Sabine Kirchmeier-Andersen
University of Odense

Bolette Sandford Pedersen
Centre for Language Technology
University of Copenhagen

Lene Schøsler
University of Odense

Combining Semantics and Syntax in Monolingual Dictionaries

Attacking the Enemy from Both Flanks

Abstract

In this paper we present a method for sense description and discrimination for Danish which combines semantic and syntactic approaches in the development of dictionaries. Dictionaries for humans are often based on intuitive semantic distinctions, whereas most dictionaries for machines are mainly based on syntactic analysis. In both cases, the dictionaries are unable to provide coherent and sufficiently fine-grained sense distinctions and descriptions. It is our claim that a cognitive frames approach to lexical semantics combined with a syntax-driven approach based on intersubjectively determinable distributional criteria provides us with a systematic and explicit method which paves the way for the development of multi-purpose dictionaries.

0. Introduction

The incorporation of consistent and meaningful methods for sense description and discrimination in the development of dictionaries, be it dictionaries for human beings or for NLP applications, presents a serious problem in most lexicographical work. Dictionaries for humans are often based on intuitive semantic distinctions, whereas most dictionaries for machines are mainly based on syntactic analysis. In both cases, the dictionaries are mostly unable to provide coherent and sufficiently fine-grained sense distinctions and descriptions; a shortcoming which becomes even more evident when dealing with small language dictionaries like the Danish.

Consider the Danish verb *dreje.* According to the standard Danish-English dictionary (Vinterberg & Bodelsen 1990), the main English translations are: 'turn', 'twist', 'dial', 'change', 'rotate', and 'lathe'. In addition, the dictionary distinguishes different syntactic constructions such as '*dreje* + reflexive' meaning 'to be about something' or '*dreje* + particle'. The standard Danish monolingual dictionary (Becker–Christensen & Widell 1990) distinguishes three senses: one corresponding to 'turn', 'dial', 'change' and 'rotate', another sense corresponding to 'to be about something', and the third meaning 'lathe'. The entries for *dreje* in the two dictionaries illustrate the need for a more systematic sense discrimination based on criteria which are explicit and less intuitive.

It is our claim that a combination of a cognitive frames approach and a syntax–driven approach based on intersubjectively determinable distributional criteria can provide a more systematic and explicit method for dictionary development.

Work in this area has been carried out by independent researchers in Denmark, partly on the development of a fully distributional approach to word description, partly in the area of corpus–based frame semantics. Thanks to financial support from the Danish Research Council of the Humanities, it has now become possible to coordinate the two lines of research. The main efforts concentrate on a comparison of the two approaches by examining a number of Danish verbs from the two different viewpoints. The two research groups have based their work on the only available large scale machine–readable corpus of modern Danish (approximately four million words) (Bergenholtz 1990).

1. A cognitive frames approach

The method applied at the Centre for Language Technology is related to the corpus–based work on cognitive frame semantics carried out by Atkins and Fillmore (Atkins & Fillmore 1991). On the basis of our corpus, we aim at an identification of the conceptual framework underlying the meaning of a word, partly as a means to express word content, partly as a method for establishing meaningful and coherent sense distinctions.

Our point of departure is a grouping of words in semantic classes, so that words that can be considered semantic neighbours with similar sets of semantic parameters (henceforth called 'frame elements') are treated coherently. If we look again at *dreje*, the corpus investigations demonstrate that it belongs to the following semantic classes and subclasses (Table 1):

Verbs of physical motion	
Simple verbs of motion	
Hjulet drejer	(the wheel rotates)
han drejer til venstre	(he turns to the left)
han drejer af/fra	(he branches off)
Causative verbs of motion	
hun drejer hovedet	(she turns her head)
han drejer filmen frem	(she winds on the film)
hun drejer op for gassen	(she turns on the gas)
Locative verbs	
vejen drejer til venstre	(the road turns to the left)
Creation verbs	
han drejer i træ	(he lathes in wood)

Table 1: semantic classes of *dreje*

Apart from these three basic groupings, the corpus demonstrates a very frequent metaphoric use of the motion senses, as in *det drejer sig om dig* (it concerns you), and finally it provides us with a list of idiomatic expressions.

In this paper, we will focus on the physical motion class. Here we will see how ambiguity can occur even within one semantic class and why a thorough semantic analysis is required in order to identify the sense-distinguishing factor(s).

Levin and Rappaport, who have undertaken some of the most recent studies of motion verbs for English (Levin and Rappaport 1991), identify three basic types of simple verbs of motion:

(1) **arrive** verbs, implying a semantic element of direction
(2) **run** verbs, implying a semantic component of manner and which have no direct external cause (or expressed positively: that have protagonist control)
(3) **roll** verbs, implying a semantic component of manner and which have direct external cause (or expressed negatively: that have no protagonist control)

Much in the line of Levin's and Rappaport's investigations for English, we have established a similar although further subdivided taxonomy for Danish motion verbs and we have identified the following central frame elements as decisive for our taxonomy:

> **Theme** (The) the item that moves
> **Direction** (Di) origin or goal for the movement, which indicates that direction is involved
> **Manner** (Ma) how movement takes place
> **Direct External Cause** (DEC) direct external factor that causes the movement

These frame elements are in essence cognitively based which means that they originate from our intuition about what the verbs in question mean and which meaning elements they evoke. However, as is also stated in Levin's recent work on alternations and semantic verb classes (Levin 1993), the presence or absence of semantic elements can to a large extent be verified by the way verbs behave. Although Danish does not make use of alternation of arguments to the same degree as English, a set of alternations relevant for Danish motion verbs has been identified in order to establish a test set which can support the affiliations of the verbs in question.[1] Verbs that are considered to belong to the same semantic type or subtype are tested upon the set of alternations in order to verify that they actually behave in a similar way.

Thus, the identification of frame elements constitutes the primary basis for the grouping and description of our verbs and we shall see that *hjulet drejer* (the wheel rotates) and *han drejer til venstre* (he turns to the left) belong to **roll** verbs and **run** verbs, respectively. This can be seen by the fact that they evoke different frame elements, as is illustrated in Table 2. The linking relations to deep syntactic functions (deep subject = argument1, deep object = argument2) and syntactic functions are also expressed in the codings, as well as selectional restrictions on fillers:

reading 1

Example: Hjulet drejer 'the wheel rotates'
Class: Verb of physical motion **Subclass:** simple motion verb
Type: roll verb **Subtype:** iterative
Definition: move in a circular movement
Comments:

frame element	argument	syntax	constituent	selec.restriction
The: +	argument2	subject	NP	inanimate
Di: -				
Ma: +				
DEC: +				

Subcategorisation: subj	**Auxiliary:** have

reading 2

Example: han drejer til venstre/op/opad vejen 'he turns to the left/upwards/up the road'
Class: Physical motion verb **Subclass:** simple motion verb
Type: run verb **Subtype:** body verb
Definition: move by changing direction
Comments: vehicle can be involved

frame element	argument	syntax	constituent	selec.restriction
The: +	argument1	subject	NP	animate or vehicle
Di: +	(modifier)	(modifier)	(PP/adv/adv PP)	(directional)
Ma +				
DEC: -				

Subcategorisation: subj (directional modifier)	**Auxiliary:** være

Table 2: codings of *dreje*

If we proceed to the causative group, which includes only transitive relations, we will see that the set of frame elements must be slightly altered; **Causator** (Ca) replaces direct external cause and is now realised as a primary complement (subject in active constructions, object in passives), and direction is excluded from the set of frame elements, as can be seen in Table 3:

reading 3

Example: hun drejer hovedet (til venstre) / Martin drejede på rattet 'she turned her head (to the left)'/'Martin turned the wheel'				
Class: verb of physical motion				
Subclass: causative motion verb				
Definition: make something move by making it change direction, single turn				
Comments:				
frame element	**argument**	**syntax**	**constituent**	**selec.restriction**
Ca: +	argument1	subject	NP	animate
The: +	argument2	object/ pobj	NP/PP	entity
Ma +				
Subcategorisation: subj obj/pobj (directional modifier)		**Auxiliary:** have		

Table 3: codings of *dreje*

Some of the senses in this group even require the introduction of a **result** frame element, as in *han drejede op for gassen* (he turned on the gas). Here the activated object (the tap) is not realised syntactically.

2. A distributional approach

Since 1991, the research team behind the Odense Valency Dictionary at the University of Odense has been working on a Danish version of the so–called Pronominal Approach, with the intention of constructing a valency database of Danish verbs. At present, the database contains approximately 1000 verb senses. The method applied is an adaptation of the constructivism expounded in the studies by the PROTON–group at the Catholic University of Leuven, Belgium (Blanche–Benveniste et al. 1987).

The Pronominal Approach has been presented as a method capable of establishing language immanent criteria for sense distinction and word description. As the pronominal valency scheme proposed by the Pronominal Approach is able to describe the core combinatoric capacities of a language (Gebruers 1991, p. 282), the Pronominal Approach has been called a "short–cut to linguistic knowledge representation" (Gebruers 1991, p. 247).

One of the basic assumptions of the Pronominal Approach is the existence of a permanent relationship, a relation of proportionality, between the pronouns and the nouns (see Table 4).

He/she/	tells	him/her/	this/that/
+animate +concrete		+animate +concrete	-animate +/-proposition. -concrete
Jim/the boy/		Ann/the girl/	a story/that he liked her/

Table 4: Relation of proportionality in the Pronominal Approach

The identification of specific pronominal paradigms for each valency slot together with a number of distributional tests[2] provide detailed semantico–syntactic characteristics of the verbs and constitute the formal criteria for sense distinction.

It is illustrated in Table 4 that the subject and the indirect object of the verb 'tell' must be a person (i.e. an element caracterized by the semantic features: +animate, +concrete), whereas 'tell' allows an abstract noun or a sentence as direct object (i.e. an element having the semantic features: –animate, +/–proposition, –concrete). These features can be derived directly from the pronominal paradigms (Table 4, top line), but not from the proportionate lexicalized elements listed in the bottom line. It is specific for the pronominal paradigms that they compress semantic and syntactic information in one single paradigm. Lists like the one in the bottom line of Table 4 are the traditional way of giving information on combinatoric possibilities in dictionaries for humans. However, this type of information is insufficient since the user will not be able to guess whether the list is complete or incomplete, which again means that he or she cannot deduce the relevant semantic restrictions from it.

Jim, the boy, Ann, the girl, are all ambiguous with respect to syntactic function, whereas he and she are unambiguous. Him and her are just ambiguous as to the functions of direct or indirect object, this ambiguity being eliminated by means of word–order rules. Thus, the pronominal paradigms provide unambiguous syntactico–semantic information. Being a distributional method, the Pronominal Approach applies only semantic features derived from the pronominal paradigms and does not use intuitively defined semantic features (case roles, theta roles ...), as there is no one–to–one–relationship between such semantic features and surface syntax.

The following semantic and syntactic features are derived directly from the pronominal paradigms which have been established in the different argument slots:

1. **semantic features**: human, concrete, abstract, proposition, location, direction, manner, quantity, countability;
2. **syntactic functions**: subject, object, prepositional object, valency bound adverbial complement;
3. **syntactic form**: noun phrase, prepositional phrase, adverbial phrase, clause (non–finite, finite).

The following semantic and syntactic features are derived indirectly by means of distributional tests:

1. **number of arguments** (including optional arguments): 1–4;
2. **type of passive**: inflectional, analytical either with the passive auxiliary *blive* ('become'), corresponding to the German 'Vorgangspassiv', or *være* ('be'), corresponding to the German 'Zustandspassiv';
3. **'Aktionsart'**: perfective, imperfective, none or both;
4. **type of auxiliary**: *have* ('have') or *være* ('be');
5. **use of 'existentials'**: *der* ('there');
6. **use of preliminary subject**: *det* ('it');
7. **control**, in the case of infinitive complements: subject–, object–, indirect object–, external control;
8. **linking phenomena**: dative alternation, causative/inchoative alter–nation, etc.;

It is important to stress that sense distinctions made in the dictionary result from the observation of differences in the pronominal paradigms combined with the result of the distributional tests. Using the Pronominal Approach the three motion senses described in section 1 would, thus, be distinguished as follows (Table 5):

dreje_1	argument slot 1
pronominal paradigm	hvad, denne her *hvem, *det[3]
semantic features	concrete
syntactic functions	subject
number of arguments	1 (obligatory)
type of passive	no passive
"Aktionsart"	imperfective
type of auxiliary	’have’
example	Kloden/vejrhanen drejer The earth/the weather cock rotates

dreje_2	argument slot 1	argument slot 2
pronominal paradigm	hvem, hvad, denne her *det	hvorhen
semantic features	human, concrete	directional
syntactic functions	subject	adverbial complement
number of arguments	2 (adverbial complement optional)	
type of passive	no passive	
"Aktionsart"	perfective	
type of auxiliary	’være’	
example	Hun/bilen drejer (til højre) She/the car turns (right)	

dreje_3	argument slot 1	argument slot 2
pronominal paradigm	hvem *hvad	hvem, hvad, denne her *det
semantic features	human	human, concrete
syntactic functions	subject	object
number of arguments	2 (both obligatory)	
type of passive	inflectional and both analytical passives	
"Aktionsart"	perfective	
type of auxiliary	’have’	
example	Hun drejer barnet/stolen She turns the child/the chair	

Table 5: Codings of *dreje*

Comparing these results with the three senses described in Section 1, we see that dreje_1, denoting a movement around an axis, can be distinguished formally from the other two senses in the Pronominal Approach by the pronominal paradigms, the number of arguments, and the 'Aktionsart' which is imperfective. Dreje_2 and dreje_3 denote a single turn with no external cause and a causator, respectively. By means of the Pronominal Approach these two senses are distinguished from dreje_1 as described above, and, furthermore, they can be distinguished from each other by the pronominal paradigms, the syntactic form of the argument, and the auxiliary which is *være* for dreje_2 and 'have' for dreje_3.

3. Combining the two approaches

To sum up, the corpus–based Cognitive Frames Approach employing frames and semantic classes is methodologically quite different from the Pronominal Approach which is based on the syntactic and semantic features derivable from a closed word class, the pronouns, as well as distributional tests. Nevertheless, having tested a set of different verbs, we can conclude that the two approaches lead to the identification of identical core sense distinctions. In the light of the ongoing discussion about the relation between syntax and semantics (see among many others Ravin (1990) and Levin & Rappaport (1991)), we find our results striking. It strongly supports the claim that these linguistic domains are in fact closely interrelated.

What still remains to be discussed, however, is exactly how the two methods supplement each other. If the two approaches are able to distinguish the same basic word senses, why then work towards a combination of the two radically different strategies? The Cognitive Frames Approach derives the fundamentals for a semantic grouping and description of words by identifying the cognitive concepts behind the words on the basis of extensive corpus investigations. When the underlying frame elements differ, a new core sense can be identified. But what is not given by the semantic frames is the full syntactic potential of the words in question. General mapping relations between frame elements and syntactic functions are specified on the basis of corpus examples, but the very expensive and time–consuming task of performing corpus–based registration of 'all' possible alternations is not carried out.[4] Here, the Pronominal Approach constitutes a 'short–cut' by means of the distributional test set described above. On the basis of his/her linguistic competence, the linguist can identify the combinatoric potential of a word, and the corpus is only consulted in problematic cases.

On the other hand, the Pronominal Approach can be enriched by the frame semantic analysis with the semantic description and categorisation of words, which is required if the dictionary is to be applied by humans or by NLP systems performing more than a superficial interpretation of word meaning. Using the Pronominal Approach alone, it is only possible to group the different senses of a verb on the basis of shared distribution–based features; semantic relatedness in the sense of semantic domains cannot be expressed. This is well illustrated by *dreje*, where the Pronominal Approach lacks the information of relatedness between i.e. *dreje til venstre* (turn to the left) and *dreje hovedet*, contrary to the creation sense *dreje i træ* (lathe in wood).

4. Concluding remarks

To conclude, the result of our comparisons leads to the following observations: First, the fact that the same core sense distinctions are established by two fundamentally different methods provides evidence for the soundness of each approach and emphasizes their linguistic relevance. Secondly, the fact that the two approaches are complementary, suggests that a combination of the two can constitute a coherent and meaningful method for the future development of multi-purpose dictionaries.

Notes:

1. Alternations considered for Danish motion verbs are, among others: causative alternations, as in: *bolden triller, jeg triller bolden* (the ball rolls, I roll the ball), induced action alternations, as in: *han kører bilen, bilen kører* (the car drives, I drive the car), locative alternations, as in: *han gik til bageren, han gik hen til bageren* (he went to the baker's, he went over to the baker's), cognate object alternations, as in: *han gik en tur* (lit: he walked a walk), unpersonal alternations, as in: *bierne sværmer, det sværmer med bier* (lit: the bees swarm, it swarms with bees).
2. To a large extent these distributional tests correspond to the typology of alternations recently presented in Beth Levins work on verb classes and alternations (Levin 1993). As was also mentioned in Section 1, Levin aims at a more or less intuitive semantic classification supported by syntactic observations, whereas in the OVD syntactico-semantic observations first and foremost are used to establish sense distinctions and a semantic classification is envisaged in a second step. Furthermore, Levins alternation typology is defined vaguely on a mixture of syntactic, semantic and lexical criteria whereas the alternations observed in the OVD are defined formally on the basis of pronominal paradigms and syntactic criteria.
3. Translations: *hvem* ('who'), *denne her* ('this one'), *hvad* ('what'), *det* ('it'), *hvorhen* ('where to')
4. It should be mentioned that Fillmore and Atkins do perform this registration in their work (Fillmore & Atkins 1991) by consulting 2213 citations of 'risk'.

References

Atkins, S. & Fillmore, C. 1992. "Starting where the dictionaries stop: the challenge of corpus lexicography." Atkins, B. & Zampolli, A. (Ed.) *Conceptual approaches to the lexicon.* Oxford University Press, Oxford.

Becker-Christensen, C. & Widell, P. 1990. *Politikens Nudansk Ordbog,* Politikens forlag, Copenhagen.

Bergenholtz, H. 1990. DK87-90 Korpus, Århus.

Gebruers, R. 1991. *On Valency and Transfer-Based Machine Translation,* Leuven.

Levin, B. 1993. *English Verb Classes and Alternations, a Preliminary Investigation.* The University of Chicago Press, Chicago and London.

Levin, B & Rappaport, M. 1991. "The Lexical Semantics of Verbs of Motion: the Perspective from unaccusativity." In: Roca, I. (Ed.): *Thematic Structure.* Berlin.

Levin, B & Pinker, S. 1991. "Introduction to special issue of COGNITION on lexical and conceptual semantics." In: Levin, B. & Pinker, S. (Ed.):*Cognition 41.* Amsterdam.

Ravin, Y. 1990. *Lexical semantics without Thematic Roles.* Oxford.