomy* and *polysemy* (Lyons (1977:p.550-569). A standard criterion for the division into homonomy and polysemy is that of *etymology*; if two senses are historically derived from the same word then they can be considered to be polysemous, otherwise we must be dealing with homonomy. Another central criterion is *relatedness in meaning*: if two senses are related in meaning then they can be considered polysemous; otherwise they are homographs.

Although these criteria may not always be fully operational, they are convenient since they allow us to distinguish between ambiguity that should be expressed by pure sense enumeration in the dictionary (homographs) and ambiguity which is a candidate for a more general treatment in terms of generative sense producing devices.

Cruse (1991:p.49-83) adopts an intermediate position between those linguists that believe that a word form is associated with a number of discrete senses and those (like Lyons) who believe that such discreteness is fundamentally illusory. Cruse defines different kinds of ambiguity, namely *sense modulation* and *sense selection*. Sense modulation cannot, according to Cruse, be expressed in discrete 'jumps', since every context in essence modulates its own 'sense' or interpretation of a word. From a lexical point of view it is therefore clearly inconvenient to try and describe such modulations as individual senses, partly for theoretical reasons, partly for practical reasons since the dictionary would explode with redundancies. In opposition to sense modulation, sense selection can, according to Cruse, to some extent be described in terms of discrete senses which can be verified by means of *distributional tests* of different kinds.

In the line of Cruse, recent studies in linguistics and lexicography have resulted in the development of several new approaches in order to find more adequate ways of defining polysemous words than that of pure enumeration techniques. Pustejovsky (1995) designs a Generative Lexicon and treats systematic polysemy by means of a set of generative devices which operate on core lexical entries containing different levels of semantic information, the central level being the Qualia Structure.

Copestake & Briscoe (1995) follow a similar line but distinguish between vagueness and ambiguity, or, in their terminology, constructional polysemy where a single sense assigned to a lexical entry is contextually specialised and thus is seen as a specialisation of an underspecified lexical entry, and sense extension which predictably relate two or more senses by means of lexical rules (this roughly corresponds to Cruse's division).

Atkins & Fillmore (1994) and Atkins (1994) view the problem rather from the traditional lexicographer's point of view and suggest a Corpus-based Frame Semantic approach to the notion of polysemy leaving the traditional 'dictionary sense' aside in favour of what they define as *situation types* described by means of frame elements. Corpus-based in its approach, an explanation of the relationship between elements of the conceptual frame and their realisations within the linguistic structures constitute an important ingredient of Frame Semantics.

3 Danish motion verbs seen from a Frame Semantic perspective

In Frame Semantics, a taxonomic approach to verbs is taken, dividing verbs into semantic classes and subclasses so that words that can be considered semantic neighbours with similar sets of semantic parameters (henceforth called *frame elements*) are treated coherently. In some variants of Frame Semantics (see Alonge et al. 1992), frame elements can be compared with semantic roles since they consist of semantic labels assigned to arguments or modifiers in the surrounding context; however, contrary to traditional semantic roles they are specific for each frame: a *Perception Frame*, thus, has another set of frame elements than a *Move Frame* and a *Speech Act Frame*.

For the purpose of machine translation we prefer to adopt a less restricted approach to frame elements, as described in Atkins & Fillmore (1994) and Fillmore (1985). In these works, frame elements can also be established across constituents in the context, i.e. on the basis of *inherent* semantic features of the verb. This means that frame elements can go beyond lexicalisation patterns and this is essential for contrastive purposes where a syntactic constituent in one language may trigger a completely different sentence structure in another, as was seen in the Danish to Spanish translation example above.

Following this line, we can establish a set of core frame elements for Danish motion events which go beyond the semantic role assignment but which on the other hand restricts itself to a minimum in order to be as operational as possible in the sense identification process (for a more detailed description, see Pedersen 1997):

FRAME: Motion
FRAME ELEMENTS:
Mover (Mo)
Direction (Di)
Manner (Ma)
Vehicle (Ve)
Protagonist Control (Procon)

In figure 1, a taxomony of Danish motion verbs is given which expresses the prototypical assignment of these frame elements[1].

4 Types of Polysemy with Danish motion verbs

Following the line of Cruse (1991) and Copestake & Briscoe (1995) we wish to distinguish between *sense modulation* and *sense extension*, since this distinction is crucial for the establishment of senses in the dictionary. Our working hypothesis is that the minimal set of frame elements that we have established for motion verbs, together with a set of distributional tests, can be applied for the identification of these two kinds of sense changes. In other words, if a new sense in the corpus evokes another set of frame elements than the one evoked prototypically (i.e. on the basis of the taxomony) then we believe to be dealing with sense selection which should also be reflected in distribution.

An examination of the corpus material available for this group of Danish verbs [2], allows us to further divide into different categories of deviation from the prototypical sense of the verbs. These deviations prove to be highly systematic in the sense that verbs belonging to the same subgroup tend to be sensitive to the same changes in meaning (but there are exceptions to this as will be illustrated below). We have subdivided these derivations into four kinds of systematic meaning shift:

1. Meaning extension, where a new semantic dimension is added to the already existing meaning components.
2. Meaning reduction, where the prototypical meaning component of a verb is reduced to only part of that meaning.

[1] Note that Manner and Direction can be subdivided almost infinitely; the taxonomy partly illustrates this, although distinguishing frame elements are only assigned to a certain depth.

[2] The corpus material used is taken from the Bergenholtz DK87-90 corpus (4 mill. words) and from the corpus of Den Danske Ordbog (40 mill. words).

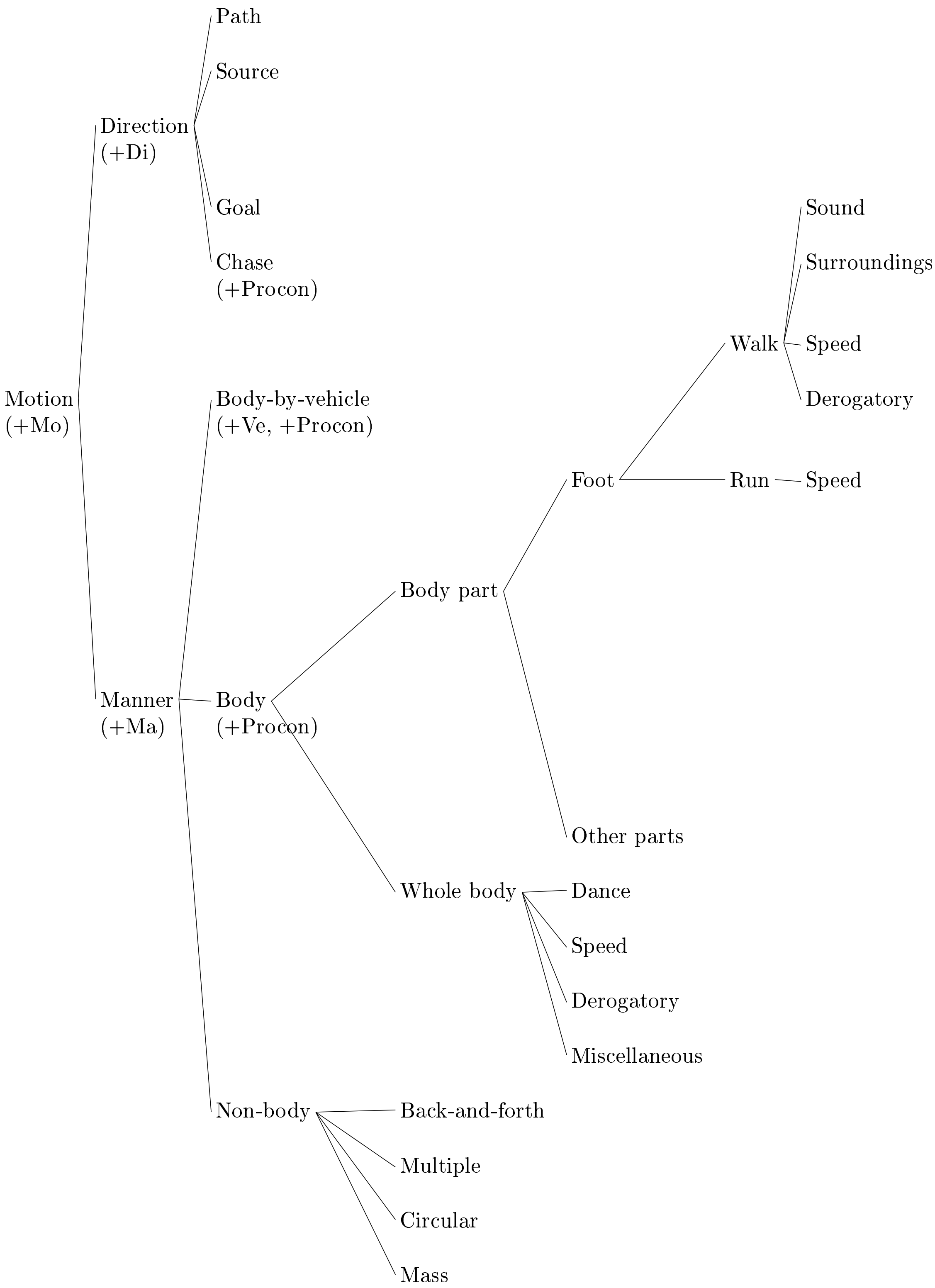

Figure 1: Taxonomy of Danish motion verbs

3. Co-composition, where a new related meaning is derived in composition with a non-prototypical argument to the verb.

4. Meaning focusing, where focus is added to one specific meaning component of the verb frame.

Examples of each meaning shift are given in figures 2-5 below.

If we revise the frame element assignment in figure 2-5 it becomes clear that meaning focussing differs from the rest since there is no deviation in frame elements between the two senses described. This leads to the conclusion that we must here be dealing with sense modulation, whereas the other three examples represent cases of sense selection.

5 Lexical rules as productive device

It has often been argued that lexical rules raise a problem of theoretical consistency since they permit an actual violation of the type assignments. However, these and other equal devices provide a handy way of treating systematic polysemy since they help us avoid actual sense enumeration in the lexicon.

With deviation in frame elements as our semantic constraint, lexical rules should be applied to meaning extension, co-composition and meaning reduction, whereas meaning focusing does not satisfy our constraint and should therefore rather be treated as a specific usage of the prototypical sense.

In the following we shall therefore leave meaning focusing aside and only concentrate on the three former kinds of sense change.

For the lexical entries, we apply an HPSG-like formalism (Pollard & Sag 1994) and add frame elements as features to a *FEG* subsign which stands for the Frame Element Group. This is again a subsign to the CONTENT part of the lexical sign, which typically contains semantic information in HPSG. Frame elements that correspond to an argument or a subcategorised adjunct are structure shared with these in the SUBJ or COMPLS lists; whereas frame elements that refer to inherent features are assigned atomic values.

A lexical rule for the systematic sense extension described in the previous section can now be expressed as depicted in figure 6.

The rule takes all lexical entries of type tbody that unify with the input part of the rule (I) and generates an output lexical entry (O) which has an extended GO-meaning. Note that the Manner element must be present in the input sign but that it is unrestricted wrt. its value. This enables us to generalise over all kinds of tbody verbs.

A lexical rule for co-composition with animal verbs is depicted in figure 7.

This rule covers animal verbs like *flyve* (fly) which changes meaning when combined with a human argument, involving then a vehicle. Similar changes of co-composition are seen with verbs that have the ability to express induced action when combined with a human Mover, such as for instance *køre* (run/drive), where we can say *bilen kører* (the car runs), but also *hun kører* (she drives) meaning that both she and the car move.

6 Shortcomings of a lexical rule approach to systematic polysemy

Applying a lexical rule approach to systematic polysemy may solve our problem of sense enumeration but it also causes us another severe problem because of the strong generative nature of this device: it introduces the problem of overgeneration. When we talk of semi-productive mechanisms in polysemy we indicate exactly the fact that there are always exceptions to these.

To take an example, let us look at the co-composition based meaning shifts that certain non-body verbs undergo when they combine with a human mover. In the cases of *daske* (flap) and *drive* (drift), the human mover obtains protagonist control of the movement, even if he moves without aim or goal. Other verbs from this group do not assign a frame element of protagonist control when combined with a human mover and should therefore not, according to our criteria, be treated as a new sense. An example of this is *dingle* (dangle) in

> *Fra lygtepælene dinglede tyve, grever og embedsmænd side om side.*
> (lit. from the lamp posts dangled thieves, counts, and officials side by side)

Here the human Mover is dead and definitely not controlling the movement.

In such cases we would not want to trigger a lexical rule because we are not dealing with a new sense. One solution to this is to elaborate a blocking mechanism that can be assigned to lexemes that cannot undergo default sense change of a given type. We do not at the current stage of our investigations see a clear pattern of preemption in relation to Danish motion verbs and although further research is called for on the topic, we are tempted to conclude that many cases of blocking can only be explained on the basis of etymological and social conventions. For the moment, we therefore establish an idiosyncratic

Prototypical sense	Extended sense
Class description: to move your body in a certain manner	Class description: to go in a certain manner
Frame elements: +Mover (animated) +Manner, +Protagonist control	Frame elements: +Mover (animated) +Manner, +Protagonist control +Direction
Example: *han kunne gå da han var 10 måneder gammel* (he could walk when he was ten months old)	Example: *han gik hen til skranken* (he went to the counter)
Syntactic constraint: none	Syntactic constraint: directional modifier
Aktionsart: imperfective	Aktionsart: perfective

Subgroups from the taxonomy sensitive to this permutation are body verbs, such as *liste* (tiptoe) and *svømme* (swim), body-by-vehicle verbs, such as *cykle* (cycle), and some non-body verbs, such as *trille* (roll).

Figure 2: Meaning extension

Prototypical sense	Reduced sense
Class description: to move iteratively up and down or back and forth	Class description: to move suddenly up or down, forwards or backwards in a single movement
Frame elements: Mover (mostly non-human) + Manner	Frame elements: Mover (mostly non-human) + Manner + Direction
Example: *han lod salmiakstangen vippe i munden* (he made the stick of liquorice flip up and down in his mouth)	Example: *han lod den umagelige højryggede stol vippe baglæns* (he let the uncomfortable high-backed chair tip backwards)
Syntactic constraint: none	Syntactic constraint: directional particle - phrasal verb construction
Aktionsart: imperfective	Aktionsart: perfective

Subgroups from the taxonomy sensitive to this permutation are back-and-forth verbs, such as *svaje* (sway), and *svinge* (swing).

Figure 3: Meaning reduction

Prototypical sense	Co-composisional sense
Class description: non-human entities moving in a certain manner without protagonist control	Class description: humans moving about casually without aim or goal
Frame elements: Mover (non human) + Manner	Frame elements: Mover (human) + Manner + Protagonist control
Example: *de tørrede gedder havde hængt og dasket vinteren over* (the dried pikes had been dangling the whole winter) *sodede skyer driver afsted under ildhimlen* (sooty clouds drift away under the burning sky)	Example: *jeg havde god tid til at daske rundt i byen* (I had time enough to saunter around in town) *han gik og drev ovre hos dem der sad på flæsket* (he hanged around with the top guys)
Selectional constraint: Mover is non-human	Selectional constraint: Mover is human

Subgroups from the taxonomy sensitive to this permutation are several non-body verbs.

Figure 4: Co-composition

Prototypical sense	Focused sense
Definition: to move about in large groups	Definition: to move (or be) in large groups with focus on location
Frame elements: Mover (no restriction) + Manner (+ Location)	Frame elements: Mover (no restriction) + Manner + Location
Example: *vagtler, agerhøns og vilde kaniner vrimlede mellem stenene* (quails, partridges, and wild rabbits were running around between the stones)	Example: *Næsgaards jorder kommer atter til at vrimle med blå spejdere* (the fields of Næsgaard will once again be teeming with blue scouts)
Syntactic constraint: none	Syntactic constraint: location is mapped to subject

Subgroups from the taxonomy sensitive to this permutation are multiple verbs, such as *myldre* (crawl/teem), *sværme* (swarm/teem), and *vrimle* (crawl/teem)).

Figure 5: Meaning focussing

$$
{}_{\mathit{tMOVEGOlexicalrule}}\begin{bmatrix}
\mathrm{O} & \begin{bmatrix} \mathrm{SYNSEM}\,|\,\mathrm{LOCAL} & \begin{bmatrix}
\mathrm{CAT} & \begin{bmatrix} \mathrm{SUBJ}\ \langle \mathrm{NP}\,\mathit{animate}\boxed{1}\rangle \\ \mathrm{COMPS}\ \langle \mathrm{PP}\,\mathit{til}\boxed{2}\rangle \end{bmatrix} \\
\mathrm{CONT} & \begin{bmatrix}
\mathrm{PSOA} \begin{bmatrix} \mathrm{ARG1} & \boxed{1} \\ \mathrm{ARG2} & \boxed{2} \end{bmatrix} \\
\begin{bmatrix} \mathrm{FEG} & {}_{\mathit{tbody}}\begin{bmatrix} \mathrm{MOVER}\ \boxed{1} & \\ \mathrm{MANNER} & \_ \\ \mathrm{DIRECTION} & \boxed{2} \end{bmatrix} \end{bmatrix}
\end{bmatrix}
\end{bmatrix} \end{bmatrix} \\
\mathrm{I} & \begin{bmatrix} \mathrm{SYNSEM}\,|\,\mathrm{LOCAL} & \begin{bmatrix}
\mathrm{CAT} & \begin{bmatrix} \mathrm{SUBJ}\ \langle \mathrm{NP}\,\mathit{animate}\boxed{1}\rangle \\ \mathrm{COMPS}\ \langle\rangle \end{bmatrix} \\
\mathrm{CONT} & \begin{bmatrix}
\mathrm{PSOA} \begin{bmatrix} \mathrm{ARG1} & \boxed{1} \end{bmatrix} \\
\begin{bmatrix} \mathrm{FEG} & {}_{\mathit{tbody}}\begin{bmatrix} \mathrm{MOVER}\ \boxed{1} & \\ \mathrm{MANNER} & \_ \end{bmatrix} \end{bmatrix}
\end{bmatrix}
\end{bmatrix} \end{bmatrix}
\end{bmatrix}
$$

Figure 6: Lexical rule for meaning extension

$$
{}_{\mathit{tCOCOMPlexicalrule}}\begin{bmatrix}
\mathrm{O} & \begin{bmatrix} \mathrm{SYNSEM}\,|\,\mathrm{LOCAL} & \begin{bmatrix}
\mathrm{CAT} & \begin{bmatrix} \mathrm{SUBJ}\ \langle \mathrm{NP}\,\mathit{human}\boxed{1}\rangle \\ \mathrm{COMPS}\ \langle\rangle \end{bmatrix} \\
\mathrm{CONT} & \begin{bmatrix}
\mathrm{PSOA} \begin{bmatrix} \mathrm{ARG1} & \boxed{1} \end{bmatrix} \\
\begin{bmatrix} \mathrm{FEG} & {}_{\mathit{tbodybyvehicle}}\begin{bmatrix} \mathrm{MOVER}\ \boxed{1} & \\ \mathrm{MANNER} & \mathrm{inair} \\ \mathrm{VEHICLE} & \mathrm{aircraft} \\ \mathrm{PROCON} & \boxed{1} \end{bmatrix} \end{bmatrix}
\end{bmatrix}
\end{bmatrix} \end{bmatrix} \\
\mathrm{I} & \begin{bmatrix} \mathrm{SYNSEM}\,|\,\mathrm{LOCAL} & \begin{bmatrix}
\mathrm{CAT} & \begin{bmatrix} \mathrm{SUBJ}\ \langle \mathrm{NP}\,\mathit{fly\_animal}\boxed{1}\rangle \\ \mathrm{COMPS}\ \langle\rangle \end{bmatrix} \\
\mathrm{CONT} & \begin{bmatrix}
\mathrm{PSOA} \begin{bmatrix} \mathrm{ARG1} & \boxed{1} \end{bmatrix} \\
\begin{bmatrix} \mathrm{FEG} & {}_{\mathit{tanimal}}\begin{bmatrix} \mathrm{MOVER}\ \boxed{1} & \\ \mathrm{MANNER} & \mathrm{inair} \\ \mathrm{PROCON} & \boxed{1} \end{bmatrix} \end{bmatrix}
\end{bmatrix}
\end{bmatrix} \end{bmatrix}
\end{bmatrix}
$$

Figure 7: Lexical rule for co-composition

blocking mechanism by means of a flag inserted in the relevant lexical signs[3].

7 Conclusion

The aim of the work presented in this paper has been twofold: firstly, we have wanted to adapt recent developments in the treatment of polysemy to machine translation where pure sense enumeration techniques in the monolingual analysis dictionary have proven to be inadequate to handle several disambiguation and translation problems. Secondly, Danish motion verbs have proven to be a scarcely studied area although they constitute a highly complex and heterogeneous group of verbs. This relates to a general aspect of Danish verb lexemes, namely that these are in broad terms semantically very ambiguous and rather difficult to classify at hand (see Herslund 1993). This fact about Danish verb lexemes stresses the importance of capturing any systematic pattern that may exist between the different senses of a verb in order to be able to treat them adequately in machine translation, as well as in other NLP applications.

References

Alonge, Bes, Braasch, Calzolari, Dini, Emele, Fillmore, Heid, Monachini, Ostler, Spanu & Vliegen, (1993). 'Some Prominent Approaches to Syntactic and Lexical-Semantic Modelling in the View of DELIS Descriptive work', in: *Report from LRE-project: DELIS*. European Commission, Luxembourg.

Atkins, S. (1994) 'Word Meaning in the Lexicographical Process'. Course material from *EURALEX 1994*, Amsterdam.

Atkins, S. & C. Fillmore, (1994). 'Starting where the dictionaries stop: the challenge of corpus lexicography', in: Atkins, B. & Zampolli, A. (eds.)*Conceptual approaches to the lexicon*. Oxford University Press, Oxford.

Copestake, A. & T. Briscoe (1995). Semi-Productive Polysemy and Sense Extension, in: *Journal of Semantics 12*, p.15-67, Oxford University Press.

Cruse, D.A. (1991). *Lexical Semantics*, Cambridge Textbooks in Linguitics, Cambridge University Press.

Fillmore, C., (1985). 'Frames and the Semantics of Understanding', in: *Quaderni di semantica VI:2*. University of Bologna.

[3]For a further documentation of the actual implementation (which is performed in the ALEP formalism), we refer to Pedersen (forthcoming).

Herslund, M. (1993) 'Transitivity and the Danish Verbs', in:*LAMBDA no. 18*, Copenhagen Business School, Copenhagen.

Lyons, J. (1977). *Semantics*, Cambridge University Press.

Pedersen, B. (1997) 'Danish Motion Verbs: Syntactic Alternations and the Hypothesis of Semantic Determination', in: U. Nikkanne (ed.)*Nordic Journal of Linguistics*, Oslo.

Pedersen, B. (forthcoming) 'A Sample Formalisation of Systematic Polysemy in a Frame Semantic Environment', in: B. Maegaard & B. Pedersen: *UDOG report no. 6*, Center for Sprogteknologi, Copenhagen.

Pollard, C. & I. Sag (1994) *Head-Driven Phrase Structure Grammar*, CSLI, The University of Chicago Press, Chicago.

Pustejovsky, J. (1995) *The Generative Lexicon*, The MIT Press, Cambridge, MAssachusetss, London.

Talmy, L. (1985). 'Lexicalization Patterns: Semantic Structure in Lexical Forms', in: T. Shopen (ed.) *Grammatical Categories and the Lexicon vol. 3*, Press Syndicate of the University of Cambridge.